

Digital Libraries: Virginia Tech Courseware

To learn about digital libraries, you may wish to visit the Self-Study course materials or the online courses listed below.

- [Self-Study Courseware](#)
- [Honors 3004](#): Digital Libraries, Fall 1997, Virginia Tech
- [CS6604](#): Digital Libraries, Fall 1997, Virginia Tech

Please send comments/suggestions to [Ed Fox](#).

Self-study Courseware on

Digital Libraries

[Contents](#)

Introduction: This WWW site has been developed to assist those interested in learning about digital libraries. It is based upon materials tested in 2 Virginia Tech courses taught Fall 1997:

- [CS6604](#)
- [Honors 3004](#)

Students in those courses especially liked Michael Lesk's "[Practical Digital Libraries: Books, Bytes & Bucks](#)" so we refer to it as a supplemental text throughout this site.

There is a set of [quizzes](#) (to be added) to test your knowledge of the chapters in Dr. Lesk's book. We also will support discussion related to these course materials through:

- Listserv (to be added)
- Hypernews (to be added)

Revisions: This site will undergo frequent changes, so do check back. The latest revision was completed 6/27/98.

Acknowledgements: This WWW site was developed in part through funding from NSF grants CDA-9312611, DUE-9752408, and DUE-9752190.

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

[Booksellers](#)[Orders & Inquiries](#)[Catalog](#)[New & Noteworthy](#)[Site Index](#)

Practical Digital Libraries: Books, Bytes, and Bucks

Michael Lesk

The Morgan Kaufmann Series in
Multimedia Information and Systems,
Edward Fox, Series Editor

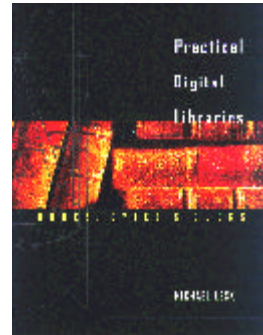
1997

320 pages

cloth

\$49.95

ISBN 1-55860-459-6



[Order This Book](#) | [Authors](#) | [Contents](#) | [Related Titles](#)

"While others speculate on the organizational dilemmas facing libraries and universities in a digital information environment or argue the relative philosophical merits of print versus digital media, Lesk has constructed an on-the-ground picture of the various working components of the digital environment. Viewed as a whole and with an engineering sense of composition, his picture is remarkable--almost astonishing--because it reveals how advanced the digital environment has truly become."

--Donald J. Waters

Associate University Librarian, Yale University

A digital library is not merely a collection of electronic information. It is an organized and digitized system of data that can serve as a rich resource for its user community. This authoritative and accessible guide for librarians and computer scientists explores the technologies behind digital libraries, the choices to be made in building them, and the economic and policy structures that affect them.

The most comprehensive book on the subject, *Practical Digital Libraries*

- offers the most wide-ranging overview of digital libraries currently available
- analyzes economic and intellectual issues in the emerging digital environment

- shows how text, images, audio, and video can be represented, distributed, used, and collected as forms of knowledge

Authors:

[Michael Lesk](#) joined the computer science research group at Bell Laboratories after receiving his Ph.D. degree in Chemical Physics in 1969. He went on to manage the computer science research group at Bellcore, where he is now a chief research scientist. He is best known for his work in electronic libraries, but has worked in document production and retrieval software, computer networks, computer languages, and human-computer interfaces as well. Past chair of the Association for Computing Machinery's special interest groups on Language Analysis and Information Retrieval, Lesk was Senior Visiting Fellow of the British Library in 1987 and is currently Visiting Professor of Computer Science at University College London.

Table of Contents:

- [1 Evolution of Libraries](#)
- [2 Text Access Methods](#)
- [3 Images of Pages](#)
- [4 Multimedia Storage and Access](#)
- [5 Knowledge Representation Methods](#)
- [6 Distribution](#)
- [7 Usability and Retrieval Evaluation](#)
- [8 Collections and Preservation](#)
- [9 Economics](#)
- [10 Intellectual Property Rights](#)
- [11 International Activities](#)
- [12 Future: Ubiquity, Diversity, Creativity, and Public Policy](#)

Instructors are invited to [request an examination copy](#).

Related Titles:

[Multimedia Information & Systems Database](#)



Copyright © 1997, Morgan Kaufmann Publishers, Inc.
Telephone: (415) 392-2665 Email: mkp@mkp.com

Contents :

- [Introduction to Digital Libraries](#): This holds general information such as definitions, glossary of digital library terms, foundations and scenarios.
 - [Topics](#): This contains information classified under various topics of/related to Digital Libraries e.g. "Metadata" etc.
 - [Resources](#): Provides other information based under more general headings such as various people involved in Digital Libraries, projects, countries and regions etc.
 - [References](#): This category contains references, links and pointers such as conferences/workshops, journals and books, and various related courses being conducted at different universities.
-

Pedagogy:

We recommend that beginners start with the Introduction and then proceed through the Topics, following along with the text by Dr. Lesk. The Resources provide alternate views of the contents, and the References should serve those desiring additional details.

[\[Main\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Introduction to Digital Libraries:

- [Definitions](#): Some of the attempts made by various people to define a digital library.
- [Foundations](#): Introductory material related to digital libraries...
- [Scenarios and Perspectives](#): Various scenarios and perspectives that arise in a Digital Library context.

[\[Main\]](#) [\[Contents\]](#)

Please send comments/suggestions to [Ed Fox](#). (c) Copyright 1998, Edward A. Fox, Rajat Gupta

Definitions :

- "The generic name for federated structures that provide humans both intellectual and physical access to the huge and growing worldwide networks of information encoded in multimedia digital formats."
([The University of Michigan Digital Library: This Is Not Your Father's Library](#), [Birmingham](#), 1994)
- "Systems providing a community of users with coherent access to a large, organized repository of information and knowledge."
([Lynch](#), 1995)
- "Digital libraries are a set of electronic resources and associated technical capabilities for creating, searching, and using information. In this sense they are an extension and enhancement of information storage and retrieval systems that manipulate digital data in any medium (text, images, sounds; static or dynamic images) and exist in distributed networks. The content of digital libraries includes data, metadata that describe various aspects of the data (e.g., representation, creator, owner, reproduction rights), and metadata that consist of links or relationships to other data or metadata, whether internal or external to the digital library."
([UCLA-NSF Social Aspects of Digital Libraries Workshop](#))
- Digital libraries are constructed -- collected and organized -- by a community of users, and their functional capabilities support the information needs and uses of that community. They are a component of communities in which individuals and groups interact with each other, using data, information, and knowledge resources and systems. In this sense they are an extension, enhancement, and integration of a variety of information institutions as physical places where resources are selected, collected, organized, preserved, and accessed in support of a user community. These information institutions include, among others, libraries, museums, archives, and schools, but digital libraries also extend and serve other community settings, including classrooms, offices, laboratories, homes, and public spaces." ([UCLA-NSF Social Aspects of Digital Libraries Workshop](#))
- "systems providing a community of users with coherent access to a large, organized repository of information and knowledge. This organization of information is characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technology" (adapted from [Inter operability, Scaling, and the Digital Libraries Research Agenda](#))
- "Digital library is a concept that has different meanings in different communities. To the engineering and computer science community, digital library is a metaphor for the new kinds of distributed data base services that manage unstructured multimedia data. To the political and business communities, the term represents a new marketplace for the world's information resources and services. To futurist communities, digital libraries represent the manifestation of Wells' World Brain. The perspective taken here is rooted in an information science tradition."
([Gary Marchionini](#))
- "A digital library is a distributed technology environment which dramatically reduces barriers to the creation, dissemination, manipulation, storage, integration, and reuse of information by individuals and groups."
([Edward A. Fox](#) , editor, Source Book on Digital Libraries, pg. 65)
- "A digital library is a machine readable representation of materials which might be found in a

university library together with organizing information intended to help users find specific information. A digital library service is an assemblage of digital computing, storage, and communicate machinery together with the software needed to reprise, emulate, and extend the services provided by conventional libraries based on paper and other material means of collecting, storing, cataloging, finding, and disseminating information."

([Edward A. Fox](#) , editor, Source Book on Digital Libraries, pg. 65)

- "an organized data base of digital information objects in varying formats maintained to provide unmediated ease of access to a user community, with these further characteristics:
 - an overall access tool (e.g. a catalog) provides search and retrieval capability over the entire data base;
 - organized technical procedures exist through which the library management adds objects to the data base and removes them according to a coherent and accessible collections policy."

([Peter Graham](#), Rutgers University Libraries)

- "A library that has been extended and enhanced by the application of digital technology. Important aspects of the digital library that may be extended and enhanced include :
 - Collections of the library
 - Organization and management of the collections
 - Access of the library items and the processing of the information contained in the items
 - Communication of information about the items "

([Smith](#), 1995)

Digital Library related terms/glossary

(by [Peter Graham](#), Rutgers University Libraries):

- digital archive: a digital library which is intended to be maintained for a long time, i.e. periods longer than individual human lives and certainly longer than individual technological epochs. (Sometimes formerly also "digital research library.")
- digital preservation: preservation of artifactual information by digitizing its image (e.g. scanning a manuscript page, digitally photographing a vase, or converting a cylinder recording to digital form).
- electronic preservation: preservation of information that is in digital (that is, electronic) form, i.e. the techniques associated with refreshing, migration and assurance of integrity.

Digital Preservation techniques:

- Refresh: to copy digital information from one long-term storage medium to another of the same type, with no change whatsoever in the bit stream (e.g. from a decaying 800 bpi tape to a new 800 bpi tape, or from an older 5 1/4" floppy to a new 5 1/4" floppy).
- "Modified refreshing" is the copying to another medium of a similar enough type hat no change is made in the bit pattern that is of concern to the application and operating system using the data, e.g. from an 800 bpi tape to a 1600 bpi tape or to a "square", cartridge, tape; or from a 5 1/4" floppy disk to a 3 1/2" floppy disk.
- Migrate: to copy data, or convert data, from one technology to another, whether hardware or

software, preserving the essential characteristics of the data; generally forward in time. (At the moment, it is recognized, this final qualifier begs many questions.) Examples: conversion of XyWrite w/p files to Microsoft Word; conversion of ClarisWorks v3 spreadsheet files to Microsoft Excel v4 files; conversion of binary tape images of survey research multi-punched tab cards to a data base format; copying an 800 bpi tape file to a sequential disk file; converting a DOS FoxPro data base to a Visual Basic database for Windows 95; converting a PICT image to a TIFF image; converting a ClarisWorks for Windows v4 w/p file to a Macintosh ClarisWorks v4 file.

Examples can be given, as here, for cases known to be required; the longer term preservation problem is to prepare for forward migrations when the future technologies are unknown.

- Emulate: (find and use better Comp SCI terms here, probably) in hardware terms, the creation of software for a computer that reproduces in all essential characteristics (as defined by the problem to be solved) the performance of another computer of a different design. Computers may emulate earlier computers in order to provide backward compatibility, or may emulate a future computer in order to provide a software development environment while the newer computer is still being fabricated.

In software preservation terms, the creation of software that analyzes the software environment of a document such that it can provide a user interface to the document that substantially reproduces the essential characteristics of the document as it was created by its originating software.

- Document: (use sense that Apple began to use, with Macintosh; anything manipulated by an application; find their definition and build on it. Note Dublin Core [and other] use of "document like object").
- Authenticate: of users, to verify that network users are in fact who they identify themselves to be; of documents, to validate the integrity of a document with respect to its original authorized creation.
- Authentication: (of a resource--i.e. of data, not people)
- Authenticity: (of a resource--i.e. of data, not people)
- Integrity: synonym of authenticity (of a resource--i.e. of data, not people)

[\[Main\]](#) [\[Introduction\]](#) [\[Contents\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Foundations (see Lesk Ch. 1, 8):

- [As We May Think](#) by Vannevar Bush - the visionary article that helped motivate early work on digital libraries, hypertext and information retrieval
 - What is a "[digital library](#)"? (vs. a virtual library)
 - UCLA workshop (focusing on user perspectives):
 - [Introduction](#)
 - [information life cycle](#)
 - [Artists](#)
 - [Business Records as Artifacts](#)
 - [Health-Information Systems](#)
 - IITA workshop: [Definitions and Roles of Digital Libraries](#)
 - [Digital Libraries: Issues and Architectures](#)
 - [Digital Library: Gross Structure and Requirements: Report from a March 1994 Workshop](#). Also available in [PDF](#).
-

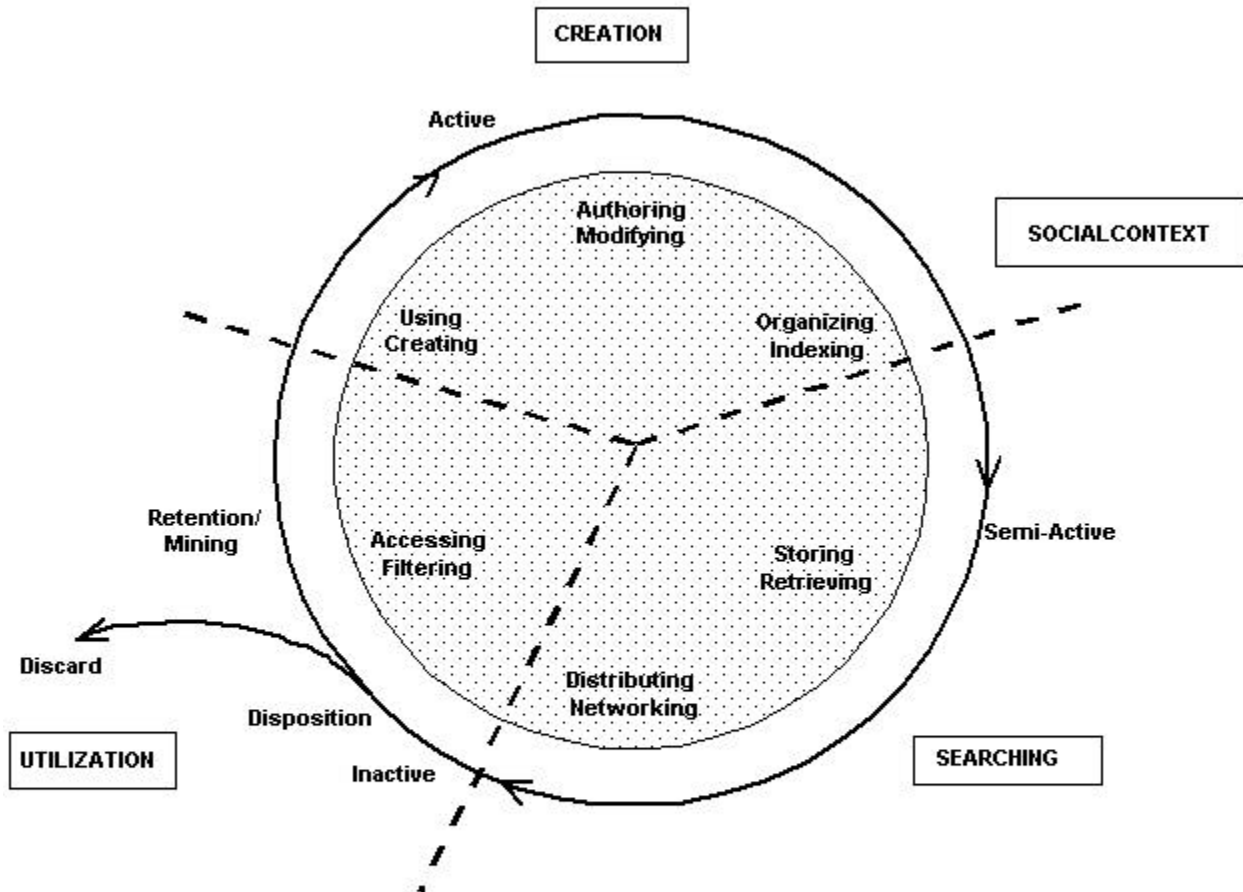
Pedagogy:

We recommend that the above items be skimmed to obtain a general background regarding digital library research, development, and practice. Please also read chapters 1 and 8 of Dr. Lesk's book.

[\[Main\]](#) [\[Contents\]](#) [\[Introduction\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Information Life Cycle

NOTE: The outer ring indicates the life cycle stages (active, semi-active, and inactive) for a given type of information artifact (such as business records, artworks, documents, or scientific data). The stages are superimposed on six types of information uses or processes (shaded circle). The cycle has three major phases: information creation, searching, and utilization. The alignment of the cycle stages with the steps of information handling and process phases may vary according to the particular social or institutional context.

Interoperability, Scaling, and the Digital Libraries Research Agenda:

A Report on the May 18-19, 1995

IITA Digital Libraries Workshop

August 22, 1995

Clifford Lynch (clifford.lynch@ucop.edu)
Hector Garcia-Molina (hector@db.stanford.edu)

Converted to HTML using GradStudentWare 2.2

Contact [Christian Mogensen](#) with bug reports.

[Introduction](#)

[Definitions and Roles of Digital Libraries](#)

[Defining Interoperability in the Digital Library Environment](#)

[Infrastructure Requirements for Digital Library Research](#)

[Research Issues and Priorities](#)

[1. Interoperability](#)

[2. Description of Objects and Repositories](#)

[3. Collection Management and Organization](#)

[4. User Interfaces and Human-Computer Interaction](#)

[Conclusions](#)

[Executive Summary](#)

[Appendix 1 - List of Participants](#)

[Appendix 2 - Strawman Report](#)

[Appendix 3 - Report of the working groups](#)

[3-1 - The Publishing Perspective](#)

[3-2 - The Commercial Perspective](#)

[3-3 - The Library Perspective](#)

[3-4 - The Internet Perspective](#)

[3-5 - The Multimedia Perspective](#)

Introduction

This report summarizes the results of a workshop on Digital Libraries held under the auspices of the U.S. Government's Information Infrastructure Technology and Applications (IITA) Working Group in Reston, Virginia on May 18-19, 1995. The objective of the workshop was to refine the research agenda for digital libraries with specific emphasis on issues of scaling and interoperability, and to identify the infrastructure developments needed to make progress on these issues.

While there have been a number of workshops and other meetings examining the broader questions of support for applications in the National Information Infrastructure (NII), we believe this was the first workshop that focused specifically on Digital Libraries in this context. In the past year, Digital Libraries

have emerged as one of the central and most compelling applications enabled by the NII; numerous digital library research projects are underway, including six large-scale pilot projects that have been funded jointly by ARPA, NASA, and NSF. While Digital Libraries are now a vibrant research area, and also a field in which considerable commercial development is taking place (presaging the future economic importance of Digital Library technology to the United States), many new questions are emerging as a result of this flowering of research activity. Informed by insights gained from current research, this workshop offered an opportunity to consider questions such as interoperability objectives that might be defined among projects now underway.

The workshop was organized by Hector Garcia-Molina of Stanford University and Clifford Lynch of the University of California Office of the President. The IITA working group, which sponsored the meeting, reports to the National Science and Technology Council (NSTC) through the High Performance Computing, Communications, and Information Technology subcommittee of the Committee on Information and Communication. The workshop was attended by some 60 leading digital library researchers and developers and by representatives from a wide range of federal government organizations concerned with research and development and policy formulation related to digital libraries (see [Appendix 1](#) for a roster of attendees).

Workshop attendees were asked to consider the following questions as a point of departure in developing the research agenda:

1. What is a Digital Library? How does it differ from an information repository or from today's World Wide Web? How many Digital Libraries will there be, and how will they interlink? How might this look to users?
2. What Digital Library infrastructure is needed? What does "infrastructure" consist of in this context and how does it differ from the broader applications support infrastructure for the emerging NII? What is the relationship between infrastructure and standards? Who will use this infrastructure? When must it be defined, and what parts are most urgently needed? How does the infrastructure relate to intellectual property management and publisher concerns?
3. How can a Digital Library be evaluated? How will we know in three to four years if current research projects have been successful in developing effective digital library services for their user communities?

To further frame and stimulate discussion, Hector Garcia-Molina prepared a position paper discussing the issues and distributed it prior to the workshop (see [Appendix 2](#)).

Participants spent the majority of the workshop in one of five groups; unlike many workshops, in which each group is assigned a different set of issues, here each group approached the full spectrum of questions from a specific, unique viewpoint and generated a summary of their discussions that reflected that viewpoint. After a presentation from the five group leaders representing each group's approach to the issues, each participant selected his or her group. The five groups and their leaders were

Bill Arms,
Corporation for National Research Initiatives:
The Publishing Perspective

Michael Lesk,

Bellcore:
The Commercial Perspective

Bruce Schatz,
University of Illinois Urbana Champaign:
The Library Perspective

Mike Schwartz,
University of Colorado:
The Internet Perspective

Terry Smith,
University of California, Santa Barbara:
The Multimedia Perspective

The reports of these five groups appear in [Appendix 3](#). This summary of the workshop extracts common themes and also key points of disagreement from the work of the five groups and places them in broader context. The report is not a consensus document; while it draws heavily on the five group reports and has also benefited greatly from comments from attendees, it does not attempt to reflect completely any of the five group reports.

This report addresses responses to the first two questions posed to the attendees (the definition of a digital library and infrastructure needs to support digital libraries and discusses the research agenda. The third question posed to the attendees -- how to evaluate Digital Library projects -- did not receive much attention from most of the groups; it is to be the subject of a separate workshop on User Evaluation Methods to be held October 29-31 at the Allerton Center under the auspices of The University of Illinois Urbana-Champaign and NSF. Some groups did identify the need for consistent instrumentation and data gathering across projects to facilitate evaluation. In addition, several groups stressed the need to make the transition from a systems technology framework to one driven by user access and collection organization in developing future digital library technology and systems. This view is perhaps most eloquently stated in the reports of the Internet working group and the Library working group.

Definitions and Roles of Digital Libraries

Considerable work has already been done on operational definitions of Digital Libraries and their relationship to traditional library institutions, as well as to the broader systems of scholarly and commercial publishing (see, for example, Communications of the ACM, April 1995). Much of the discussion in this workshop was motivated by questions of scaling, interoperability and needed support infrastructure.

Digital libraries were viewed as systems providing a community of users with coherent access to a large, organized repository of information and knowledge. One group made the provocative proposal that this organization of information was characterized by the absence of prior detailed knowledge of the uses of the information. The ability of the user to access, reorganize, and utilize this repository is enriched by the capabilities of digital technology; the Multimedia group provided particularly vivid examples of these possibilities.

Several groups pointed out that, in fact, digital libraries would, for the foreseeable future need to span both print and digital materials and that the central issue was to provide a coherent view of a very large

collection of information. In this sense, an emphasis on content solely in digital format is too limiting. Really, the objective is to develop information systems providing access to a coherent collection of material, more and more of which will be in digital format as time goes on, and to fully exploit the opportunities that are offered by the materials that are in digital formats. Additionally, the comprehensiveness and value of the collection accessible through a digital library system can be strengthened by the ability to integrate materials in digital formats that have not been well-represented, easy to access, or effectively usable in traditional library collections, such as multimedia, geospatial data, or numerical datasets. There is, in reality, a very strong continuity between traditional library roles and missions and the objectives of digital library systems.

Participants in the workshop repeatedly underscored this continuity, and emphasized that the traditional library institutional missions of collection development, collection organization, access, and preservation must extend to the digital library environment. Digital libraries will be a component in the broader range of future library services, and librarians will play a central role in developing and managing digital libraries.

While there would be many digital repositories, a given digital library system should provide a coherent, consistent view of as many of these repositories as possible. From the user's perspective, to the extent possible, there should appear to be a single digital library system. Users increasingly have access to various types of digital collections and information systems: personal information resources, workgroup and organizational information collections and collaboration environments, and more public digital libraries. Defining the boundaries and characteristics of these information spaces and exploring ways in which they can be fused into a coherent whole is a central problem that cuts across all aspects of the research agenda. From the user's perspective, the digital library system needs to extend smoothly from personal information resources, workgroup and organizational systems, and out to personal views of the content of more public digital libraries.

Some groups raised, but did not resolve, the question of the extent to which the digital library system should incorporate support for publishing, annotation, and integration of new information, and the extent to which additions to repositories within the digital library system should be mediated by librarians. It is clear that the development of digital libraries is closely linked to the changes that are occurring in modes of scientific and scholarly communication; the extent to which the digital library should actively embrace -- and perhaps even drive -- these changes remains to be fully explored.

Libraries -- digital or traditional -- exist to serve diverse purposes and constituencies. To some extent, each discipline, constituency, and collection creates its own organization of information. In the digital library world this differentiation among library collections, organization, and services may become more visible. One of the key challenges is to retain this diversity, which is responsive to unique constituencies, and at the same time permit information to be effectively shared across disciplines and constituencies. This is an essential component of the interoperability questions that formed a major focus for the workshop. Workshop participants represented many of these diverse perspectives: university research libraries, archives, libraries supporting teaching, public libraries, and libraries of the performing arts.

Defining Interoperability in the Digital Library Environment

Defining interoperability proved difficult. It is clear that this is still a central research problem in its own right, and one that merits continued attention. Discussions of infrastructure focused on common tools, enabling technologies and standards that would provide a basis for further exploration of interoperability issues, particularly by encouraging and facilitating the growth of digital libraries on the Internet.

Considerable effort was spent on identifying infrastructure that was either unique or particularly critical to progress in digital libraries, as opposed to more general-purpose infrastructure that a range of NII applications, including digital libraries, might share. One clear theme was that an understanding of interoperability issues required operational experience which could only be gained by large-scale deployment of digital library systems. Speculation about interoperability in the abstract is of very limited value.

Participants expressed a full spectrum of views on interoperability. At one end of the spectrum is the use of common tools and interfaces that provide a superficial uniformity for navigation and access but rely almost entirely on human intelligence to provide any coherence of content. At the opposite end of the spectrum is deep semantic interoperability. The precise definition of deep semantic interoperability was the subject of some debate, but deals with the ability of a user to access, consistently and coherently, similar (though autonomously defined and managed) classes of digital objects and services, distributed across heterogeneous repositories, with federating or mediating software compensating for site-by-site variations. It also extends beyond passive digital objects to actual services offered by specific digital library systems. Deep semantic interoperability is a "grand challenge" research problem; it is extraordinarily difficult, but of transcendent importance, if digital libraries are to live up to their long-term potential. An intermediate position between these two extremes advocates primarily syntactic interoperability (the interchange of metadata and the use of digital object transmission protocols and formats based on this metadata rather than simply common navigation, query, and viewing interfaces) as a means of providing limited coherence of content, supplemented by human interpretation.

Note that the term "digital object" here is intended only to describe, in the broadest sense, the type of information objects that may comprise a digital library -- textual, audio, video, numeric, computer programs, or multimedia composites of such components. It is not intended either to endorse or preclude an object-oriented architectural framework for digital library systems (in the sense of object-oriented programming or object-oriented databases, for example).

Infrastructure Requirements for Digital Library Research

The most urgent infrastructure need is to establish common schemes for the naming of digital objects, and the linking of these schemes to protocols for object transmission, metadata, and object type classifications. The consensus of the groups was that naming schemes for digital objects that allow global unique reference represented perhaps the most immediate infrastructure deployment priority in order to facilitate resource sharing, linkages, and interoperation among digital library systems and to facilitate scale-up of digital library prototypes. It was recognized that the design of large-scale naming systems and their integration into the larger digital library framework will continue to be an important research area, but that infrastructure support needs to be put in place quickly for at least an interim system, and that in fact experience with such an interim system would inform further research.

The deployment of a public key cryptosystem infrastructure -- including the development of a system of key servers and the definition of standards and protocols -- was also identified as essential to progress in digital libraries; this is necessary to support digital library needs in areas such as security and authentication, privacy, rights management, and payments for the use of intellectual property. While the need for public key cryptosystem infrastructure is hardly unique to digital libraries, the importance of the digital library services and components which depend on this infrastructure mean that its absence represents a significant barrier. In particular, until these problems are addressed, it seems unlikely that we will see commercial publishers and other information suppliers making large amounts of high-value copyrighted information broadly available to digital library users. This in turn will constrain the

development of research prototypes and may be a distorting factor in studies of user behavior.

Research Issues and Priorities

The working groups outlined a wide range of important research issues; most groups were less successful at prioritizing them, beyond the immediate infrastructure needs already discussed. The five key research areas that emerged from the workshop are described below; arguably, the first three are of most central and immediate importance, specifically to the development of digital libraries, though the long-term importance of research in the fifth area (economic, social, and legal issues) cannot be overemphasized. The distinctions among the five areas are to some extent arbitrary; for example, progress on interoperability (the first area) depends critically on progress in our ability to describe successfully objects and repositories (the second area).

1. Interoperability

The difficulty in defining the objectives for interoperability have already been discussed; clarifying these objectives, mapping the spectrum of interoperability, and establishing the key challenges at points along this spectrum are key research issues in their own right.

The more technical interoperability research involve protocol design that supports a broad range of interaction types, inter-repository protocols, distributed search protocols and technologies (including the ability to search across heterogeneous databases with some level of semantic consistency), and object interchange protocols. Interoperability is not simply a matter of providing coherence among passive object repositories. Digital library systems offer a range of services, and these services must be projected in an interoperable fashion as well. One particular issue that emerged was that existing Internet protocols (such as HTTP, the basis of the World Wide Web) are clearly inadequate. Research must move beyond the current base of deployed protocols and systems. This raises complex questions about how to deploy prototype systems and the tradeoffs between advanced capabilities and ubiquity of access.

The practical question of the nature of the installed technology base and the need to support this installed base will increasingly frame and influence interoperability research. Access to digital libraries is not an end in itself for most users, but rather a support service; many will be willing to sacrifice advanced functionality for consistency, stability, and ability to use familiar, common access tools. Just as the installed base has become the greatest barrier to meaningful large-scale trials of new approaches that improve existing services (as opposed to providing entirely new services which do not compete with an installed base) in the overall Internet environment, user expectations and the installed base will ultimately impede progress in fundamental technology research within the large-scale experiments necessary to gain insights into interoperability among digital libraries. Managing this tension will be a critical element in the continued development of the community's research agenda.

It should be noted that, at this relatively early stage in the evolution of digital library technology, it is of vital importance that projects strive for approaches that incorporate high functionality and extensibility. A high level of functionality in the standards and protocols used, even if not fully exploited initially, will postpone the time when the inertia of the installed base begins to confine research opportunities. Careful design of extensibility in digital library systems will facilitate continued research progress and understanding of the impact of new approaches on the user community without the need to attempt to displace an installed base.

2. Description of Objects and Repositories

In order to provide a coherent view of collections of digital objects, they must be described in a consistent fashion which can facilitate the use of mechanisms such as protocols that support distributed search and retrieval from disparate sources. Research in description of objects and collections of objects provides the foundation for effective interoperability. Interoperability at the level of deep semantics will require breakthroughs in description as well as retrieval, object interchange, and object retrieval protocols.

Issues here include the definition and use of metadata and its capture or computation from objects, the use of computed descriptions of objects, federation and integration of heterogeneous repositories with disparate semantics, clustering and automatic hierarchical organization of information, and algorithms for automatic rating, ranking, and evaluation of information quality, genre, and other properties. Other key issues involved knowledge representation and interchange, and the definition and interchange of ontologies for information context. The idea of active "information matchmaking" emerged in several group reports.

Research is also needed to understand the strengths and limitations of purely computer-based technologies for describing objects and repositories, and the appropriate roles for the efforts of human librarians and subject experts in the digital library context as a complement to these technology-based approaches.

3. Collection Management and Organization

Collection management and organization research is the area where traditional library missions and practices are reinterpreted for the digital library environment. Progress in this area is essential if digital library collections are to meet successfully the needs of their user communities.

Policies and methods for incorporating information resources on the network into managed collections, rights management, payment, and control issues were all identified as central problems in the management of digital collections. Approaches to replication and caching of information and their relationship to collection management in a distributed environment need careful examination. The authority and quality of content in digital libraries is of central concern to the user community; ensuring and identifying these attributes of content calls for research that spans both technical and organizational issues. Research is also needed to clarify the roles of librarians and institutions in defining and managing collections in the networked environment.

With the enhanced potential to support nontextual content effectively in the digital library environment, issues in nontextual and multimedia information capture, organization, and storage, indexing and retrieval are clearly key research areas. However, textual digital documents remain a vitally important research area in their own right, and are far from fully understood. The role of knowledge bases in digital libraries remains a poorly explored but potentially important question.

The preservation of digital content for long periods of time, across multiple generations of hardware and software technologies and standards is essential in the creation of effective digital libraries. This is an extraordinarily difficult research problem which has not received sufficient attention.

4. User Interfaces and Human-Computer Interaction

While user interfaces and human-computer interaction issues are an extensive field of research in their own right, there are some specific problems that are central to progress in digital libraries.

Display of information, visualization and navigation of large information collections, and linkages to information manipulation/analysis tools were identified as key areas for research. The use of more sophisticated models of user behavior and needs in long-term interactions with digital library systems is a potentially fruitful area for research. The necessity for a more comprehensive understanding of user needs, objectives, and behavior in employing digital library systems was stressed repeatedly as a basis for designing effective systems. Finally, it was observed that digital library systems must become far more effective in adapting to variations in the capabilities of user workstations and network connections (bandwidth) in presenting appropriate user interfaces; new technologies such as personal digital assistants and nomadic computing models will emphasize this need.

5. Economic, Social, and Legal Issues

Digital libraries are not simply technological constructs; they exist within a rich legal, social, and economic context, and will succeed only to the extent that they meet these broader needs. Rights management, economic models for the use of electronic information, and billing systems to support these economic models will be needed. User privacy needs to be carefully considered. There are complex policy issues related to collection development and management, and preservation and archiving. Existing library practice may shed some light on these questions. The social context of digital documents, including authorship, ownership, the act of publication, versions, authenticity, and integrity require a better understanding. Research in all of these areas will also be needed if digital libraries are to be successful.

Conclusions

This workshop has made substantial progress in refining and focusing a research agenda for digital libraries, as well as in developing insights into questions about interoperability among digital libraries and the infrastructure necessary to support such interoperability. Interoperability is likely to continue to be a useful organizing theme in refining this agenda in the coming years. The outcomes of the workshop also suggest that a focus on broad architectural issues in digital libraries will be fruitful. Several working groups commented on the need to develop component software strategies that would facilitate the transfer of technology among the current digital library pilot projects and from these projects to other new digital library research efforts. The Internet working group went further in suggesting that the development of a broadly available software base for the digital library community would contribute to rapid progress, and we believe that this suggestion deserves careful consideration.

Scaling was identified as a major area of concern. The common vision is one of tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system. Accommodating this very large number of repositories -- a very different environment than that in which today's handful of pilot projects operate -- will clearly have major implications for infrastructure definition and design. We must move rapidly towards an infrastructure that can support and facilitate research towards this common vision. The full range of issues here are unclear. Some immediate needs are evident; these are reflected in the emphasis on establishing naming systems for digital objects as a high priority, for example.

We don't know how to approach scaling as a research question other than to build upon experience with the Internet. However, attention to scaling as a research theme is essential and may help in further clarifying infrastructure needs and priorities, as well as informing work in all areas of the research agenda outlined above. For example, reliability questions are poorly understood; in a sufficiently large system, some components will inevitably be out of service during the processing of any given query. The need to support large-scale deployment projects (in terms of size of user community, number of objects, and number of repositories) and to study subsequently the effectiveness and use of such systems was emphasized repeatedly. It is clear that limited deployment of prototype systems will not suffice if we are to understand the research questions involved in digital libraries.

Research in scale-up is very difficult to perform except by building and deploying a large-scale digital library system. Establishing infrastructure and tools to facilitate experimentation with large-scale systems is essential, as is funding to study use and behavior of large-scale systems once deployed through this infrastructure. The Internet as a context for deploying digital library systems offers an unprecedented opportunity -- not only technically by providing connectivity to an enormous potential user base but also culturally, given the Internet community's models and traditions of technology diffusion through the distribution of publicly available prototype software -- to move ahead large-scale experiments. Research efforts should exploit these opportunities.

Finally, it seems clear that the inevitable presence of large amounts of commercially valuable, proprietary information in the future -- which can be viewed as another form of scale-up in digital libraries -- will also shape the research agenda in new ways. The near-term focus is on overcoming the infrastructural barriers to supporting proprietary information (such as authentication, billing, and rights management). There are research issues in the design of such an infrastructure, but also operational and policy problems impeding deployment. While some of the research issues are complex and will require ongoing exploration, putting at least the first steps towards the necessary infrastructure in place to accommodate such commercially valuable information is a high priority in advancing the research agenda and addressing scale-up issues. It will also stimulate commercial developments that will complement existing research initiatives. The development of an increasingly rich marketplace of information resources under a wide range of economic and legal constraints will create new opportunities in all areas of the research agenda presented above, and will allow us to explore vital new research questions in the development of description, navigation, access, and resource discovery technologies and systems that can function in this broader environment.



Digital Libraries: Issues and Architectures

Peter J. Nürnberg

Richard Furuta

John J. Leggett

Catherine C. Marshall

Frank M. Shipman III

Center for the Study of Digital Libraries

Texas A&M University

College Station, TX 77843

USA

{pnuern, furuta, leggett, marshall, shipman}@bush.cs.tamu.edu

ABSTRACT

The research field of digital libraries must be viewed as a union of subfields from a variety of domains combined with new research issues in order to realize its full potential. A clear exposition of the research issues involved has not yet been given. Most approaches to building digital library systems have thus far been limited to addressing specific digital library problems as variations of problems from other fields. This paper presents a taxonomy of digital library elements. Consideration of the elements in this taxonomy helps suggest a variety of issues. Example elements and some issues they suggest are used to populate the taxonomy. The paper continues by presenting a general digital library system architecture. Issues suggested by the taxonomy are shown to have implications at many levels of digital library system architectures for both design and implementation. This is illustrated by considering the implications of one issue (personalizing presentations) at several architectural levels and in the context of a set of current technologies.

Keywords: digital library issues, digital library architecture, databases, physical libraries, World Wide Web

INTRODUCTION

The emerging field of digital libraries brings together participants from many existing areas of research. Currently, the field lacks a clear agenda independent of these other areas. It is tempting for researchers to think that the field of digital libraries is a natural outgrowth of an already known field. From a database or information retrieval perspective, digital libraries may be seen as a form of federated databases. From a hypertext perspective the field of digital libraries could seem like a particular application of hypertext technology. From a wide-area information service perspective, digital libraries could appear to be one use of the World Wide Web. From a library science perspective, digital libraries might be seen as continuing a trend toward library automation. There is some truth to these perspectives (as well as others) but none address the field as a whole and its research agenda. The field of digital libraries will be

limited if viewed only as a subfield of prior research interests. To realize its full potential, the field must be viewed as a union of subfields from a variety of domains combined with additional goals, and thus new research issues. Digital library research must both respect the existing tradition of our physical libraries and transcend current practice in developing a new, broader research agenda.

What are the research issues central to digital libraries? One issue might be how to digitize objects and put them on-line. A second might be how to include new forms of information that do not have temporal or tangible representation necessary for inclusion into physical libraries. Another could be how to locate materials in the new digital library. Yet another would be when to use and when to transcend the existing technologies and traditions of the physical library in its digital form. Still other issues stem from the problems of information overload created by new information technologies. This paper presents a framework for thinking about the field of digital libraries and the research issues that are part of it and demonstrates how these issues affect digital library systems.

The next section gives an analysis of the digital libraries field by positing that the digital library can be modeled to some degree after the physical library, and discussing the relationship between the two. In order to show the breadth of the research agenda in digital libraries, a taxonomy of the elements of the digital library, and some issues raised by considering these elements is then presented. Following this, a general system architecture for digital library systems is presented. Issues suggested by considering the prior taxonomy are shown to affect many layers of these systems.

PHYSICAL AND DIGITAL LIBRARIES

Why is a digital library called a library at all? This question has been addressed by various members of this research community. Miksa and Doty [1994] discussed the notions of collection, information sources, and place with respect to physical libraries and how these notions might carry over into the digital realm. Levy and Marshall [1995] considered how work practices in physical libraries might be used in the design of digital libraries. The physical library can provide the starting point for discussing the elements and domains of digital libraries. An element of a library is a constituent part of the library. A domain of the library is the universe from which the library materials are drawn.

Elements

It is helpful to consider three broad classes of library elements: data, metadata, and processes. *Data* are library materials. *Metadata* are information about the library and its materials. *Processes* are active functions performed over library elements. For example, a book in a library may be thought of as being data of that library. An index over book titles (in a card catalog, for example) may be thought of as library metadata. The act of a librarian helping a patron find a book by suggesting the use of the card catalog may be thought of as a process.

This classification is vague, in the sense that it may be difficult or impossible to classify any given library element as distinctly belonging to a particular class. It may be possible to view a single element as belonging to all three classes. However, this classification is useful since it provides a framework for discussion about library elements. Physical library elements often fulfill some role for a given library user at a given moment. These roles often can be assigned in specific cases in a meaningful way.

Because this classification concerns elements in the library, it ignores differences in roles played by people interacting with the library, the various ways in which these roles are being reassigned in the digital library, and the different high-level tasks people fulfilling these roles perform. These are of course

all important issues, but will not be considered here.

This classification of physical library elements can be applied to digital library elements as well, with the same understanding that a given element may be thought of differently by different users at different times.

Domains

A physical library deals primarily with physical data, whereas a digital library deals primarily with digital data. Of course most modern libraries deal with both, but it is useful for sake of discussion to consider hypothetical "all-physical" and "all-digital" libraries as foils.

If physical libraries primarily contain physical data and digital libraries primarily contain digital data, then how can digital libraries preserve and disseminate the vast amounts of existing physical data? Instead of containing the physical data itself, digital libraries will contain digital translations of this data. The term translation is used, because the process of generating these digital representations of physical data is not necessarily a completely meaning-preserving process. The product may not be perceived by users in the same way that the source is perceived since their media of presentation are necessarily different [McLuhan 1964].

It might be tempting to think that if there are differences between analogous physical and digital objects, they have no practical consequence. This would imply, however, that all such differences are already known. Not only is this not the case, but it is not even clear that all such differences can ever be known, because one cannot know, a priori, all the important characteristics of an object in any situation [Suchman 1987]. Without knowing all of the differences between physical and digital objects, how could one claim that these differences are insignificant?

The magnitude of differences between physical and digital analogs may be related to the accuracy of the physical/digital translation. A spectrum of translation quality certainly exists. Without more research into the effects of translating material between physical and digital form, it is difficult to know the accuracy of such translation.

The difference between the physical and digital domains also has implications for translating the metadata and processes of physical libraries. Some of the metadata and processes of a physical library (e.g. card catalogs and shelving) are themselves physical elements, and thus, the discussion of translations as formulated above applies. However, even those elements of the physical library that have no direct physical reality (e.g. the Library of Congress classification scheme) are often inextricably tied to the physicality of data and the library itself. These abstractions, also, need to be translated into the digital realm.

In summary, though both physical and digital libraries may be thought of as sharing certain goals and of consisting of elements that may be classified similarly, the domains of the two types of libraries differ. Digital libraries will deal with translated physical elements, conceptual elements of the physical library adapted to the digital realm, and completely new digital elements with no apparent physical library analog (e.g. hypertexts). Differences between physical library and digital library elements have created many open problems concerning how to adapt the tradition of the physical library into the digital realm.

TAXONOMY OF DIGITAL LIBRARY ISSUES

Given the above discussion, it is reasonable to classify the elements in digital libraries along two axes. Firstly, elements may be classified as data, metadata, or processes. Secondly, these elements may be translations of physical library elements or new digital library elements with no clear physical library analog. This results in the grid shown in Figure 1.

	Data	Metadata	Processes
Translations of Physical Library Entities	Book Journal Movie	Static index Classifications Spatial arrangement	Acquiring data Suggesting sources Helping locate sources
New Digital Library Entities	Hypernovel Scientific visualization Computer program	Dynamic index Personalized structure Annotations	Full-text searching Personalizing presentation Retrieving by agents

Figure 1: Taxonomy of Digital Library Elements.

Each section of the grid is discussed below. Examples of elements that may be thought of as belonging to the section in question are given, followed by an issue particularly relevant to that section. These issues and their positions in the grid are shown in Figure 2. As stated earlier, a given element may be thought of as being classified in many different sections on the grid, but elements are placed so that some typical use of that element is highlighted. Also, problems raised in each section may (and often do) apply to other sections as well, but may be thought of as having special significance in their respective sections.

	Data	Metadata	Processes
Translations of Physical Library Entities	What to translate?	How to translate metadata that is dependent on data physicality?	How to provide tools for human involvement in these processes?
New Digital Library Entities	How to account for the continual rapid evolution of new data types?	How to insure consistency of separately maintained metadata?	How to distribute computation?

Figure 2: Issues Raised by Considering the Taxonomy of Digital Library Elements.

Translations of Physical Library Data

It is easy to find examples of physical library data that are translated into digital form routinely. For example, books, journals, and movies are all examples of physical library data that are scanned, digitized, or otherwise translated and put on-line [Lesk 1991].

A central problem in translating physical library data is deciding which aspects of the original merit consideration in the translation process. When translating a book into digital form, when does an ASCII representation of the text suffice? When must each page be scanned as a photograph would be? How are such decisions to be made? These questions involve many tradeoffs, and answers cannot be known in the general case [Løkken 1993].

It is not even clear which characteristics of an object are most meaningful. Many characteristics of physical data, such as size and shape of a book, may be meaningful only to some people or in only some

circumstances. Consider how grease smudges on the sides of auto parts manuals aid people in finding desired pages [Hill and Hollan 1992]. It is impossible to include every characteristic of a physical data object that may ever be deemed meaningful to any person, but ignoring meaningful aspects of an object during translation has important implications for the preservation of function in a digital library.

Translations of Physical Library Metadata

Examples of physical library metadata are plentiful. Long-lived indexes (such as those in card catalogs), classification schemes (such as the Library of Congress classification scheme) and spatial arrangement of library materials are three examples.

A problem with translating such physical library metadata is that often either the metadata itself or its application is influenced by the physicality of the data. For example, the spatial arrangement of data objects in a physical library conveys meaning and is a form of metadata. Spatial arrangement of objects is meaningful because the objects have some physical presence. How can this be translated into the digital realm? Is a virtual reality approach, in which digital objects are associated with some virtual physical presence in a virtual physical place, the correct way to translate this metadata? Or, is the correct approach one that spatially arranges abstract images in an abstract space?

While spatial arrangement of library materials is a physical library metadata element with physical presence, other metadata with no direct physical reality must also be translated, or adapted in its application, if it is to be used in a digital library. For example, the Library of Congress classification scheme may not have any physical reality itself, but its application is sometimes constrained by the physicality of the objects it classifies. For example, such a classification scheme is often used to guide the physical location of data in a library, because placing like-classified objects in physical proximity can aid patrons in locating data. If a library has one copy of a book, but the book could be classified in more than one category, how is the book to be located? It can effectively only be co-located with sources of one classification. This same limitation does not hold for digital objects located in a virtual space.

Translations of Physical Library Processes

Many kinds of physical library processes exist. Three examples of such processes are acquiring data, suggesting the usefulness of elements, and aiding in the location of elements. An example of acquiring data is choosing new books to add to a library. Suggesting the usefulness of elements might take the form of a patron identifying potentially helpful data and metadata sources to a colleague who might otherwise not have known about nor used these sources. An example of aiding in the location of elements is a library worker helping a patron locate an object given incomplete information.

One characteristic shared by many physical library processes is that they are performed by human beings. A key problem in translating such physical library processes into the digital library realm is how to provide human beings with tools to assist them in performing these often informal processes, especially since digital library patrons and librarians cannot rely on co-location with people likely to be helpful. This problem is particularly important given the inherently collaborative nature of many tasks performed in the library [Ehrlich and Cash 1994, Marshall et al. 1994, Schnase et al. 1994].

New Digital Data

Hypernovels, scientific visualizations, and active computer programs are all examples of new digital library data that do not have clear physical library data analogs. It could be claimed that novels on paper

are clear predecessors to hypertexts, but hypertexts have many characteristics that qualitatively differentiate them from their paper counterparts [Moulthrop 1991]. It is certainly conceivable to build a library of active computational objects. Also, many physical objects (usually) not currently included in the physical library due to space or other restrictions (e.g. transcripts of radio programs or videos of television shows) may have digital analogs in the digital library.

One problem faced by digital library designers and implementers when considering new digital library data is that new types of this data are constantly and rapidly evolving. While it is true that new physical types of data are constantly evolving, the pace of change in the digital realm is currently greater, because of immaturity of new digital data types. New potentials are constantly being recognized and used. It is particularly difficult to design or implement a digital library if the types of data to be included in the library are not yet known.

New Digital Metadata

Many new kinds of metadata are possible in a digital library. Three examples are dynamically generated indexes, personalized structures over library elements, and annotations. Dynamically generated indexes may have relatively short life-spans compared to the long-lived indexes of the physical library. One example of personalized structures are user- or group-specific sets of hypertext links over some set of library elements. Annotations are virtual modifications of data objects by patrons - these modifications exist separately from the data but may be always displayed with the data for a particular user or group, thereby effecting a "virtual" modification [Løkken 1993].

A problem with new digital library metadata is that much of it is personal, and thus may be stored separately from the data over which it applies, leading to possible consistency errors. If many users build structure over certain data in a library, and that data changes, what should be done with all of the metadata that is in some way invalidated by this change? This is certainly a problem in the physical library. Because most physical library metadata resides in the library itself, however, it may be easier to modify the metadata to reflect any changes in data. With personal digital library metadata, all such copies of metadata may not be known. To what degree is the digital library system responsible for propagating changes to patrons with metadata that relies on the changed material? How can this propagation be effected?

New Digital Processes

Finally, the digital library allows new processes not found in the physical library. Specifically, processes such as full-text searching, personalizing presentations, and retrieving by agents are new digital library processes. Full-text searching refers to querying a full-text index. Personalizing presentations involves access control issues as well as tailored screen layouts. Retrieving by agents involves programs that search data autonomously and report findings to users.

One problematic aspect of these new processes is that they involve computation that may access large amounts of library data or metadata. A central problem is how to distribute the computation needed to maintain these processes. For example, how much of the computation involved in personalizing presentation of information should be done by the server and how much should be done by the client? If such processes are computationally expensive, how can this load be fairly distributed? What is the optimal mix of client / server communication, server-side computation, and client-side computation for effecting these processes?

DIGITAL LIBRARY SYSTEMS

The taxonomy of issues presented in the previous section illustrates the wide range of problems to be considered when designing and implementing a digital library. This section presents a conceptual template of a general digital library system architecture and illustrates by example how issues identified in the previous taxonomy can have implications in several areas of this architecture. The section closes by considering what role is played by some of today's current technologies when constructing a digital library system.

Digital Library System Architecture

Conceptually, a digital library system may be thought of as mediating certain kinds of interactions among people and computing systems. Figure 3 shows some relationships and interactions among several parts of the digital library and several people and systems external to the library. To help clarify the interactions occurring in these relationships, the computing resources in this figure have been partitioned into server resources and client resources. This allows the classification of computer-supported relationships into human/human, human/client, human/server, and client/server classes.

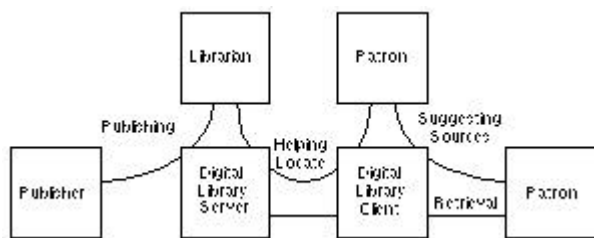


Figure 3: Conceptual Role of a Digital Library System with Example Relationships.

The real relationships are often more complicated than shown. For example, publishing in the digital library is not strictly a relationship between publisher, librarian, and the digital library server. Patron needs, budgetary constraints, limitations of library computing resources, and a number of other factors may be involved. Any robust digital library system should provide support for these complex relationships.

The client and server computing systems may each be further subdivided. Each may be thought of as consisting of three parts: the back-end, the "middle-end", and the front-end. Both the back-end and the front-end of a system define interfaces between the system itself and some external entity. A system front-end normally provides services to external clients, while the back-end is provided with services from external servers. The middle-end provides some intermediate mapping between the front- and back-ends. Figure 4 illustrates the same entities as shown in Figure 3, but with the divisions of the client and server into their respective back-, middle-, and front-ends.

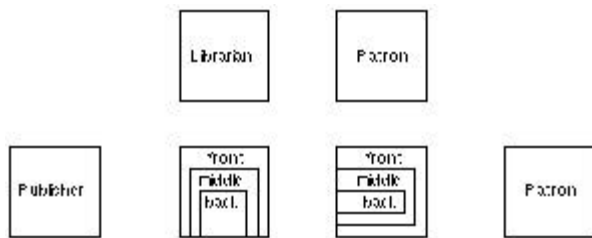


Figure 4: Digital Library System Architecture.

Mapping Issues to Solutions

The issues identified in the taxonomy may have implications in several areas of the digital library system. This section illustrates this point by taking one issue raised previously and identifying the areas of the digital library system that are affected.

Consider the issue raised in the discussion of new digital processes - how can the computational and storage load be equitably divided between client and server for these new processes. Specifically, consider the new digital process of personalizing the presentation of material.

Addressing this issue cannot be confined to any one part of the digital library system. The publishers of digital library data must consider *how to format* their data stored in the server back-end so that it may be presented in a personalized way on the client side. The server middle-end must address *how much preprocessing* should be done, which involves a tradeoff between possibly sending too much unprocessed data versus spending too much computing time on the server side. The server front-end and the client back-end must agree on *which protocol* to use to send the semi-processed data. The client middle-end must address *how to distribute* data retrieved from the server among many displays on the client front-end processes. Finally, the client front-end must address *how to make personalization of presentation a usable feature* for library patrons. These points are just some examples of what must be considered at different levels of a digital library system to address one element or issue raised in the above discussion on the taxonomy of elements.

Current Technologies

This section closes by considering how one set of current technology maps to the general digital library system architecture, and how the example of personalized presentations is addressed by this current technology. The technology considered is a set of WWW clients communicating with httpd servers that use Common Gateway Interface (CGI) scripts and/or binaries to access a database [Berners-Lee et al. 1992]. This system and its mapping to the terminology presented above are shown in Figure 5.

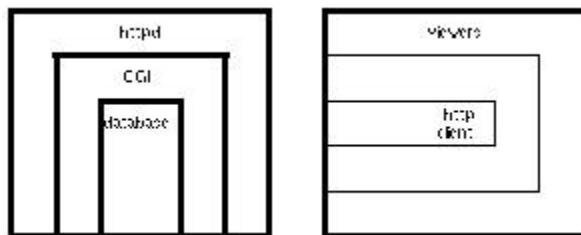


Figure 5: Current Technology Mapping to Digital Library System Architecture. Distinct processes are

separated by heavy lines. Divisions that may or may not imply separate processes are marked by medium lines. Hypothetical intra-process divisions are marked with light lines.

Consider how this technology answers just the questions raised in the above section. There are many ways for publishers to answer the question of *how to format* their data. Several popular formats exist for digital data translated from the physical realm, such as Graphics Interchange Format (gif) for still video images or ASCII for plain text. Publishers of database data may choose any of these popular formats appropriate for their needs, since many of the more popular formats can be handled on the client front-end. Formats for new digital data types are still forming, such as the evolving HyperText Markup Language (HTML) for hypertextual documents [Berners-Lee and Connolly 1995]. There are no generally agreed upon formats for more exotic digital elements such as process-based dynamic hypertexts.

The question of *how much server-side preprocessing* of the data can be done by CGI scripts is difficult to answer. On the one hand, these scripts are capable of arbitrary computation, and can be passed meaningful strings appended to URL's. However, the scripts themselves are static. In current practice, because presentations are rarely personalized at the client front-end, CGI scripts rarely do much preprocessing of the retrieved data before passing it to the server front-end.

The question of *what protocol* is to be used between the server front-end and the client back-end seems to be temporarily resolved in favor of a mix of http, ftp, gopher, and a handful of other protocols. New protocols can clearly be and will need to be added to support new data types by adding new URL access methods. However, the fact that the same object referenced by two URL's with different access methods may have different (non-access method) identifiers does not allow easy dynamic negotiation of protocols between server and client. One research issue to consider is the effects this dependence of the identifier has on the access method.

Currently, most Web clients do not support multiple front-ends in any meaningful way. This means that multiple front-ends require the back-end to replicate server calls even if they are displaying the same data. Thus, the current technology does not address *how to distribute* client-retrieved information to multiple client front-ends.

Finally, current Web clients only allow a small degree of personalization of presentation. This is essentially limited to specifying viewers for non-inlined data, specifying some parameters for how to display in-lined data, and possibly providing information to the server via an HTML forms interface about what kind of data should be retrieved. Thus the only personalization of data in the client front-end concerns display of data and not access to data. Web clients need to provide more tools to patrons of digital libraries to *allow easy personalization* of data with respect to both presentation and access.

In summary, Web clients communicating with httpd servers using CGI scripts to access databases has technology in several of the areas of the general digital library system architecture outlined above, with the exception of an identifiable client middle-end to handle multiple front-ends corresponding to one client back-end. Some issues, such as how to format new data types, and what protocols to use to communicate this data, can be addressed somewhat independently and solutions can be integrated at a later time. Other issues, such as client-side filtering of information that allows personalization with respect to access, are not currently addressed.

CONCLUSIONS

Physical libraries provide a good starting point for discussion of digital libraries. Elements of both the

physical and digital libraries may be categorized as data, metadata, or processes; these categories are determined in specific instances by the intended use of elements by librarians, patrons, or others. Data, metadata, and processes of the physical library must be translated into the digital domain if they are to be used in the digital library. Additionally, there are types of library elements with no clear physical library analog - wholly new digital library elements. These observations led to the development of a taxonomy of digital library elements.

Issues raised by the taxonomy of digital library elements have implications at several levels of digital library systems. Examining the problem of personalizing presentations identifies sample issues at all levels of the architecture. Specifically, considering personalizing presentations led to identifying issues of data format (server back-end), server-side preprocessing (server middle-end), protocols (server front-end to client back-end), client-side distribution (client middle-end), and user tools (client front-end). By first identifying a digital library issue, and then considering the implications for system design and implementation, the myopia of considering issues at one architectural level isolated from issues at other levels is avoided. Also, by applying this approach from *digital library* issue to *digital library system* solutions, system designers and implementers can better understand that decisions made at one architectural layer about seemingly low-level issues (e.g. how to format data) can affect high-level capabilities (e.g. personalizing presentations) provided to the end-user.

The field of digital libraries presents a set of complex issues, and solutions to these problems will require a blending of approaches from a variety of fields. Claims that any one technology has solved all of the issues posed in the design and implementation of digital libraries fail to address the entire problem. For example, proponents of the view that federated databases solve the technical issues of digital libraries have only considered technology at the server back-end to handle already made translations of physical library data and metadata. Even augmenting such databases with other current technologies such as Web clients, httpd's and CGI scripts does not provide a fully functional digital library system. Instead, any successful attempt at constructing a digital library system will need to address issues raised by considering the many different kinds of digital library elements throughout the various levels of the general digital library system architecture.

ACKNOWLEDGEMENTS

This work was supported in part by the Texas Advanced Research Projects Agency under grant number 012345.

REFERENCES

Berners-Lee, T. J. and Connolly, D. W. 1995. HyperText Markup Language Specification - 2.0 (IETF Draft).

Berners-Lee, T. J., Cailliau R., Groff, J. F., Pollermann B. 1992. World-Wide Web: The information universe. *Electronic Networking: Research, Applications and Policy* 2(1) (Spring), pp. 52-58.

Ehrlich, K., and Cash, D. 1994. Turning information into knowledge: Information finding as a collaborative activity. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 119-125.

Hill, W. C., and Hollan, J. D. 1992. Edit wear and read wear. *Proceedings of the Human Factors in*

Computing Systems '92 Conference, (Monterey, CA, May 3-7), pp. 3-10.

Lesk, M. 1991. The CORE electronic chemistry library. *Proceedings of the Fourteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, (Chicago, IL).

Levy, D. M., and Marshall, C. C. 1994. Going digital: a look at assumptions underlying digital libraries. *Communications of the ACM* 38(4) to appear.

Løkken, S. 1993. Text Representations In Digital Hypermedia Library Systems. M.S. Thesis. Department of Computer Science, Texas A&M University. College Station, TX (Dec).

Marshall, C. C., Shipman, F. M., and McCall, R. J. 1994. Putting digital libraries to work: Issues from experience with community memories. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21, pp. 126-133.

McLuhan, M. 1964. Understanding media; the extensions of man. Mc-Graw-Hill. New York.

Miksa, F., and Doty, P. 1994. Intellectual realities and the digital library. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 1-5.

Moulthrop, S. 1991. Beyond the electronic book: A critique of hypertext rhetoric. *Proceedings of the Third ACM Conference on Hypertext (Hypertext '91)*, (San Antonio, TX, Dec), pp. 291-298.

Schnase, J. L., Leggett, J. J., Metcalfe, E. S., Morin, N. R., Cunnius, E. L., Turner, J. S., Furuta, R. K., Ellis, L., Pilant, M., Ewing, R. E., Hassan, S. W., and Frisse, M. 1994. The CoLib project - Enabling digital botany for the 21st century. *Proceedings of the Digital Libraries '94 Conference*, (College Station, TX, Jun 19-21), pp. 108-118.

Suchman, L. A. 1987. Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge University Press. New York.

Defining Scenarios & Perspectives:

- [Publishing](#)
 - [Commercial](#)
 - [Library](#)
 - [Internet](#)
 - [Multimedia](#)
-

Pedagogy:

We recommend that the scenarios given be examined, especially for the group in which the reader fits.

[\[Main\]](#) [\[Contents\]](#) [\[Introduction\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. FOX, Rajat Gupta

Contents :

- [Introduction to Digital Libraries](#): This holds general information such as definitions, glossary of digital library terms, foundations and scenarios.
 - [Topics](#): This contains information classified under various topics of/related to Digital Libraries e.g. "Metadata" etc.
 - [Resources](#): Provides other information based under more general headings such as various people involved in Digital Libraries, projects, countries and regions etc.
 - [References](#): This category contains references, links and pointers such as conferences/workshops, journals and books, and various related courses being conducted at different universities.
-

Pedagogy:

We recommend that beginners start with the Introduction and then proceed through the Topics, following along with the text by Dr. Lesk. The Resources provide alternate views of the contents, and the References should serve those desiring additional details.

[\[Main\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Resources:

- [Projects](#)
- [People](#)
- [Countries and regions](#)
- [Centers, sites and organizations](#)

[\[Main\]](#) [\[Contents\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. fox, Rajat Gupta

Projects:

DLI

- [home page](#)
- [information & resources](#)
- [publications](#)
- [Carnegie Mellon University](#)
- [Stanford University](#)
- [University of California at Berkeley](#)
- [University of California at Santa Barbara](#)
- [University of Illinois](#)
- [University of Michigan](#)

[American Memory Project \(Library of Congress\)](#)

[UK Electronic Library Programme](#)

Virginia Tech Projects:

- [Interactive Courseware on Digital Libraries](#) (this site itself is a part of it)
- **Interactive Learning with a Digital Library in CS** <http://ei.cs.vt.edu/>
 - Interactive Learning with a Digital Library in CS arch <http://ei.cs.vt.edu/~cs5604/Adv/Adv-ILDLCS.html>
 - Courseware <http://ei.cs.vt.edu/courses.html>
 - [Project Overview \(for FIE'96, in PDF\)](#)
 - [Project Interim Report, Oct. 1996](#)
 - [Project Report for NSF EI PI Meeting, Nov. 1996](#)
- **Envision (CS literature)** <http://ei.cs.vt.edu/~cs5604/Adv/Adv-Envision.html>
 - Envision report <http://ei.cs.vt.edu/papers/ENVreport/final.html>
- **CODER** <http://ei.cs.vt.edu/~cs5604/Adv/Adv-CODER.html>
- **MARIAN**
 - overview <http://ei.cs.vt.edu/~cs5604/Adv/Adv-MARIAN.html>
 - system <http://opac3.cc.vt.edu/htbin/marian>

Singapore Network: [SINGAREN](#)

Some Extra Virginia Tech resources on various projects:

- Build upon existing electronic materials
 - Netlib (numerical analysis) <http://www.netlib.org/>
 - Attribute/value search http://www.netlib.org/utk/misc/netlib_query.html
- Build upon publishers collections
 - AAAS - Science Online <http://www.aaas.org/>
 - ACM DL <http://www.acm.org/dl/>
 - ACS (Chemistry) - Online <http://www.acs.org/>
 - CORE Overview <http://ei.cs.vt.edu/~cs5604/DL/DL2.html>

- D-Lib Magazine, Dec. 1995, Making a Digital Library, Chemistry Online Retrieval Experiment <http://www.dlib.org/dlib/december95/briefings/12core.html>
 - CORE at OCLC <http://www.oclc.org:5047/oclc/research/projects/core/>
- Elsevier
 - Science Direct <http://www.elsevier.nl/>
 - TULIP (material science & engineering) homepage <http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml>
 - With universities + OCLC
- Highwire Press
- IEEE
- IEEE CS
- JSTOR
- Commercial services and systems
 - IBM <http://www.software.ibm.com/is/dig-lib/>
 - Version 2 <http://www.software.ibm.com/is/dig-lib/v2factsheet/>
 - collection treasury <http://www.software.ibm.com/is/dig-lib/treasury/>
 - images - QBIC <http://www.qbic.almaden.ibm.com/>
 - news archive <http://www.software.ibm.com/is/dig-lib/newsarchive/>
- Enhance WWW (hypertext):
 - HyperWave <http://www.hyperwave.de/>
 - HyperWave server features
 - HyperWave author <http://www2.iicm.edu/hyperwave/author>
 - HyperWave author features <http://www2.iicm.edu/hyperwave/author/features.html>
 - HyperWave author specs <http://www2.iicm.edu/hyperwave/author/specifications.html>
 - Harmony <http://www2.iicm.edu/harmony>
 - Harmony orientation
 - Harmony screens <http://ei.cs.vt.edu/~cs5604/Adv/Adv-Harmony.html>
 - Harmony information structuring
 - Harmony document viewers
 - Amsterdam model <http://ei.cs.vt.edu/~mm/gifs/Amsterdam-hm.html>
- Community network multimedia history
 - BEV <http://www.bev.net>
 - BEV History <http://history.bev.net/bevhist/>
 - Timeline <http://history.bev.net/bevhist/historyBase/mainTimeline.html>
 - Screen for 1992
 - Screen for Article
- Discipline - Greek Literature <http://www.perseus.tufts.edu/>
 - Evaluation - TOIS
- Discipline - Computer Science
 - Technical reports
 - WATERS - through 1995
 - CSTR <http://WWW.CNRI.Reston.VA.US/home/cstr.html>
 - NCSTRL <http://www.ncstrl.org/>
 - Search results, Search results abstract
 - Doc. thumbnails, Doc. page 1
 - Ptrs
 - DLs for CS <http://fox.cs.vt.edu/DLCS.html>
 - Dienst <http://researchsmp2.cc.vt.edu:8090/>
 - Results page, document page from search

- Genre - ETDs - electronic theses and dissertations
 - Virginia Tech <http://etd.vt.edu/>
 - Submission form <http://scholar.lib.vt.edu/cgi-bin/etd.cgi>
 - Approval form <http://etd.vt.edu/submit/approval.htm>
 - Letter to students <http://etd.vt.edu/submit/letter.htm>
 - Standards <http://etd.vt.edu/submit/mm.htm>
 - Collection <http://www.theses.org>
 - Project - Networked Digital Library of Theses and Dissertations <http://www.ndltd.org>
 - Brief description <http://www.ndltd.org/info/descr.htm>
 - D-Lib Magazine Overview September 1996
<http://www.dlib.org/dlib/september96/theses/09fox.html>
 - D-Lib Magazine Update September 1997
<http://www.dlib.org/dlib/september97/theses/09fox.html>
 - FIPSE proposal
 - abstract <http://www.ndltd.org/support/fipseabs.htm>
 - full-text <http://www.ndltd.org/support/fipse10.pdf>
- [PDF PART 1 - PROJECTS](#)
 - Fig. 1: Timeline of Recent Information & DL Systems
 - Fig. 2: NCSTRL Architecture
 - NETLIB (numerical analysis)
 - CORE (chemistry)
 - TULIP (material science & engineering, with Elsevier, OCLC)
 - IBM digital libraries products and projects
 - Hyper-G/HyperWave (clients and servers)
 - BEV HistoryBase
 - CS technical reports (CS-TR, WATERS, NCSTRL) and related efforts
 - CS education (ACM literature, courseware on IR, multimedia, hypertext, history)
 - Digital Library Initiative (CMU, Michigan, Stanford, UC Berkeley, UC Santa Barbara, University of Illinois Urbana-Champaign)
 - ETD (electronic theses and dissertations - NDLTD)
- [PDF PART 2 -SOURCES, RESEARCH](#)
 - Digital Library conferences
 - IITA meetings (e.g., May 1995 workshop)
 - Allerton Institutes (from U. Illinois, NSF)
 - D-Lib (research, magazine, working groups)
 - D-Lib research articles (architecture, metadata, URNs, use)
 - Virginia Tech information (DL page, Sourcebook)
 - Virginia Tech projects (Envision, ILDLCS, WWW traffic analysis)
 - Z39.50 (overview, OCLC, CNIDR)
 - Library of Congress
 - CNRI (architecture, handles)
 - UMBC agents
 - LIS: preservation, TEI, ...



AVAILABLE RESEARCH

University of California at Berkeley

Environmental Planning and
Geographic Information Systems

University of California at Santa Barbara

The Alexandria Project:
Spatially-referenced Map
Information

Carnegie Mellon University

Informedia Digital Video Library

University of Illinois at Urbana-Champaign

Federating Repositories of
Scientific Literature

University of Michigan

Intelligent Agents for Information
Location

Stanford University

Interoperation Mechanisms
Among Heterogeneous Services

DLI Project [Contacts](#)

[DLI Workshop Series](#)

[DLI Publications](#)

The Initiative's focus is to dramatically advance the means to collect, store, and organize information in digital forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.

Digital Libraries basically store materials in electronic format and manipulate large collections of those materials effectively. Research into digital libraries is research into network information systems, concentrating on how to develop the necessary infrastructure to effectively mass-manipulate the information on the Net.. The key technological issues are how to search and display desired selections from and across large collections. Summaries of the six DLI projects from the May 1996, [Special Issue on Digital Libraries](#) in the Institute of Electrical and Electronics Engineers, IEEE Computer Magazine.

The magazine of digital library research, the [D-Lib Magazine](#), including the July/August 1996 issue [The DLI Testbeds: Today and Tomorrow.](#)

Digital Library conference information, publications, related projects and resources to the DLI, [Digital Library Related Information and Resources.](#)

[NSF Digital Libraries Contact](#)

National Synchronization for the Digital Library Initiative is being coordinated by the University of Illinois at Urbana-Champaign, and supported by a supplemental grant by the National Science

Digital Library Information and Resources

Research on digital libraries encompasses a range of intertwined technical, social and political issues. One of the better descriptions of digital libraries comes from the Santa Fe Workshop on Distributed Knowledge Work Environments. "[T]he concept of a "digital library" is not merely equivalent to a digitized collection with information management tools. It is rather an environment to bring together collections, services, and people in support of the full life cycle of creation, dissemination, use, and preservation of data, information, and knowledge." I have made my selections for this page on the basis of their breadth, depth, ingenuity and availability of content online.

Table of Contents:

1. [The Digital Libraries Initiative \(DLI\)](#)
2. [Select Digital Library Related Projects](#)
3. [Upcoming Digital Library Conferences](#)
4. [Previous Digital Library Conferences](#)
5. [Previous Digital Library Related Conferences with Online Proceedings](#)
6. [Full Text of Other Digital Library Related Publications](#)
7. [Other Digital Library Related Resources](#)
8. [Digital Library Funding, Coordination and Policy Organizations](#)
9. [Intellectual Property](#)
10. [Human Computer Interaction \(HCI\)](#)
11. [Computer Supported Cooperative Work \(CSCW\)](#)

The Digital Libraries Initiative

The [Digital Libraries Initiative](#) is comprised of six projects in the [joint initiative](#) of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (DARPA) and the National Aeronautics and Space Administration (NASA) for digital libraries. These projects are developing the next generation of tools for information discovery, management, retrieval and analysis. A mostly comprehensive list [DLI publications](#) and the [DLI Workshop Series](#) are online.

The DLI projects are: [University of Illinois Urbana-Champaign](#), [Carnegie-Mellon University](#), [Stanford University](#), [University of California at Berkeley](#), [University of California at Santa Barbara](#) and [University of Michigan](#).

Select Digital Library Related Projects

The [Interspace](#) is a long term information infrastructure research project which seeks to unify disparate distributed information resources in one coherent model. The Interspace, is a collection of interlinked information spaces where each component space contains the knowledge of a community or a subject domain.

The [Networked Computer Science Technical Reports Library](#) at Cornell University Department of Computer Science. NCSTRL is a distributed technical report library developed by the ARPA-sponsored Computer Science Technical Report Project. "NCSTRL (pronounced "ancestral") is an international collection of computer science technical reports from CS departments and industrial and government

research laboratories. The NCSTRL collection is distributed among a set of interoperating servers operated by participating institutions."

The [Networked Digital Library of Theses and Dissertations](#) is a project which aims to increase the availability of theses and dissertations by placing them online with the content in an accessible form. The works may be accessed through the [Electronic Thesis and Dissertation Library](#).

The Los Alamos National Laboratory(LANL) [Library Without Walls](#) is a broad based digital library project to make information available to researchers no matter where their desktops are located. The [LANL e-Print archive](#) "has already supplanted traditional research journals in some fields of physics. It is a formal mode of communication in which each entry is archived and indexed for retrieval at later times."

The [The CURIA Project's Thesaurus Linguarum Hiberni](#), "is a joint project of the Royal Irish Academy and the University College Cork to provide an interactive on-line searchable database archive of literary and historical materials in the various languages of early, mediæval and modern Ireland. The documents are being scanned from authoritative printed editions, or keyboarded from fresh manuscript transcriptions and encoded in SGML according to the recommendations of the Text Encoding Initiative."

The [Perseus Project](#) centered at the Department of Classics of Tufts University is a well known and respected collection which focuses upon the ancient Greek and Roman world. Perseus contains texts in Greek and in translation. The major authors of the classical period are represented, as well as some later authors from the fifth century B.C. Perseus also contains images of vases, sculptures and sculptural groups, coins, buildings, as well as color maps of Greece taken from satellite images, annotated with place names.

The [RYHINER-Project at the University Library of Berne](#) "consists of more than 15,000 maps, charts, plans and views from the 16th to the 18th century, covering the whole globe. Together with the 20,000 manuscript maps of the Public Records Office, the Canton of Berne owns not only a local, but a worldwide geographical memory. Work on this project includes conservation, microfilming and building up a generally accessible catalog."

[Project Bartleby](#) from Columbia University seeks to be the public library of the Internet. It reproduces classic literature in hypertext and maintains a strong emphasis on the quality and integrity of the text.

[Project Gutenberg](#) is the granddaddy of literary content on the Net. The goal of it's director and founder, Michael Hart, is no less than putting 10,000 works online by the year 2001. All works are in plain ASCII and in the public domain. In making the texts available to the lowest common denominator Project Gutenberg attempts to reach the most people and thus have the greatest impact.

Xerox has created a collection called [Digital Libraries and Xerox](#) with papers discussing digital libraries and their research efforts. Xerox also has a number of interesting related projects including the [Digital Tradition Folk Song Database](#) which contains the words and music to thousands of folk songs. Additionally, the [Xerox PARC Map Viewer](#) uses public geographic data to render sections of the world on the fly.

The [Visible Human Project](#) from National Library of Medicine (NLM) produced "a complete, anatomically detailed, three-dimensional representations of the male and female human body. The current phase of the project is collecting transverse CT, MRI and cryosection images of representative

male and female cadavers at one millimeter intervals. The long-term goal of the Visible Human Project is to produce a system of knowledge structures that will transparently link visual knowledge forms to symbolic knowledge formats such as the names of body parts."

The [Digital Library Collection from the Library of Congress](#) is the beginnings of a National Digital Library which includes: the American Memory project, Special American Collections at the Library of Congress and Country Studies.

[The Institute for Advanced Technology in the Humanities](#) has [research reports](#) about computing in the humanities at the University of Virginia, their online journal [Postmodern Culture](#), [technical reports](#) and a forms based demonstration of the Institute's [Image Annotation Tool for Humanists](#).

The [IBM Digital Library](#) was an early commercial entry into the digital library arena. A major focus is on technical enforcement to copyright management.

Upcoming Digital Library Conferences

[Digital Libraries '98](#). The Third ACM International Conference on Digital Libraries. June 23-26, 1998. Pittsburgh, PA.

[International Summer School on the Digital Library 1998](#). The third International Summer School for librarians. August 16-21, 1998. Tilburg University, The Netherlands.

[ISIC 98](#). Information Seeking in Context: an International Conference on Information Needs, Seeking and Use in Different Contexts. August 13-15, 1998 in Sheffield, UK.

[ECDL '98](#). Second European Conference on Research and Advanced Technology for Digital Libraries. September 19-23, 1998 in Heraklion, Crete, Greece.

[CoLIS3](#). The Third International Conference on Concepts in Library and Information Science with the theme of Digital Libraries: Interdisciplinary Concepts, Challenges and Opportunities Inter-University Centre Dubrovnik (IUC) Dubrovnik, Croatia, May 23-26, 1999. The primary [CoLIS3 page is in Croatia](#), however the initial link is a much faster US based mirror (at least for those in the US).

[LIBRES: Conferences and Meetings](#) is an up to date list of conferences and meetings from the Library and Information Science Research Electronic Journal. The list has many items of interest to the digital library community.

Previous Digital Library Conferences

This archives digital library conferences which have information online, but not full text of the proceedings.

[ADL '98](#). Advances in Digital Libraries Conference. April 22-24, 1998. Fess Parker's Doubletree Resort Santa Barbara, California, USA.

[Digital Libraries Asia 98 Conference and Exhibition](#). The Digital Era: Implications, Challenges and Issues. March 17-20 1998, The Westin Stamford and Westin Plaza, Singapore.

[ECDL '97](#) First European Conference on Research and Advanced Technology for Digital Libraries. September 1-3 1997, Pisa, Italy.

[International Summer School on the Digital Library 1997](#). August 10-22, 1997. Tilburg University, The Netherlands.

[AI in Digital Libraries](#). Part of the International Joint Conference on Artificial Intelligence Workshop Series. August 23-29, 1997, Nagoya, Japan.

[Digital Libraries '97](#). The Second ACM International Conference on Digital Libraries. July 24-26, 1997. Philadelphia, PA.

[ELVIRA4](#). The 4th UK Digital Libraries Conference (Electronic Library and Visual Information Research.) May 6-8, 1997. Milton Keynes, UK.

[ADL '97](#). A Forum on Research and Technology Advances in Digital Libraries. May 7-9, 1997. Library of Congress, Washington, D.C.

[Visualizing Subject Access for 21st Century Information Resources](#) is the 34th Annual Clinic on Library Applications of Data Processing at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. March 2-4, 1997. Urbana, IL.

[ADL '96](#). Forum on Research and Technology Advances in Digital Libraries May 13-15, 1996, Washington, D.C.

[ELVIRA3](#). The UK Digital Libraries Conference. Third International Conference, Electronic Library and Visual Information Research. Hilton National Hotel, April 30-May 2, 1996, Milton Keynes, UK.

[ADL '95](#). Research and Technology Advances in Digital Libraries. May 15-19, 1995. McClean Hilton at Tysons Corner, VA.

[Digital Libraries Conference](#). Singapore Information Technology Institute. March 27-28, 1995. Raffles City Convention Centre, Singapore.

Previous Digital Library Related Conferences with Online Proceedings

The following conferences and workshops have made all, or at least a substantial selection, of the full text of the proceedings available online.

[Successes and Failures of Digital Libraries](#) is the 35th Annual GSLIS Clinic formerly known as the Annual Clinic on Library Applications of Data Processing at the Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign. March 22-24, 1998. Urbana, IL.

[ISDL'97](#). The International Symposium on Research, Development & Practice in Digital Libraries is sponsored by University of Library and Information Science. November 18-21, 1997. Tsukuba Science City, Japan.

[IEEE Metadata 97](#). The Second IEEE Metadata Conference. September 16-17, 1997, Silver Spring, Maryland.

[Beyond the Beginning: The Global Digital Library](#) an international conference organized by UKOLN on behalf of JISC, CNI, BLRIC, CAUSE and CAUL was held June 16-17, 1997 at The Queen Elizabeth II Conference Centre, London, UK.

[Information Technology Workshop](#) was held for the Goddard research community to learn about ASA sponsored activities in new information technologies. March 11-13, 1997. ASA Goddard Space Flight Center, Greenbelt, Maryland.

[Santa Fe Planning Workshop](#) on Distributed Knowledge Work Environments: Digital Libraries was held to discuss issues surrounding a follow on initiative to the Digital Libraries Initiative. March 9-11, 1997 Santa Fe, New Mexico.

[Allerton '96](#). Libraries, People and Change: A Research Forum on Digital Libraries. The 38th Allerton Institute of the Graduate School of Library and Information Science University of Illinois at Urbana-Champaign. October 27-29, 1996. Allerton Park, Monticello, Illinois.

[SIGIR-96 Workshop on Networked Information Retrieval](#). The workshop was held during the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. August 22, 1996. ETH, Zurich, Switzerland.

[Institute on Digital Library Development](#). July 15-19 and July 29-August 2, 1996 Berkeley, California.

[IATUL 1996](#). The International Association of Technological University Libraries. The overall theme of the conference will be "Networks, Networking and Implications for Digital Libraries." June 24-28, 1996 at the University of California, Irvine.

[OGDL II](#). Organizing the Global Digital Library II and Naming Conventions May 21-22, 1996. Library of Congress, Washington, D.C.

[IEEE Metadata 96](#). The First IEEE Metadata Conference. April 16-18, 1996, Silver Spring, Maryland.

[ACM DL'96](#). The First ACM International Conference on Digital Libraries. March 20-23, 1996, Bethesda, MD. The conference proceedings may be found at: [DL '96. Proceedings](#) of the 1st ACM international conference on Digital libraries. *Note: this is part of the [ACM Digital Library](#) and requires a subscription to access.*

[Social Aspects of Digital Libraries](#). February 16-17, 1996, University of California, Los Angeles.

[OGDL](#). Organizing the Global Digital Library Conference December 11, 1995. Library of Congress, Washington, D.C.

[ISDL'95](#). International Symposium on Digital Libraries 1995. August 22-25, 1995. Tsukuba Science City, Ibaraki 305, Japan.

[Digital Libraries '95](#) (DL'95). The Second International Conference on the Theory and Practice of Digital

Libraries, held June 11-13, 1995. Austin, Texas.

[Building the Digital Library: Content Issues](#). Proceedings of the Library of Congress Network Advisory Committee. June 4-6, 1995. Library of Congress, Washington, D.C.

[IITA Digital Libraries Workshop](#). Interoperability, Scaling and the Digital Libraries Research Agenda. May 18-19, 1995, Reston, Virginia.

[Allerton '95](#). How we do user-centered design and evaluation of Digital Libraries: A methodological forum. The 37th Allerton Institute conference of the Graduate School of Library and Information Science University of Illinois at Urbana-Champaign. October 29-31, 1995. Allerton Park, Monticello, Illinois.

[Information Gathering from Heterogeneous, Distributed Environments](#), the American Association for Artificial Intelligence (AAAI) Spring Symposium Series. March 27-29, 1995 Stanford University, Stanford, California.

[Seminar on Cataloging Digital Documents](#) October 12-14, 1994 sponsored by the University of Virginia Library, Charlottesville and the Library of Congress.

[Digital Libraries '94](#) (DL '94). The First Annual Conference on the Theory and Practice of Digital Libraries June 19-21, 1994. College Station, Texas.

[TREC-6](#) the sixth Text REtrieval Conference (TREC) held in Gaithersburg, Maryland, November 19-21, 1997. The articles are in Postscript.

[TREC-5](#) the fifth Text REtrieval Conference (TREC). November 20-22, 1996. Gaithersburg, Maryland. The articles are in Postscript.

[TREC-4](#), the fourth Text REtrieval Conference (TREC). November 1-3, 1995. Gaithersburg, Maryland. The articles are in Postscript.

[TREC-3](#), the third Text REtrieval Conference (TREC). November 2-4, 1994. Gaithersburg, Maryland. The articles are in Postscript.

[WWW6](#), the Sixth International World Wide Web Conference April 7-11, 1997, Santa Clara, California.

[WWW5](#), the Fifth International World Wide Web Conference. May 6-10, 1996, at CNIT-Paris La Défense, France.

[WWW3](#), the Third International World-Wide Web Conference: Technology, Tools and Applications April 10-14, 1995, Darmstadt, Germany.

[WWW2](#), the Second International World-Wide Web Conference: Mosaic and the Web. October 17-20, 1994, Chicago, IL.

[WWW1](#), the First International World-Wide Web Conference May 25-27, 1994, CERN, Geneva Switzerland.

Full Text of Other Digital Library Related Publications

These are pieces as well as collections that have been placed online in their full an unabbreviated form.

[D-Lib Magazine](#), an on-line, monthly magazine coordinated by CNRI and sponsored by DARPA on behalf of the IITA Working Group of the HPCC program, covers articles, news and commentary on advanced research and implementation projects in digital libraries.

[Buildings, books, and bytes](#) is the November 1996 by the Benton Foundation which reports on what library leaders and the public have to say about the future of libraries and communities in the digital age.

ERCIM - the European Research Consortium for Informatics and Mathematics has placed its [ERCIM News special theme on digital libraries](#) online. ERCIM News No.27 - October 1996.

The IEEE Computer Society's has placed the full text of related articles online for their [May 1996 theme issue of Computer on the US Digital Library Initiative](#).

[Solaris](#) is an annual review of research in information science and communications, including digital libraries from the Groupe interuniversitaire de recherche en sciences de l'information et de la communication (GIRSIC). The 1994, 1995 and 1996 are available in French with some English.

Many of the articles in the [SIGLINK Newsletter Special issue on Digital Libraries](#) are online. The articles are in a mix of HTML and Postscript. September, 1995 (Volume 4, Number 2).

An online edition of [Communications of the ACM - August 1995](#) Special Issue on Designing Hypermedia Applications.

The Association for Computing Machinery (ACM) has placed the full text of the Volume 38, No. 4 (April 1995) online for the [Communications of the ACM issue on Digital Libraries](#). *Note: this is part of the [ACM Digital Library](#) and requires a subscription to access.*

The [Digital Library Source Book](#), 1993, edited by Edward Fox. The articles are in Postscript and PDF.

Other Digital Library Related Resources

These sites contain contain well rounded and or unique selections of information and resources about digital libraries.

The [Berkeley Digital Library SunSITE](#) is dedicated to gathering and publishing information about digital library projects and other digital content. It will also provide a platform for digital research and development as well as promote discussions on topics related to digital libraries, museums and archives.

The International Federation of Library Associations and Institutions or (IFLA) maintains a set of references for [digital libraries resources and projects](#), [metadata resources](#), [cataloging and indexing of electronic resources](#) and [interlibrary loan, document delivery and resource sharing information](#). IFLA also runs a number of mailing lists including the [DIGLIB](#) mailing list.

[New Horizons in Scholarly Communication](#) maintained by the Librarians Association of the University

of California deals broadly with the use of new media in teaching and research, new publishing models and access issues. The section on access issues includes an [introduction to the digital library](#).

An [annotated bibliography of digital library related sources](#) maintained Steven Ketchpel contains a wide array of annotated entries along with rankings for relevance and suggestions for intended audience .

The [Digital Libraries Resource Page](#) maintained by Karin L. Trgovac.

References on [Building Digital Libraries](#) from TexShare.

[WWW Library Resources - Discussion Lists](#) maintained by Randy D. Ralph contains descriptions and subscription information for many mailing lists related to digital libraries.

[Pointers to national and international library projects](#) from the BELNET User Forum Workgroup on Libraries.

Digital Library Funding, Coordination and Policy Organizations

The following organizations all provide explicit support or help contribute on a coordination or policy level to digital library related projects.

The [National Science Foundation](#) is involved in funding and coordinating a large portion of digital library research in the United States. They have taken the lead role in funding the Digital Libraries Initiative (DLI).

The [Corporation for National Research Initiatives](#) (CNRI) "is a non-profit organization dedicated to formulating, planning and carrying out national-level research initiatives on the use of network-based information technology." Many of their projects are digital library related.

The [Digital Library Technology](#) project from the Information Sciences and Technology Branch Space Data and Computing Division NASA Goddard Space Flight Center. The DLT Project supports the development of new technologies to facilitate public access to NASA data via computer networks.

National Coordination Office for Computing, Information, and Communications (CCIC of NCO) formerly the National Coordination Office for High Performance Computing and Communications (HPCC) has made digital libraries a National Challenge Application in since 1993. "Blue Books" are annual reports presenting CCIC Program plans and accomplishments. Here are pointers to the relevant sections.

Digital Libraries in the Blue Book

CCIC: Computing, Information, and Communications Technologies for the 21st Century ([FY 1998 Blue Book](#))

HPCC: Advancing the Frontiers of Information Technology ([FY 1997 Blue Book](#))

HPCC: Foundation for America's Information Future ([FY 1996 Blue Book](#))

HPCC: Technology for the National Information Infrastructure ([FY 1995 Blue Book](#))

HPCC: Toward a National Information Infrastructure ([FY 1994 Blue Book](#))

The [Digital Library Federation](#) is an organization constructed from fifteen of the nation's largest research libraries and archives.

Intellectual Property

Intellectual property rights and intellectual property rights management systems will be key issues and components of future digital libraries.

The [Intellectual Property Center](#) contains daily news with coverage of patents, copyright, trademark, Internet law, etc.

The [Information Law Web](#) is a collection of links of people, place and things geared to helping people understanding their rights in terms of online information.

The [EFF Intellectual Property Online Archive](#) includes topics such as patents, trademarks and copyright contains a wide array of articles, legal documents and links to other resources in the area of intellectual property.

The [Online Law Library](#) contains many high quality references to legal materials including [journals dealing with intellectual property issues](#).

The [WWW Multimedia Law](#) site producers and publishers of multimedia are oriented to legal liabilities faced on a number of platforms, not necessarily the Internet.

Human Computer Interaction (HCI)

The importance of user interfaces and human-computer interaction in general should not be underestimated with regard to digital libraries. Major advances in usability will come from innovation in the interfaces and not the underlying databases or processing engines.

There are a number of good Human-Computer Interaction related sites on the web, one of these is the [Human-Computer Interaction Virtual Library](#) maintained by [Keith Instone](#).

Another is the [HCI resources](#) list maintained by [Mikael Ericsson](#).

As well as the [HCI Index](#) maintained by [Hans de Graaff](#).

The [ACM SIGCHI Home Page](#). SIGCHI is the ACM special interest group on Computer-Human Interaction. Conference proceedings from 1995 onward are available online. *Note: this is part of the [ACM Digital Library](#) and requires a subscription to access.*

[The HCI Bibliography](#) is a large and broad bibliographic database on Human-Computer Interaction.

Computer Supported Cooperative Work (CSCW)

Enhancement of collaborative and Cooperative forms of searching, communicating and creating are great advantages of the online medium and thus must be included into digital libraries.

A number of [CSCW](#) references including a [CSCW Bibliography](#), a large list of CSCW projects and products the [CSCW Yellow Pages](#) and [CSCW Related Links](#) have been compiled by [Michael Koch](#).

Contributors to the the USENet news group, [comp.groupware](#) have produced a number of FAQs which include the [comp.groupware FAQ hierarchy](#).

[ACM SIG GROUP](#) concentrates on applications which have a team or group focus.

The [WWW Collaboration Projects](#) is a well rounded comprehensive site for applications on the Web that support some type of collaboration.

Please reference this URL: --> <http://www.uiuc.edu/ph/www/bgross/dl/>

(C) Ben Gross 1995-1998

Last update 5/13/98

[Ben Gross](#) -- [Contacting me](#)

DLI - Carnegie Mellon:

- [Informedia](#)
- [NetBill](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



The Informedia Digital Video Library is a research initiative at Carnegie Mellon University funded by the NSF, DARPA, NASA and others that studies how multimedia digital libraries can be established and used. Informedia is building a multimedia library that will consist of over one thousand hours of digital video, audio, images, text and other related materials.

Informedia's digital video library is populated by automatically encoding, segmenting and indexing data. Research in the areas of speech recognition, image understanding and natural language processing supports the automatic preparation of diverse media for full-content and knowledge-based search and retrieval. Informedia is one of six [Digital Libraries Initiative](#) projects.

[Project Description](#)

[Sponsors and Partnerships](#)

[Publications & Reports](#)

[Project Team](#)

[Project News](#)

[NSF Digital Library Projects](#)

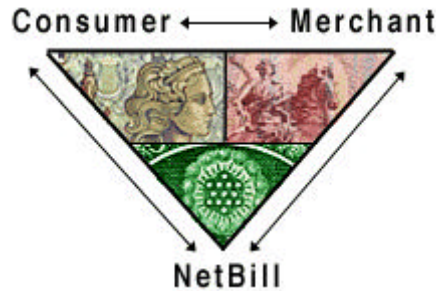
Copyright © 1997 Carnegie Mellon University
All Rights Reserved



The NetBill Project

- ◆ Overview
- ◆ News
- ◆ Publications
- ◆ Technical Partners
- ◆ Project Members
- ◆ Commerce Resources

A dependable, secure, and economical payment method for purchasing digital goods and services through the Internet.



The NetBill electronic commerce project at Carnegie Mellon's [Information Networking Institute](http://www.ini.cmu.edu/NETBILL/) is researching design issues of highly survivable and secure distributed transaction processing systems, as well as accounting and access control for digital libraries. NetBill is addressing these issues by developing the protocols and software to support network-based payment for goods and services over the Internet.

These protocols and software have been implemented in a test system, currently in its Alpha trial, on the Carnegie Mellon campus. This system enables consumers and merchants to communicate directly with each other, using NetBill to confirm and ensure security for all transactions.

We invite you to take a look at this test system at:

<http://www.netbill.com>

NetBill is publicly available to United States residents. For those not in the US, there is plenty of information about NetBill for you to explore.

For more information about the NetBill project, please explore this web site using the links on the left of each page.

If you require further information, please contact us at support@netbill.com



All contents copyright © 1995,1996,1997 Carnegie Mellon University.
All rights reserved.
Last revision: Fri Oct 10 11:54:34 EDT 1997

DLI - Stanford:

- [Home Page](#)
- [IEEE Computer article](#)
- [testbed development](#)
- [info finding](#)
- [user interfaces](#)
- [DLITE \(task env\)](#)
- [DLITE comps](#)
- [DLITE screens](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



Stanford University Digital Libraries Project

The Stanford Digital Libraries project is one participant in the 4-year, \$24 million Digital Library Initiative, started in 1994 and supported by the [NSF](#), [DARPA](#), and [NASA](#). In addition to the ties with the [five other universities](#) that are part of the project, Stanford also has a large number of [industrial partners](#). Each university project has a different angle of the total project, with Stanford focusing on **interoperability**.

Our collection is primarily computing literature. However, we also have a strong focus on networked information sources, meaning that the vast array of topics found on the World Wide Web are accessible through our project as well. At the heart of the project is the [testbed](#) running [the "InfoBus" protocol](#), which provides a uniform way to access a variety of services and information sources through "proxies" acting as interpreters between the InfoBus protocol and the native protocol.

With the InfoBus protocol running under the hood, a variety of user level applications provide powerful ways to [find information](#), using cutting-edge [user interfaces](#) for direct manipulation or through [Agent technology](#). A second area of focus for the Stanford Digital Library Project is the [legal and economic issues](#) of a networked environment.

[PROJECTS](#)

See the entire list, or jump directly to projects related to [information finding](#), [user interfaces](#), [legal and economic issues](#), [the testbed](#), or [agents](#).

[DOCUMENTS](#)

A collection of introductory information, and our [publications](#), our [working papers](#), our [presentations](#), our [mailing archives](#), and our [project reports](#).

[INFO RESOURCES](#)

A collection of pointers to digital library-related resources, both at Stanford and elsewhere.

[SEMINARS](#)

A schedule of our weekly Digital Library seminar, which meets Mondays at 4:30 in Gates B08.

[SOFTWARE](#)

A collection of software developed for and used by the Stanford Digital Library project.

[PEOPLE](#)

A list of the Stanford faculty, staff, and student participants and industrial partners.



February 4, 1998

Andreas Paepcke's [overview presentation](#) gives a good "big picture" view of the project. Gerard Rodriguez put together an [introductory page about using the InfoBus](#). There's also a [presentation](#) by Andreas Paepcke on how to build a proxy. The [slides from the most recent DLI all-project meeting](#) are online. Plus, congratulations to Martin Roscheisen, our newest Ph.D. His dissertation is entitled [A Network-Centric Design for Relationship-Based Rights Management](#).



Quick Tabs to Projects: [GLOSS](#) -- [Query Translation](#) -- [SenseMaker](#) -- [FAB](#) -- [STARTS](#) -- [Grassroots](#) -- [SONIA](#) -- [BackRub](#) -- [Metadata Architecture](#) -- [ComMentor](#) -- [R-Manage](#) -- [InterPay](#) -- Distributed Transactions -- [InterOp Protocol](#) -- Z Server -- Proxy Generator -- [Infobus Socket Interface](#) -- [JYLU](#) -- [DLITE](#) -- [Audio HTML Access](#) -- WebWriter -- [Interbib](#) -- [SCAM](#) -- [COPS](#)

From *Computer* theme issue on the US Digital Library Initiative, May 1996

Using Distributed Objects for Digital Library Interoperability

Andreas Paepcke, Steve B. Cousins, Hector Garcia-Molina, Scott W. Hassan, Steven P. Ketchpel, Martin Röscheisen, and Terry Winograd, *Stanford University*

Distributed object technology can provide interoperability among emerging digital library services. This project uses CORBA objects as wrappers to handle differences in service interaction models.

Information repositories are just one of many services tomorrow's digital libraries might offer. Other services include automated news summarization, trend analysis across news repositories, and copyright-related facilities. Traditional library services such as archiving and collection building will continue to be relevant as well. Archiving issues in the digital world include, for example, dangling hyperlinks and storage media obsolescence.

This distributed collection of services has the potential to be enormously helpful in performing information-intensive tasks. It could also turn such tasks into confusing, frustrating annoyances by forcing programmers and users to learn many interfaces and by confronting users with the bewildering details of fee-based services that were previously only accessible to professional librarians.

The Stanford Digital Library project has undertaken work to address the problem of interoperability, which is particularly important because standardization efforts are lagging behind the development of digital library services. We used CORBA,[\[1\]](#) the distributed-object standard developed by the Object Management Group, to implement information-access and payment protocols. These protocols are designed to provide the interface uniformity necessary for interoperability, while leaving implementers a large amount of leeway to optimize performance and to provide choices in service performance profiles.

We have implemented an experimental version of our information-access protocol for Knight-Ridder's Dialog information service, various World Wide Web information sources, Z39.50 servers (one of the best-known information-access protocols),[\[2\]](#) Oracle's ConText summarization tool, and others. Our implementation is based on Xerox PARC's ILU (InterLanguage Unification) facility, a public-domain implementation of CORBA. It is supported on common platforms, such as Microsoft Windows 3.1 and NT, Linux, and the Unix implementation of Sun Microsystems, IBM,

Hewlett-Packard, and Silicon Graphics. Language bindings include C, C++, Common Lisp, Python, Modula-3, and Java. We are using several vendor platforms and languages in our experiments. The availability of ILU helps our experimentation with the wider community.

Our initial experience indicates that a distributed object framework--and our access protocol in particular--do give clients and servers the flexibility to manage their communication and processing resources effectively. Distributed objects let our protocols access existing services without requiring changes in the services. The [sidebar](#) describes the broader focus of the Stanford project.

Distributed object technology

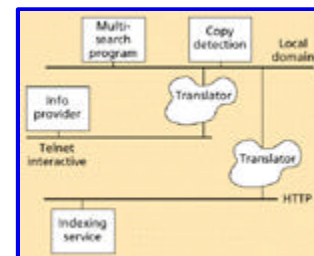
In an ideal world, clients and service providers that are part of a digital library would be created independently, on the basis of implementation choices the respective consumers and providers deemed appropriate. Then everyone would plug their components into a virtual software bus that would take care of all the protocol-level interoperability issues. Within this *information bus* (which we call *InfoBus*), library services would transparently translate formats, broker services, and support financial transactions. If all services conformed to one standard, the developers of digital libraries could easily realize this vision.

Unfortunately, protocol convergence has not occurred, even in the long-standing area of information retrieval. An overly simple solution would call for cross-translations among all standards. This would be a formidable effort. Distributed object technology may help achieve the long-term goal of an InfoBus without requiring all participants to agree on a single standard mode of interaction.

Interoperating across protocol domains

To explain how this vision might be achieved, we start with a very simple example. Figure 1 shows three protocol domains. The first domain depicted, the local domain, is a local network used by an information-services provider such as a company, a university, or even an individual. The second domain employs the Telnet protocol, in which clients log in to remote machines. HTTP, the protocol used for the WWW, is shown as the third domain. Each additional protocol, such as Z39.50, introduces another domain.

Figure 1. *Interoperating across protocol domains.*



All the domains are populated with services accessible through their respective protocols. The service-interaction protocols in the local domain are under local control. The Dialog information service is an example of a Telnet-based information provider. The WebCrawler, which indexes documents on the WWW and returns their URLs in response to queries, is an example of an HTTP-based service.

Because information repositories are the best-known digital library service, we will use Dialog and WebCrawler as examples. We anticipate that many services will eventually conform to some of the emerging standards, such as HTTP, Z39.50, and SQL, or to standards yet to be developed. We use Dialog's current, minimally standardized, human-oriented teletype interface to illustrate the breadth of diversity that remains today.

In Figure 1, the local domain includes a multisearch program. This kind of program accepts a query and multiplexes it to several information sources. In this example, it also uses a copy-detection service to check the retrieved documents for substantial overlap with a database of other documents and eliminates near-duplicates. This kind of service illustrates why interoperability is a base requirement for digital libraries. Without an interoperability infrastructure, the multisearch program would be very cumbersome to write. The programmer would have to learn the interaction models and search languages of both Dialog and WebCrawler. To avoid this, two translators are needed to link the local domain to the two remote ones.

Translators

Figure 2 shows a very simplified view of interactions with both Dialog and WebCrawler.

Figure 2. *Unification of simplified service-interaction models.*



Dialog presents a teletype interface, through which the user first follows a standard login sequence (**P**lease **l**ogon:), then selects one of the many databases offered through Dialog (**b**egin 245). Using a proprietary query language (**s**elect **L**ibrary/**t**i), the user searches the database, examines the results, and terminates the session (**l**ogout). On the left is one possible abstraction of this process: An **open session** operation is followed by **open database**, **search**, and **quit** operations. Of course, this abstraction would be more elaborate for a full-scale system, possibly including parts of the Z39.50

protocol or variants of other related resources.[\[3\]](#)

At first, the WebCrawler model looks very different from Dialog. At WebCrawler's home page, the user clicks to open a search form, fills it out, views the results, and leaves the home page. Yet the abstraction of this interaction model is the same as Dialog's.

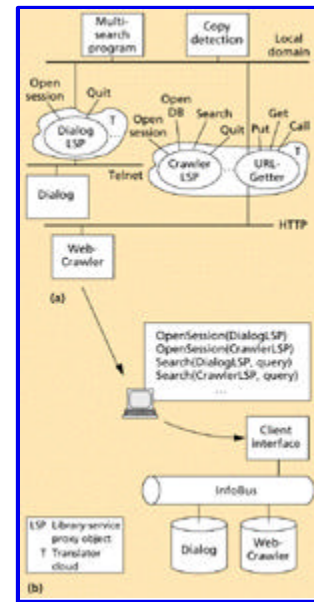
If we could program an interface that presented this common abstraction, it would be much easier to write a multisearch program. Object technology is ideally suited for this.

Objects for interface unification

Object orientation's polymorphism can be used to present a unified interface like the abstraction in Figure 2. For example, we created a *library-service proxy* (LSP) object. Method calls on an LSP object invoke each interface element (`open session`, `open database`, and so on), and the method performs the appropriate operation on the corresponding service. For example, the `open session` method for a Dialog LSP starts a telnet session and logs into the Dialog service. The implementation of the same method for the WebCrawler LSP contacts an HTTP demon with the proper URL.

Figure 3a shows how LSPs can be used as the building blocks for the translators in Figure 2. The translator clouds are filled with LSPs, each of which represents one service. A common interface thus makes two very different services accessible from the local domain. Figure 3b shows the effect of this arrangement on a digital library programmer. The LSPs and their polymorphic implementations act as a wrapper, providing the beginnings of an InfoBus abstraction. The URL-Getter object in Figure 3a offers a pure bridge functionality that can suffice when the development of a full LSP is not justified.

Figure 3. (a) Service proxy objects implement translation; (b) programmers experience the illusion of an InfoBus.



Requirements for information flow

Object technology can help provide extensible interfaces for information access. However, the implementation of every LSP method requires the resolution of several important information-flow issues. Consider **search** methods. Some services implement a single-interaction model: The client calls the **search** method once (including a query as a parameter) and waits; when the server has assembled the result set, it returns the complete answer. Other services implement a piecemeal method: The user receives a steady stream of information that slowly builds up, rather than a complete set after a longer wait. This gives the perception of faster response time and lets users manipulate the early results while the later ones are still being transmitted. (An example of this can be observed in some Web browsers when pictures are being loaded. The picture appears first in coarse granularity and is refined slowly as more information arrives.)

Because we cannot dictate how clients and services operate, the LSP search method must be as general-purpose as possible. A client that wants to wait for complete results should be able to do that. If the information service (or its proxy) can give piecemeal information, and the client can handle it, then the **search** method should support that too.

There are other dimensions along which we would like to have flexibility:

- *Caching.* It should be possible to cache the set of search results or some of the information for future use.
- *Processing.* It should be possible to off-load related processing tasks to other machines, including the client computer.
- *Messaging.* It should be possible to minimize the number of message exchanges in the event we have to operate across a slow link.
- *Instantiation.* It should be possible to instantiate and materialize objects

(documents) at various times and locations. Instantiation creates an empty object; materialization fills it with information from the provider. When and where these activities occur can affect efficiency. A prefetching strategy, for example, would materialize documents at the client side before their contents are requested while an on-demand strategy would wait until an application asks for the document's contents.

This aspect of protocol design arises in a distributed object environment because these systems generally package documents into objects as well. The alternative would be to maintain documents as strings. One advantage of the object approach is that document structure, which is painstakingly provided by repositories, can be preserved and accessed more easily. Methods on document objects (**title**, **author**, **abstract**, for example) can be used to extract the corresponding document pieces. For example, to search SGML documents through method calls, client programs do not need to contain code for parsing out pieces of marked-up text. This presents a clean interface to programmers, but it raises the question of when and where document objects are instantiated and materialized. A simple-minded protocol would have the LSP instantiate and materialize all retrieved documents as objects at the remote site. The client would then access these documents through remote method calls. But this would be wasteful because users often discard query results as they narrow their search, and because local method calls are cheaper than remote ones. A protocol that takes advantage of an object-based architecture should allow implementations to determine when a document is needed, to shift its raw information to the site where it will be used most, and then to cast it into an object.

Existing information-access protocols do not provide this new level of flexibility. For example, Z39.50 requires that result sets be maintained on the server and delivered to clients on request. The protocol we are developing uses a distributed object infrastructure to provide such flexibility.

Sketch of sample protocol

Our protocol, developed in cooperation with researchers at the University of Illinois at Urbana-Champaign, the University of Michigan, and the University of California at Santa Barbara, provides a uniform search interface and preserves flexibility. We have implemented several variants of the protocol in our testbed, and we plan to use it initially to exchange information between those universities and Stanford.

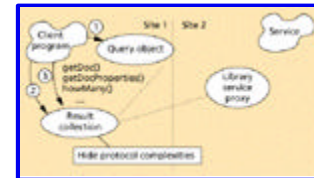
Figure 4 shows how we present the querying process to client programmers. The process has three steps:

1. Create a Query object that contains the query string and any other search details. The query string could be of a form native to one source or it could be of a more standard form that is later translated to a native form--we are not concerned with this aspect of interoperability here.
2. Create a local Result Collection object, specifying the Query object and

the intended LSP.

- From now on, the client program interacts with this Result Collection, as if it was immediately filled with document objects. For example, the client may invoke a **howMany** method to find out how many documents are in the result or a **getDoc** method to fetch a particular document. When these methods are called, the Result Collection object may or may not have the necessary information, so the client calls may be blocked.

Figure 4. *Clients program to a very simple interface.*



Before or after the client tries to fetch documents, the Result Collection retrieves them from the LSP. The protocol for this, which has four steps, is illustrated in Figure 5.

Figure 5. *Moving information.*



- Client collection asynchronously requests query execution. Here, "client collection" refers to the Result Collection object on the client side; the server side may choose to create a Server Collection object to aid processing. The client collection initiates the query using an asynchronous LSP method invocation, passing its own object identifier as the return address for the query results and indicating how many result documents it wants to access initially. As in Z39.50, the LSP may be requested to return selected "teaser" fields, such as title, author, or cost. This allows the earliest possible delivery of some useful information, without having to transfer the entire document. In contrast to Z39.50, the client does not need to wait for the server to complete its result collection--the method call is asynchronous.

In response to the call, the LSP causes the query to be executed on its associated service. When it receives the results, it may delegate further handling of requests to its own Server Collection object. If the service is session-based, the Server Collection object can either maintain a session with the remote service in anticipation of requests for full documents or it can pull the

documents out of the service and cache them. Because distributed objects may be created anywhere, the Server Collection object may be located on a different machine, freeing the LSP machine to handle more requests. Or the LSP can decide not to create a Server Collection object and manage the follow-up requests itself.

- The service asynchronously delivers document references, either as they arrive or all in one method call. Depending on whether the delivery of results was delegated or not, this step is executed by the LSP or by the Server Collection object. The service delivers the number of documents found, some or all of the teasers, and document references so the client can obtain the complete content of documents. This step can be repeated many times as results are accumulated, so the implementation can deliver access to documents before it has found all the results. The key elements of this step are as follows:
- Method calls to the client collection are asynchronous and include contact information for the client to use when requesting access to more documents than were indicated in the original request. This is how the delegation to server collection objects is accomplished.
- When the LSP or server collection returns teasers for a document, it includes the document's access capability, which describes how the full document can be found. Each capability is made up of one or more access options, each specifying one alternative way to get the document.
- The server side objects send information to the client side via "callback" methods on the client collection. Each of these callbacks includes the object ID of the server side object to contact for additional information. Thus, at any time, a server object (like the LSP) can delegate the responsibility of future interactions with this client collection to other objects (like the server collection). Similarly, each document access option received by the client collection contains the ID of the server object to contact to obtain that document.
- Client collection repeatedly requests more document references (optional).
- Client collection asynchronously requests document contents, using the references of steps 2 and 3 (optional). If necessary, object documents are instantiated on the server or client side.

Each option in an access capability contains the ID of the object to contact to get the document, plus a "cookie" that identifies the document. From the client's point of view, a cookie is simply an uninterpreted bit string that must be given to the server object from which the document is being fetched. From the server object's point of view, the cookie contains information necessary for accessing the document to deliver. For example, a cookie could be an index into a memory cache where the document was placed earlier; it could be a file name for a local file containing the document; it could be a call number in some information retrieval system; or it could be a permanent document handle.[\[4\]](#)

The reason for allowing multiple access options within a capability is that the mechanisms for getting a document may vary over time. For example, consider a Dialog search. While the LSP (or the server collection) maintains an open session with the service, it can refer to a particular document by an index into a Dialog-generated result set. Thus, one possible cookie for an access option would be the result set identifier and the index. However, once a session with Dialog is terminated, this access mechanism no longer works. Instead, the document's unique record identifier needs to be used as the cookie. By providing both options in the access capability, the LSP is free to serve document contents quickly while sessions with the service are open, but to close down sessions without losing the ability to deliver documents for which it handed out access capabilities. The holder of an access capability tries the easier options first. As they fail, it tries more expensive ones.

As the Client Collection object receives document-access capabilities, it can wait until the client program actually requests them. If it instantiates document objects, it can fill in any teaser fields it received but wait to materialize the rest on demand. Alternatively, it can begin to materialize immediately in anticipation of impending demand. The decision may, for example, be made dependent on statistical user behavior or on an evaluation of the likelihood that the remote site will crash or disconnect.

If the client result collection needs teasers and access capabilities for more documents than it initially requested in step 1 in Figure 5, it initiates step 3, using the contact information received in step 2. The client collection does not know if this request for additional information is handled by the LSP, the Server Collection, or any other helper object. The result of this request is another round of step 2a/b activity that delivers the teasers and capabilities, as shown in Figure 5.

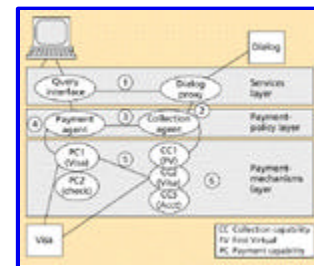
Fee-for-service as an interoperability problem

This sample protocol provides only base-level functionality for searching diverse information services, only one of the many aspects of interoperability. We are developing an architecture to address other interoperability problems, such as fee-based services. Several on-line payment mechanisms have been suggested, and some are beginning to be deployed.[5] To users of digital libraries that include some fee-based services, the differences in payment scheme are one more potential source of frustration. Our *InterPay* architecture is designed to ease this problem.[6] We have implemented a prototype that accesses several services, each with a different payment scheme.

Layered InterPay architecture

Figure 6 shows our three-layer InterPay architecture:

Figure 6. *Interactions among InterPay components.*



- The *services* layer provides all the task-related interactions with users. For information services, these interactions include login, query submission, result transmission, and so on--all the activities supported by the protocol we just described.
- The *payment-policy* layer controls and enforces payment-related preferences and rules. The policies are implemented by payment agents on the payer side and collection agents on the payee side. For example, a

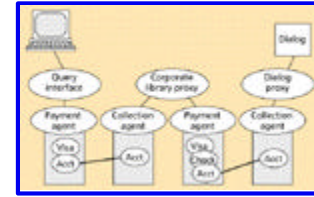
payment agent may enforce a policy such as "pay charges of \$1 or less without conferring with the human operator, but notify the operator when total charges exceed \$30." On the service side, a collection agent may include rules about delayed payment for trusted clients or limitations on the use of particular payment mechanisms.

- The *payment-mechanisms* layer comprises elements that implement the mechanics of particular payment schemes. On the payer side, these are *payment capabilities*; on the payee side, they are *collection capabilities*. Each payment capability is programmed to interact with one particular payment agency or payment scheme. Each collection capability is programmed to verify receipts or otherwise interact with one agency or scheme. New payment capabilities can easily be added to the system because all elements of InterPay are objects. A new payment scheme is added by implementing a payment and collection capability pair that may even be installed and removed dynamically.

Figure 6 also shows how InterPay components interact in a typical transaction:

1. Set up the session and make a request. The client entity and service entity have an interaction, such as the submission of a query. During the interaction, the client's payment agent is included as a parameter. Depending on the service, charges might be initiated immediately, after a search, or at the end of a session.
2. Initiate a charge. Once the service decides to charge, it delegates this task to its collection agent.
3. Send an invoice. The collection agent sends the payment agent an invoice that identifies the service, the charge, and the acceptable payment mechanisms.
4. Validate the invoice and agree on a payment mechanism. The payment agent verifies the legitimacy of the charge and picks one of the payment mechanisms.
5. Initiate the fund transfer. The payment agent delegates the mechanics of payment to the proper payment capability. The payment capability interacts with the respective financial service and the server-side collection capability to transfer the funds and a receipt. In the case of an account-based service, the currency tendered could simply be the user's account number.
6. Verify the payment and complete the transaction. The collection capability verifies payment and notifies the collection agent, which in turn notifies the LSP, which releases the information to the client.

One activity InterPay needs to accommodate is payment through third parties. For example, research libraries generally have bulk discount accounts at commercial information providers. When patrons of the library's local community access these providers, they do it under the library's bulk contract, with expenses sometimes billed to the patron's department. Figure 7 sketches an example of how third-party payment is accomplished in InterPay.

Figure 7. Example of a third-party payment.

Conclusion

Distributed object technology helps us deal with some of the interoperability problems that arise in a digital library comprising numerous independent services, each potentially presenting a different interface and interaction model. And we demonstrated how this technology can be used to help with the specific heterogeneity problem of multiple on-line payment schemes.

References

1. *The Common Object Request Broker: Architecture and Specification*, Object Management Group, Framingham, Mass., 1993.
2. *Information Retrieval: Application Service Definition and Protocol Specification*, ANSI/NISO, Bethesda, Md., 1994.
3. R. Rao, B. Janssen, and A. Rajaraman, *GAIA Tech. Overview*, tech. report, Xerox PARC, Palo Alto, Calif., 1994.
4. R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services," Tech. Report cnri.dlib/tn95-01, Reston, Va., 1995; <http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>.
5. D. Chaum, "Achieving Electronic Privacy," *Scientific American*, Aug. 1992, pp. 96-101; <http://www.digicash.com/publish/sciam.html>.
6. S. Cousins et al., "InterPay: Managing Multiple Payment Mechanisms in Digital Libraries," *Proc. 2nd Ann. Conf. Theory and Practice of Digital Libraries*, Hypermedia Research Laboratory, College Station, Tex., 1995, pp. 9-17; <http://diglib.stanford.edu/diglib/pub/reports/cousins-dl95.ps>.

Andreas Paepcke is a senior research scientist at Stanford University and director of the Digital Library Project. While at Hewlett-Packard Laboratories, he designed and implemented one of the early persistent object systems and an object view over a large collection of text databases. At Xerox PARC he participated in the development of a tutorial on open implementations. His current research interests include object-oriented programming, open implementations, and metaobject protocols applied to problems of information access. Paepcke received a BA and MS in applied mathematics from Harvard University and a PhD in computer science from the University of Karlsruhe, Germany.

Steve B. Cousins is a PhD candidate in computer science at Stanford University, in the area of user interfaces to digital libraries. Previously he was a research associate in the medical informatics laboratory at Washington

University. Cousins received a BS and an MS in computer science from Washington University.

Hector Garcia-Molina is professor of computer science and electrical engineering at Stanford University and one of the principal investigators of the Digital Library project. He was previously on the faculty of the computer science department at Princeton University. His research interests include distributed computing and database systems. Garcia-Molina received a BS in electrical engineering from the Instituto Tecnológico de Monterrey, Mexico, and an MS in electrical engineering and a PhD in computer science from Stanford University.

Scott W. Hassan is a designer and implementer for the Stanford Digital Library project. His research interests are using distributed object technologies, hypermedia, and wide-area computer networks as infrastructure for future digital library systems. Hassan received a BS in computer science from State University of New York at Buffalo.

Steven P. Ketchpel is a PhD candidate in computer science at Stanford University. His research interests are distributed artificial intelligence and electronic commerce. Ketchpel received a BA in computer science from Harvard University and an MS in computer science from Stanford University.

Martin Röscheisen is a PhD candidate in computer science at Stanford University. He is currently working on content and access control, privacy, and intellectual property issues. Röscheisen received an MS in computer science from Munich Technical University and Stanford University.

Terry Winograd is professor of computer science at Stanford University, where he directs the Project on People, Computers, and Design, and the teaching and research program on Human-Computer Interaction Design. He is also one of the principal investigators in the Digital Library project. He has done extensive research and writing on the design of human-computer interaction. His early research on natural-language understanding by computers was a milestone in artificial intelligence, and he has written two books and numerous articles on that topic. Winograd received a BS in mathematics from The Colorado College and a PhD in applied mathematics from MIT. He is on the national board of Computer Professionals for Social Responsibility, of which he is a founding member and past president, on the national advisory board of the Association for Software Design, and on the editorial board of several journals.

Address questions about this article to the authors at Stanford University, Gates Information Science, Rm. 426, Stanford, CA 94305; paepcke@cs.stanford.edu.

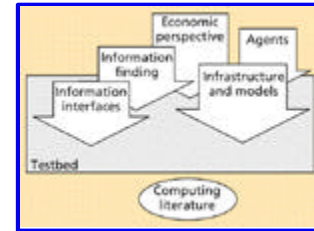
Sidebar

Stanford Digital Library Project

[Return to the main text](#)

Our digital library testbed will comprise a variety of computing literature sources, including Knight-Ridder's Dialog service, MIT Press, the ACM, the World Wide Web, and Stanford's libraries. Figure A shows the five areas that are driving the development of the testbed.

Figure A. *Stanford's approach to digital library development.*



The research on information interfaces seeks to help users interact with information in diverse formats and with various interaction models. Work in this thrust also explores uses of digital libraries as places for users to communicate about documents. For example, we have built the prototype of a wide-area annotation service.[\[1\]](#) It allows users to annotate pages on the WWW without modifying the original documents. Annotations are organized into sets, each with its own permission facility. Annotation sets may be located on servers other than those housing the documents with which the annotations are associated. Users may choose to view documents with no annotations, or with annotations from any of the sets they have permission to access. The many uses of this facility include independent product reviews and document content ratings: Users can view the ratings produced by the organization they happen to trust and rely on for guidance.

The second thrust of the project is concerned with technologies for locating appropriate library services and relevant information. For example, we have prototyped [Gloss](#), (Glossary-of-Servers Server) a service that efficiently maintains enough meta-information about a set of repositories that it can point users to the most promising sources for a particular query.[\[2\]](#) The [SIFT](#) (Stanford Information Filtering Tool) service is a prototype that explores efficient algorithms for matching large numbers of user-interest profiles with large numbers of documents.[\[3\]](#) Other efforts address the problem of query integration across multiple services.

Technologies supporting the evolving economic aspects of digital libraries are at the core of the third project thrust. Our SCAM and COPS (Copyright Protection System) efforts develop algorithms and a prototype for the efficient comparison of a text document against a large number of reference documents to detect partial overlap.[\[4\]](#) This service can be used to protect authors against illegal use of their intellectual property. Another effort in this third thrust is the

development of an architecture to manage interaction with the many emerging payment schemes. This InterPay mechanism is described in the main text.

The fourth thrust is developing models and a supporting infrastructure for the interaction with documents and services. These models form the basis for the protocols and architecture of our testbed. They include the models for meta-information about documents and repositories, to be used to search and visualize results. They also include protocols for the effective use of client-server models when potentially large amounts of information need to be moved among sites. The access protocol described in the main text is part of this effort.

The fifth thrust, finally, examines how agent technology can be employed to help operations throughout the system. We use very simple agent technology to help monitor on-line payment transactions. More substantial agent technologies are being used to retrieve information from the WWW on the basis of user-interest profiles that are successively refined.[\[5\]](#)

All five thrusts of the Stanford Digital Library project's work leave room for a wide variety of future work, some of which is currently in preliminary stages. At the user-interface level we are working on the problem of interactively configuring the use of library services to accomplish a task, and of reusing and sharing the results of such efforts. In the information-finding thrust, current work focuses on the problem of users' needing to query multiple services for the same information, without having to contend with disparate query languages and result schemata. In the area of support for economic activity, problems of security and privacy are being considered. In the infrastructure thrust, we continue to develop protocols that allow highly flexible distribution of information among machines, while providing satisfactory response time. Agent work is being pursued in the area of profile-based information filters.

References

1. M. Röscheisen, C. Mogensen, and T. Winograd, "Shared Web Annotations As A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples," tech. report, Stanford University, 1995;
<http://www.diglib.stanford.edu/diglib/pub/reports/commentor.html>.
2. L. Gravano, H. Garcia-Molina, and A. Tomasic, "The Effectiveness of [GLOSS](#) for the Text-Database Discovery Problem," *Proc. SIGMod Conf.*, ACM Press, New York, 1994, pp. 126-137;
<http://www-db.stanford.edu/pub/gravano/1994/stan.cs.tn.93.002.sigmod94.ps>.
3. T.W. Yan and H. Garcia-Molina, "[SIFT](#)--A Tool for Wide-Area Information Dissemination," *Proc. Usenix Tech. Conf.*, Usenix, Berkeley, Calif., 1995, pp. 177-186.
4. N. Shivakumar and H. Garcia-Molina, "SCAM: A Copy Detection Mechanism for Digital Documents," *Proc. Second Annual Conf. Theory and Practice of Digital Libraries*, Hypermedia Research Laboratory, College Station, Tex., 1995, pp. 155-163.

5. M. Balabanovic, Y. Shoham, and Y. Yun, "An Adaptive Agent for Automated Web Browsing," *J. Visual Communication and Image Representation*, Dec. 1995.

[Return to the main text](#)

Guest Editors' Introduction

University of Illinois

University of California at Berkeley

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

[Computer](#) | [Computer Society home page](#)

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.

Stanford Digital Library Testbed Development

Department of Computer Science
Stanford University
Stanford, CA



What is the Stanford Digital Library Testbed?

The Stanford Digital Library testbed is our platform for experimentation with interoperation among online services. Our basic approach is to use **distributed objects** to allow integrated access to heterogenous services across networks. The distributed approach allows the interaction of processes on different machines, with different architectures, implemented in different languages. We use **CORBA** to provide communication between remote processes. In particular, we use Xerox PARC's **ILU**, a free implementation of a CORBA superset. It offers language bindings for C++, C, CommonLisp, Python and Modula-3. We use the interpreted, object-oriented language Python for most of our development work.

For more information, see:

CORBA

- Information from the [OMG](#), including a [Manual](#)
- [Common Object Services](#) developed at Stanford

ILU

- [Xerox PARC's ILU Home Page](#)
- The current [ILU Manual](#)
- Information about the [Stanford installation](#)
- A [technical performance evaluation](#) of ILU, HTTP, and basic TCP

Python

- [The Python Language Home Page](#)
- Information about the [Stanford installation](#)

What Protocol does the Testbed Use?

We have developed the **Digital Library Interoperation Protocol (DLIOP)** for information access and retrieval. It is an asynchronous protocol, providing robustness in the face of network or server outages. Moreover, it also gives the programmer a high degree of control over where and when information objects are materialized, affecting tradeoffs of space and cost vs. time. This protocol has been adopted by other participants of the Digital Library Initiative, including University of Michigan and University of California at Santa Barbara.

For more information, see:

- A PostScript version of the [Full Protocol Specification](#)
 - A PostScript version of a [presentation describing DLIOP](#)
 - [A full CORBA IDL specification of DLIOP](#), or the [specification in ILU's ISL](#)
-

How Can I Use the Stanford Testbed?

Even if you're not local to Stanford, there are two simple ways of accessing the InfoBus from a remote site. Both use the DLIOP protocol. [IBClient](#), the first method, accesses the InfoBus through ILU calls and is therefore a full-functionality client. It requires client sites to have ILU or another CORBA implementation installed. The second alternative, the [InfoBus Socket](#) delivers the DLIOP calls via an ASCII stream over a socket. It does not require a CORBA implementation, but it is of limited functionality.

We include [code for the IBClient](#) and [code for the InfoBus Socket](#). The IBClient example is written in Python and thus requires your machine to have [Python](#) installed. If you do not have Python, you can use the example to build your own client in the language of your choice. The example represents a minimum, bare-bones client-side implementation. See the DLIOP documentation for additional facilities that can be added to clients. As an example, first download all of the files from the IBClient directory (the subdirectory CVS is not required). Then, try typing `ibclient.py WebCrawler 'digital library'` to see the titles of WebCrawler searches for those keywords. The InfoBus Socket is written in C++. It mimics the DLIOP calls in syntax, but it delivers them through UNIX sockets. See the [description of how it works](#).

We also have CORBA interface specification files for the DLIOP protocol.

- [IDLInterchange.idl](#)
 - [IDLInterchange.isl](#)
-

Information of Interest to local Stanford Developers

- [ILU -- Our installation and examples.](#)
 - [Python -- Our installation of Python programming language](#)
 - [CVS -- Our use of CVS in the testbed.](#)
 - [Various Manuals \(CVS, Python, ILU\)](#)
-



Digital Libraries Webmaster



Information Finding Projects in the Stanford Digital Library

One of the major research thrusts of the Stanford Digital Library project is helping users to find information. We have initiated a number of projects in this area, most related to our over-arching theme of interoperability. We have looked at ways that search tools can be used across multiple sources that use different syntaxes or languages. We have also looked at tools to provide statistical or collaborative filtering to locate relevant articles.

FAB

FAB is an adaptive multi-agent information retrieval system which finds interesting pages on the web.

"[An Adaptive Agent for Automated Web Browsing](#)"

- [Marko Balabanovic](#)

GLOSS

The Glossary Server of Servers (GLOSS) project is designed to locate relevant information sources for your query.

"[Generalizing GLOSS to Vector-Space Databases and Broker Hierarchies](#)"

- [Luis Gravano](#)

Query Translator

Databases have different query syntax and different capabilities, even for simple Boolean queries. Translation allows a single query to be mapped into the native format appropriate for each database.

- [Chen-Chuan K. Chang](#)

SenseMaker

SenseMaker helps users iteratively reformulate their information needs through multi-dimensional organizing and active gathering of search results.

"[SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests](#)"

- [Michelle Q Wang Baldonado](#)
-

Grassroots

Groupware for information finding, combines mail, news, and web in a single environment with distribution lists

" [Grassroots: A System Providing a Uniform Framework for Communicating, Structuring, Sharing Information, and Organizing People](#)"

- [Kenichi Kamiya](#)
 - [Martin Röscheisen](#)
-

The Stanford Digital Library Metadata Architecture

Services need to provide

- metadata about their offerings to help users decide when they should be invoked
- protocol metadata to figure out how they should be invoked, and
- collection metadata for what they should be invoked upon.

The metadata architecture provides a system organization to provide these metadata in a uniform, scalable way.

Metadata for Digital Libraries: Architecture and Design Rationale

- [Michelle Q Wang Baldonado](#)
 - [Chen-Chuan K. Chang](#)
 - [Luis Gravano](#)
 - [Andreas Paepcke](#)
-

STARTS: Stanford Protocol Proposal for Internet Retrieval and Search

A set of informal standards negotiated among the major search vendors and users to facilitate interoperation.

- [Chen-Chuan K. Chang](#)
 - [Hector Garcia-Molina](#)
 - [Luis Gravano](#)
 - [Andreas Paepcke](#)
-

Machine Learning for Information Retrieval

Statistical AI techniques allow the extraction of minimal sets of meaningful search terms

"[Toward Optimal Feature Selection](#)"

- [Mehran Sahami](#)
 - [Daphne Koller](#)
-

[BackRub](#)

BackRub is a web crawler which is designed to store the connection graph for the web. In other words BackRub stores which pages every web page links to. Currently we are developing techniques using this link data to improve web search engines as well as understand the structure of the web.

- [Larry Page](#)
-

[ComMentor](#)

Third-Party Annotations on web pages provide for ways to share information, rate content, and keep notes

["A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples"](#)

- [Martin Röscheisen](#)
 - [Christian Mogensen](#)
 - [Terry Winograd](#)
-

[InterOp Protocol](#)

The heart of the "InfoBus", this protocol describes access methods to search collections, acquire results, and find out about sources.

- [Steve Cousins](#)
 - [Prof. Hector Garcia-Molina](#)
 - [Scott Hassan](#)
 - [Andreas Paepcke](#)
-

[SCAM: The Stanford Copy Analysis Mechanism](#)

Making a perfect digital copy of a copyrighted work is easy in a networked world. How can the intellectual property rightsholders be protected? By detecting attempted distribution of illegal copies. Duplicate detection has other uses in information finding as well. An earlier, related project was known as COPS: The Copyright Protection Scheme.

["Building a Scalable and Accurate Copy Detection Mechanism"](#)

- [Prof. Hector Garcia-Molina](#)
 - [Narayanan Shivakumar](#)
-

[InterBib](#)

InterBib is a tool for maintaining bibliographic information. Capable of reading from and writing to many different formats, it acts as a unified, searchable repository of bibliographic records.

[Information on InterBib](#)

- [Andreas Paepcke](#)



User Interface Projects in the Stanford Digital Library

Too often the power of a search engine goes untested because users don't know how to exploit the advanced (or even basic) features. The use of a browser front-end has eased platform independent rapid prototyping, allowing a wide variety of services such as information clustering, annotating, and re-distributing via the WWW. One project even uses a web application to help create web applications! But the web does have drawbacks, such as being largely inaccessible to blind users (hear our audio interface!) and limiting the types of possible interaction. Therefore, our DLITE interface uses a direct manipulation metaphor of iconic representations, rather than relying on CGI forms.

[SenseMaker](#)

SenseMaker helps users iteratively reformulate their information needs through multi-dimensional organizing and active gathering of search results.

["SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests"](#)

- [Michelle Q Wang Baldonado](#)
-

[DLITE: A Digital Library Interface](#)

A direct manipulation user interface designed to support user tasks, to smoothly integrate the results of many services, to handle services of widely-varying time scales, to be extensible, and to support sharing and reuse.

["The Digital Library Integrated Task Environment \(DLITE\)"](#)

- [Steve Cousins](#)
-

[Grassroots](#)

Groupware for information finding, combines mail, news, and web in a single environment with distribution lists

["Grassroots: A System Providing a Uniform Framework for Communicating, Structuring, Sharing Information, and Organizing People"](#)

- [Kenichi Kamiya](#)
 - [Martin Röscheisen](#)
-

[ComMentor](#)

Third-Party Annotations on web pages provide for ways to share information, rate content, and keep notes

["A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples"](#)

- [Martin Röscheisen](#)
 - [Christian Mogensen](#)
 - [Terry Winograd](#)
-

Audio Interfaces to HyperText

The structure of a document is captured in HTML/SGML tags which most browsers map to visual display characteristics. We are seeking ways in which this structural information can be conveyed in audio format for blind users or users connecting via telephone.

[CSLI Annual Report](#)

- [Frankie James](#)
 - [Prof. Terry Winograd](#)
-

WebWriter

WebWriter is a direct manipulation Web page editor that allows users to create new web pages, including advanced features such as tables, without knowing HTML or CGI.

["WebWriter: A Browser-Based Editor for Constructing Web Applications"](#)

- [Arturo Crespo](#)
-

[RManage/FIRM](#)

Interoperable rights management is one of the service layers that the current Internet is still lacking. FIRM defines a platform for "smart contracts" that is based on a computational reification of contract law; it is realized as part of a novel, network-centric architecture for managing control information that generalizes previous models centered around clients or servers.

["A Network-Centric Design for Relationship-based Rights Management"](#)

- [Martin Röscheisen](#)
 - [Prof. Terry Winograd](#)
-





Digital Library Integrated Task Environment



The Digital Library Integrated Task Environment (DLITE) is an experimental, direct-manipulation interface to information objects and services. Information services are accessed via the InfoBus, and are presented to the user as components in workcenters.



For More Information...

- ▶ [Interface Details](#)
- ▶ [Interface Architecture paper](#)
- ▶ [Summary of interface goals \(CHI '96 paper\)](#)
- ▶ [List of DLITE Components](#)
- ▶ [Screen Shots](#)



DLITE is implemented as a distributed, client/server application. The server is written in Python, and clients have been written in Python/Tk and Java/AWT. DLITE makes extensive use of the Stanford InfoBus for search and query translation. We have completed a pilot study of the interface, and are continuing to test various aspects of it as well.



Credits...

DLITE is the PhD project of Steve Cousins. The following people have helped to build or design various aspects of the system:

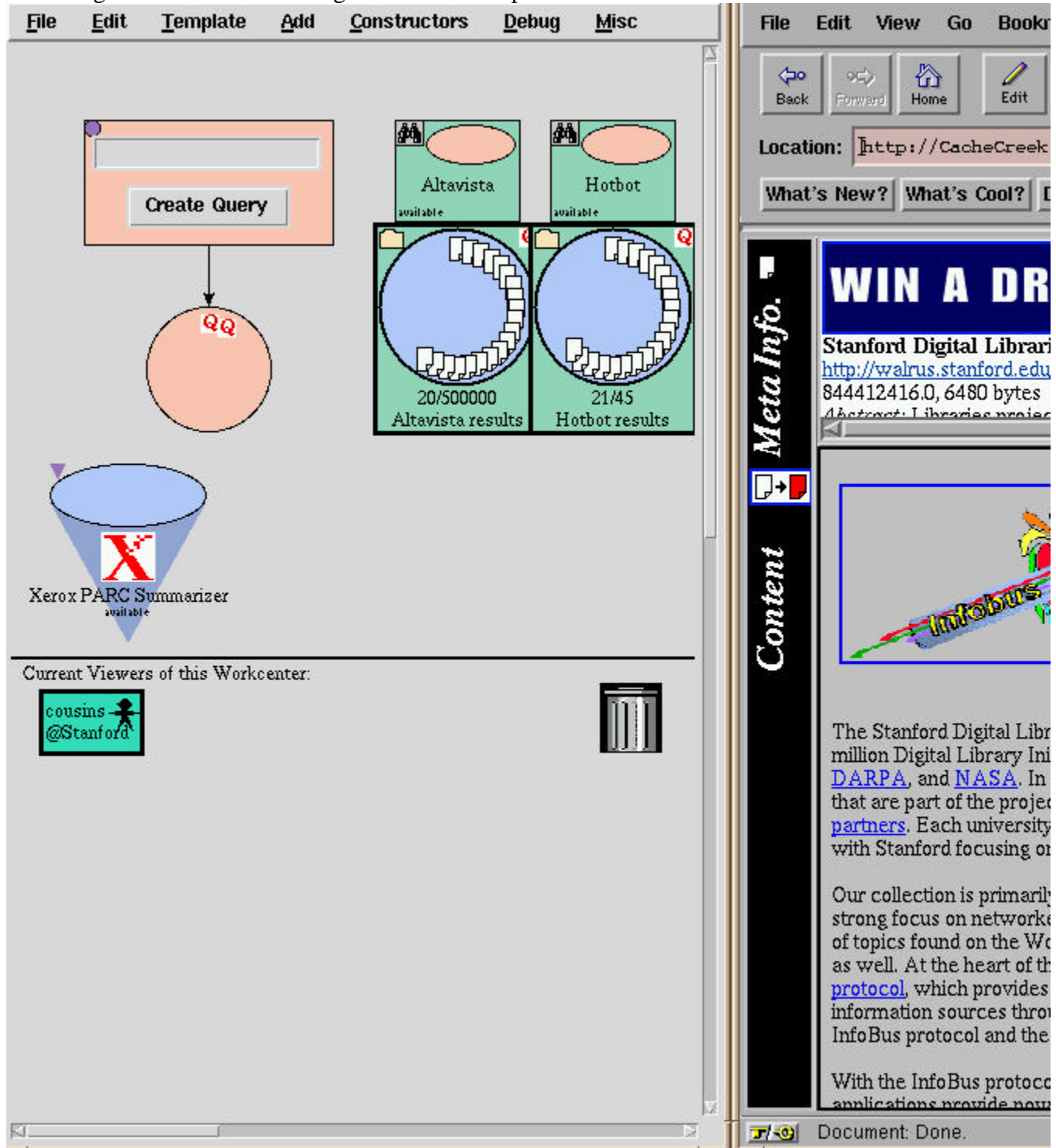
- Scott Hassan
- Alan Steremberg
- Terry Winograd
- Ken Pier
- Eric Bier
- Andreas Paepcke
- Mark Mortensen



DLITE Screen Shots

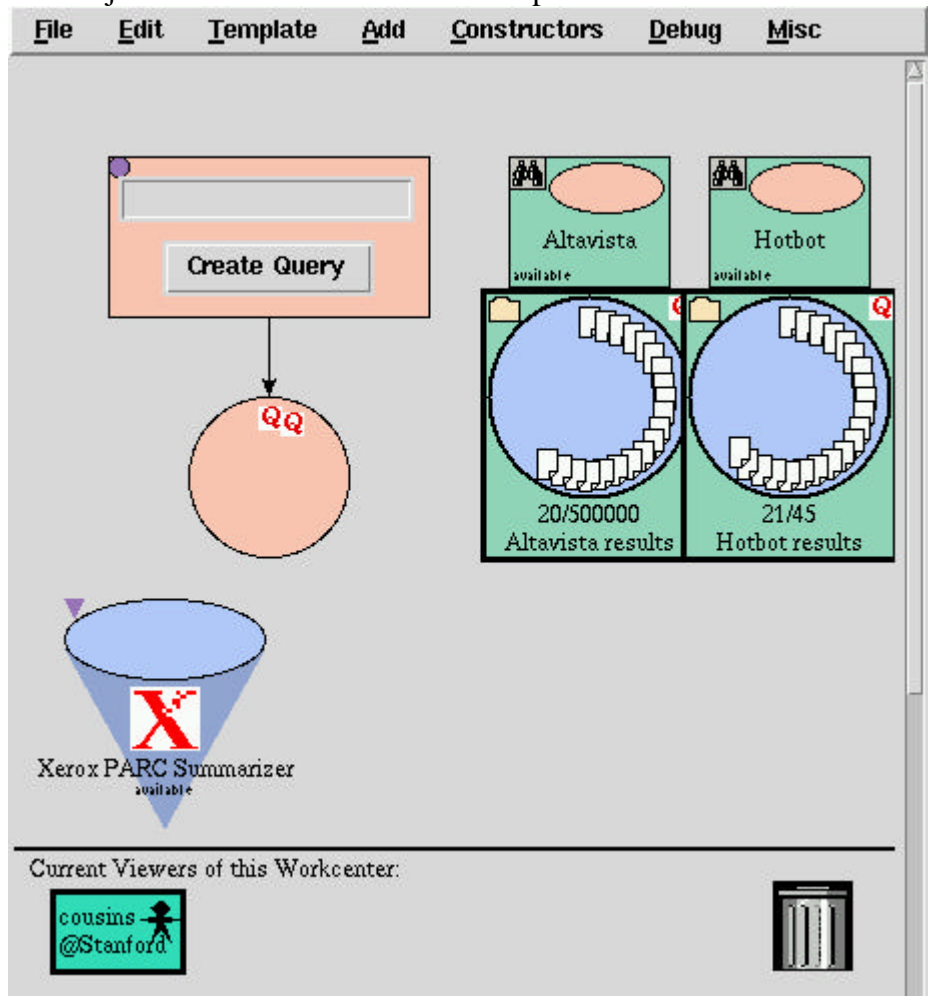
The whole screen

This image shows DLITE running next to a Netscape browser.



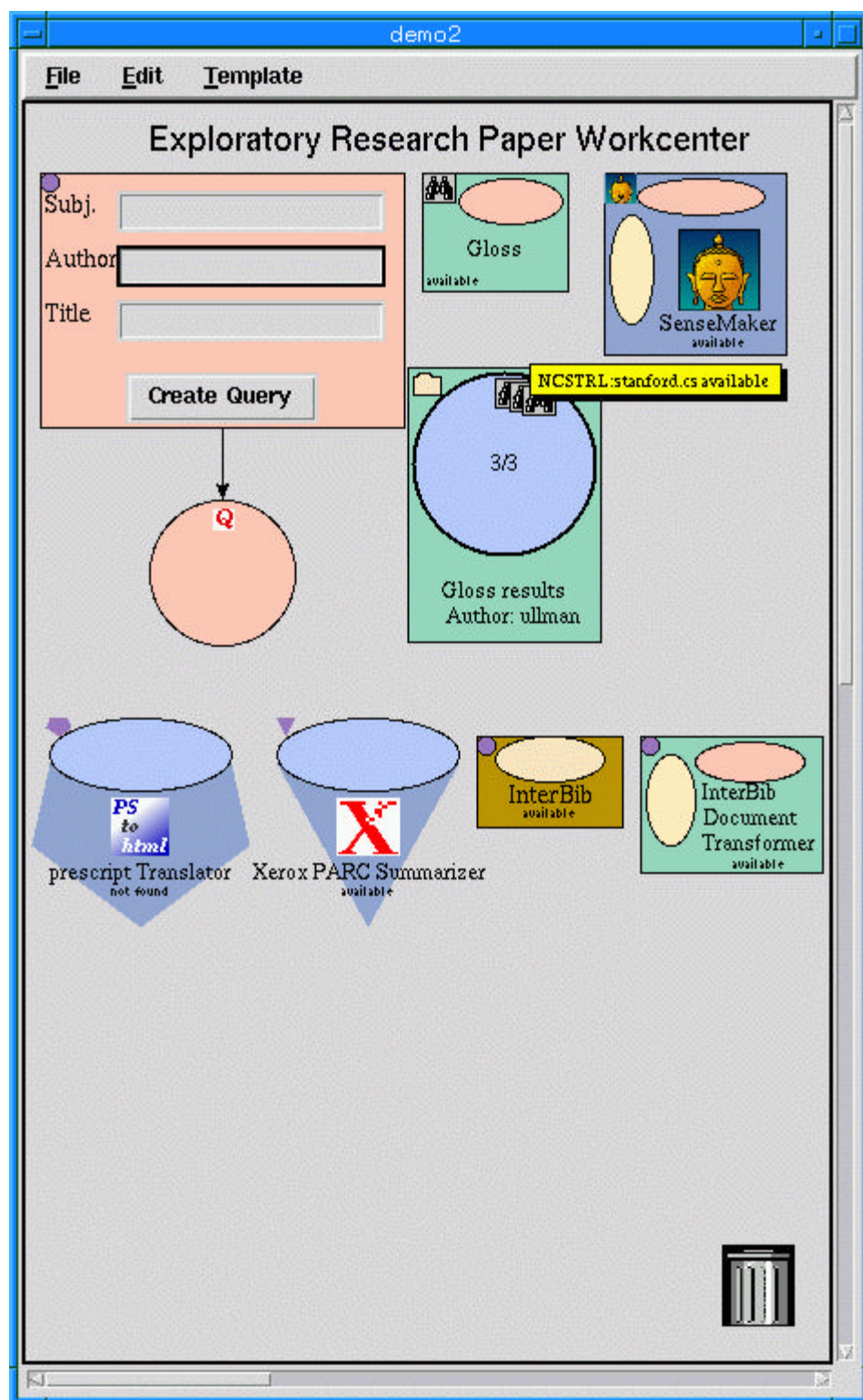
Simple search

Here is just the DLITE window for a simple search task.



Exploratory Research Workcenter

Here is a workcenter for doing exploratory research. Services are present for source-finding, sense-making, postscript-to-ascii translation, text summarization, and bibliography generation.



[Steve Cousins](#)

Last modified: Tue Jan 28 23:47:12 PST

DLI - Berkeley:

- [Home Page](#)
 - [IEEE Computer article](#)
 - [Tours](#)
 - [Collections](#)
 - [Source Code](#)
 - [Document-specific image decoders](#)
 - [GISviewer](#) (needs latest browser)
 - [Photos](#) and demos
 - [Context-based image queries](#)
 - [Blobworld](#)
 - [Image classification](#)
 - [California Aerial Photos](#)
 - [United States Department of Agriculture PLANTS Photo Gallery](#)
-

Pedagogy:

We recommend that the reader study these materials as part of work to answer the following questions:

- MVD
 - How well does [MVD 0.9](#) work for you? Could you get the links on that page to work (use 2 windows of browser, one for the instructions, and one for testing)? What do you like most about it?
 - Did you use it on video or a PC or Mac with Netscape 4?
 - Did you work out Lens overlaying, such as OCR and then Magnify?
 - For the TableSort example, could you under Anno view the note?
 - Could you get the special behaviors to work: Biblio, where you Select a type of format, use the mouse to select an entry, use Edit and Copy to get a version in that format, and then paste elsewhere?
 - Could you get Doublespace in the View menu to work?
- Cheshire
 - Can you find interesting environmental documents using Cheshire II?
- TileBars
 - What happens with TileBar search of "document" and "retrieval"?
 - What happens with TileBar search of "fault" and "dam"?
 - When is TileBar searching useful on a single document?
- Collections
 - What is the name of the DBMS used?
 - What is a database "schema"? How does it relate to "metadata"?
 - How many documents and how many images are in their collection?
 - How good is the OCRing? What research is underway to improve OCRing beyond that of ScanWorX and how well does it work? What is the main idea behind it?
 - How can you find the dams for a county?
 - How does the database table information for Almond dam relate to the page about it? To the OCR output about that page?

- What is a VLURL? How do you construct it? Can you build one and show results for getting pictures of California wildflowers that have the string "rose" in their common names?
- Display a distribution map for your favorite flower in California.
- Can you tell the direction of flight from the aerial photos?
- How do layers help with managing GIS information with the [GIS viewer](#)? Can you zoom in and out and pan around?

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



The UC Berkeley Digital Library project is part of the [NSF/ARPA/NASA](#) Digital Library Initiative and part of the California Environmental Resource Evaluation System ([CERES](#)). Research at Berkeley includes faculty, staff, and students in the Computer Science Division, the School of Information Management & Systems, and the Research Program in Environmental Planning & Geographic Information Systems, as well as participation from government agencies and industrial partners. The project's goal is to develop the technologies for intelligent access to massive, distributed collections of photographs, satellite images, maps, full text documents, and "multivalent" documents.



Welcome to the Berkeley Digital Library Project.

[Click here to browse all our collections.](#)

Our Collections:

- ♦ [Environmental Documents](#)
- ♦ [Photographs](#)
- ♦ [Aerial Photos](#)
- ♦ [Geographic Data](#)
- ♦ [Botanical Datasets](#)

Participants:

- ♦ [People](#)
- ♦ [Organizations](#)
- ♦ [Other DL Sites](#)
- ♦ [Related Projects](#)

About the Project:

- ♦ [Tours](#)
- ♦ [Papers](#)
- ♦ [Presentations](#)
- ♦ [Source Code](#)
- ♦ [Database Info](#)
- ♦ [Data Statistics](#)
- ♦ [Web Statistics](#)

Administrative Files:

- ♦ [Calendar](#)
- ♦ [Mailing Lists](#)
- ♦ [Seminar Schedule Winter '98](#)

What's New: new pictures, ARIA web reports, DLI 98 talks

Contact Us:



[Email to: \[www@elib.cs.berkeley.edu\]\(mailto:www@elib.cs.berkeley.edu\)](mailto:www@elib.cs.berkeley.edu)

[Sign Our Guestbook](#)

This server is powered by a [SUN Microsystems](#) Enterprise 3000 Server, backed by an [IBM](#) 7013 RS 6000 and 3494 Tape Library Dataserver running AMASS software by [EMASS](#). For additional information, see [About Our System](#).

Digital Library Initiative University of California at Berkeley

From *Computer* theme issue on the US Digital Library Initiative, May 1996

Information retrieval becomes an increasing challenge as comprehensive image databases emerge alongside traditional text databases. Here, a set of digital information services offers intriguing new retrieval possibilities.

Toward Work-Centered Digital Information Services

Robert Wilensky, *University of California, Berkeley*

Work-centered digital information services are library services that address a work group's information retrieval needs. These services differ in several ways from those required of digital libraries or information systems that meet, for example, education- or entertainment-related needs.

First, work groups frequently want to retrieve information, rather than documents per se. Because the answer to a query may be in more than one document, or even in textual form, users require information systems that can perform powerful, complex retrieval and analysis of heterogeneous objects.

Second, a work group must be able to access its own collections of varying data types, including legacy documents, in addition to external data collections. Work groups also continually create new materials, which are subject to differing degrees of external access. This requires flexible authoring, structuring, and delivery mechanisms.

Third, users must be able to integrate an information system into their established work practices, even as the system augments those practices. System interoperability is thus essential and may require custom interfaces. Information system evaluation must consider the system's contribution to the work group's goals, its support of existing work group practices, and its contribution to work practice innovations.

Realizing work-centered digital information systems requires a broad technical agenda that includes

- document image analysis, natural language analysis, and computer vision analysis for effective information extraction;
- new user interface paradigms and authoring tools for better accessing of multimedia information; and
- improved protocols for client program interaction with repositories.

We need to better understand the database issues involved in managing these distributed collections so that digital information services can be used by tens

of thousands of multiterabyte servers. We also must develop new types of documents to exploit these capabilities. At the University of California, Berkeley, we are researching these topics and developing associated technologies.

The testbed

To develop the appropriate technologies, we are creating a prototype set of information services called the California Environmental Digital Information System, which includes many different kinds of environmental data. Meanwhile, we have formed a consortium of data providers and users. We want our services to become a national resource and our prototype to serve as the basis for a California environmental information system.

We focused on environmental information because the data sets are large and diverse, highly motivated and technically sophisticated users will want to access the resources we make available, and the repositories we create will be a valuable national resource.

In particular, we want to collect California environmental information pertinent to the evolving needs of our consortium partners, including

- about 1 million pages of environmental technical reports;
- all county general plans;
- aerial and ground photography;
- US Geological Survey topographic, land use, and other special-purpose maps;
- computer models that simulate such environmental factors as traffic and water use;
- California Resources Agency videos; and
- California plant resources classification and distribution databases.

So far, we have scanned about 450 documents (roughly 100,000 pages) from the California Department of Water Resources (DWR). In addition, we have scanned about 200 air photos, about 100 of which are currently on line, including images of the California Delta and the California Water Project. We also have 11,643 ground photographs on line.

User needs

Because our project is work-centered, we have concentrated on user-needs assessment, and we have adapted and extended existing user-assessment research methods to the emerging digital information services technology. Recently, the Xerox PARC Work Practices and Technology Department joined our user evaluation effort.

We have frequently met with our initial user group, in the DWR's Sacramento offices, to learn about their work practices, information needs, information

products, and preferences for our testbed. We demonstrated our prototypes and installed different versions on a workstation so that users could gain experience with them.

We interviewed many people involved in state water planning to ascertain their needs and preferences. Some are consultants, and some work for state and local agencies, in environmental groups, and for water contractors. All are potential corpus contributors and users.

To collect data, we also observed meetings of, for example, the California Biodiversity Council. In addition, we investigated the contents, information retrieval needs, and current image retrieval methods of the DWR Graphics Services Unit film library. We also tested various data collection methods and plan to use the more successful ones extensively in coming months. Furthermore, our evaluation team has met with users to evaluate some of our systems, such as Cypress, an image retrieval system we discuss later.

In our iterative design processes, we have exploited information we learned about which data our users value and how to best display retrieved images and documents. For example, we produced custom interfaces for our DWR users based on the way they want information to be presented in Cypress. Users were also enthusiastic about our TileBars idea, which led us to implement Java-based TileBar access to our document collection.

Van House^[1] provides more information on user-needs assessment.

Architecture

Our system has a simple architecture, consisting of repositories, clients, indexing and searching, interoperability, and protocols.

Repositories

Any number of repositories, or information servers, are possible. Each is implemented as a database that supports user-defined functions and user-defined access methods. Building a repository on top of a true database system leverages the database community's work on distributed, scalable systems. Using a database management system (DBMS) that supports user-defined functions and access methods lets us easily incorporate new object analysis, structuring, and indexing technology into a repository.

Clients

We have developed several interoperable clients. These can be considered browsers designed for different document data types, such as images, geographic information system (GIS) data sets, and traditional (paged) documents. A GIS browser, for example, simplifies information requests about a geographic region. On a map, such a browser may display icons corresponding to documents, referenced in geographic terms, that pertain to

each location. A user activates a document browser to see a document. If the document contains a map, viewing the map will activate the GIS browser on it.

Indexing and searching

The repositories act as their own indexing servers. Much of our research involves the use of natural-language processing, computer vision, and GIS techniques to improve access to textual, image, and map-oriented information. We are experimenting with distributed search techniques for multiple repository searching.

Interoperability

We are experimenting with several forms of interoperability. One is repository-level interoperability, wherein we provide low-level access to our collections, as proposed by Kahn and Wilensky.[\[2\]](#) At a higher level, we perform schema-level interoperability in which we can apply our clients to another project's repository, and vice versa. For our two interoperability experiments, we are working with the Alexandria Project at the University of California, Santa Barbara (UCSB). Thus far, we have created views of both projects' metadata schemata sets that let us run our clients against the union of both projects' aerial photograph databases.

Protocols

The repositories communicate with clients via several protocols, most notably the widely used HyperText Transfer Protocol (HTTP). However, some clients communicate directly in the Structured Query Language (SQL). We are designing our own protocol, called ZQL, whose name we derived by combining the names of the Z39.50 protocol standard and the SQL. Moreover, we plan to implement the repository access protocol described by Kahn and Wilensky.[\[2\]](#)

Our client-server proposal

Our document collection includes traditional documents, images, and map-oriented data. Each document type may contain multiple data for which indexing is useful. For example, our ground photographs have textual captions by the photographer. These photographs are also preclassified into specific categories. For example, each photograph pertains to a specific geographic area, although, unlike our aerial photographs, its location is not explicitly indicated. Similarly, our traditional documents consist largely of text, which originates as paper and is made available as scanned images. Moreover, much of the important information in these documents is in tabular form, rather than English text. In addition, the documents contain maps and photos.

We develop access methods (often more than one) for each primary data type-text, image, and map-oriented data-and index each document by all applicable methods. For example, we index our photographs by image content, preassigned categories, location, and so forth. Generally, an access method

requires data analysis to provide the basis for an index. Analysis, usually performed at data acquisition, results in the assignment of additional features to the data, an index, or both. Various client programs let us enter queries about the data, generally by filling out forms or making geometric gestures, such as clicking on a map image. The client displays the analyzed information, which is used to service the query.

Example 1: Image retrieval subsystem

Some examples will illustrate our approach. Cypress, a client that provides access to our ground photographs, was derived from Chabot,[3] which used a custom Tcl/Tk client and a Postgres database back end. Like Chabot, Cypress yields photos that contain text and other metadata. At data acquisition, we run various image analysis processes and compute derived data. In particular, the processes perform color and texture analysis on each photo, generate an overall color histogram, and perform several kinds of object recognition, such as finding images with horizons. The metadata, computed data, and photo are then inserted into an Illustra object-relational database. Access in Cypress, unlike in Chabot, is via HyperText Markup Language (HTML) forms.

As shown in Figure 1, users making queries can select various color textures, such as a cluster of small orange blobs. Users can also select one of several color descriptors, such as "SomeYellow." In addition, the user can specify a geographic region; a subject and category, each from a fixed vocabulary; and text, which can be used to find a text caption. The client translates a form into an SQL query and submits it to the database manager.

Figure 1. A query designed to find American flags by looking for photos of ceremonies with horizons and thick red swatches.



External functions previously registered in the database are used to determine, for example, whether a particular photo's histogram matches the color descriptor, and to control text-matching details. The resulting relation is returned and formatted into one of several presentations, such as a table featuring each photo and its text caption. Figure 2, the result of Figure 1's query, illustrates this format. We created a custom form that lets our DWR users enter internally known data, such as photo CD number, and also access more metadata.

Figure 2. *The result of the "American flag" query.*



Example 2: Geographic browser subsystem

Our Napa browser is a geographically oriented database client that communicates directly to the DBMS and that primarily displays map-based information. The client lets users select data sets for map display, zoom and pan perspectives, and easily specify the altitude at which each data set should be displayed. For example, at very high altitudes, only state boundaries might be worth displaying. As one zooms in, increasing detail, such as roads and bridges, is useful. Panning, zooming, or clicking on an icon that represents a particular data object sends a query to the database server, and the resulting data set updates the display. In the process, the database must respond to geographic queries, such as what parts of a given data set are within the region to be displayed.

Our architecture philosophy is also evident in the interfaces to our document collection. The scanned documents, along with ASCII text produced by optical character recognition (OCR) of the images, are stored in a modified Cornell's Dienst server.^[4] This server lets users search for and access documents by their attributes, and then browse their page images. In addition, the document server lets the user browse each document's HyperOCR format—an ASCII version of the document produced by an OCR process that preserves page layout. Because the HyperOCR is one ASCII file, it is convenient for quick file browsing, searching, and performing other character-based operations, such as select-and-paste. However, it is produced by a lossy OCR process, so the user should refer back to the image for authentication. We make it easy for the user to switch between the HyperOCR and the images.

Client-server functionality

The use of Web clients promotes access but limits client-side functionality, unduly straining the network and servers. For example, we also have a Web-based, map-oriented access to our collections, specifically our geographically referenced aerial photographs. However, the Napa browser is a custom client and has considerably more client-side functionality than a Web

browser.

For example, by clicking on a given data set's display, a user can make the Napa browser change the display without consulting the server. Similarly, zooming and panning only the client's cache data can be done locally. In all such cases, the Web browser would have to consult the server to compute and transmit a new display. Likewise, Cypress's progenitor, Chabot, used a custom Tcl/Tk client, which gave the user additional functionality, such as defining and saving named queries for future use. Client availability however is limited by software distribution overhead.

We believe that the ubiquity of Web clients makes them an overwhelming force for client services. We migrate as much functionality to our Web clients as possible and extend client-side functionality when necessary, relying on scriptable browsing capabilities, notably Sun Microsystems' Java language.

Improving information access

Natural-language processing

We apply statistical natural language processing techniques to augment more traditional keyword approaches. Specifically, we want to provide a TileBars-style text interface, perform automatic text categorization, and provide an automated facility for locating geographically referenced text.

We have implemented a TileBars interface to our document collection. TileBars^[5] was introduced as a way to exploit what we call TexTiles--meaningful, automatically computed, multiparagraph, topically coherent text segments. A tile graphically reflects the relevancy of a text unit to a query so that the varying relevance of document segments is displayed in a bar with one tile corresponding to one segment. The user can specify multiple term sets and inspect the results.

We implemented a Java version of TileBars. The server uses a standard relevance metric (currently FreeWAIS 5.0) to determine each document's relevance to each of two term sets. The server then transmits the relevant figures to the client. The client lets the user dynamically choose how to display this result set by clicking on the appropriate screen button. In particular, the user can choose to see documents with segments that are highly relevant to both term sets, highly relevant to one and somewhat relevant to the other, somewhat relevant to both, or highly relevant to one and irrelevant to the other.

For example, Figure 3 shows the result of a query with the term sets "Berkeley" and "Santa Barbara." The second button is selected, meaning that the two TileBars above the solid line correspond to our documents that are highly relevant to "Berkeley" and somewhat relevant to "Santa Barbara." Clicking on individual tiles will activate our multivalent document browser (discussed later) on the appropriate documents, with the matching term sets highlighted in the display.

Figure 3. The result of a TileBars query contrasting "Berkeley" and "Santa Barbara." The row of buttons under the term sets allows different result-set sortings. In this case, the second button is selected, showing documents highly relevant to the first term set and somewhat relevant to the second. Documents below the bar are beneath the given sort's threshold. The "X" in various documents indicates there are many pages with no relevant terms. The arrows at the sides of some TileBars scroll the bar and are used for long documents.



Currently, we are indexing individual document pages rather than TextTiles. While this should work reasonably well on our collection, the mapping of TextTiles to page images is complex. We plan to create a TextTile version of our corpus and extend the interface to map TextTile boundaries on top of the other document representations.

To perform automatic categorization, we are further developing the topic assignment techniques begun earlier.^[6] We used a thesaurus automatically constructed from Wordnet to define the constituent categories. Since then, we have improved our algorithms and obtained *Roget's International Thesaurus* on line. This thesaurus gave us 1,073 assignment categories.

We trained our algorithms with 10 million words of text on a 10-Sparcstation network-of-workstations cluster and are assessing their accuracy. Our ultimate goal, of course, is to assign categories to our environmental document collection. While the assignment process ranks each category's relevance to a document, the categories can be used as a controlled vocabulary to index the documents. In addition, because multiple categories will typically be assigned to each document, the assigned tuples will fit into a large abstract lattice. We can semantically navigate the document collection by moving through this lattice. We are devising a user interface for such a navigation method.

We mentioned earlier how geographic location is an important way to access our information. For text, we have developed the Georeferenced Information Processing System (GIPSY),^[7] which hypothesizes the geographic regions pertinent to a document. GIPSY contains information about all California locations on US Geological Survey topographic maps and information about California agriculture and geography. This lets GIPSY indirectly hypothesize locations based on references to general agricultural and geographic features, as well as on direct references to proper nouns. With GIPSY, we have located geographically referenced photographs based on their text captions. We can then use a map-oriented browser, such as Napa, to access the photographs by location. Recently, we've developed a new GIPSY implementation expressed as

user-defined functions in our DBMS environment.

Document recognition

The content of scanned documents is a key research area because many documents' authoritative versions are still the ones in print, despite document processing software's widespread availability. Also, images are one of the few forms in which documents can be accessed across platforms.

We have developed page recognition and parsing tools, including tools for deskewing a page, separating it into connected components, and clustering it into characters. Unlike tools in commercial systems, our tools are modular, which facilitates experimentation. For example, we implemented a system for learning character template bitmaps from whole-page document images and unaligned transcriptions. As a result, our system lets users easily develop document-specific character models.

Compared with typical OCR devices, which are not tuned to a particular font, document-specific models offer much lower OCR error rates. However, training an OCR system for a particular font typically involves considerable manual effort. With our system, a user prepares several document pages containing document font character samples. From the transcription and page images, the system generates a set of document-specific character templates, which are used to recognize the remaining pages.

A quantitative system performance evaluation showed that the OCR error rate improved by a factor of seven to 20, depending on the language model used with the scanned document. We based the evaluation on a 406-page environmental bulletin in our collection, for which a source file was available to assess OCR performance. Templates were generated from 20 nontabular page images, using the corresponding source file as the training transcription. The resulting templates were used to decode 375 pages of document tables.

Following Kopec and Chou's approach,[\[8\]](#) we developed document-specific decoders for two environmental bulletins that our DWR users designated as very valuable. The decoder analysis recognizes document structure, and this can be exploited in various ways. For example, we used the decoder output to produce HTML document versions.

We built a relational database from information on 1,395 California dams under state jurisdiction that was in one of the two bulletins with which we were working. A user can now interrogate this database via a form to respond to such commands as "find all the dams located on the Sacramento River." Because each dam is geographically referenced in the document, we could easily create a map interface that displays the dams as points on a map of California. The display lets the user click on a point to access information about the corresponding dam. Finally, the dam display has a button that lets us determine whether we have photos of the dam.

A new document model--multivalent documents

We are developing a general digital document model called multivalent documents. Multivalent documents begin to exploit digitization's possibilities and offer much greater functionality than existing document models. In this new paradigm, complex documents comprise multiple layers of distinct but intimately related content that may be geographically distributed. Small, dynamically loaded program objects called behaviors activate the content. The behaviors work with each other and layers of content to support arbitrarily specialized, complex document types. Behaviors bind together the disparate pieces of a multivalent document to present the user with a unified conceptual document. Such behaviors are crucial in meeting the needs of group work and achieving our work-centered goal.

OCR-select-and-paste is an example of multivalent documents' diverse functionality. The user first selects a geometric region on a printed page's scanned image. The OCR-generated corresponding text is then copied into the window system's cut buffer. The user thus interacts with an image, but the actions may reference other content layers, such as the page's OCR, in an intuitive and natural manner.

Table manipulation is another kind of multivalent document functionality. Through document recognition or direct authoring, a user could identify a document's important tables and insert them into a database. The user then might "ask" the document to sort the table by an arbitrary column or to perform more complex transformations that take advantage of the underlying database functionality.

User annotations that augment, or possibly transform, the conceptual document's content are another example of multivalent document functionality. A multivalent perspective is particularly appropriate for geographic data, as geographic information systems already take a layered view of their data.

More complex forms of functionality are also possible. For example, aligned layers of language translations could be consulted to transparently "translate" the pasted characters into the preferred language. A structural map that associates geometric regions on the page with semantic labels could enable more sophisticated operation by other behaviors. For instance, rather than laboriously hard-code links for references, a behavior could register interest in a specific user interaction on a semantically labeled region, then carry out a hyperlink or other action. By operating through the higher-level semantic layer, users can add or reprogram behaviors easily and efficiently.

Video subtitling is a much different multivalent document type. In subtitling, a video clip is aligned with a script and with language translations so that a video can be presented simultaneously in different languages while remaining searchable with text-based techniques.

A document management infrastructure built around a multivalent perspective

provides an extensible, networked system that supports incremental content addition; incremental interaction addition, with the user and with other components; content reuse across behaviors; behavior reuse across document types; and efficient network bandwidth use. These functions are essential to work group support.

We have implemented a multivalent document infrastructure, along with several types of document functionality, including OCR-select-and-paste, table sorting, hyperlink layers, and alternative pasting. The implementation is in Java, with powerful client-side functionality that facilitates multivalent documents.

Figure 4 depicts an example of our initial implementation. Although the figure shows a scanned page image, an underlying layer holds the images' OCR. Another layer-constructed on the fly across the network at a Xerox PARC server-contains information that maps the words to the image positions. The user can thereby search the image for textual matches.

Figure 4. Multivalent document example with a scanned image and an OCR layer. The terms in the search window are highlighted where they appear in the display. The region selected by a click-and-drag mouse motion is highlighted in yellow. The corresponding OCR-derived text is available in the select buffer.

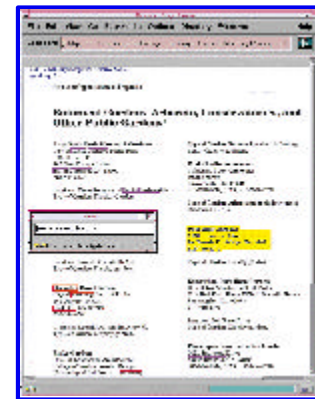


Figure 4, in fact, results from the user's selecting a page from our earlier TileBars example that was relevant to UCSB and to us. That interface called the multivalent browser and passed it the disjunction of topic terms. The image regions corresponding to the matching text are highlighted in different colors. The user can add or remove search terms using the small window. In addition, the user can select text from the image with a click-and-drag mouse motion. The region selected is shown by the yellow background highlight. The corresponding OCR-derived text is now available for pasting.

The multivalent browser will run on all 100,000 of our scanned page images and is available on our server to Java-compliant browsers. The distributed annotation facility is currently being implemented. For more details on multivalent documents, see Phelps and Wilensky.[\[9,10\]](#)

Image understanding

We have implemented a few object detectors that find objects in images. In

particular, we can find horizons with reasonable accuracy. We currently have a tree detector prototype and detectors that can detect clothed and nude humans. We soon expect to perform automatic recognition of several dozen kinds of things, such as canals and roads. We are developing learning methods to automatically construct detectors from a sample training set of our collection. Meanwhile, we also expect to implement automated image segmentation for use by these processes.

Conclusion

Our system has several other interesting aspects, such as tertiary storage management and scalable multiresolution compression to enhance the use of networked resources. We think that tertiary storage management in particular will acquire new significance as multiterabyte information needs become commonplace.

It is premature to reach major conclusions about user needs. For example, we are still learning how users want to retrieve images by content so that we can develop the appropriate technology. Moreover, we expect that user-needs assessment research will continue to evolve.

Meanwhile, we are still just beginning our work. As suggested here, the very notion of a digital document is rudimentary and will no doubt develop into something whose form and function we can now only dimly imagine. Certainly, our concept of digital libraries or digital information systems must also undergo transformation.

To learn more

We invite readers to obtain more information from our Web site, <http://elib.cs.berkeley.edu>. Most of the clients discussed in this article are available for experimentation from the project server at our site, where readers can also find most of our source code and examine our access methods.

Acknowledgments

The work described here is a joint effort of many people, including Ken Arneson, Paul Brown, Michael Buckland, Mark Butler, Chad Carson, Isaac Cheng, Gary Darling, Richard Fateman, David Forsyth, Howard Foster, Kenn Gardels, Hayit Greenspan, Jon Hull, Gary Kopec, Ray Larson, Thomas Leung, Jitendra Malik, Ray McDowell, Greg McKean, Ginger Ogle, Tom Phelps, Lisa Schiff, Mike Schiff, Mike Stonebraker, Taku Tokuyasu, Richard Troy, Robert Twiss, and Nancy Van House. The work was supported in part by National Science Foundation grant IRI-9411334 in connection with the NSF/NASA/ARPA Digital Library Initiative.

References

1. N. Van House, "User Needs Assessment and Evaluation for the UC Berkeley Electronic Environmental Library Project," *Proc. Digital Libraries '95: Second Ann. Conf. Theory and Practice of Digital Libraries*, Texas A&M Univ. Hypermedia Research Laboratory, College Station, Tex., 1995.
2. R. Kahn and R. Wilensky, "A Framework for Distributed Digital Object Services," CNRI Tech. Report TN95-01, Corp. for Nat'l. Research Institutions, Reston, Va., 1995; also see URL <http://www.cnri.reston.va.us/home/cstr/arch/k-w.html>.
3. V.E. Ogle and M. Stonebraker, "Chabot: Retrieval from a Relational Database of Images," *Computer*, Vol. 28, No. 9, Sept. 1995, pp. 40-48.
4. J.R. Davis and C. Lagoze, "A Protocol and Server for a Distributed Digital Technical Report Library," Tech. Report CS:TR94-1418, Dept. of Computer Science, Cornell Univ., 1994; also see URL <http://cs-tr.cs.cornell.edu:80/Dienst/UI/2.0/Describe/ncstrl.cornell%2fTR94-1418?abs>
5. M.A. Hearst, "Context and Structure in Automated Full-Text Information Access," Tech. Report UCB:CSD-94-836, Computer Science Dept., Univ. of California, Berkeley, 1994.
6. D.E. Fisher, "Topic Characterization of Full Length Texts Using Direct and Indirect Term Evidence," Tech. Report UCB:CSD-94-809, Computer Science Dept., Univ. of California, Berkeley, 1994.
7. A. Woodruff and C. Plaunt, "GIPSY: Georeferenced Information Processing System," *J. American Soc. for Information Science*, Vol. 45, No. 9, 1994, pp. 645-655.
8. G.E. Kopec and P.A. Chou, "Document Image Decoding Using Markov Source Models," *IEEE Trans. PAMI*, Vol. 16, No. 6, June 1994, pp. 602-617.
9. T.A. Phelps and R. Wilensky, "The Case for Multivalent Document Decomposition," *Proc. 29th Hawaii Int'l Conf. System Science*, IEEE CS Press, Los Alamitos, Calif., Order No. PR07336, 1996.
10. T.A. Phelps and R. Wilensky, "Toward Active, Extensible, Networked Documents: Multivalent Architecture and Applications," *Proc. Digital Libraries '96*, ACM, New York, 1996, pp. 100-108.

Robert Wilensky is a professor and the Computer Science Division chair at the University of California, Berkeley, where he has been on the faculty since 1978. Wilensky founded the Berkeley Artificial Intelligence Research Project and the Berkeley Cognitive Science Program. He directs the UC Berkeley/Hewlett-Packard Science Center and is the UC Berkeley Digital Library Project's principal investigator. Wilensky has published numerous articles on artificial intelligence, planning, knowledge representation and natural language processing, and he has authored two computer programming books. He received a BA in mathematics in 1972 from Yale College and a PhD in computer science in 1978 from Yale University. He is a fellow of the American Association for Artificial Intelligence.

Readers can contact Wilensky at Computer Science Division, 389 Soda Hall, Univ. of California, Berkeley, CA 94720; phone (510) 642-0930; e-mail wilensky@cs.berkeley.edu.

Guest Editors' Introduction

University of Illinois

University of California at Berkeley

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

[Computer](#) | [Computer Society home page](#)

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.



Our Collections

Berkeley Digital Library Project

Other Information: [About the Database](#) | [About the Digital Library project](#) | [Data Statistics](#)

	description	query form	browse a list	static map	active map*	more information
Documents	Cal. environmental documents: plans, ordinances, EIRs, etc.	search TileBars *	X			about the collection advanced structured docs about MVD about TileBars tour the documents
Databases	California dams	X	X	X	X	about the dams
	CalFlora	X				about Calflora
	CalFlora Occurrences	X				about Calflora Occurrences
	Fishes of the Sacramento-San Joaquin Estuary		X			
	Bay Area Streets				X	
	California Gazetteer				X	
Aerial Photos	Sacramento R. Delta, Suisun		X	X		about air photos
Geographical Layers	Northern California				X	about GIS Viewer tour the GIS Viewer
	Russian River				X	
	Sacramento R. Delta Fish Flow				X	
Photographs	Cal. flora and habitats, Dept. of Water Resources photos, etc.	X	X			about the photos computer vision research tour the images
External Collections	BIOSIS dictionary: water subdomain	X				about OASIS
	INSPEC dictionary	X				about OASIS

* java is required

UC Berkeley Digital Library Project | www@elib.cs.berkeley.edu | last updated June 26, 1997



GIS Viewer: North Coast of California

Berkeley Digital Library Project

Please be patient while this Java applet loads.

For more information about the GIS Viewer check out our [tour](#) or our help pages. If you click on the help button within the GIS Viewer applet the help pages will come up in a separate browser window. If you click [here](#) the help pages will come up in this window.

[Berkeley Digital Library](#)

www@elib.cs.berkeley.edu



Photographs

Berkeley Digital Library Project

The Berkeley Digital Library collection includes a large number of digitized images from many different sources. We add new images to the collection on a weekly basis. As of April 1998 there were over 58,000 images available for online searching. For a current count of images, see [Data Statistics](#).

Search all the images in our collection

- Fill out a [query form](#) for all the images
- Take a [tour of the image collection](#)
- View [sample queries](#) for all the pictures

Browse individual photo collections

- [California Department of Water Resources](#) DWR film library collection of over 15,000 images.
- [Brousseau Collection of California Flora](#) Brother Alfred Brousseau (1908-1988) of St. Mary's College in Moraga, California made a collection of 35mm color slides of native wildflowers of California which consists of over 11,000 slides of over 2,000 species, as well as pictures of trees and mushrooms.
- [California Wilderness Scenes](#) Brother Alfred Brousseau's photographs taken between 1954 and 1984 of California's natural beauty.
- [California Habitats](#) A collection of 158 photographs of California habitats taken by Marc Hoshovsky of the State of California Department of Fish and Game.
- [Russian River](#) A small collection of photographs from the area around the Russian River.
- [Corel Stock Photos](#) This is a collection of images from [Corel](#) that we use for computer vision research. These images may not be downloaded or saved.

Computer vision research

Click on the heading above to learn more about how we use these pictures for our computer vision work, including demos, papers, faculty and researchers involved in this work.

NEW Check out our [Blobworld](#) image retrieval system

UC Berkeley Digital Library Project | www@elib.cs.berkeley.edu



Demos: Content-based Queries

Berkeley Digital Library Project

The following queries use image content information alone to retrieve pictures from a collection of 50,000 images. The database query that was generated will be shown at the bottom of each page of pictures. For more information about image analysis techniques used, see [Computer Vision Research](#). To construct your own query, see [Content-based Query on all Images](#).

Finding Objects in Pictures



Horses (14)

see [Finding horses using body plans](#)

Colored Blobs and Color Percentages



Sailing and Surfing (17)

blue-green % > 30 and very sm. yellow dots > 0 and collection = corel or DWR



Pastoral Scenes (93)

green % > 25 and lt. blue % > 25

[Pastoral Scenes: non-Corel pictures only](#)



Purple Flowers (114)

sm. purple dots > 3



Fields of Yellow Flowers (75)

very sm. yellow dots > 15



Pink People (69)

lg. or very lg. pink dots > 0 and orange % > 1 and collection = corel or DWR



Animals (229)

very lg. brown dots > 0 and very sm. black dots > 1 and green % > 20

Welcome to Blobworld!

Why Blobworld?

Very large collections of images are growing ever more common. From stock photo collections to proprietary databases to the Web, these collections are diverse and often poorly indexed. Unfortunately, image retrieval systems have not kept pace with the collections they are searching. The shortcomings of these systems are due both to the image representations they use and to their methods of accessing those representations to find images:

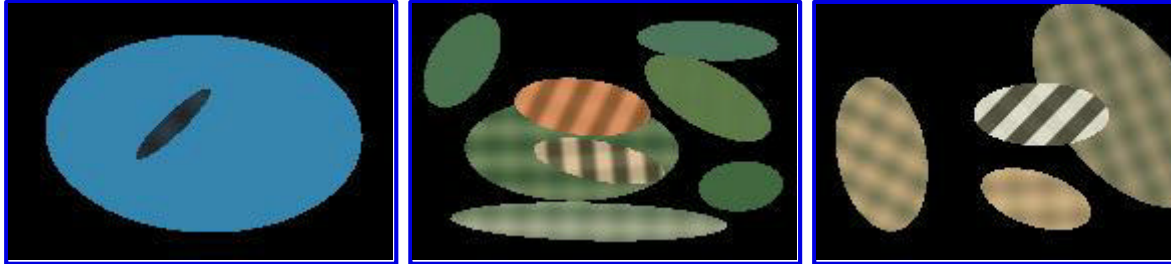
- While users often would like to find images containing particular objects ("things"), most existing image retrieval systems represent images based only on their low-level features ("stuff"), with little regard for the spatial organization of those features.
 - Systems based on user querying are often unintuitive and offer little help in understanding why certain images were returned and how to refine the query. Often the user knows only that he has submitted a query for, say, a bear and retrieved very few pictures of bears in return.
 - For general image collections, there are currently no systems that automatically classify images or recognize the objects they contain.
-

What is Blobworld?

We have developed a new image representation, "Blobworld," and a retrieval system based on this representation. While Blobworld does not exist completely in the "thing" domain, it recognizes the nature of images as combinations of objects, and querying and learning in Blobworld are more meaningful than they are with simple "stuff" representations.

We use the Expectation-Maximization (EM) algorithm to perform automatic segmentation based on image features. EM iteratively models the joint distribution of color and texture with a mixture of Gaussians; the resulting pixel-cluster memberships provide the segmentation of the image. After the image is segmented into regions, a description of each region's color, texture, and spatial characteristics is produced.

Here is a visualization of the Blobworld representation. We show each image region as an elliptical blob; each blob's two dominant colors are shown in the plaid patterns. The orientedness of the pattern indicates the texture's anisotropy, the orientation of the plaid indicates the orientation of the texture, and the sharpness of the plaid indicates the texture contrast.

**Original
images:****Blobworld:**

What can we use Blobworld for?

In a querying task, the user can access the regions directly in order to see the segmentation of the query image and specify which aspects of the image are central to the query. When query results are returned, the user sees the Blobworld representation of the returned images; this assists greatly in refining the query. You can see the [results](#) of several image queries using Blobworld, or [try your own query](#) on the images in the Digital Library collection.

Because Blobworld often encodes the objects in an image, we can also classify images automatically using an algorithm that learns the distributions of categories in Blobworld. You can see the [results](#) from such a system.

Want to learn more?

- [Try a Blobworld query!](#)
- Check out the [query results](#) or [classification results](#).
- Read our most recent [paper about Blobworld](#) or [other papers](#).

The original images are copyright [Corel](#). They are for viewing only and may not be saved or downloaded.

Last updated January 29, 1998, by Chad Carson



Image Classification

Berkeley Digital Library Project

The 14 categories shown below were chosen from the [Corel](#) image collection. About 90 pictures from each category were used for training and testing an algorithm that classifies images using [regions of coherent color and texture](#). The images used for testing are available [here](#). Use the table below to see all the images in each category and the classification of each image in a given category. For comparison, we also show the classification using color histograms.

All images in a category	Classified into a category using Blobworld	Classified into a category using color histograms
Airplanes	Classified as airplanes by Blobworld	Classified as airplanes by color histograms
Bald eagles	Classified as bald eagles by Blobworld	Classified as bald eagles by color histograms
Brown & black bears	Classified as brown & black bears by Blobworld	Classified as brown & black bears by color histograms
Cheetahs	Classified as cheetahs by Blobworld	Classified as cheetahs by color histograms
Deserts	Classified as deserts by Blobworld	Classified as deserts by color histograms
Elephants	Classified as elephants by Blobworld	Classified as elephants by color histograms
Fields	Classified as fields by Blobworld	Classified as fields by color histograms
Horses	Classified as horses by Blobworld	Classified as horses by color histograms
Mountains	Classified as mountains by Blobworld	Classified as mountains by color histograms
Night scenes	Classified as night scenes by Blobworld	Classified as night scenes by color histograms
Polar bears	Classified as polar bears by Blobworld	Classified as polar bears by color histograms
Sunsets	Classified as sunsets by Blobworld	Classified as sunsets by color histograms
Tigers	Classified as tigers by Blobworld	Classified as tigers by color histograms
Zebras	Classified as zebras by Blobworld	Classified as zebras by color histograms



[Berkeley DL](#)



[AccessMatrix](#)



[Information](#)



[Photographs](#)



[Comments](#)



California Aerial Photos

Berkeley Digital Library Project

Click on a **Flightline** to see thumbnail images for that flightline.

Some of the available flightlines are positioned on a map. Click on a **Description** to see the map for that area, with links to images.

Description	Contractor's ID	Elib ID	Type	Flightlines	Date	Contractor	Source
California Aqueduct: East Branch	WR-BED-C	aqd_east	color	1 2 3 4 5 6 7 8 9	Aug 03, 1994	I.K.Curtis Services, Inc.	DWR
North Bay Aqueduct	WR-AXY-C	aqd_nbay	b&w	1 2 3 4	Oct 02, 1990	Radman Aerial Surveys	DWR
South Bay Aqueduct: Livermore to Terminal Facilities	WR-AXX	aqd_sbay	b&w	1 2 3 4 5 6 7	Oct 02, 1990	Radman Aerial Surveys	DWR
North Delta Flood Plain Environmental Study	WR-BBG-C	delta_nflood	color	1 2 3 4 5 6 7	Feb 14, 1993	Radman Aerial Surveys	DWR
Statutory Delta	WR-BCM-CIR	delta_stat	colorIR	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22	Jun 22-23, 1993	Radman Aerial Surveys	DWR
Suisun Marsh Vegetation Study Low Tide	WR-BDW-C	suisun	color	1 2 3 4 5 7 9 11 13 16 19 20 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48	Jun 10-14, 1994	Radman Aerial Surveys	DWR

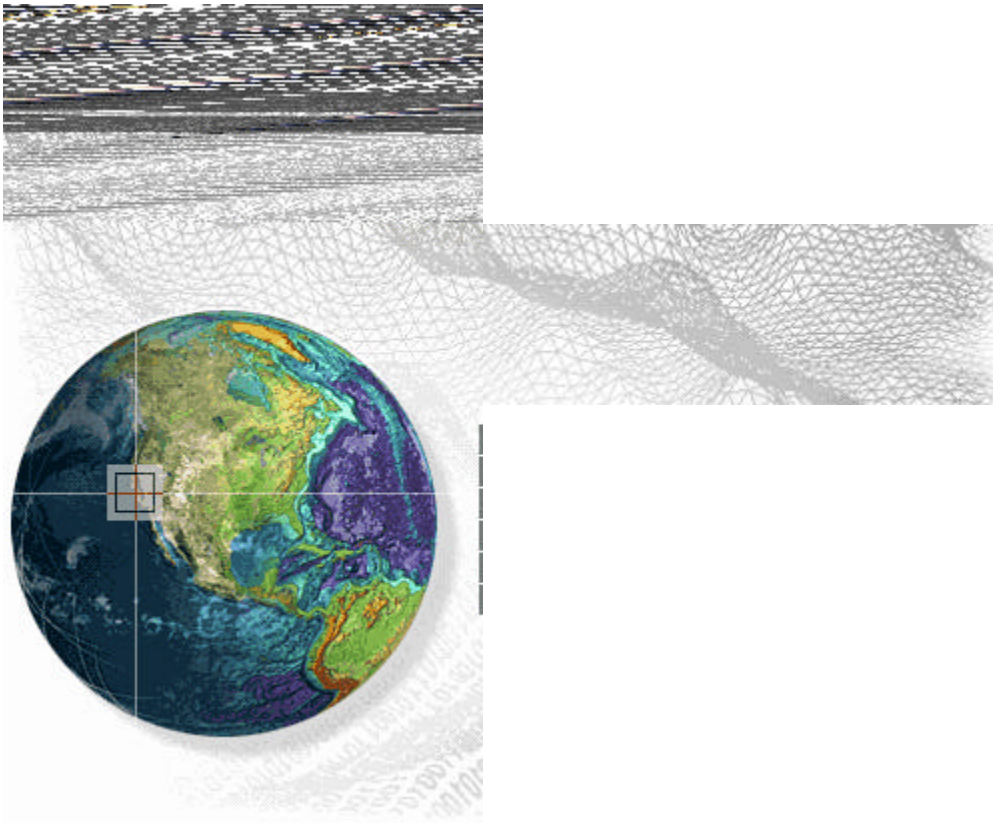
DLI - Santa Barbara:

- [Home Page](#)
- [Tutorial](#)
- [World Spatial Data](#)
- [Annual Report](#)
- [H. Chen's work](#) (with "cool DL, Web, agent, visualization, and multilingual IR demos"), and [GIS work](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



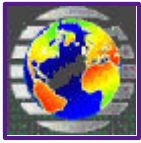
THE ALEXANDRIA DIGITAL LIBRARY

University of California, Santa Barbara
1205 Girvetz Hall
Santa Barbara, CA 93106, USA
TEL: 805.893.7665 **FAX:** 805.893.3045
URL: www.alexandria.ucsb.edu



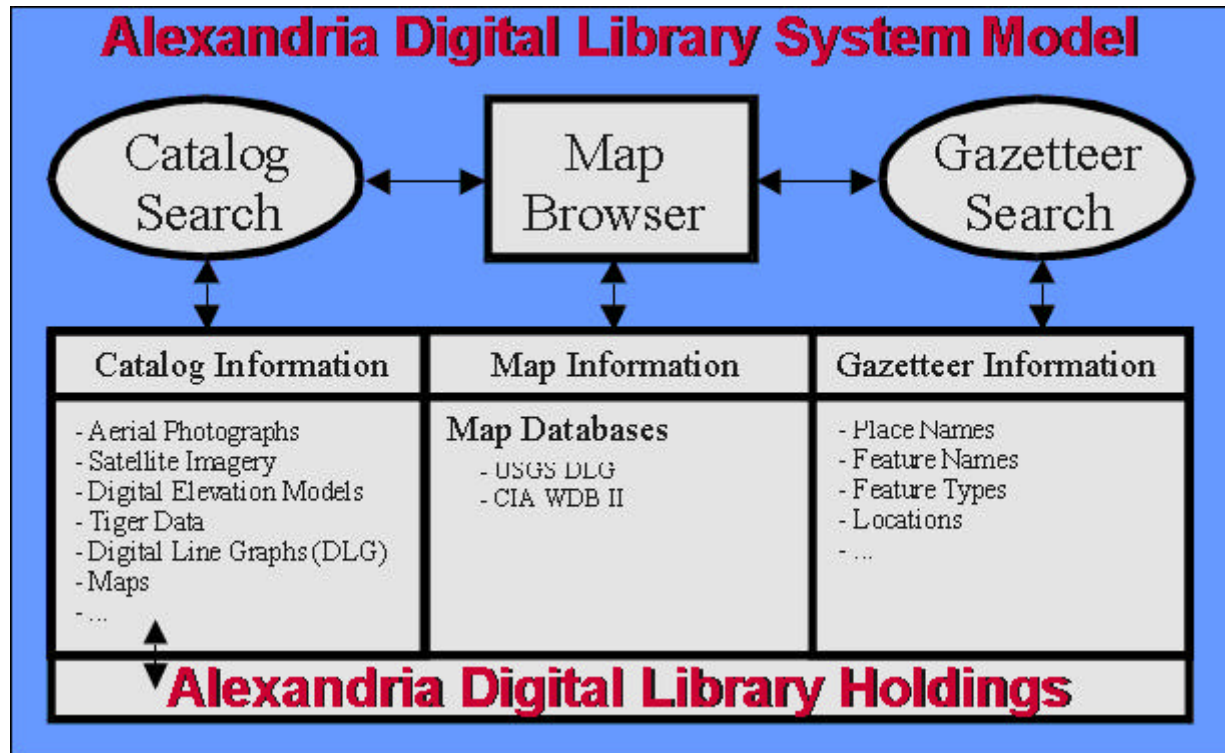
Last Modified: May 30, 1998

Our website is being redesigned and now uses some high-end browser features that may not be available in the browser you are currently using. We utilize frames and tables extensively in the new design and require a compatible browser, such as **Netscape Navigator 3.0** or **Microsoft Internet Explorer 3.0**.



Tutorial Table of Contents

◀ Prev Next ▶



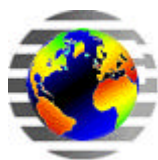
Conceptual model of the Alexandria Web interface

- [Conventions](#)
- [Session / System Setup](#)
- [Map](#)
- [Gazetteer](#)
- [Catalog](#)
- [Overview of Current Holdings](#)
- [Walkthroughs \(Example Sessions\)](#)
- [Feedback](#)
- [Technical Reference](#)
- [Acknowledgements](#)
- [Access ADL](#)

- [World Wide Web Help](#)
(How to use a web browser. That's what you're using right now.)

◀ Prev Next ▶

Universe



Alexandria Digital Library: [ADL](#)

[\[comment\]](#) [\[suggestions\]](#) [\[information\]](#) [\[add a URL\]](#)

Universe

[\[UNIVERSE\]](#) [\[EARTH\]](#) [\[AFRICA\]](#) [\[AMERICAS\]](#) [\[ANTARCTICA\]](#) [\[ASIA\]](#) [\[EUROPE\]](#) [\[OCEANIA\]](#)
[\[By Subject\]](#) [\[By Title\]](#)

Earth	Jupiter	Mars	Moon
Neptune	Saturn	Sun	Uranus
Venus			

Universe

Aerial photographs

- [Regional Planetary Image Facility](http://ceps.nasm.edu:2020/rpif.html)::http://ceps.nasm.edu:2020/rpif.html
- [Sources of Earth and Planetary Photography](http://ceps.nasm.edu:2020/RPIF/RPIFsources.html)::http://ceps.nasm.edu:2020/RPIF/RPIFsources.html

Artificial satellites

- [Mission and Spacecraft Library](http://leonardo.jpl.nasa.gov/msl/home.html)::http://leonardo.jpl.nasa.gov/msl/home.html

Astronomical - Observations

- [The Web Window to the Invisible Universe](http://wwwpks.atnf.csiro.au/databases/surveys/aitoff/aitoff.html)::http://wwwpks.atnf.csiro.au/databases/surveys/aitoff/aitoff.html

Astronomical photometry

- [JPL Public image archive](http://www.jpl.nasa.gov/archive/images.html)::http://www.jpl.nasa.gov/archive/images.html
- [Latest HST Observations. \(Hubble\)](http://www.stsci.edu/pubinfo/Latest.html)::http://www.stsci.edu/pubinfo/Latest.html
- [NASA JSC Digital Image Collection](http://images.jsc.nasa.gov/)::http://images.jsc.nasa.gov/
- [Space Image Libraries](http://www.okstate.edu/aesp/image.html)::http://www.okstate.edu/aesp/image.html
- [Stereoscopic Maps of Nearby Stars](http://www.clockwk.com/stars/index.html)::http://www.clockwk.com/stars/index.html
- [Today@NASA](http://www.hq.nasa.gov/office/pao/NewsRoom/today.html)::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html

Astronomy

- [Astronomical Data Center](http://adc.gsfc.nasa.gov/)::http://adc.gsfc.nasa.gov/
- [CyberAstronomy](http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html)::http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html
- [NASA/IPAC Extragalactic Database \(NED\)](http://ned.ipac.caltech.edu/)::http://ned.ipac.caltech.edu/
- [Planet Finder](http://www.calweb.com:80/~mcharvey/planet_all.html)::http://www.calweb.com:80/~mcharvey/planet_all.html
- [SEDs Internet Headquarters](http://sed.s.lpl.arizona.edu/)::http://sed.s.lpl.arizona.edu/

Astrophysics

- [HEASARC/GSFC Home Page](http://guinan.gsfc.nasa.gov/)::http://guinan.gsfc.nasa.gov/

- [NASA Data Archive and Distribution Service \(NDADS\)](http://nssdca.gsfc.nasa.gov/)::<http://nssdca.gsfc.nasa.gov/>

Cartography

- [Atlas celeste](http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg)::<http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg>
- [Ptolemy's Geography](http://www.oir.ucf.edu/wm/map/ancient-world-map-1482.jpg)::<http://www.oir.ucf.edu/wm/map/ancient-world-map-1482.jpg>

Comets

- [Comets and Meteor Showers](http://medicine.wustl.edu/~kronkg/index.html)::<http://medicine.wustl.edu/~kronkg/index.html>

Directories

- [JPL Organizational Home Pages](http://www.jpl.nasa.gov/orgs/)::<http://www.jpl.nasa.gov/orgs/>

Earth sciences

- [Windows to the Universe](http://www.windows.umich.edu/)::<http://www.windows.umich.edu/>

Education

- [Quest: NASA K-12 Internet Initiative](http://quest.arc.nasa.gov/)::<http://quest.arc.nasa.gov/>

Gazetteers

- [GEOnet Names Server](http://164.214.2.59/gns/html/index.html)::<http://164.214.2.59/gns/html/index.html>

Glossaries

- [Remote Sensing Glossary](http://lamont.lidgo.columbia.edu/rspage/glossary.html)::<http://lamont.lidgo.columbia.edu/rspage/glossary.html>

Meteors

- [Comets and Meteor Showers](http://medicine.wustl.edu/~kronkg/index.html)::<http://medicine.wustl.edu/~kronkg/index.html>

Remote-sensing images

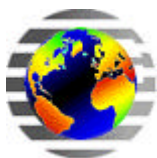
- [American Society for Photogrammetry and Remote Sensing \(ASPRS\)](http://www.us.net/asprs/)::<http://www.us.net/asprs/>
- [BATSE Image of the Galactic Center](#)

[Region](http://heasarc.gsfc.nasa.gov:80/cossc/descriptions/batse_galcen.html)::http://heasarc.gsfc.nasa.gov:80/cossc/descriptions/batse_galcen.html

- [EROS Selected Image Gallery](http://edcwww.cr.usgs.gov/bin/html_web_store.cgi)::http://edcwww.cr.usgs.gov/bin/html_web_store.cgi
- [JPL Public image archive](http://www.jpl.nasa.gov/archive/images.html)::<http://www.jpl.nasa.gov/archive/images.html>
- [Latest HST Observations. \(Hubble\)](http://www.stsci.edu/pubinfo/Latest.html)::<http://www.stsci.edu/pubinfo/Latest.html>
- [NASA JSC Digital Image Collection](http://images.jsc.nasa.gov/)::<http://images.jsc.nasa.gov/>
- [NASA's Image Access Homepage](http://www-pdsimage.JPL.NASA.GOV/PIA/)::<http://www-pdsimage.JPL.NASA.GOV/PIA/>
- [NASA's Observatorium](http://www.rspac.iov.nasa.gov/nasa/core.shtml)::<http://www.rspac.iov.nasa.gov/nasa/core.shtml>
- [NASA's Planetary Photojournal. Other](http://www-pdsimage.jpl.nasa.gov/cgi-bin/PIAGenPlanetPage.pl?Other)::<http://www-pdsimage.jpl.nasa.gov/cgi-bin/PIAGenPlanetPage.pl?Other>
- [NSSDC Photo Gallery](http://nssdc.gsfc.nasa.gov/photo_gallery/photogallery.html)::http://nssdc.gsfc.nasa.gov/photo_gallery/photogallery.html
- [Planetary image finders](http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/)::<http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/>
- [Regional Planetary Image Facility](http://ceps.nasm.edu:2020/rpif.html)::<http://ceps.nasm.edu:2020/rpif.html>
- [Sources of Earth and Planetary Photography](http://ceps.nasm.edu:2020/RPIF/RPIFsources.html)::<http://ceps.nasm.edu:2020/RPIF/RPIFsources.html>
- [Space Image Libraries](http://www.okstate.edu/aesp/image.html)::<http://www.okstate.edu/aesp/image.html>
- [Today@NASA](http://www.hq.nasa.gov/office/pao/NewsRoom/today.html)::<http://www.hq.nasa.gov/office/pao/NewsRoom/today.html>

Space sciences

- [Windows to the Universe](http://www.windows.umich.edu/)::<http://www.windows.umich.edu/>



Alexandria Digital Library: [ADL](#)

Last modified on 1998-06-17 at 00:46 GMT by [the Alexandria Web Team](#)

Lab Info

Personnel
Recognition
Tech Reports
Facilities

Project Groups

COPLINK
GIS
Interspace
MedInfo

Technologies

Analysis & Viz
Multilingual IR

Public Demo

CSQuest
ET-Map
GS-Map
Internet Spider
WormSpace

Private Demo

Cancer Space
Cat-Map
Geo-Map

**Java Agent
Workshop**

BFS Spider
GA Optimizer
Itsy Bitsy Spider



Artificial Intelligence Lab

Department of MIS, University of Arizona.

Head, [Dr. Hsinchun Chen](#)

University of Arizona, Department of MIS, Artificial Intelligence Lab, McClelland Hall 430, Tucson, AZ 85721.

TEL: (520) 621-2748, FAX: (520) 621-2433, <http://ai.bpa.arizona.edu>

For comments, please send to: [Webmaster](#)

Welcome

Lab Information

The UA/MIS Artificial Intelligence Lab, headed by Dr. Hsinchun Chen, consists of 3-5 Ph.D.-level researchers and 15-20 research scientists and assistants. It specializes in database integration, digital libraries, knowledge discovery, internet/intranet technologies, and intelligent information retrieval.

It has received multi-million-dollar funding from various government agencies including National Science Foundation (NSF), Advanced Research Projects Agency (ARPA), National Aeronautics and Space Administration (NASA), and National Institutes of Health (NIH).

Selected Ongoing Projects

- DARPA-funded ``The Interspace Prototype: An Analysis Environment for Semantic Interoperability" project, 1997-2000.
- NSF/NASA/ARPA-funded Digital Library Initiative (DLI) ``[Building the Interspace](#)" project, 1994-1998.
- NSF-funded ``[Internet Categorization and Search](#)" project, 1995-1998.
- NSF-funded ``Supplement to Alexandria DLI Project: [A Semantic Interoperability Experiment for Spatially-Oriented Multimedia Data](#) ," 1996-1998.
- [National Center for Supercomputing Applications](#) High-performance Computing Resources Grant ``[Information Analysis and Knowledge Discovery for Digital Libraries](#) " project, 1995-1997.
- [National Library of Medicine](#) funded ``Semantic Retrieval for Toxicology and Environmental Health Databases" project, 1996-1997.
- [National Cancer Institute](#) funded ``Information Analysis and Visualization for Cancer Literature" project, 1996-1997.

Projects Featured In

- ``Information Retrieval in Digital Libraries: Bringing Search to the Net," Featured in Volume 275 of *Science Journal*, January 17, 1997 (cover article).
- `` [Digital Libraries Computation Cracks Semantic Barriers Between Databases](#), " Featured in Volume 272 of *Science Journal*, June 7, 1996.
- `` [The Ultimate Indexing Job](#), " Featured in *Business Week*, Developments to Watch Column, August 12, 1996.
- `` [DLI Breaks the Semantic Barrier](#), " Featured in Volume 10, No. 2 of *Access High-performance Computing Magazine*, National Center for Supercomputing Applications, Summer 1996.



Home

[Copyright](#) © 1996. [Artificial Intelligence Lab](#), [The University of Arizona](#).

Last Updated: 10/31/97

DLI - Illinois:

- [Home Page](#)
- [IEEE Computer article](#)
- [Glossary](#)
- [SGML/XML Home Page](#), [SD Unit Notes in CS5604](#), [SoftQuad Products](#)
- Collections: [Publishers](#), [Software Companies](#)
- [Interspace](#), [concept extraction](#), [concept spaces](#), [term suggestion](#)
- [UIUC DLI Spring 1997 Partners Workshop](#)
- [Social Science Home Page: Fn Reqts](#)
- [DeLiver](#)
 - Before using DeLiver you should get one of the following 2 files and install it on your Windows 95/NT system. Be sure to have any version of Netscape closed after the download, when you do the install. These files are local to VT to save you the time of downloading as per the U. Ill. instructions. The Panorama versions each take about 1.9M for the install package but less than 1M for the C: drive installed version: [for Netscape 3](#), [for Netscape 4](#)
 - Explore the DeLiver pages, and try to answer the following questions.
 - What does the Help tell you about the system?
 - What is the coverage?
 - What are unusual services not provided by similar systems?
 - What is Panorama and what does it do to enhance WWW capabilities?
 - Can you use browsing to find the IEEE-CS articles (i.e., v. 29 n. 5) we looked at for this course?
 - Can you use searching to find the IEEE-CS articles we looked at for this course?
 - How does the presentation using WWW and Panorama differ from that you are familiar with (HTML, PDF)? What benefits are there from having Panorama?
 - What other interesting articles about digital libraries did you find?
 - Is the field specific searching of help?
 - Is the interface for DeLiver easy to understand? How could it be improved?

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Projects\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



ACHIEVEMENTS

[Progress Reports](#)
[Overview Papers & Talks](#)
[Publications & Reports](#)
[Workshops](#)
[Nat'l Synchronization Effort](#)

RESEARCH GROUPS

[Repositories \(testbed\)](#)
[Social Science \(User Studies\)](#)
[Semantic Research](#)
[Interspace Prototype](#)
[System Evaluation](#)

PARTNERSHIPS

[Publishers](#)
[Software Providers](#)

PEOPLE

[Contact Information](#)

TECHNOLOGY HIGHLIGHTS

[The 5 other DLI Projects](#)
[UIUC Digital Libraries](#)
[DL Related Information](#)
[Global Cultural Memory Project](#)

[Computing the Future](#)
[A National Research Council Report](#)



Note: DeLiver can be accessed by UIUC faculty, staff, and students.

DeLiver
USAGE STATISTICS: updated daily

The Digital Libraries Initiative (DLI) project at the [University of Illinois at Urbana-Champaign](#) has developed the information infrastructure to effectively search technical documents on the Internet. We have constructed a large [testbed](#) of scientific literature, are [evaluating](#) its effectiveness under significant use, and [researching](#) enhanced search technology. We are building repositories (organized collections) of indexed multiple-source collections and federating (merging and mapping) them by searching the material via [multiple views](#) of a single virtual collection.

Our testbed of [Engineering and Physics journals](#) is based in the [Grainger Engineering Library](#). We are placing article files into the digital library on a production basis in Standard Generalized Markup Language ([SGML](#)) from engineering and science [publishers](#). The [Research](#) section of the project is using NCSA supercomputers to compute indexes for new search techniques on large collections, to simulate the future world, and to provide new technology for the Testbed section.

The UIUC DLI is a recipient of a grant in the [NSF/DARPA/NASA Digital Libraries Initiative](#).

ONLINE SUMMARIES

From *Computer* theme issue on the US Digital Library Initiative, May 1996

A University of Illinois project is developing an infrastructure for indexing scientific literature so that multiple Internet sources can be searched as a single federated digital library.

Federating Diverse Collections of Scientific Literature

Bruce Schatz, William H. Mischo, Timothy W. Cole, Joseph B. Hardin, Ann P. Bishop, *University of Illinois*
Hsinchun Chen, *University of Arizona*

The most important recorded information medium on the Internet, and in the world at large, is the document. Although text might seem prosaic in contrast to multimedia objects, it is still the major medium for communicating information. Internet document retrieval can draw upon years of research results and practical experience in on-line information access as well as from traditional physical libraries. The technology for text information retrieval is far more mature than that for other media. Therefore, documents are also the best vehicle for investigating problems specific to digital libraries, such as the federation problem of making distributed collections of heterogeneous materials appear to be a single integrated collection.

The Digital Library Initiative (DLI) project at the University of Illinois at Urbana-Champaign is developing the information infrastructure to effectively search technical documents on the Internet. We are constructing a large testbed of scientific literature, evaluating its effectiveness under significant use, and researching enhanced search technology. We are building repositories (organized collections) of indexed multiple-source collections and federating (merging and mapping) them by searching the material via multiple views of a single virtual collection.

Developing widely usable Web technology is also a key goal. Improving Web search beyond full-text retrieval will require using document structure in the short term and document semantics in the long term. Our testbed efforts concentrate on journal articles from the scientific literature, with structure specified by the Standard Generalized Markup Language (SGML). Our research efforts extract semantics from documents using the scalable technology of concept spaces based on context frequency. We then merge these efforts with traditional library indexing to provide a single Internet interface to indexes of multiple repositories.

Our project focuses on developing a large-scale infrastructure adequate for solving real-world problems. The Testbed part of the project is based in the

University Library in a new facility that showcases engineering and science information and literature. We are placing article files into the digital library on a production basis in SGML directly from major engineering and science publishers. The National Center for Supercomputing Applications (NCSA) is developing software for the Internet version in an attempt to make server-side repository search as widely available as its Mosaic software made client-side document browsing.[\[1\]](#) The Research section of the project is using NCSA supercomputers to compute indexes for new search techniques on large collections, to simulate the future world, and to provide new technology for the Testbed section.

Federating distributed repositories

A traditional physical library is a single repository for materials from many sources to which a user comes seeking information. A repository is just an organized collection in which documents and other objects are indexed for effective search. The Net situation is quite different, since users can directly access the sources themselves. A digital library is a group of these distributed repositories that users see as a single repository.

It is difficult to support the federation of multiple physical sources into a single logical source. Part of the difficulty is in handling the text: Documents have differing structures and styles. Handling searches is also difficult. They must support different classification schemes so that sources can be indexed in various ways at different levels of detail.

Once Net retrieval is transparent, the digital library becomes similar to a typical physical library. Reference librarians help users locate information in a large collection by examining various indexes (search) and sources (display). Traditional libraries should naturally want to support digital libraries, since the range of indexes and sources available already far exceeds what a library building can physically house.

Figure 1 illustrates the major efforts in the Illinois Digital Library Project, (one of six participants in the overall DLI). The publishers, our partners, are filtering scientific literature and collecting it into repositories. Our Testbed is developing index technology for effective search and display of SGML repositories. The Internet part of our project is developing interface technology to support multiple indexes for multiple Internet repositories. This will let us evaluate the Testbed's effectiveness for thousands of users on thousands of documents. Our Research effort is developing semantic technology to support federated search across multiple repositories, using document content rather than structure.

Figure 1. Major efforts within the Illinois Digital Library Initiative project. (Click on the thumbnail to view a 18K GIF image.)



The Internet interface will incorporate Research technology that provides semantic federation of distributed repositories for scientific literature. The Testbed is the middle ground of our large-scale experiment, where we deploy the technology and evaluate the sociology.

Repositories for scientific literature (testbed)

Our Testbed provides enhanced access over the Internet to the full text of selected engineering journals, using SGML document structure to facilitate search. The Testbed is based in the Grainger Engineering Library Information Center, a \$30 million facility opened in March 1994 to showcase emerging information technologies. The Testbed was formally deployed in February 1996, with the production stream consisting of *Applied Physics Letters* from the American Institute of Physics. The production Testbed will gradually encompass the full collections of all publisher partners (listed below). Students and faculty at the University of Illinois, and then the other Big Ten universities, will be able to access the experimental digital library in accordance with our partner agreements.

Publishers and collection development

The testbed collection gathers articles directly from publishers in SGML format. These articles include the text and all figures, tables, images, and mathematical equations. Our publisher partners are committed to providing us with materials in the same time frame that they produce the print versions. That way we can place articles into our digital library before they reach the shelves of our physical library. We have chosen to manipulate SGML to the fullest extent possible, foregoing, for example, PDF (Portable Data Format), HTML (HyperText Markup Language), and ASCII, as later discussed. We are thus engaged in finding effective, scalable methods for the processing, indexing, retrieval, and display of structured document articles.

The testbed collection presently comprises over 4,000 articles from journals in computer science, electrical engineering, physics, civil engineering, and aerospace engineering. Publishers represented in this initial collection are

- the IEEE Computer Society,
- the Institute of Electrical and Electronics Engineers (IEEE),

- the American Physical Society (APS),
- the American Institute of Physics (AIP),
- the American Society of Civil Engineers (ASCE), and
- the American Institute of Aeronautics and Astronautics (AIAA).

Thus, for example, this issue of *Computer* will be in our collection before you read this article. Other professional societies (such as the American Association for Advancement of Science, which publishes *Science*) and commercial publishers (such as John Wiley & Sons) have committed to supply us with articles in SGML.

We believe that SGML will become the premier language of open document systems. SGML enables a system to treat documents as fine-grained objects to view, manipulate, and output. Tags delineate header (such as author, title, affiliation, and journal) and body (such as chapter, figure, table, and equation) structures. SGML's strength, in terms of retrieval, is that it reveals such deep document structure. SGML is becoming ubiquitous, but publishers are still mostly generating it as a byproduct of their production process, rather than as an integral part. In many cases we have been the first to actually display the SGML version of the published articles.

In the first phases of this project, we developed procedures for generating collections of SGML materials.[\[2\]](#) We process the heterogeneous SGML we receive from publishers into a federated repository of structured documents. Tags differ from one publisher to another. For example, every publisher has several author tags, which differ across publishers. We can federate some differences with simple syntactic transformations, such as AU or AUT or AUTHOR for the author tag. However, others reflect semantic differences and conventions. Yet the user wants to merely issue a query for **author**. We settled on an extension of the ISO 12083 Article Document Type Definition (DTD) for the project's canonical DTD. We are writing heuristic software for each DTD that maps publisher tags into our canonical set for indexing and retrieval. This tag normalization is our approach to structure federation.

To display journal articles, the testbed team has been working with SoftQuad to test and evaluate its Panorama SGML viewer. Figure 2 shows a portion of an SGML document as received from a publisher and displayed in this viewer. The bottom window is an American Institute of Physics document with federated tags, and the top window is how Panorama displays the SGML. Panorama can display all tagged parts directly: the text itself, titles (in this case, **PACS**), and equations. Style definitions for each DTD associate particular fonts and other aspects of display style with particular tag structures. At present, we are specifying these styles, but eventually publishers must define the styles just as they define the tags. Preserving the "look and feel" of the magazine layout is just as important as maintaining the article structure.

Figure 2. *Testbed SGML sample: (top) "cooked," after styles; and (bottom) "raw," with tags. (Click on the thumbnail to view a 72K GIF image.)*



Repositories and federated search

After adding an SGML document to the collection, we must index it for efficient retrieval. Our indexing techniques utilize the fine-grained structure of the documents so that, for example, users can search for a phrase solely within figure captions. We experimented with full-text retrieval using an SQL (Structured Query Language) engine before we settled on Open Text's Open Text Index search engine for indexing and accessing the DLI Project documents. This engine, tailored to SGML processing and retrieval, has the scalability to index large document stores (the Open Text Web Search Server presently indexes over 3 billion words and over 30 million links).

To evaluate database structures and retrieval effectiveness, we implemented a prototype client (written in Visual Basic) under Microsoft Windows. Figure 3 illustrates this prototype, which is our currently functioning testbed. The search query, shown in the upper overlaid window of the composite screen dump, finds **nanostucture** appearing only in figure captions. Selecting a retrieved article and viewing its short entry version shows that the caption of its Figure 2 contains that word. This figure, labeled as F2, can be viewed within the full article as shown in the window at the bottom of the screen dump. This service of our Engineering Library lets users access SGML document search within the context of other electronic retrieval services. Integrating bibliographic databases, on-line catalogs, local and remote periodical index databases, and the full-text SGML collection is vital to the Illinois digital library system.

Figure 3. *Current Testbed client prototype within the context of the Engineering Library. (Click on the thumbnail to view a 70K GIF image.)*



The information science literature shows that providing different search interfaces tuned to each search need helps users find information. In the current testbed interface, for example, users can use Boolean connectors to specify a phrase with different amounts of proximity or specify multiple phrases, and employ SGML tags to restrict the search to particular subparts of documents or

to selected information sources. They can also use a "word wheel" list to choose possible terms appearing in the collection and use preselected lists of "classic" documents to choose documents directly.

At present, we are placing the sources into a single repository maintained at our home site at the University of Illinois. There we process the SGML articles into a single index with federated tags. That index drives the search engine and the document store. Concurrently, we are training our publisher partners to build their own repositories. They can then process and index their own materials and run their own servers for searching across the Net. We expect a number of our publisher partners to establish such repositories, using our federated tag schema. Uniform searching across these will then provide a true testbed for distributed repositories of professional materials.

User and usage evaluation

To evaluate testbed users and usage, we combined a broad study of use with a deep study of social phenomena.^[3] Throughout the DLI project, we will observe how engineering work and learning activities intersect with using distributed, digital information. We will interview, individually and in groups, a range of potential and actual testbed users from the engineering community. We will conduct usability tests of various testbed components and versions, and experiment with economic models and charging mechanisms. Large-scale user surveys and testbed transaction logs will also yield extensive data.

Our sociological research has already yielded some valuable results. We asked focus groups of engineering students and faculty members how they use journals to support research and educational activities. The groups also discussed the biggest problems they have in identifying, retrieving, and using journal material. For example, focus group responses supported the relationship between journal structure and information needs and strategies. Many professors noted that figures, rather than abstracts or conclusions, were accurate indicators of whether they would be interested in the entire paper. They claimed that figures revealed what the authors had really done, as opposed to what they wished they had done. Several also reported that sometimes the equations were the only part they really needed to support certain work tasks. Several graduate students reported that the paper's bibliography indicated the paper's utility better than its title or author. In fact, sometimes they used the bibliography without reading the paper at all. The introductory paragraph was how most undergraduates decided whether an article was interesting, relevant, and written at the right level. These findings provide preliminary evidence that flexible interaction with document structure will enhance digital library effectiveness.

Inadequate information retrieval because of shallow semantics was a universal observation. Virtually all participants reported major difficulties in "getting the right words" to perform topic searches. This suggests a critical mismatch between the users' and the library's vocabulary systems. Students reported asking professors what "old" or "weird" term a particular database used to refer to the concept they wanted. They also searched their word processor's thesaurus

for suggestions of alternative terms and asked other library patrons if they could think of "better words."

Multiple views for distributed repositories (Internet)

The typical entry into a digital library is a specific search query, which matches some selected documents. The user can display these documents at different levels of detail and issue another search, so that a session gradually retrieves documents relevant to the user's needs. We are developing a multiple-view interface that enables transparent drag-and-drop between multiple indexes for multiple repositories. In addition, we are developing gateway technology to maintain the state and protocols for heterogeneous distributed repositories.

Interfaces to multiple indexes

Multiple views means that different searching techniques are available concurrently. We have built a prototype multiple-view interface, which will be used in the Internet version of the DLI Testbed to be introduced this summer. This interface incorporates a number of different view types, dynamically loading the actual data. We discuss this interface and compare the effectiveness of its views for different information retrieval purposes elsewhere.[\[4\]](#)

The view types integrate into a single framework many indexing styles and the major results from our projects. The primary views are subject thesauri, co-occurrence lists, and full-text search. A human indexer-a professional librarian-generates each subject thesaurus. The thesaurus arranges important terms in a subject area into a semantic hierarchy. A machine indexer-an automatic program-generates each co-occurrence list, which contains a more extensive list of terms, arranged by contextual frequency. Users can employ either one to interactively discover alternative search terms. They can then enter the new terms into a search engine for full-text search of the document collection.

Many studies, including our own, show that users have difficulty generating search terms that appear within the document collection. That is why our interface offers different types of term suggestion, then provides a high-quality search based on these new terms. Our experiments indicate that a typical user session is as follows. First the user consults the subject thesaurus for coarse-grained suggestions to identify the general subject area. Then the user accesses the co-occurrence lists for fine-grained suggestions to gather a list of desired terms. Finally, a full-text search retrieves the documents containing these terms.

Interactive term suggestion

The primary index for our initial testbed collection is INSPEC, from the Institution of Electrical Engineers (the British IEEE). It offers extensive coverage of electrical engineering, computer science, and physics. Our subject thesaurus is the INSPEC thesaurus, which has 10,000 terms in a

broader-narrower term hierarchy. The co-occurrence list is 200,000 terms from the INSPEC abstracts collection, which we arrange in concept graphs by co-occurrence frequency. The prototype interface lets users drag-and-drop suggested terms into the full-text search system constructed as part of testbed efforts.

The left side of Figure 4 illustrates the INSPEC thesaurus interface,[5] which provides a graphical display of the subject hierarchy of important concept terms. The user can specify a term and see broader and narrower terms, as well as graphically examine related ones. The graph is traversed by specifying **computer applications**, showing the narrower terms such as **deductive databases**, whose broader term is **database management systems**, and whose related terms include **logic programming**. The prototype interface enables terms so located to be passed into a search query.

Figure 4. *Current Internet prototype, showing the multiple-view interface. (Click on the thumbnail to view a 66K GIF image.)*



The right window in Figure 4, marked "Search Wizard," illustrates the co-occurrence lists.[6] Unlike the subject classifications, which are generated by professional librarians, these are automatically generated directly from the document content. The automatic generation employs co-occurrence analysis, which records how often a term occurs within the same sentence as another. The list of terms in the figure thus reflects terms that appear frequently with deductive databases. The concept graph, which relates term co-occurrence, is the collection of all lists. This approach is based on document content, rather than structure. Thus even in domains where the materials are unstructured, such as the Net, it captures more of the underlying concept semantics.

Stateful gateways and distributed repositories

To implement complete search sessions, we need techniques for providing state information within the Web. The Web is essentially stateless, with each transaction fetching a document, then stopping. Complete searching requires levels of stateful gateways to provide session history. First, each individual CGI-style gateway must maintain the state of the requests made to each server. Next, a higher level gateway must route queries to the appropriate servers and route results back to the appropriate clients. Finally, a search history must be kept for each user to record the session requests to each gateway. This function logically belongs in the client, which is where it is placed in our current design. However, it could potentially exist in any combination of client, server, or gateway depending on their functionalities.

Our distributed repositories prototype implements the levels of stateful gateways across a variety of protocols. The primary testbed search is an Open Text engine, with a custom protocol built on sockets. We implemented suggestion indexes using a Microsoft SQL engine. The SGML documents themselves reside in files accessed by an HTTP server. The interfaces to external search engines, such as the on-line catalog, follow the Z39.50 protocol. We even have an initial publisher repository, the experimental American Astronomical Society (AAS) server, connected via the CNIDR (Center for Network Information Discovery and Retrieval) Z39.50 software to test distributed repository protocols.

Our DLI project is providing major input to the next-generation server that NCSA is building. The server will move from a WWW document server using HTTP to a distributed repository host using multiple protocols. The server version 2.0, due the summer of 1996, will feature a modular protocol design and integrated security. We will later incorporate the work on stateful gateways into the server on the output end. The input end will incorporate the work on collection development. Thus the new server will eventually support session history and metadata checking. Later versions will also support security measures such as token passing, which our economic charging trials involving the NetBill software from the Carnegie Mellon DLI project will use.

We expect that during the course of the DLI project many of our publisher partners will create their own repositories. This will help the testbed evolve into a multiple-view reference system to distributed repositories. The repository management package will let other organizations and individuals make their organized collections searchable via a multiple-view interface.

Semantic federation across repositories (research)

The holy grail of information retrieval has always been deep semantics across heterogeneous sources. This is clearly expressed in the recent report^[7] on the research agenda for digital libraries from a workshop sponsored by the Information Infrastructure Technology and Applications (IITA) committee (the primary technical committee for setting National Information Infrastructure (NII) directions for federal government R&D investment). The report said that "deep semantic interoperability is the grand challenge for digital libraries." At its base, information retrieval technology matches terms specified by the user to terms occurring in documents in a digital collection. This term-matching is most effective when specialists access materials in their own subject area with precise terminology.

Concept spaces for scalable semantic retrieval

Broadening access requires different techniques to extend effective support to nonspecialists or specialists working outside their area of expertise. Specialists in even a closely related subject area usually cannot find relevant materials using current information systems. They know the concepts, but not the right

terms. Artificial intelligence and natural-language approaches that parse deep document structure to deduce semantics are usually effective only in narrow subject domains. The broad subject domains in our testbed in particular and the Net in general call for a different approach.

Our research focuses on methods that interactively provide the user with conceptual maps that offer alternative search terms. Interactive term suggestion, where the system suggests terms for the user to choose, can significantly enhance retrieval effectiveness. Although traditional library indexes provide some degree of term suggestion, effective Net searching requires automatic indexing. Many Net repositories are too small or specialized for a human indexer to provide the required level of fine-grained indexing. In addition, most digital repositories are "fluid," containing concepts and vocabularies too new or dynamic for controlled-vocabulary-based human indexing.

We have developed algorithms to extract concepts from documents so as to provide automatic indexing for semantic retrieval. The automatic indexing we are investigating generates concept spaces, which are concept graphs based on co-occurrence analysis.[\[8\]](#) Concept spaces lead to an approach for semantic federation across digital repositories, in particular towards solving the "vocabulary problem."[\[9\]](#) The vocabulary problem is the version of the semantic interoperability problem for text documents, the Grand Challenge of digital library research.

When digital libraries become widespread, every specialized community will have its own digital library of documents. This is already true for large professional communities. The increasing maturity of Net publishing will soon make it increasingly true for small amateur communities as well. The vocabulary problem will increasingly become an obstacle to the propagation of digital libraries.

Solving the vocabulary problem involves mapping a community library's specialized terms into the corresponding terms of other libraries being searched. Intersecting co-occurrence graphs from different domains provides an approach to concept-mapping across community libraries. Two graphs from different subject domains can be intersected by having the user specify a term common to both domains and displaying the graph around that term for both domains. This creates two term suggestion lists that can be compared for terms that are different in each subject domain but represent the same concept. In practice, the user needs to interactively cull the lists, but often discovers vocabulary that can be switched across domains.

Vocabulary-switching experiments

We are running large-scale experiments to investigate using co-occurrence graphs for vocabulary switching. These experiments build on smaller successful experiments for vocabulary switching in molecular biology.[\[10\]](#) Since part of our project is based at NCSA, we can use their supercomputers to perform experiments with realistic-scale collections. The experiments use algorithms for

vocabulary switching across subject domains based upon the co-occurrence frequency of phrases within documents to generate concept spaces.

Last year we generated the concept space used as the co-occurrence list for the term suggestion above from a collection of 400,000 computer engineering abstracts extracted from the INSPEC database.[\[6\]](#) By using one day (24.5 hours) of CPU time on the 16-node Silicon Graphics Power Challenge, we created a comprehensive concept space of about 270,000 terms and 4,000,000 links. During this two-week period, our application was the single largest user of NCSA supercomputers, beating out even the physicists and biologists.

This year we performed an order-of-magnitude-larger computation to generate multiple concept spaces for a large-scale vocabulary-switching experiment. We used some 4,000,000 abstracts from the Compendex database covering all of engineering as the collection. We partitioned it along classification code lines into some 600 community repositories. For example, (400) is civil engineering, (401) is bridges and tunnels, and (401.1) is bridges. We then generated a concept space for each individual repository and intersected the spaces to provide semantic mapping. This covers engineering fairly well and provides a large-scale test of mapping similar concepts across related domains with different terms. We used time during the testing phase of the new 64-processor Convex Exemplar at NCSA. The computation took roughly four days of CPU time over two weekends of dedicated machine usage, proving a good match for the shared-memory multiprocessor (SMP) architecture.

The scale of a repository in the Compendex experiment is, for example, on **bridges** rather than on **civil engineering**. This means that our prototype can realistically support dialogues across community repositories. Our system can display a list centered around a term like **fluid dynamics** in several domains. The user can then choose which terms in one domain to map into which terms in another. The user can thus interactively navigate between the spaces (see discussion of Interspace below). We are also experimenting with the concept space approach to semantic interoperability for other data types. For example, we will be switching texture images in spatial maps through a collaboration with the University of California at Santa Barbara DLI project. (This finds the co-occurrence frequency of textures in maps instead of phrases in documents.)

An example of vocabulary switching in our prototype might be:

I'm a civil engineer who designs bridges. I'm interested in using fluid dynamics to compute the structural effects of wind currents on long structures. Ocean engineers who design undersea cables probably do similar computations for the structural effects of water currents on long structures. I want you [the system] to change my civil engineering fluid dynamics terms into the ocean engineering terms and search the undersea cable literature.

Building the interspace

The encouraging results with concept spaces lead us to believe that we could build a complete information system supporting semantic retrieval. Since supercomputers can be used as a "time machine" to simulate future ordinary processing, ordinary personal computers will be able to generate similar concept spaces in years hence. This will provide essential infrastructure for the information systems possible on the Net of the twenty-first century. We are designing prototypes for community repositories on the Net that researchers outside the community can readily search. These prototypes will demonstrate the technological feasibility of "analysis environments," where researchers solve problems by correlating information from multiple sources across the network.

In the next century, information systems will directly support correlation of information across community repositories. Thus a user will deal with the Interspace rather than the Internet.[\[11\]](#) (The term Interspace indicates interconnection of spaces, just as Internet indicates interconnection of networks.) The fundamental interaction is intersecting concept spaces of related terms across subject domains, extracted from information spaces of interlinked objects comprising community repositories. Each individual and each community will have their own spaces. The Net will then enable information analysis, rather than merely document transfer as it does now.

The DLI project's prototype Interspace environment embeds concept spaces into the infrastructure of a network information system. Basic retrieval employs semantic matching to support information analysis. The user selects navigation paths of relevant objects, which the system records. The system then matches the user path to related paths across community repositories using semantic retrieval on concept spaces. We have completed the preliminary design and are beginning to implement the first prototype.

The Interspace prototype concentrates on the scalable technology for concept spaces:[\[12\]](#)

- semantic retrieval (using concept spaces for term suggestion),
- semantic interoperability (vocabulary switching across subject domains),
- semantic indexing (concept identification of document content),
- information representation (information units for uniform manipulation),
- and
- collaboration support (paths and grouping operations).

Since we are prototyping future Net functionality, we assume that distributed objects and syntactic interoperability have already become a mass infrastructure. Our choice of software tools--Smalltalk, CORBA, and ObjectStore--enables us to simulate building upon the future Internet-wide operating system. We are collaborating with research projects like the Stanford DLI project (object interoperability) and the CNRI (Corporation for National Research Initiatives) repository project (object naming). This will help us track and influence the object infrastructure necessary to support the concept

infrastructure we are prototyping.

Conclusion

In the coming years, we will continue to investigate whether concept spaces are a generic protocol that supports semantic interoperability across subject domains. We plan to construct complete analysis environments based on these protocols as prototypes of fundamental information infrastructure for the next wave of the Net. These future network information systems will support cross-correlation of information across distributed repositories.

We are optimistic that the Testbed efforts of the Illinois Digital Library project will influence the facilities for searching information on the Net with the help of technology evolved in our Internet version. We are also hopeful that the Research efforts will influence the facilities for analysis of information after the Internet becomes the Interspace.

Acknowledgments

Many people have contributed to the ideas and the prototypes discussed here. In particular, we thank Larry Jackson, Beth Frank, Eric Johnson, Jason Ng, Pauline Cochrane, Leigh Star, Roy Campbell, Charlie Catlett, Dorbin Ng, Kevin Powell, and Susan Harum. We also thank our many publishing partners for making their materials available to us on an experimental basis. This project is funded by NSF/ARPA/NASA Digital Library Initiative DLI grant to the University of Illinois IRI-94-11318COOP.

For further information on the Illinois DLI project, see
<http://www.grainger.uiuc.edu/dli/>.

References

1. B. Schatz and J. Hardin, "NCSA Mosaic and the World Wide Web: Global Hypermedia Protocols for the Internet," *Science*, Vol. 265, Aug. 12, 1994, pp. 895-901.
2. T. Cole and M. Kazmer, "SGML as a Component of the Digital Library," *Library Hi Tech*, Vol. 13, No. 4, 1995, pp. 75-90.
3. A. Bishop et al., "Building a University Digital Library: Understanding Implications for Academic Institutions and Their Constituencies," *Proc. Monterey Conf. on Higher Education and the NII: From Vision to Reality*, Coalition for Networked Information, Washington, D.C., 1995.
4. B. Schatz et al., "Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-Occurrence Lists for Information Retrieval," *Proc. First ACM Int'l Conf. Digital Libraries*, ACM Press, New York, 1996, pp. 126-133.
5. E. Johnson and P. Cochrane, "A Hypertextual Interface for a Searcher's Thesaurus," *Proc. Digital Libraries '95 Conf.*, 1995, available at <http://csdl.tamu.edu/DL95>.

6. H. Chen et al., "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project," *IEEE Trans. Pattern Analysis and Machine Intelligence* (special issue on digital libraries: representation and retrieval), to appear 1996.
7. "Interoperability, Scaling, and the Digital Libraries Research Agenda," report of IITA Digital Libraries Workshop, May 1995; available at <http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html>.
8. H. Chen et al., "Automatic Thesaurus Generation for an Electronic Scientific Community," *J. American Soc. Information Science*, Vol. 46, No. 3, Apr. 1995, pp. 175-193.
9. H. Chen, "Collaborative Systems: Solving the Vocabulary Problem," *Computer* (special issue on computer-supported cooperative work), Vol. 27, No. 5, May 1994, pp. 58-66.
10. H. Chen et al., "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Processing: An Experiment on the Worm Community System," *J. American Soc. Information Science*, to appear 1996.
11. B. Schatz, "Information Analysis in the Net: The Interspace of the Twenty-First Century," America in the Age of Information: A Forum on Federal Information and Communications R&D, sponsored by Committee on Information and Communications, National Science and Technology Council, 1995; http://www.hpcc.gov/cic/forum/CIC_Cover.html.
12. B. Schatz et al., "Building the Interspace: Overview and Architecture," <http://csl.ncsa.uiuc.edu/interspace.html>.

Bruce Schatz is principal investigator of the Digital Library Initiative project at the University of Illinois and a research scientist at the National Center for Supercomputing Applications, where he is the scientific advisor for digital libraries and information systems. He is also an associate professor in the Graduate School of Library and Information Science, the Department of Computer Science, and the Program in Neuroscience. He holds an NSF Young Investigator award in science information systems. Schatz has worked in industrial R&D at Bellcore and Bell Labs, where he built prototypes of networked digital libraries that served as a foundation of current Internet services (Telesophy), and the University of Arizona, where he was principal investigator of the NSF National Collaboratory project that built a national model for future science information systems (Worm Community System).

His current research in information systems is building analysis environments to support community repositories (Interspace), and in information science is performing large-scale experiments in semantic retrieval for vocabulary switching using supercomputers. Schatz received an MS in artificial intelligence from Massachusetts Institute of Technology, an MS in computer science from Carnegie Mellon University, and a PhD degree in computer science from the University of Arizona.

William H. Mischo is the director of the Grainger Engineering Library

Information Center at the University of Illinois at Urbana-Champaign and professor of library administration. He has been responsible for the design and development of several client-server information retrieval systems and has written several articles on interface design, including a benchmark 1987 ARIST (Annual Review of Information Science and Technology) chapter. He is the principal designer and supervisor of the Illinois Digital Library Initiative Testbed.

Timothy W. Cole is the system librarian for digital projects in the University of Illinois Library. From 1989-1994 he held the position of assistant librarian at the UIUC Engineering Library. While there, he helped to develop the microcomputer interface for end-user searching of bibliographic databases currently used at the UIUC Library. Cole is responsible for the acquisition, processing, and indexing of the SGML materials in the UIUC DLI database. Cole received both a BS in aeronautical and astronautical engineering (1978) and the MS in Library and Information Science (1989) from the University of Illinois at Urbana-Champaign.

Joseph B. Hardin has been the associate director for software development at NCSA since 1992. Previously he was the manager of the software development group and a visiting research associate at NCSA. He has taught in the department of Speech Communication at the University of Georgia at Athens. Hardin has received a number of grants and awards in the area of scientific visualization and network-based software development, and has spoken extensively on workstation tools for computational science, technologies for networked information systems, and the human dimensions of collaboration technologies in cyberspace. He served as cochair of the Second International World Wide Web Conference 94: Mosaic and the Web. He is also a founder and cochair of the International World Wide Web Conferences Committee, which is coordinating future WWW conferences.

Ann P. Bishop is an assistant professor in the Graduate School of Library and Information Science at the University of Illinois. On the DLI project, she heads the testbed evaluation and social science team. She is currently studying the impact of electronic networking on engineering work and communication and on community life. Recently completed collaborative research projects include a study of federal information inventory/locator systems (sponsored by the US Office of Management and Budget), and an assessment of the impact of high-speed networks on scholarly communication and research (sponsored by the US Office of Technology Assessment). In 1990, Bishop was a cowinner of the American Library Association's Jesse H. Shera Award for research.

Hsinchun Chen is an associate professor of Management Information Systems at the University of Arizona and director of the Artificial Intelligence Group. He is the recipient of an NSF Research Initiation Award, the Hawaii International Conference on System Sciences Best Paper Award, and an AT&T Foundation Award in Science and Engineering. He has published more than 30 articles about semantic retrieval and search algorithms. Chen received a PhD in information systems from New York University.

Readers can contact the authors at Digital Library Initiative Project, Grainger Engineering Library Information Center, 1301 W. Springfield Ave., University of Illinois, Urbana, IL 61801; e-mail dli@uiuc.edu. Hsinchun Chen's address is Dept. of Management Information Systems, McClelland Hall, University of Arizona, Tucson, AZ 85721; hchen@bpa.arizona.edu.

Guest Editors' Introduction

University of Illinois

University of California at Berkeley

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

[Computer](#) | [Computer Society home page](#)

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to webmaster@computer.org

Copyright (c) Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution must be obtained from the IEEE. For information on obtaining permission, send a message to whagen@ieee.org.

By choosing to view this document, you agree to all provisions of the copyright laws protecting it.



Glossary

ARPA (DARPA)

The Defense Advanced Research Projects Agency (DARPA) is the central research and development organization for the Department of Defense (DoD). It manages and directs selected basic and applied research and development projects for DoD, and pursues research and technology where risk and payoff are both very high and where success may provide dramatic advances for traditional military roles and missions and dual-use application.

Broad System of Ordering (BSO)

A general subject classification scheme, commissioned by UNESCO, intended to be a switching language among existing classification schemes and thesauri to make them mutually compatible on a general level. It provides about 4,000 subdivisions.

Collection Interface Agent

A program which interacts with the Collection Registry. For searchable collections (Z39.50, FTL, ...) it takes care of talking to the remote collection, submitting searches, fetching and processing results. It is also referred to as a CIA or a collection agent.

Collection Registry

The database in which descriptions of collections are stored.

Concept Space

Graph of terms occurring within objects linked to each other by the frequency with which they occur together.

Corporation for National Research Initiatives (CNRI)

A non-profit organization dedicated to formulating, planning, and carrying out national-level research initiatives on the use of network-based information technology. CNRI is concentrating on research and development for the National Information Infrastructure, working collaboratively with industry, academia, and government.

Derived Data

Data that was originally supplied in one form, but was converted to another form using some automated process.

DID

Document Image Decoding, a methodology for document recognition founded on statistical communication theory.

Digital Libraries

Digital libraries basically store materials in electronic format and manipulate large collections of those materials effectively.

Digital Library Federation

The Federation is comprised of leaders of fifteen of the nation's largest research libraries and archives and the Commission on Preservation and Access ([CPA](#)). A primary goal of the Federation is the implementation of a distributed, open digital library accessible across the global Internet. The library will consist of collections expanding over time in number and scope to be created from the conversion of digital form of documents contained in founding member and other libraries and archives, and from the incorporation of holdings already in electronic form.

DLI

Digital Libraries Initiative. Six research projects developing new technologies for digital libraries -- storehouses of information available through the Internet, -funded through a joint initiative of the National Science Foundation (NSF), the Department of Defense Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA). The projects' focus is to dramatically advance the means to collect, store, and organize information in digital

forms, and make it available for searching, retrieval, and processing via communication networks -- all in user-friendly ways.

ESRI

Environmental Systems Research Institute

European Digital Library Consortium (ERCIM)

The European Research Consortium for Informatics and Mathematics aims to foster collaborative work within the European research community and to increase cooperation with European industry. Leading research establishments from fourteen European countries are members of ERCIM.

Federated Repositories

Organized collections (heterogeneous databases) located in different places but searched transparently as one database via merging and mapping (federating).

HTML

Hypertext Markup Language. An [SGML](#)-based text markup language used on the WWW (World Wide Web).

IETF

Internet Engineering Task Force - an all volunteer organization responsible for publishing RFCs and Internet Standards.

IIPA

International Intellectual Property Alliance.

ITA

Information Infrastructure Technology and Applications

ITF

Information Infrastructure Task Force.

Information Visualization

A method of presenting data or information in non-traditional, interactive graphical forms. By using 2-D or 3-D color graphics and animation, these visualizations can show the structure of information, allow one to navigate through it, and modify it with graphical interactions.

Intellectual Property Usage License

The authority to employ a particular intellectual work in a designated way, possibly associated with other specifications of scope.

Intellectual Work

The object requiring an intellectual property usage license (i.e., an authored document). This object has an associated individual or agent with authority to grant such licenses.

Interoperability

The ability of software and hardware on multiple machines from multiple vendors to communicate.

Interspace

The Interspace is a vision of what the Internet will become, where users cross-correlate information in multiple ways from multiple sources. It is an applications environment for interconnecting spaces to manipulate information, much as the Internet is a protocol environment for interconnecting networks to transmit data. Navigating information paths and grouping related items is a fundamental operation. So is semantic retrieval and community classification, with interactive support for vocabulary switching across domains and subject indexing for amateur classifiers.

IR

Information Retrieval

ISO 12083

The new international standard for electronic manuscript preparation and markup. ISO 12083

speeds computerized text from author to publisher to typesetter without retyping and transforms the document into a searchable database.

JAVA

Java is a simple, object-oriented, distributed, interpreted, robust, secure, architecture-neutral, portable, high-performance, multithreaded, dynamic, buzzword-compliant, general-purpose programming language.

Machine Learning

The ability of a machine to improve its performance based on previous results.

Magic Lenses

This is an idea out of [Xerox PARC](#) where a region of the display (the "lens"), positioned by the mouse, is rendered in a special way. Lenses are specialized local views which might show labels where none were before, or handles on objects, or highlight certain subsets of items.

Metadata

Data about data. Includes information describing aspects of actual data items, such as name, format, content, and the control of or over data.

Middleware

Software that mediates between an applications program and a network. It manages the interaction between disparate applications across the heterogeneous computing platforms. The Object Request Broker (ORB), software that manages communication between objects, is an example of a middleware program.

Multiple View User Interface

Multiple views means that phrases can be drag-and-drop across each individual interface for each information source.

Multivalent Document (MVD)

A single document made of multiple layers of difference but intimately related material. Each layer is of homogeneous content, but is of a relatively limited scope and functionality. Layers have dynamically loaded program objects associated with them called behaviors, that manipulate the content, often communicating with other layers and other behaviors to achieve a desired effect.

NASA

National Aeronautics and Space Administration. NASA's mission is to advance and communicate scientific knowledge and understanding of the Earth, the solar system, and the universe and use the environment of space for research.

NetBill

The NetBill project at CMU's Information Networking Institute is designing the protocols and software to support network-based payment for goods and services delivered over the Internet. NetBill acts as a third party to provide authentication, account management, transaction processing, billing, and reporting services for network-based clients and users.

NI

National Information Infrastructure.

NSF

National Science Foundation. An independent agency of the U.S. government with the mission of promoting science and engineering.

NTIA

National Telecommunications and Information Administration. Responsible for the Information Superhighway.

OCR

Optical Character Recognition

Ontology

An explicit formal specification of how to represent the objects, concepts, and other entities that

are assumed to exist in some area of interest and the relationships that hold among them.

PAD++

Software which provides a virtual infinite extent, infinitely zoomable work surface, being developed under an ARPA grant at the University of New Mexico. Its multiscale interface, allowing interaction at many scales, is expected to allow the visualization of large scale information structures, and the organization of large and complex work activities. It is integrated with the Tcl/Tk prototyping environment and is being used as the development platform for the University of Michigan's Advanced User Interface ([AUI](#)).

PAT

Indexing software developed by the OpenText Corp. which serves as the basis for its products used for searching the WWW, intranets, etc.

Portals

Windows on a zooming work surface which can be used to bring distant regions close, to give simultaneous views at multiple scales, or, when given special active functionality, to create Magic lenses.

Query Planning Agent

A kind of Task Planning Agent. In many contexts, this means task planners who specialize in query tasks. Some select only from a library of existing plans for executing queries, others construct new plans.

Registration

The process of adding new descriptions to the registry database.

Registry Database

The database in which descriptions of agents (including collections) are stored. Also called the Conspectus database or the registry.

Remora Agents

An agent which, given a URL, will check the links of a homepage at a specified interval of time, check a specified homepage for any changes in the homepage at a specified interval and notify the user of any changes, and/or search a specified homepage for key phrases, results of which are emailed to the user.

Scaffolding

This concept is based on the idea that at the beginning of learning, students need a great deal of support, gradually, this support is taken away to allow students to try their independence. Providing support takes place in a number of ways - the way in which the selections are organized in a theme, the amount of prior knowledge activation that is provided, the way in which the literature is read by students, and the types of responses students are encouraged to make.

Semantic Retrieval

Searching for words within a concept space (graph of terms occurring within objects linked to each other by the frequency with which they occur together).

Semantic Zooming

In a multiscale interface like PAD++, normal, geometric zooming simply changes the size of objects in the view. In semantic zooming, objects change appearance or shape as they change size. For example, a growing dot will become a simple box, then a box with a one-word label, then a box with a longer label, then a rectangle filled with text and pictures. The goal is to give the most meaningful presentation at each size.

SGML

Standard Generalized Markup Language. SGML is a platform-neutral standard for creating documents and information archives--it's a series of rules that everyone can follow in order to make their documents publishable in different media (print, CD-ROM, the Web) and to make their documents readable with different kinds of computers. SGML is also a structure for storing

information which eases information-management and manipulation. It supports very powerful searching and allows large information repositories to be repurposed, broken down, and rearranged intelligently into individual documents. For more information, see [SGML info](#).

Testbed

A platform on which an assortment of experimental tools and products may be deployed and allowed to interact in real-time. Successful tools and products may be identified and developed in an interactive, evolutionary, interdependent process.

TestTiles

TextTiling is a method for partitioning full-length text documents into coherent multi-paragraph units.

Thesaurus

A controlled vocabulary with a syndetic structure within a circumscribed subject field used to organize material or information.

TileBars

An interface for document that allows the user to make informed decisions about which documents to view based on the distribution of search terms in the document.

URC

Uniform Resource Characteristic

Uniform Resource Citation

A collection of attribute/values about an object. Some of the values may be URIs. URCs are not formally defined, yet.

URI

Universal Resource Identifier - an address of some sort. See [IETF URI-WG](#) and the [W3.org](#).

URL

Uniform Resource Locator. URLs are a particular kind of URI.

URN

Uniform Resource Name. URNs are another kind of URI. Names are more persistent than Locations. A location may change, but a name rarely will.

Vocabulary Switching

The mapping of vocabulary from one discipline onto the vocabulary of another discipline.

[Z39.50](#)

The American National Standard Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection. The National Information Standards Institute (NISO), an American National Standards Institute (ANSI) accredited standards developer that serves the library, information, and publishing communities, approved the original standard in 1988 (referred to as Z39.50-1988 or Version 1). NISO published a revised version of the standard in 1992 (Z39.50-1992 or Version 2). ANSI/NISO Z39.50 defines a standard way for two computers to communicate for the purpose of information retrieval. Z39.50 makes it easier to use large information databases by standardizing the procedures and features for searching and retrieving information. Specifically, Z39.50 supports information retrieval in a distributed, client and server environment where a computer operating as a client submits a search request (query) to another computer acting as an information server. Software on the server performs a search on one or more databases and creates a set of records that meet the criteria of the search request as a result. The server returns records from the resulting set to the client for processing. The power of Z39.50 is that it separates the user interface on the client side from the information servers, search engines, and databases. Z39.50 provides a consistent view of information from a wide variety of sources and offers client implementers the capability to integrate information from a range of databases and servers.

[Index](#)[News](#)[Software](#)[Biblio](#)[XML](#)[XSL](#)[XLL](#)

The SGML/XML Web Page

Copyright (c) Robin Cover 1994-98. Last modified June 21, 1998. The SGML/XML Web Page lives at <http://www.sil.org/sgml/sgml.html>. Support for the development and maintenance of this SGML/XML Web Page is provided in part by [SoftQuad, Inc.](#) and by [the Summer Institute of Linguistics](#), to whom gratitude is acknowledged.

[[Search the entire SGML/XML database](#)] - [[Submit a Contact Address Form](#)]

The SGML/XML Web Page is a comprehensive online database containing reference information and software pertaining to the Standard Generalized Markup Language (SGML) and its subset, the Extensible Markup Language (XML). The database features an [SGML/XML news column "What's New?"](#) and a cumulative annotated [bibliography](#) with over 2000 entries. The collection contains over 2400 documents explaining and illustrating the application of the SGML/XML family of standards, including HyTime, DSSSL, XSL, XLL, XLink, XPointer, SPDL, CGM, ISO-HTML, and several others. These documents are accessible from the topical overview presented below, or from the [fully expanded contents listing](#) (Site Index) in a separate document.

Overview

The SGML/XML Web Page	<ul style="list-style-type: none"> • Site Index • Site Description
News	<ul style="list-style-type: none"> • What's New in the SGML/XML Web Page? • XML News Articles • XML Press News • Earlier News Highlights: [1997] [1996] [1995]
Introductions	<ul style="list-style-type: none"> • General Introduction to SGML • General Introduction to XML • SGML Frequently Asked Questions (FAQs) • XML Frequently Asked Questions (FAQs)
XML, XSL, XLL	<ul style="list-style-type: none"> • XML (Extensible Markup Language) • XSL (Extensible Style Language) • XLL (XLink and XPointer Languages)
Related Standards	<ul style="list-style-type: none"> • DSSSL • HyTime • Other Standards Related to SGML/XML
Applications	<ul style="list-style-type: none"> • General SGML/XML Applications • Academic Applications • Government and Industry Applications • Proposed XML Applications
Publications	<ul style="list-style-type: none"> • Essential SGML/XML Books • Comprehensive SGML/XML Bibliography • Journals, Newsletters and other Serials

	<ul style="list-style-type: none">• XML Books• XML Articles• XML Article Archive
Software	<ul style="list-style-type: none">• Public Software Tools for SGML/XML/DSSSL• XML Software Tools• XSL Software Tools• Commercial SGML/XML Software
Support	<ul style="list-style-type: none">• Industry Consortia, SIGS, Working Groups• SGML/XML Mailing Lists and Discussion Groups• Special Lists and Groups for XML and XSL• Commercial XML Support
Events	<ul style="list-style-type: none">• Conferences, Seminars, Tutorials, Workshops
Special Topics	<ul style="list-style-type: none">• SGML/XML Grammar• Architectural Forms and SGML/XML Architectures• Groves, Grove Plans, Property Sets• SGML/XML and (La)TeX• Miscellaneous
Contacts	<ul style="list-style-type: none">• Contact Addresses - Corporate Entities• Personal Home Pages - Some SGML/XML Experts

[**Note:** This Home Page is experimental/provisional. Readers are welcome to [send comments and criticisms](#) to the author. The [original home page](#) is still available for use by anyone who prefers it.]

UNIT SD

Course Notes on SD Unit --- SGML, Document Processing/Translation

SGML and Document Processing

Word Processing

Document Management

Markup, OHCO

SGML

Summary - SGML and Document Processing

- Word Processing - providing data
 - Document Management - bigger issue than IS&R (e.g., OIS)
 - Markup Approaches - use last 3
 - SGML - brief introduction
 - Advantages of SGML -> adoption
 - Document modeling - open problem
-

Document Translation

Electronic Publishing

Document Translation

<PRODUCTS>



SoftQuad Product Catalogue

▶ HoTMetal PRO 4.0
▶ HoTMetal App. Server
▶ HoTMetal Power Pack
▶ Panorama CDWeb
▶ Panorama Publisher
▶ Panorama Viewer
▶ Author/Editor 3.5
▶ Rules Builder
▶ Apex
▶ Sculptor

Pricing

\$129 US
\$169 CDN
£99 UK (+VAT)
995 FHT (France)
DM 259 +MwSt
(Germany)

System Requirements

Windows 95/NT
486/33 (Pentium Optimal)
16MB RAM
40MB available hard disk
space

Sales Inquiries

Email: sales@sq.com
Phone: (416) 544-9000
Toll-free Order Line
(North America only):
(800) 387-2777



SoftQuad

HoTMetal[™] PRO 4.0

Introducing HoTMetal Pro 4.0, the award-winning power tool for creating professional-quality Web pages, featuring:

- **Easy-to-use** features including the HoTMetal Site Maker
- **Great Flexibility** with three authoring environments to choose from
- **Quick use** with hundreds of sample images and files to get you started
- **Incredible Power** with a number of top of the line tools included for Web authors
- **Guaranteed perfect syntax** with automatic rules checking

Order

[Order HoTMetal Pro 4.0](#)

Order

[Upgrade to HoTMetal Pro 4.0 for only
US\\$49.95](#)

Free!

[Download a free Eval version of HoTMetal
Pro 4.0](#)More
Info[View the HoTMetal Pro 4.0 Feature sheet](#)More
Info[Learn more about the HoTMetal Power Pack](#)



[Academic Press, Inc.](#)

[American Association for the Advancement of Science \(AAAS\)](#)

[American Astronomical Society \(AAS\)](#)

[American Chemical Society \(ACS\)](#)

[American Institute of Aeronautics and Astronautics \(AIAA\)](#)

[American Institute of Physics \(AIP\)](#)

[American Physical Society \(APS\)](#)

[American Society of Agricultural Engineers \(ASAE\)](#)

[American Society of Civil Engineers \(ASCE\)](#)

[American Society of Mechanical Engineers \(ASME\)](#)

[Institution of Electrical Engineers \(IEE\)](#)

[Institute of Electrical and Electronics Engineers \(IEEE\)](#)

[IEEE Computer Society](#)

[John Wiley & Sons](#)

[DLI Home](#) | [DLI National Synchronization](#) | [DL Related Information](#)
[Glossary](#) | [Information Science](#) | [Interspace](#) | [Testbed](#) | [User Evaluation](#)

University of Illinois at Urbana-Champaign Digital Libraries Initiative
Comments to: External Relations Coordinator, [Susan Harum](#)
01/18/98



KEY PAPERS

Information Analysis in the Net: The Interspace of the Twenty- First Century.

Bruce R. Schatz, refereed White Paper for *America in the Age of Information: A Forum on Federal Information and Communications R & D*, July 6-7, National Library of Medicine. sponsored by CIC (Committee on Information and Communications) reporting to the Science Advisor to the President of the United States.

Federating Diverse Collections of Scientific Literature

Bruce R. Schatz, William H. Mischo, Timothy W. Cole, Joseph B. Hardin, Ann P. Bishop, and Hsinchun Chen, University of Illinois DLI Paper from *IEEE Computer Magazine*, theme issue on the US DLI, May 1996

Information Retrieval in Digital Libraries: Bringing Search to the Net

Bruce R. Schatz, *Science*, invited cover article, January 1997

Computation Cracks 'Semantic Barriers' Between Databases

Bruce R. Schatz, *Science*, about Interspace, june 1996

[[showcase](#)] - [[proposal](#)] - [[talks](#)] - [[papers](#)] - [[architecture](#)] - [[reports: technical / quarterly](#)] - [[the team](#)]

[[interspace home](#)] - [[canis home](#)]

comments/questions to webmaster@canis.uiuc.edu

last updated 01-27-98



Concept Extraction in the Interspace Prototype

Nuala A. Bennett, Qin He, Conrad Chang, Bruce R. Schatz

Digital Library Initiative (DLI) Project
CANIS : Community Systems Laboratory
University of Illinois at Urbana-Champaign
704 S. Sixth Street, Champaign, IL 61820
E-mail: {nabennet, hqin, t-chang2, schatz}
<http://www.canis.uiuc.edu>

Abstract

This paper describes the concept extraction for the Interspace Research Project. A comparison was undertaken of four parsers for noun phrase extraction - FastNPE, NPtool, Chopper, and AZ Phraser. FastNPE was found to be the fastest of the parsers, and NPtool the most correct in extracting noun phrases. Both were subsequently implemented into the Concept Extractor module of the Interspace Prototype, which is described in detail. Future work on the Concept Extractor will include image concept extraction and this is described in the final section.

FULL-TEXT PAPERS 

[[showcase](#)] - [[proposal](#)] - [[talks](#)] - [[papers](#)] - [[architecture](#)] - [reports: [technical](#) / [quarterly](#)] - [[the team](#)]

comments/questions to webmaster@canis.uiuc.edu
last updated 02-04-98



A logo for the "INTER SPACE ARCHITECTURE CONCEPT SPACE GENERATOR". The text "INTER SPACE ARCHITECTURE" is in blue, and "CONCEPT SPACE GENERATOR" is in red. A blue arrow points to the right from the end of the red text.

The purpose of the **Concept Space service** is to automatically generate domain-specific thesaurus subsets which represent the concepts and their associations in the underlying information corpus. Concept Space generation is based on a statistical co-occurrence analysis which captures the similarity between each pair of concepts (1).

The greater the similarity between concepts the more relevant they are to one another. Concept Spaces are used in a retrieval environment to assist users in performing functions such as term suggestion (2,3).

(1) Chen, H. and Lynch, K. J., "**Automatic Construction of Networks of Concepts Characterizing Document Databases**", *IEEE Transactions on Systems, Man and Cybernetics*, 22(5) 885-902, Sept./Oct., 1992.

(2) Schatz, B.R., and Chen, H., "**Building Large-Scale Digital Libraries**", *IEEE Computer*, Special Issue on Building Large-scale Digital Libraries, 29(5) 22-27, May 1996.

(3) Schatz, B.R., **Information Retrieval in Digital Libraries: Bringing Search to the Net**, *Science*, 275(5298), 327-334, cover story and lead article, January 17, 1997.

[[showcase](#)] - [[proposal](#)] - [[talks](#)] - [[papers](#)] - [[architecture](#)] - [[reports](#)] - [[the team](#)] - [[interspace home](#)] - [[canis home](#)]

comments/questions to webmaster@canis.uiuc.edu
last updated 1-5-98

Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval

Bruce R. Schatz
 Eric H. Johnson, ejohnson@uiuc.edu
 Pauline A. Cochrane
 Digital Library Initiative
 Grainger Engineering Library Information Center
 University of Illinois at Urbana-Champaign
 Urbana, IL 61801, USA

Hsinchun Chen
hchen@bpa.arizona.edu
 Department of Management Information Systems
 University of Arizona, Tucson

Abstract

The basic problem in information retrieval is that large-scale searches can only match terms specified by the user to terms appearing in documents in the digital library collection. Intermediate sources that support term suggestion can thus enhance retrieval by providing alternative search terms for the user. Term suggestion increases the recall, while interaction enables the user to attempt to not decrease the precision.

We are building a prototype user interface that will become the Web interface for the University of Illinois Digital Library Initiative (DLI) testbed. It supports the principle of multiple views, where different kinds of term suggestors can be used to complement search and each other. This paper discusses its operation with two complementary term suggestors, subject thesauri and co-occurrence lists, and compares their utility. Thesauri are generated by human indexers and place selected terms in a subject hierarchy. Co-occurrence lists are generated by computer and place all terms in frequency order of occurrence together. This paper concludes with a discussion of how multiple views can help provide good quality Search for the Net.

This is a paper about the design of a retrieval system prototype that allows users to simultaneously combine terms offered by different suggestion techniques, not about comparing the merits of each in a systematic and controlled way. It offers no experimental results.

Introduction to search terms

Effective information retrieval on an on-line document collection closely resembles the problem of effectively searching a library catalog by subject. As opposed to a known-item search, where you know what you want from the start and can provide precise title and/or author information, at the start of a subject search you only know that you want documents "about" something. The set of documents you come away with depends on the set of words you provide to the retrieval system and the ways in which it allows you to apply those words to the database. And even when you have a set of documents that appear relevant to your problem, you can never be sure that there are not more documents in the collection that you might find useful. This illustrates the completeness problem inherent in all information retrieval systems.

To attempt greater completeness of a set of retrieved documents, you might combine into one larger set the results of several different searches, each with a different search term. But here the problem of

precision arises: as you use more search terms to retrieve a set of documents (assuming a Boolean "or" or set union between each term used), the proportion of documents in that set that you would consider relevant to your problem tends to decrease. Even a retrieval set based on only one search term may contain a lot of irrelevant documents while still excluding many of the relevant documents in the collection. Doing effective information retrieval, then, largely depends on picking search terms that, in your own judgment, yield retrieval sets that contain a high proportion of relevant documents while excluding few, if any, of the other relevant documents in the collection. In short, you need to specify search terms that retrieve relevant documents with completeness and precision.

Picking the "right" search terms for your problem depends on how well you know the vocabulary used in the documents you want to retrieve. Therefore, you can typically get useful results when searching a collection of documents within your own field of expertise, but outside of that you will not have as much success even if you know the desired concepts, because you will not always know the correct terms to use for your search. Despite user knowledge that several terms within a particular domain may have the same meaning, known retrieval technology can only match terms provided by the searcher to terms literally occurring in documents or indexing records in the collection. While techniques like word stemming can improve retrieval somewhat, retrieval based on synonyms and other latent content of documents requires access to auxiliary search term databases outside of the actual document collection.

The use of term suggestion for information retrieval

Information specialists have long worked to bring the vocabularies used by searchers closer to those of the collections they maintain. Conventional library catalogs typically have the benefit of human indexers who assign "aboutness" to documents in the form of subject terms assigned to their bibliographic records. These come from collections of preferred subject terms, called subject thesauri, provided to indexers and searchers alike. The Library of Congress Subject Headings (LCSH) is probably the most widely known example, even though by many criteria it is not a particularly good one [6].

Indexing organizations, such as IEEE or the National Library of Medicine, that concentrate on specialized areas of knowledge tend to produce detailed subject thesauri that present terms in highly organized ways that reflect how subject experts in those fields understand those terms. Subject thesauri also provide synonym control, which reduces the number of different phrases used for a subject search to those used by the indexers of the document collection. The idea is to collapse a set of semantically equivalent terms into one preferred term that you can then use to actually retrieve bibliographic records. Otherwise, you may have the right concept in mind, which in principle should retrieve documents with sufficient completeness and precision, but in practice does not because the authors and indexers used a different term for that concept.

Besides providing vocabulary control for retrieval based on subject headings, thesauri also provide tracings between preferred terms that can suggest broader, narrower, and non-hierarchically related alternatives to the initial search term. Thesauri thus provide a dual function: they help you avoid search terms not used by indexers while they suggest other search terms which have distinct and precise meanings within a number of conceptual schema.

The idea of a thesaurus "suggesting" terms to the searcher is just that: the onus of selecting suitable terms for effective retrieval still rests on the searcher, but the thesaurus reveals much about the indexing

of the collection and thus makes the searcher's job much easier. Effective suggesting of terms can come from other mechanisms as well: browseable keyword and keyword-in-context lists, classification schemes, co-occurrence lists, and even bibliographic records with multiple subject term assignments. All of these mechanisms give you external structural cues when searching document collections. But they can only suggest terms to use; you must decide for yourself whether to use them or not.

Each of the term suggestion mechanisms listed above present to the searcher, in their own unique ways, the content of the document collection. Thesauri are constructed over time and change as the collection grows and the terminology of the fields it covers changes. Classification schema (e.g. the Dewey Decimal System) evolve in the same way, but have a more rigid structure in that they try to lay out all terms within a single grand hierarchical sequence. Bibliographic records cluster both thesaurus terms and classification terms around a single document to describe what it is about. All three of these are the result of intellectual effort by professional indexers, and provide indispensable term suggestion mechanisms for locating documents with the same kinds of "aboutness."

Professional indexers only include a term in a thesaurus or a classification scheme if it occurs in the literature that they index. Even then, it must endure a sort of canonization ritual, where it lives as a free text identifier (rather like a blessed free-text term) for a time. If it demonstrates enough usefulness there, and can also fill a gap in meaning in the present version of the thesaurus or classification scheme, only then will the lexicographer add it in the appropriate place. This tendency towards conservatism in thesaurus construction keeps the structure of thesauri stable and the terms within them viable over extended periods.

Co-occurrence lists, in contrast, are the result of intensive statistical calculations on how terms in documents in the collection occur together. The co-occurrence lists for a document collection are selected from a matrix containing the frequency of all pairs of terms occurring within, for example, the same sentence. Given a term, the list of all terms co-occurring with it can then be displayed in frequency order for use in interactive term suggestion. See [3] for a description of algorithms for term co-occurrence analysis. Currently, supercomputers are required to do the necessary computations to create such lists for large collections in a reasonable amount of time.

Different views of the same collection

Each of these term suggestion mechanisms is useful in its own way. A subject thesaurus presents "meaning," which terms are conceptually related to which, while co-occurrence lists present "context," which terms appear in context with which. They are both useful but for different purposes -- the thesaurus for precision, since the hierarchy is "correct," and the co-occurrence lists for recall since they show many more closely "associated" terms. Thus the thesaurus reflects "real" semantics at a gross level while the co-occurrence reflects "real" documents at a finer level (since all the words from the documents are included giving recent coverage but without human discretion as to their meaning).

By providing easy access to these various term suggestion mechanisms, we hope to encourage searchers to use them before attempting to access bibliographic records. This would reverse the current state of bibliographic as well as full-text retrieval, in which thesauri and other means of term suggestion (assuming they are even available) are typically accessed only after an initial bibliographic query yields either too few or too many hits.

In this paper we compare the use of two of the term-suggestion mechanisms described above, subject

thesauri and co-occurrence lists, and in doing so show how each complements the other. For our research, we have been using the INSPECTM Thesaurus as a sample thesaurus, and the "concept space" generated from 400,000 INSPEC indexing and abstracting records as a sample database of co-occurrence lists. These reflect roughly the same document collection indexed by humans and by computer respectively. (INSPEC is the indexing and abstracting service covering most of the research literature in physics, electrical engineering, and computer science. It is maintained by the Institution of Electrical Engineers, the British equivalent of the IEEE.)

Our research thus far has involved constructing a prototype system to provide interactive term suggestion to searchers of digital libraries, to be incorporated into the University of Illinois DLI testbed. Usability studies are planned during the next six to eight months to test the effectiveness of this system. Here we describe the possibilities that thesaurus and co-occurrence list browsing in particular and simultaneous use of multiple auxiliary views of a collection in general can offer to users of information retrieval systems. The examples which follow are taken from sessions with our prototype of such a system.

Documents and bibliographic records

Different term suggestion mechanisms present to the user very different views of a bibliographic collection. Their creation, as well as their use, are dictated by the very different requirements and expectations of various types of indexing, or, in terms of retrieval, by the different ways in which controlled vocabularies and natural language are used.

4/02/99/
An efficient infiniteness inference scheme in indefinite deductive databases
Journal Paper Principal: Theoretical/Mathematical English
Fu, F. S.; Kim, H. D.; Hanashiro, I. I.
Bellare, Piscataway, NJ, USA
IEEE Transactions on Knowledge and Data Engineering
Vol. 6 Iss. 5 p. 713-22
Date: Oct. 1994
Country of Publication: USA
ISSN: 1041-4347
CODEN: TKDE
Abstract: We introduce an inference scheme, based on the compilation approach, that can answer 'true,' 'provable-false,' 'indefinite,' or 'assumable-false' to a closed query in an indefinite deductive database under the generalized closed world assumption. The inference scheme proposed in this paper consists of a representation scheme and an evaluation process that uses one of two groups of positive indefinite ground
Database theory; Deductive databases; Inference mechanisms;
Knowledge representation; Query processing; Uncertainty handling
C6160K (Deductive databases); C4250 (Database theory);
U617U (Expert systems)

Figure 1. Sample INSPEC bibliographic record.

Figure 1 illustrates a typical INSPEC bibliographic record as displayed by our prototype interface. Document surrogates such as this are what we search for when doing retrieval, and what we would like thesauri and co-occurrence lists to help us find. Like the actual document it represents, the bibliographic record contains full title and author information, as well as the abstract as it appears in the document. Below the abstract in figure 1 (in the scrollable area) appear the indexing terms taken from the INSPEC Thesaurus (e.g. Database theory, Deductive databases, Inference mechanisms, etc.). At the bottom of the bibliographic record appear the classification codes and captions, which together constitute another term suggestion mechanism we plan to use but do not discuss in this paper.

The value added to a bibliographic record by having human indexers assign indexing terms to it can be seen in figure 1. The article is about deductive databases, and the term "deductive databases" appears in the title, subject terms (from the thesaurus), and the classification terms. The term "deductive database", a stem of "deductive databases", appears in the abstract. We could thus retrieve this record with the search term "deductive databases" with a title or subject search, or with a text search if the system we use supports word stemming. But such strong concurrence between title, text, and subject terms is rare. Assuming that the indexer did a good job of determining the aboutness of the article, it is also about database theory (an admittedly broad term), inference mechanisms, knowledge representation, query processing, and uncertainty handling. The list of classification terms adds expert systems as well. None

of the phrases "database theory", "knowledge representation", "uncertainty handling", or "expert systems" actually occur in the title or abstract, so a title or text search using any of those terms would not retrieve this record. Only a subject index search would. Similarly, "inference mechanisms" does not occur in the abstract or title, but "inference scheme" does; searching on the single keyword "inference" would retrieve this record, but would reduce the precision of the result set. The same is true for "query processing": "query evaluation" occurs in the abstract, but as with "inference" searching on the single keyword "query" would reduce the precision of the result set.

By using the controlled vocabulary supplied by a thesaurus, search-broadening techniques like word stemming and proximity searching are not necessary to retrieve records that cover the same concept, because indexers use the same term for that concept throughout the bibliographic database. This prevents records with other terms that represent different concepts, but match a stemmed or a word proximity query based on the desired term, from being retrieved, thus helping preserve the precision of the retrieved set of records.

A thesaurus, however, does not perfectly cover its subject domains, nor does it control all synonyms for the terms within them. "Inference scheme" and "query evaluation" from the sample bibliographic record in figure 1 are two such terms. This is due in part to the lag time involved in the canonization process described above, as well as the inability of even the most thorough lexicographer to catch every synonym for a concept in the literature. This is where computer-generated term suggestion mechanisms are most helpful. What they lack in semantic substance and conceptual precision they make up for in completeness and currentness.

Differences in both form and use between subject thesauri and co-occurrence lists are best illustrated with an example of what each might present during the term selection process. As we have already suggested, these should offer complementary yet comparable ways of retrieving records from a database of bibliographic records or full-text documents.

Subject thesaurus display

Figure 2 shows the INSPEC Thesaurus record for the preferred term "deductive databases", one of the terms used to index the sample bibliographic record in figure 1. Notice that it lists a number of Use For references, indicated by the "UF" tracing label; these correspond to references in the thesaurus from the terms "intelligent databases", "KBMS", and "knowledge base management systems" to the preferred term. Terms indicated by the NT tracing label are Narrower Terms (in this case none) of the term; by BT, Broader Terms; by TT, Top Terms; by RT, Related Terms (considered associated but not in a discernibly hierarchical way); and by PT, Prior Terms (like UFs, but used at some time previously to index items now indexed in some cases with the current preferred term).

The thesaurus record shown in figure 2 is essentially how it appears in the current printed edition of the INSPEC Thesaurus. It provides links to other terms, but the overall scheme into which it fits is difficult to discern. By flipping pages to other entries you can reconstruct the hierarchy and perhaps find RT tracings that interest you, but this is tedious and time-consuming.

deductive databases	
UF	intelligent databases KBMS knowledge base management systems
NT	database management systems
BT	database management systems
TT	computer applications file organisation
RT	active databases DATALOG knowledge based systems logic programming
PT	database management systems

Figure 2. INSPEC Thesaurus record for the preferred term "deductive databases".

The prototype thesaurus browser we have developed provides a visual representation of a thesaurus by reconstructing the disembodied conceptual schema scattered among the thousands of entries in a typical thesaurus. The lefthand side of figure 3 shows a partial view of the INSPEC thesaurus entry for "deductive databases" as displayed by our prototype thesaurus browser. (Figure 5 shows the thesaurus browser display as it might appear along with other windows on the computer screen.)

To construct a visual display for a typical entry, the thesaurus browser must use data from other thesaurus entries as well. Specifically, the thesaurus display for "deductive databases" on the lefthand side of figure 3 requires the browser to use NT tracings from the entry for "database management systems" as well as the entry for "information systems" and a number of other entries not shown.

To briefly explain the layout and function of the thesaurus browser display, scope notes and related term tracings (RTs) for the current term are taken directly from the thesaurus record and shown in the display under the phrase "Terms related to...". The hierarchy (compiled from BT, NT, and TT tracings) in which the current term occurs is the principal section of the display. A little triangle next to an entry in the hierarchy (here called an "expansion triangle") indicates that it has narrower terms, and whether the triangle is upright (pointing to the right) or tipped over (pointing down) indicates whether the hierarchy under the term is collapsed or expanded, respectively.

The thesaurus browser is completely hypertextual. You can click on any term you see on the display to see the entry for that term (which then becomes the current term). When the browser displays the entry for a thesaurus term, it automatically expands appropriate parts of the hierarchy and displays the term in boldface to clearly show you where it occurs, while leaving other parts of the hierarchy unexpanded. This yields a "fisheye" view of the term in the hierarchy, with the parts of the hierarchy near it expanded and the parts away from it left unexpanded.

See [5] for a more detailed description of how the thesaurus browser works and how it relates to the structure of the particular thesaurus it displays.

Co-occurrence list display

The right-hand side of figure 3 shows the co-occurrence list for the term "deductive databases" in the concept space generated from the INSPEC indexing and abstracting records mentioned above. The terms listed at right appear in decreasing order of the weight of their co-occurrence with the term "deductive

databases": "database theory", "logic programming", and "query languages" are the three highest weighted co-occurring terms. The further down the co-occurrence list a term appears, the less often it occurs along with "deductive databases" in the INSPEC database.

Like the thesaurus browser, the co-occurrence list browser is hypertextual: clicking on a term in a co-occurrence list yields the weighted list for that term. Both browsers allow you to navigate their respective "spaces" by following links in this way.

See [2] for a more detailed description of co-occurrence lists and an explanation of the algorithms used to generate them.

Comparing subject thesauri with co-occurrence lists

Unlike a thesaurus, there is no structure to the relationships in a co-occurrence list; only the weights of the links between the co-occurring terms. The statistical procedures employed to generate co-occurrence lists cannot discern a term's meaning and scope of application as humans can, and thus cannot discern whether a term is "broader" or "narrower" than another and assign a BT or NT relationship, respectively. Considering only the kinds of relations expressed in a subject thesaurus, the best that a co-occurrence list can manage is something like an RT (related term), where there is no discernible hierarchical relationship between the terms, though they are still considered to be associated in one way or another. Nor do co-occurrence lists make any attempt at synonym control or other kinds of vocabulary restriction.

These appear as shortcomings only if you try to use co-occurrence lists the same way you would use a thesaurus. A thesaurus gives precision to the meanings of the terms you use for retrieval, while a co-occurrence list aids in recall by revealing the context in which terms are used in the collection, be they thesaurus terms or other "uncontrolled" terms. In this way they offer terms to use in searches of fields of bibliographic records containing unrestricted vocabularies, such as the title or abstract, or even to full-text searching in systems that offer it. They may also aid in searching the thesaurus itself by suggesting terms not visible in the currently displayed hierarchy.

Thesaurus display (partial) for deductive databases:

computer applications
 ▶ engineering computing
 ▶ expert systems
 handicapped aids
 ▶ humanities
 ▶ information science
 ▼ information systems
 ▼ database management systems
 active databases
 database machines
deductive databases
 ▶ distributed databases
 object-oriented databases
 relational databases
 statistical databases
 temporal databases
 very large databases
 visual databases
 engineering information systems
 geographic information systems
 ...
 .
 .

Terms related to deductive databases:

active databases
 DATALOG
 knowledge based systems
 logic programming

Co-occurrence list (partial) for deductive databases:

database theory
 logic programming
 query languages
 query processing
 knowledge based systems
 relational databases
 deductive database
 object-oriented databases
 inference mechanisms
 formal logic
 knowledge representation
 data integrity
 logic programs
 integrity constraints
 DATALOG programs
 knowledge bases
 query evaluation
 knowledge base
 Prolog
 deductive database system
 expert systems
 database system
 logic programming languages
 distributed databases
 deductive database systems
 transitive closure
 very large databases
 query language
 class 0
 active databases
 recursive queries
 ...
 Ramakrishnan, R.
 Henschen, L.
 Han, J.
 Subrahmanian, V.

Figure 3. Comparative views of thesaurus and co-occurrence list content for the term "deductive databases" (arrangement of terms has been altered somewhat to facilitate comparison between lists). Lines connect terms that occur in both.

Earlier we suggested that thesaurus terms appearing in a co-occurrence list for a given term were akin to related terms rather than broader or narrower terms, and this example illustrates such a case. All RTs (terms listed at lower left under the heading "Terms related to deductive databases") are in the co-occurrence list (arguably, the term "DATALOG" occurs in the co-occurrence list as part of "DATALOG programs".)

Recall that in the sample bibliographic record in figure 1, "database theory" and "inference

mechanisms" appear with "deductive databases" as subject terms used to index the record. The appearance of the former two terms in the co-occurrence list for "deductive databases" is in part due to their co-occurrence in this indexing record.

The suggestion of other terms, in taking you to different parts of the thesaurus, can suggest yet more search terms. Figure 4 illustrates the result of entering "inference mechanisms", suggested by the co-occurrence list, into the thesaurus. You can play the thesaurus and the co-occurrence lists together this way to get as many search terms related to your problem as you want. This is explored in the next section.

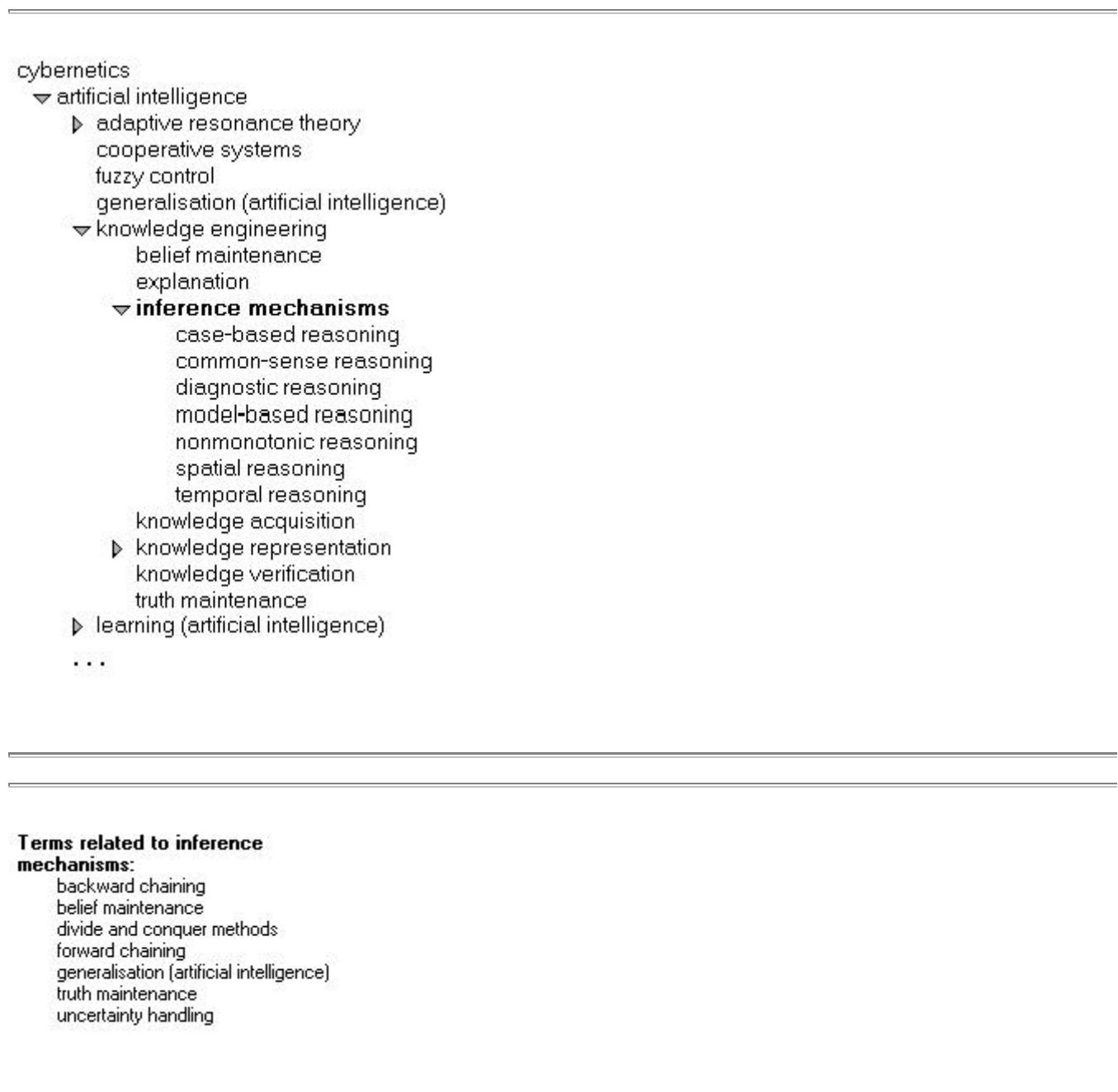


Figure 4. Partial thesaurus browser display for the term "inference mechanisms".

Another term near the top of the co-occurrence list for "deductive databases" is "query evaluation", which appears in the abstract of the sample bibliographic record in figure 1 (though it is not visible because it is in the bottom part of the abstract, scrolled out of view). "Query evaluation" is not a thesaurus term, probably because "query processing" covers the same concept, even though there is no USE tracing in the thesaurus from the former to the latter. "Query evaluation" is therefore a free-text term suggested by the co-occurrence list which can be used in a free-text search to retrieve the bibliographic record.

This in turn suggests a role for co-occurrence lists in thesaurus construction, in that they can complement the work of the human lexicographer mentioned earlier. In the "query evaluation / query processing" example above, the co-occurrence list for the preferred term "query processing", by suggesting to the lexicographer a USE tracing from "query evaluation", could help improve the synonym control provided by the thesaurus.

deductive databases
database theory
query processing
logic programming
formal logic
recursive queries
deductive database
compiled formula
query languages
generalized closed world assumption
proof procedure
allowed databases
negative rules
ground clauses
annotated logics
 . . .

Nam, Y.
Lu, J.
Han, J.
Barback, M.
Toroslu, I.
Dong-Hoon Choi.
Da Costa, N.
Franzen, M.

Figure 5. Partial co-occurrence list for the author Henschen, L.

Yet another bonus of a co-occurrence list is that, because it lists the context of all text items in a collection, it also lists the context of author names, as well as the author names that co-occur with conceptual terms, as shown at the bottom of figure 3. This is outside of the scope of a thesaurus, which at best offers named phenomena and devices, e.g. "de Broglie waves", "van de Graaff accelerators". Names include more than authors (e.g. programming languages such as "Prolog") and other proper nouns. Human indexers, in attempting to include only concepts in subject thesauri, leave out proper nouns, which are very useful search terms indeed! The computer programs that generate co-occurrence lists include personal names and other proper nouns because they can recognize strings but not the

human-understood meanings of them.

Having an author name in the conceptual area you are searching gives you a powerful search term, with which you can quickly gather a set of fruitful bibliographic records, from which in turn you can gather additional subject terms and keywords, as well as other authors. Assuming that the author has a somewhat narrow research field, which is typically the case in academic research, the retrieved record set will have good precision as well.

One of the authors in the co-occurrence list in figure 3, "Henschen, L.", is one of the authors of the article whose bibliographic record is in figure 1. Figure 5 illustrates the result of entering this author into the co-occurrence list display. Compare this with figure 3. In this way, and in ways demonstrated above, the semantic looseness and lack of structure of co-occurrence lists provides a powerful way to move among and between more structured information spaces.

Using multiple simultaneous views

Much can be inferred from the previous example about the general usefulness of having different kinds of views of a collection, as well as the usefulness of views that offer different degrees of semantic substance and structure. Recall that the manual thesaurus browser arranges terms according to human conceptual relationships, while co-occurrence lists show terms (whether in the thesaurus or not) that are contextually related to a given term. They are both useful but for different purposes -- the thesaurus for precision since the hierarchy is "correct" and the co-occurrence lists for recall since they show many more terms.

When these and other term suggestion mechanisms are combined in our prototype multiple view interface, they allow you to quickly drag and drop terms from one view to another. [4] The techniques for searching offered by each view are thus available simultaneously. It is well known in the information science literature that there are different kinds of search needs and that user behavior is best facilitated by providing different search interfaces tuned to each particular need. Figure 6 illustrates part of the session from which we gathered the previous five figures.

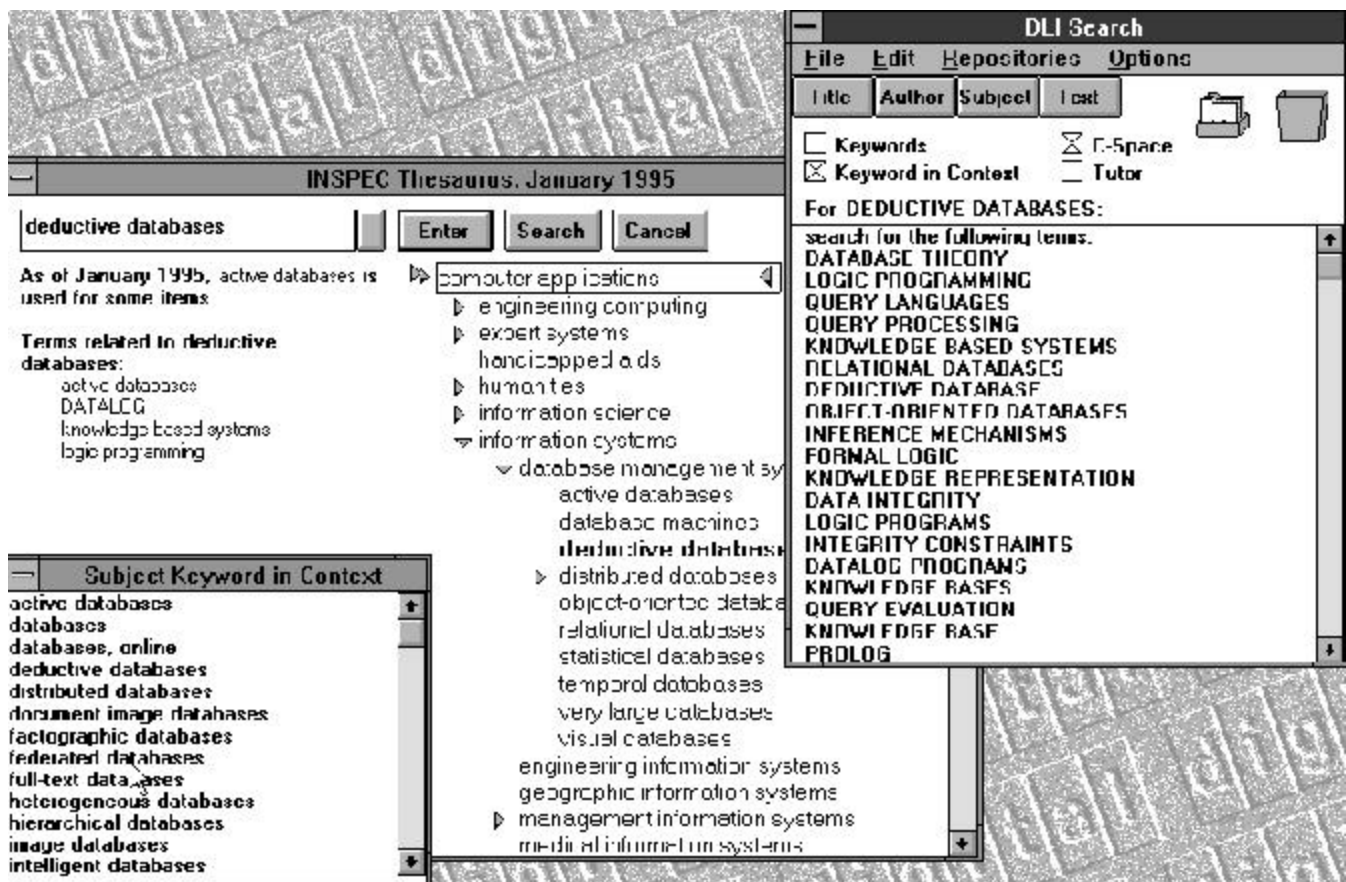


Figure 6. Session screen showing actual appearance of thesaurus and co-occurrence list displays. The keyword in context display is also shown, in the lower lefthand corner.

Multiple views are the user interface principle around which we are building the prototype described in this paper. In the coming year we are extending it to become the "Web interface" for the Digital Library Initiative (DLI) project at the University of Illinois at Urbana-Champaign. This large-scale digital library testbed is building a collection of SGML documents, consisting of articles from magazines and journals in engineering and science obtained in a direct pipeline from major technical publishers. The multiple view interface will support easy combination of term suggestion from different sources followed by full-text search of the document collection. As the SGML collection primarily covers computer science, electrical engineering, and physics, we are using the INSPEC subject thesaurus and parts of the Dewey Decimal Classification supplemented with co-occurrence lists from the bibliographic areas being covered. Since the plans in the coming three years are to build a testbed with 100,000 documents from many publishers and 100,000 users across the Big Ten universities, the extensive sociological evaluation will provide a large-scale test of the utility of the multiple view principle for information retrieval. More details on the DLI project are contained in [7].

The true utility of multiple views will only become apparent when we have many sources to combine seamlessly. The DLI testbed efforts will experiment with 2 or 3 sources of 2 or 3 kinds. An on-going experiment in the DLI research efforts (the longer-term portions of the project) will greatly extend this to an experiment with hundreds of term suggestion sources rather than just a few as we have now. Over the next few months, we plan to generate co-occurrence lists for all of engineering. Some 3 million abstracts from Compendex (Engineering Index), which has broad coverage across all of engineering, will be used as materials for the generation of fine-grained co-occurrence lists. These materials can be thought of as

15 broad areas with 200,000 abstracts each (roughly the size of the INSPEC collection which covers only 2 areas). However, we plan to divide these along Compendex class code lines to get much smaller areas covered by subjects like "bridges" rather than the much larger areas covered by all of civil engineering, for instance. This will yield hundreds of co-occurrence lists relevant for term suggestion across our user population (faculty and students in engineering at the University of Illinois). Since the INSPEC experiment showed that computing a list for an area required roughly a day (24 hours) of supercomputer time, our experiment is possible only because the NCSA (National Center for Supercomputing Applications) is granting us special time on their newest computer (Convex Exemplar) during its testing phase.

During the coming year we will rewrite the DLI Web interface from its current implementation in Microsoft Visual Basic to a production version implemented in JavaTM. Java is a new protected execution environment

being bundled into Web browsers such as NetscapeTM that enables dynamic loading of programs and data across the Internet. The plan is for our Web interface to have generic support for term suggestion in general and for subject thesauri, co-occurrence lists, and full-text search in particular. When organized collections become more common on the Web, this leads to the possibility of dynamically loading the term suggestors for all the collections desired by the user on a demand basis. Transparent Multiple Views will be a necessity when users casually perform search sessions spanning hundreds of thousands of fine-grained subject domains. The experiments in the Illinois DLI project on interactive term suggestion will give the first taste and develop the first good technology for Search in the Net.

Acknowledgments

We wish to thank Jim Ashling of IEE England and Michele Day, Manager of the INSPEC US Office, for making the INSPEC Thesaurus available to us in machine-readable form, and for permission to use the INSPEC A&I records for generating term co-occurrence lists.

This project is funded by NSF/ARPA/NASA Digital Library Initiative DLI grant to University of Illinois IRI-94-11318COOP.

References

- [1] R. Allen (1994). Navigating and Searching in Hierarchical Digital Library Catalogs, Digital Libraries '94 Proceedings, College Station, TX June 19-21, 1994, pp. 95-100.
- [2] H. Chen, B. Schatz, T. Ng, J. Martinez, A. Kirchhoff, C. Lin (1995) A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project, submitted to IEEE Trans Pattern Analysis and Machine Intelligence, Special Issue on Digital Libraries: Representation and Retrieval, 15pp.
- [3] H. Chen, B. Schatz, T. Yim, D. Fye (1995) Automatic Thesaurus Generation for an Electronic Community System, Journal American Society Information Science 46 (3): 175-193, April 1995.
- [4] E. Johnson (1995) Extending an Interactive Thesaurus by Dragging, ACM SIGLINK Newsletter, pp. 16-17 (September).
- [5] E. Johnson, P. Cochrane (1995) A Hypertextual Interface for a Searcher's Thesaurus, Digital Libraries '95 Proceedings, Austin, TX June 11-13, 1995, pp. 77-86.
- [6] M. Kirtland, P. Cochrane (1981) Critical Views of LCSH: A Bibliographic Essay. ERIC Document ED 208900.

[7] B. Schatz, B. Mischo, T. Cole, J. Hardin, L. Jackson, A. Bishop, L. Star, P. Cochrane, H. Chen (1996) Digital Library Infrastructure for a University Engineering Community: Towards Search in the Net via Structure and Semantics, submitted to IEEE Computer, Special Issue on Large-Scale Digital Libraries, 12pp.

[Go back to the May 2-3, 1996 Partners Workshop Page](#)

[Return to the DLI Homepage](#)

University of Illinois Digital Library Initiative

Partners Workshop

May 2-3, 1996

Fifth in the Digital Libraries Initiative Workshop Publications Series

[Letter from NSF Program Official](#)

[DLI Project titles and homepages](#)

[Testbed Technologies for the Distributed Repository Model - Flow Chart](#)

[UIUC DL Testbed Database & Client Technologies](#)

[UIUC DLI Testbed Processing Customization](#)

[Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval](#)

[Observations of Bibliographic Tools Use at the Grainger Engineering Library](#)

[Software Implementation and User Registration](#)

[Object Worlds and Shifting Infrastructure: Building a Digital Library for Engineers](#)

[Summary of DLI Prototype Usability Tests](#)

[Targeted Study of the First Group of DLI Users: The Physicists](#)

[Future Research: Linking Social and Technical Aspects of Digital Libraries](#)

[HPCwire: The Text-on-Demand E-zine for High Performance Computing](#)

[Agenda](#)

[List of Participants](#)

[Minutes](#)

[Partners Address List](#)

Welcome to the DLI Social Science Team Home Page

[Index](#)[Diary](#)[Internal Reports](#)[Completed Papers](#)[Papers in Progress](#)[Conference
Presentations](#)[Site Visit and
Quarterly Reports](#)[Main DLI Page](#)[Web Client-
DeLiver](#)

send comments or questions to:
l-neuma1@uiuc.edu

This page consists of links to working papers and [a brief overview of the social science team](#) projects that we have been working on as the social science team for the NSF/ ARPA/ NASA [Digital Library Initiative project](#) being conducted at the University of Illinois.

Our subgroup of the Illinois Digital Library Initiative (DLI), the Social Science Team, has a mandate to study potential and actual use of prototype systems that other subgroups of the DLI build. In addition, we study the web more generally, and how the work of engineers and other scientist will be impacted by and will impact the growth of the information infrastructure.

Our Social Science Team has articulated, from the beginning, a commitment to a three way relationship between users, designers and social scientists, following in a general way the principals of participatory design. We are especially concerned with trying to fit our formative evaluation work to the ideal of this method: close contact and communication between designers and users via a series of mutually generated, iterative prototypes. To this end, we have conducted usability studies with the emergent testbed; observations of current users of electronic systems in the traditional library and beyond; focus groups, interviews and observations with faculty and staff who are potential users; and as use of the testbed continues to grow, transaction log analyses. One of our major concerns is finding a means to fit these all together.

Members of the team include: Ann Bishop, primary investigator; [Leigh Star](#), investigator; Emily Ignacio, graduate assistant; Laura Neumann, graduate assistant; [Cecelia Merkel](#), a graduate assistant; [Bob Sandusky](#), graduate assistant; and Eric Larson, graduate assistant.

Toward Functional Requirements for the Digital Library

Toward Functional Requirements for the Digital Library (based on focus group interviews with faculty and students)

20 December, 1994 - Draft

*** = most important**

1. *--DL should allow the user to follow citation links forward and backwards (preferably to full documents; otherwise to location information)
2. *--DL should include an online meta-thesaurus that users can search and browse. The meta-thesaurus should integrated existing thesaurae across disciplines. It should also allow users to incorporate their own terms and edit existing terms. The thesaurus should allow users to type in a few letters of a word and see corresponding terms, should suggest or reference alternatives to users' terms. Users should be able to view no. and type of documents associated with terms and link automatically from thesaurus terms to documents.
3. --DL should include an acronym list to help users identify and search for terms.
4. --Users should be allowed to save a record of their searches and what each search retrieved.
5. *--Users should be able to search and view individual components of a document (e.g., author/title, abstract, figures, references) in a dynamic manner, specifying for each search which elements should be searched and which displayed.
6. *--Users should be able to customize their interfaces so that search options, procedures are presented in the manner they like best.
7. --Users should be able to view an overview description of the contents of the testbed.
8. *--Display of full documents should mimic the look and feel of the article's print version in both page layout and page "flipping" (i.e., users should be able to view multiple pages at once and in quick succession)
9. *--Users should be able to design and launch their own user profiles for any particular search session, defining what they want and how they want to get it.
10. *--Users should be able to move easily from query to results and back, rather than moving in the linear fashion common in online systems today, revising a query upon viewing results without having to lose sight of the results or start a query over.
11. *--Users should be able to easily create personal electronic article collections as a subset of the DL, manipulate and share that collection.

12. --Users should be able to define and set their own access points for searching personal collections derived from the DL
13. --DL should allow on-screen highlighting, bookmarking to help in reading full articles.
14. --Users should have access to DL from home and office.
15. --DL should allow printing of full documents
16. *--Users should be able to jump to and view individual document components. They should be able to skim, open, or skip individual document components.
17. --DL should facilitate colleague networks: allow users to view list of contact info for authors, construct mailing list of colleagues to send documents to.
18. *--DL should provide complete and intuitive online help: help balloons, full documentation, help with basic computing, gripe button, sample searches.
19. --DL should allow users to make own links to commonly used external network resources (e.g., pre-print databases, listservs)
20. --Search parameters should include physical location of material not available online
21. --DL should facilitate browsing at shelf, ToC, and article levels: users need overview and zoom capabilities.
22. *--Interface should resemble a "natural topography" of the information landscape... with a physical layout, dynamically defined (topic, material type, author, etc.)
23. *--DL should allow serendipitous discovery of "other books on the shelf," "other articles in the journal." Perhaps set browse mode as a purposeful search option: by call no., journal title, etc.

University of Michigan Digital Library Activities

DLI General Information

- [Home Page](#)
- [IEEE Computer article](#)
- [Introduction](#)
- [Current Status](#)
- [Technologies](#)
- [Agents, Ontologies](#)

Campus Strategy

- [Strategies for DL Development](#): partnership of
 - [University Library](#)
 - [Information Technology Division](#)
 - [School of Information](#)
- combine: R&D; technology infrastructure; content access & user services; outreach
- shift to 21st century library model
 - user-centric, collaborative teams, global reach
 - distributed collections, heterogeneous access protocols, just-in-time information delivery
 - mixed funding models, value = access + services
- Registries: [SGML database](#), [subject specialist librarian created digital libraries across Internet](#)
- [Electronic Reserve Shelf](#)
- [Knowledge Navigation Center](#): develop and support teaching and learning projects
- Questions:
 - How does the infrastructure at U. Michigan compare to that at your university?
 - How does this strategy relate to previous services of libraries?

Projects

- [JSTOR](#): Journal Storage: over 1.2M pages
- [Making of America](#): with Cornell - 5K volumes, [D-Lib article](#): scanning, OCR, SGML encoding, [tif2gif](#), interface
- [Museum Educational Site License Project](#): see also V. 5 N. 8 Oct. 1996 [Information Technology Digest](#)
- [Humanities Text Initiative](#) and [Collaboratory for the Humanities](#)
- [Pricing Electronic Access to Knowledge](#) (PEAK) - 1100 Elsevier journals with flexible pricing
- [Papryology](#)
- [Middle English Compendium Demo](#)
- [American Verse](#)
- [TULIP](#)
- [NDLF](#)
- Questions:
 - Which of these projects do you find most interesting? Why?
 - Which of these projects should your university become involved in?

Technical Approaches

- [see especially 1996 Ann Arbor Conf. on Electronic Records R & D](#)
 - Problem scenarios (see bullet list under **The Importance of Digital Preservation**)
 - Research questions (see **The 10 Research Questions**)
 - Research results: possible, requires changes and new types of efforts (see bullet list under **Research Projects and Results**)
 - [International Council on Archives](#): see **Guide for Managing Electronic Records from an Archival Perspective**, survey, literature review
 - [Australian Council of Archives statement](#)
- [Advanced Interfaces](#)
- [Pad++](#)
- [Ontology - Concept Descriptions](#) and [May 1997 slides](#)
- [Learning Agents](#)
- [Teaching and Learning Project](#)
- [Artemis Java Interface to UMDL Production System](#)
- [SGML creation and delivery](#)
 - enormous collection: 2M pages
 - [flowchart](#)
 - [SGML Server Program](#): middleware, training
 - cross collection searching
 - multiple representations
 -
- [Leveraging rich document formats](#)
 - patterns of use
 - ease of changing delivery: new standards (HTML), new rendering/packaging
 - collection management
 - Panorama, XML support by W3C
- Questions:
 - Will the agent and ontology approach work? Soon? For production DLs?
 - What is the support needed for establishing a digital library following the UMDL approach? Training?
 - What interfaces for DLs will be usable?

THE NSF/DARPA/NASA SPONSORED
UNIVERSITY OF MICHIGAN
DIGITAL LIBRARY
PROJECT



Digital Library Initiative University of Michigan

From *Computer* theme issue on the US Digital Library Initiative, May 1996

In the University of Michigan Digital Library, interacting software agents cooperate and compete within a virtual information economy to provide library services to students, researchers, and educators.

Toward Inquiry-Based Education Through Interacting Software Agents

Daniel E. Atkins, William P. Birmingham, Edmund H. Durfee, Eric J. Glover, Tracy Mullen, Elke A. Rundensteiner, Elliot Soloway, José M. Vidal, Raven Wallace, and Michael P. Wellman, *University of Michigan*

Providing true access to the human record means offering relevant information without prohibitive search time or an overwhelming choice among sources. Conventional libraries provide such access through two mechanisms: information organization and librarian services. Librarians themselves often rely on services like information systems or bibliographic databases to do their jobs.

Digital libraries must likewise provide organizational schemes and a wide variety of services. Most observers focus on the vast amount of information digital libraries will offer, delivered in new and interesting ways. However, we believe it is the bounty of services that will ultimately demonstrate the potential of digital libraries.

The University of Michigan Digital Library (UMDL) project^[1] is creating an infrastructure for rendering library services over a digital network. When fully developed, the UMDL will provide a wealth of information sources and library services. Of course, we cannot anticipate all the services that will eventually constitute a digital library. We therefore designed the UMDL to let third-party developers expand the library with new services and collections.

We are deploying the UMDL in three arenas: secondary-school science classrooms, the University of Michigan library, and space-science laboratories. Computer skills, information demands, and level of subject knowledge vary greatly among these user populations. Addressing the needs of high school students within a general-purpose digital library particularly stresses the flexibility of our underlying architecture. The UMDL must support services quite distinct from those that other digital libraries and the World Wide Web offer.

Many researchers and policy groups argue that students should engage in

sustained inquiry to develop an in-depth understanding of science. Digital libraries provide an outstanding opportunity to vitalize science education in public schools through inquiry-based education. However, we must avoid the inflated expectations typical of technology in the schools. Technology is only one element of a complex educational environment. Students, teachers, and curriculum planners must work together for a digital classroom library to succeed.

We are addressing the UMDL's ambitious scale and heterogeneity requirements by designing an open, distributed environment for interacting software agents. Features such as automated team formation, information search-space structuring, and market-based resource allocation help coordinate agent activities that provide library services. We are deploying the UMDL in Ann Arbor high schools.

Distributed agent architecture

Because digital-library technology is changing rapidly, user interfaces, search engines, and the structure of information sources must accommodate future innovations. Rather than adopt specific standards, we require the UMDL architecture to perform generic management operations, such as allocating resources and brokering connections. For instance, a language and protocol for communicating informational or processing capabilities and interests connects users and collections appropriately. However, determining how they interact to accomplish their task is beyond our architecture's scope.

Distributing tasks to numerous specialized, fine-grained modules promotes modularity, flexibility, and incrementality. It lets new services come and go without disturbing the overall system. We call these modules *agents*, emphasizing their local knowledge about specific tasks and their autonomy. Limiting the complexity of an individual agent simplifies control, promotes reusability, and provides a framework for tackling interoperability problems. Each agent performs a highly specialized library task and has a generic communication interface. This combination lets an agent apply specialized task competence to a wide variety of situations with other agents.

For example, an agent could generate synonyms for specified query terms and thereby produce variants likely to unearth relevant documents. Alternatively, an agent could use synonyms to assess how well some text matches an already formulated query. Encapsulating a general synonym service within a specialized thesaurus agent provides component functionality without committing to how it's employed systemwide.

Agent types

Figure 1 depicts the three classes of agents populating the UMDL: user interface agents, mediator agents, and collection interface agents. *User interface agents* (UIAs) manage the interface that connects human users to UMDL resources. Among other things, UIAs, perhaps with assistance from other

agents,

Figure 1. *Three agent types populate the University of Michigan Digital Library, performing a variety of specialized tasks.*



- express user queries in a form that search agents can interpret,
- maintain user profiles based on specified, default, and inferred user characteristics,
- customize presentation of query results, and
- manage the user's resources available for fee-for-service activities.

Mediator agents, which come in many types, provide intermediate information services.^[2] In the UMDL, mediators deal exclusively with other software agents, rather than end users or collections. They perform such functions as

- directing a query from a UIA to a collection,
- monitoring query progress,
- transmitting results,
- translating formats, and
- bookkeeping.

A subclass of mediators, called *facilitators*, exists expressly to team up other agents to accomplish a given task.

Collection interface agents (CIAs) manage the UMDL interface for collections, which are defined bodies of library content. Among other communication tasks, the CIA publishes the contents and capabilities of a collection in the registry (described below).

The agent architecture lets us develop specialized capabilities and add them to the UMDL as needed. For example, through new UIAs we can customize interfaces to user classes, rather than to collections or access mechanisms. These UIAs, in turn, can access any mediator services available in the system.

Agent teams

Complex UMDL tasks require the coordination of multiple specialized agents working together on behalf of users and collection providers. To form teams, agents must be able to describe their capabilities to each other in ways all can understand.

Levels of agent communication

UMDL agents communicate at three distinct levels of abstraction. At the lowest level, agents employ network protocols such as TCP/IP to transmit messages among themselves. Task-specific protocols dictate how the agents interpret and process these messages. For example, agents could use SQL to convey a request to perform a data-retrieval task. UMDL generally doesn't restrict task-specific protocols: Whoever designs and introduces the agents can freely choose the language(s) those agents speak.

Of course, agents are more likely to be used frequently if they communicate in widely adopted languages. In particular, a desire for broad interoperability provides an incentive to support standards like Z39.50, which libraries often use. This increases the scope of collections accessible to an agent posing a given query. While standardization has significant benefits, and many UMDL agents do use Z39.50, it is not a requirement for joining UMDL.

A specialized agent's capabilities will remain untapped unless it makes its abilities and location known and participates in team formation. We thus defined special protocols for the team formation and negotiation tasks, which all UMDL agents share. These UMDL protocols represent the third level of abstraction in agent communication.

Conspectus language

UMDL agents are defined by the information content they can deliver, the information services they can render, or both. To participate in UMDL protocols, agents need a language for describing these capabilities. Agents describe what they can contribute to an agent team and what their limitations are in the conspectus language (CL). Facilitators can also use CL to (perhaps partially) describe capabilities required for participation on a team. CL thus serves as a language for both disclosing and querying about abilities.

To ascertain a message's intent, UMDL protocols adopted a flexible notion of message types, patterned after KQML.^[3] UMDL message types, the equivalent of KQML "performatives," correspond to high-level communication acts. For example, messages intended to inform are of type Tell, and the purpose of Ask messages is to elicit information. A message can contain CL expressions, with the message type conveying what the recipient should do with the supplied content. UMDL protocols define a small number of standard message types that all agents should be able to interpret and process.

Registry agent

We designed the UMDL protocols so that agents advertise themselves and find each other on the basis of capabilities. Rather than have every agent maintain models of all others and periodically broadcast its descriptions to every other agent, we designated a registry agent. The registry is special in several respects. First, on inception, agents know how to access the registry, thus avoiding the bootstrapping problem. Second, all agents can communicate with the registry using the UMDL protocols, as further detailed below. Third, the registry

provides its services for a static price (currently free) to avoid the need to negotiate. Negotiation with the registry could lead to deadlock, since the registry contains the information identifying which agents can facilitate negotiation.

The registry agent maintains a database of all agents in the UMDL system, including descriptions of their content and capabilities. It updates the database with descriptions expressed in CL. The registry agent collects descriptions that specify the following types of characteristics:

- identification (such as name, location, and type),
- content (broad topic, audience level, language, and so on),
- capability (search engine(s) supported, translation facilities, name authority services, and so forth),
- interface (for example, task-specific languages and resource requirements), and
- economic (pricing methods, standing offers, and negotiation protocols, for example).

One simple yet representative example of a CL description is that which characterizes an author index agent (Figure 2). The agent belongs to a class of UMDL agents that search across information sources without executing the search request in each. Its CL description specifies its type and describes its service in terms of what interactions it supports. The **<Capability>** field states that the agent accepts queries with a specific author **\$A** as a bound input parameter. It then returns the associated CIAs (**\$U**) for all collections in which the author appears.^[4] It does not, however, accept requests of the reverse order-asking for authors associated with a particular collection.

```

< CL description {
  <Agent_ID AID_777>
  <Agent_type Author_index>
  <Capability
    <Author *$A> <CIA $U*> >
  <Task_Language SQL>
  <Content
    <Broad_Topics 'SCIENCES'>
    <Last_updated 12.31.1995>
    <Frequency_of_update end_of_year> >
  <Pricing fixed (1-bibliobuck-per-search) >
  <Content_Language {English,German,Latin}> }

```

Figure 2. *Conspectus language description of an author index agent.*

The registry agent communicates using UMDL protocols, translating incoming

requests into queries on the registry database. Since this service's availability and fault tolerance are critical, we employed a persistent implementation of the registry database. An SQL server provides the basic properties of consistency, concurrency, and recovery, and supports high throughput of concurrent agent requests. Our second-generation registry agent, under development, uses a more powerful distributed, open architecture. We are implementing the distributed registry using commercial database technology. Replication servers support a powerful distributed search paradigm that, while robust and scalable, is transparent to the rest of the UMDL.

The preliminary version of the distributed agent architecture contains about a hundred CIAs and spawns a UIA for each active user. In addition to the registry, we have implemented several other mediator agent types. We describe three of these--the query planner, the market facilitator, and the remora--later on.

Search types

In any UMDL context, the core task is to find the right combination of information and services to satisfy the participants' objectives. This could mean answering a user's question, finding customers for a publisher's content, or applying a sequence of format-translation services. In these cases, the fundamental activity is searching for useful content or services using minimal effort, time, and money.

Within UMDL, searching takes several forms. Once a user's UIA contacts a collection's CIA, the search concerns documents from the collection that satisfy the user's specifications. This level of search is a *collection search*. Before collection search takes place, however, the UIA must identify appropriate collections on the basis of how agents describe themselves in conspectus language. This is a *conspectus search*. Finding mediators with particular capabilities is another form of conspectus search. UMDL agents interleave these various types of search to accomplish more complex tasks.

Collection search

The UMDL architecture supports arbitrary types of collections and search engines by encapsulating them using CIAs. Thus we can accommodate even those collections that require custom browsers, such as the Blue-Skies weather service.[\[5\]](#) We extended the class of collections accessible through more standard retrieval protocols by developing Z39.50 interfaces for Mirlyn, FTL, and WAIS. (Mirlyn provides access to the University of Michigan library catalog and several abstracting and indexing databases, while FTL is a UMDL-specific search engine.) We are also investigating structuring techniques that search across complex objects such as SGML (Standard Generalized Markup Language) documents.[\[6\]](#)

There are two modes for interacting with collections: searching and browsing. In the first, the UIA knows which collection to access, perhaps because of a prior conspectus search. In this case, the user connects directly to that

collection's CIA and uses native retrieval facilities. Alternately, the UIA could conduct a search across collections. An information fusion agent then organizes the results, combining or ranking the retrieved information for presentation to the user.

Conspectus search

Conspectus search seeks to connect content providers and consumers on the basis of agents' needs and capabilities as described in conspectus language. Typical tasks include locating appropriate collections, identifying a particular work's authors, and determining the cheapest way to access certain information. This generally involves several intermediate tasks, including other conspectus searches. For example, while looking for appropriate collections, a UIA might conduct a conspectus search for a thesaurus agent.

UMDL agents formulate conspectus search tasks in terms of content or services sought and search processes by which to find them. A particular conspectus search task's description includes

- conspectus language specifications for the content or capabilities sought,
- deal parameters (such as acceptable cost ranges and delivery constraints),
- search-effort parameters (allowable search time, number of sources, and so forth), and
- search modification guidelines (for example, preferences toward using particular agents and trade-offs among the other parameters).

A conspectus search returns a set of agent deals. Each deal represents an agent's offer to provide the desired services or content, and the terms of the offer. The initiating agent can accept deals on the basis of criteria such as price and reputation. It then works with the chosen agent(s) in a task-specific language. If no deals are acceptable, the initiating agent can reinitiate conspectus search to find alternative deals.

Conspectus searches can be as simple as retrieving relevant entries from the registry as a direct result of the user's request. Other searches require the combined abilities of a team of agents to reformulate the request and balance thoroughness against cost. A query-planning mediator coordinates this kind of search.

Query-planning mediators

Agents capable of accomplishing conspectus search tasks are classified as task planners. As noted above, a task planner might require additional information or services from other agents to accomplish its task. Query-planning mediators, a subclass of task-planning agents, specifically tackle conspectus search tasks that seek collections to satisfy a query. Our initial query planner uses the UM version of the Procedural Reasoning System (UM-PRS), which provides facilities for flexible procedure specification and execution.[\[7\]](#) Our UM-PRS task planners communicate using UMDL protocols. They are goal-driven,

persistent, independent, and proactive.

Query-planning mediators embody specialized knowledge about how to seek out information sources in response to a user's query. Based on interviews with librarians, these procedures specify the control flow among various resources within the UMDL. Depending on user characteristics, library load, and desired completeness and timeliness of the search, the query planner invokes different procedures. These procedures in turn can post subtasks that could be accomplished in a variety of ways, depending again on context. Thus, query-planning mediators provide a flexible mechanism for performing conspectus search.

Figure 3 illustrates the kinds of activities the query planner might invoke. The nodes contain the name of the task and in some cases the names of some procedures for achieving it. The arrows represent subtask relationships. The actual procedure the query planner executes depends on context, in ways specified by our consulting librarians. The task requires capabilities that are distributed among various agents within the UMDL. Thus, by elaborating the procedures, the query planner dynamically builds a team of agents that together accomplish the task. See the later section "[Example queries](#)" for a brief description of this procedure.

Figure 3. *The query-planner procedure can be elaborated to build a team of agents for accomplishing search tasks.*



Market-based resource allocation

The digital library creates a potentially unbounded demand for computational resources. For example, any preprocessing of collection data--indexing, metadata gathering, or caching--might improve system response to subsequent user requests. With only finite resources, however, we cannot take advantage of all such opportunities. Neither can we try every method for accomplishing a given task. Rather, we must choose among available methods on the basis of resource requirements and prospects for success.

Information service economy

We model alternative information services as economic activities that compete to provide the highest service level for minimal computational resources. The goal of UMDL as a whole is to allocate resources efficiently to optimize user services.

To organize processing activities within an economic framework, we treat agent

interactions as supplier-producer relationships. Each agent creates value-added information products from the input products others provide.[8] Agents connect dynamically as opportunities arise for mutually beneficial exchanges. The collections provide "raw materials" in this process, whereas end users are the ultimate consumers of the "finished goods." The mediators ("middlemen") improve the value of information along the way using knowledge, processing, storage, or other computational resources.

Market facilitators

Market facilitators, or auctions, operate by collecting offers and determining agreements among agents. One simple kind of auction collects bids and settles them by some market-clearing process. Others perform a more complicated matching and search process. In our basic UMDL market protocol, one auction agent represents each good. A good could be delivery of digital objects, translation services, or other agent product. Each auction agent accepts offer messages from agents interested in buying or selling that good. Offers include a demand schedule that specifies the amount (quantity or quality) of information good the agent will transact at various prices. The auction finds a price that balances supply and demand, reports the price to the agents, and executes the transaction.

Describing goods and services

To design a market in library services, we must determine the goods and services and how to represent them in the system.[9] However, in large-scale dynamic markets, the set of goods and their important distinctions change over time. A structured, expressive good description language (part of our conspectus language) defines goods as variations and combinations of primitive concepts. From these descriptions, agents can automatically determine how to perform the necessary transformations.

For example, if the language contains the concepts NPR and Broadcast, we can construct the concept NPR Broadcast. Since one operation that agents can perform on Broadcasts is to make Transcripts, we have a meaningful notion of NPR Transcript. Parameterization provides extra degrees of freedom; for example, descriptions can qualify NPR Transcripts by date and topic.

Intellectual property usage licenses

In an information and information services market, the essence of goods is information content, not realization in some physical medium. This suggests that an exchange in information goods should distinguish between the intellectual property and its physical manifestations. Having a copy of an intellectual work does not imply the authority to do anything with the information that work represents. We refer to such authority generically as intellectual property usage licenses. Licenses are the primary type of information good exchanged in the system.

Supporting inquiry-based education

Merely wiring a classroom to the Internet-or even to a digital library-will not make students learn through inquiry.[\[10\]](#) Existing Internet-based tools do not effectively support access to digital resources or address the special constraints of a secondary-school classroom for sustained inquiry. For example, 50-minute class periods are very confining for students and teachers trying to engage in inquiry. Our strategy is to understand the real challenges in the classroom and design UMDL services that explicitly address these needs.

Teacher challenges

Developing good curriculum materials is a time-consuming task under any circumstances. The search for motivating, engaging, content-filled on-line materials is particularly so. Moreover, our experiences with on-line curriculum delivery suggest that a teacher should seed the Web pages with a few jump-start collections. Students need to find something quickly and have some immediate success to maintain their motivation and engagement.

At least two types of UMDL agent services can assist teachers in developing and managing curriculum materials. First, we are developing a customized version of the query-planning agent called QuickScan. Its specialized knowledge of pedagogical relevance helps a teacher quickly search and retrieve material useful to high school science classes. The QuickScan agent focuses on collections that are age-appropriate and have a range of nontextual media types (video, images, audio). Students, too, will be able to use QuickScan to find relevant information in a timely manner.

Second, remora agents (see "[The remora agent](#)" sidebar) provide a time-saving way for teachers to monitor the development of on-line materials. The Web contains many potentially relevant sites. However, a large percentage of them are still not sufficiently developed to permit effective classroom use. Also, while many Web sites provide information about current events, like volcanic eruptions, checking sites manually is tedious and time-consuming. Remora agents help teachers monitor the evolution of these sites and incorporate the materials into an on-line curriculum.

Student challenges

Teachers are often reluctant to have their students "waste precious classroom time" searching for materials. They would rather just show the students sites that provide answers. However, the inquiry-based approach, by definition, requires students to engage in on-line search. Finding and evaluating sites for relevance is an intrinsic component of inquiry. The tension is real: Current search technology, particularly keywords, is time-consuming, frequently unproductive, and fosters a random approach to searching.

Our strategy is to provide UMDL interfaces and agents that support students' learning through the search process. For instance, the UMDL search interface

will provide tools like spell-checking and content-specific thesauri to help sharpen query formulation. We are also developing a UIA with an interface designed to scaffold query reformulation. This will help students who find "re-searching" and following a coherent line of exploration difficult.

A second real problem in the classroom is the lack of collaboration among students. Substantive classroom conversation is a key component of learning.[\[11\]](#) Professionals continually engage in discourse to invent, explicate, and refine their ideas; students need dialogue for the same reasons. We are developing interface, registry, and search agents that let students share the fruits of their on-line searches. This encourages classroom interaction by providing artifacts for students to discuss. For example, a group of students could register in the UMDL their collection of on-line materials regarding a specific topic. The search agents will direct other groups of students in the class to that collection first.

Fast, simple registry of student-generated work is also allowing students to publish their findings more easily in the UMDL. For example, a class of 11th-graders recently completed a six-week unit on water contaminants. Each pair of students wrote a report on a different water contaminant, then published it on the World Wide Web. These students filled a gap. Until their efforts, no site on the Web had a comparable in-depth treatment of various water contaminants. Feeling that their ideas are respected--even desired--greatly motivates students. This typically translates into more engagement and more effective learning.

UMDL Status

The first version of the UMDL is currently operational at the university and is being deployed at Ann Arbor high schools. The earth and atmospheric sciences collections include material from the popular press, academic journals, encyclopedias, the World Wide Web, and local curriculum. The system is highly extensible, and we are continually expanding and enhancing content and services.

Example queries

We can illustrate a subset of the UMDL's current capabilities by summarizing its behavior for two example queries. The agents in this example include a query planner, a thesaurus agent, a BSO agent, and a remora agent. The Broad System of Ordering, or BSO, agent uses a hierarchy of terms to broaden or narrow a topical search. The remora agent has the task of persistently monitoring and summarizing message traffic in the UMDL.

For a simple task, the query planner gets a query that matches entries in the registry, requiring little interaction among the various services. The communication matrix generated by the remora, Figure 4a, shows this low level of interaction. In a more difficult query, however, the query planner must invoke the BSO and thesaurus agents. They then reformulate the query in terms

of topics about which some collections have professed capability (Figure 4b). These simple examples suggest the dynamic, flexible interactions that we rely on to fulfill our ambitious vision for the UMDL.

Figure 4. *The remora agent monitors the number of messages passed between agents during two simple tasks. (a) The query planner returns a single CIA ("MSU") that can respond to the query. (b) The query planner consults the Broad System of Ordering (BSO) and thesaurus agents before passing the query to a Web crawler.*



High school deployment

We're initially deploying the UMDL in four high schools and two middle schools in Ann Arbor, with other locations planned. Besides installing the UMDL infrastructure, we have developed a substantial body of associated curricular material that includes tutorials on searching for on-line information, and specific topics in high school earth and space science.

By May 1996, we expect that over one thousand students will have used UMDL services. Working in a handful of classrooms is an important start. However, our aim is not merely to create a successful, innovative pilot project. We want to understand the fundamental issues involved in implementing digital libraries in schools and making them relevant to today's classrooms.

Conclusion

As the previous section suggests, many challenges remain in making technologies such as the UMDL meaningful in inquiry-based education. We are only in the initial stages of deploying the UMDL in high school and middle school classrooms. However, we already find that the UMDL agent architecture provides welcome flexibility for creating technology-based strategies to meet the challenges.

Building the UMDL raises many difficult problems of scale, decentralization, interoperability, and resource allocation. Our approach has been to define very general mechanisms and then test them with specific instances of software agents and protocols that use these mechanisms to provide library services.

Although our work on the UMDL is preliminary, the first year and a half made some things clear: First, the scale and diversity of the project will test our technical ideas-distributed agents, interoperability, mediation, and economical resource allocation. Second, the UMDL project will test our theories about the

role and impact of educational technology.

Acknowledgments

Other project members contributing to the work described herein include Ken Alexander, Gene Alloway, Karen Drabenstott, Randall Frank, Olivia Frost, George Furnas, Daniel Kiskis, Wendy Lougee, Jeffrey MacKie-Mason, Greg Peters, John Price-Wilkin, and Amy Warner. This work was supported by the NSF/ARPA/NASA Digital Library initiative. Further information is available at <http://www.si.umich.edu/UMDL/>.

References

1. W.P. Birmingham et al., "The University of Michigan Digital Library: This Is Not Your Father's Library," *Proc. Digital Libraries 94*, Hypermedia Research Laboratory, Texas A&M University, College Station, Tex., pp. 53-60.
2. G. Wiederhold, "Mediators in the Architecture of Future Information Systems," *Computer*, Mar. 1992, pp. 38-49.
3. T. Finin et al., "KQML as an Agent Communication Language," *Proc. Third Int'l Conf. Information and Knowledge Management*, ACM Press, New York, 1994.
4. A. Rajaraman, Y. Sayiv, and J.D. Ullman, "Answering Queries Using Templates with Binding Patterns," *Proc. ACM Symp. Principles of Database Systems*, ACM Press, New York, 1995, pp. 105-112.
5. P.J. Samson, K. Hay, and J. Ferguson, "Blue-Skies: Curriculum Development for K-12 Education," *Proc. Conf. Interactive Information and Processing Systems*, American Meteorological Soc., Boston, 1994.
6. A. Nica and E.A. Rundensteiner, "Uniform Structured Document Handling Using a Constraint-Based Object Approach," in *Advances in Digital Libraries*, N.R. Adam, B.K. Bhargava, M. Halem, and Y. Yesha, eds., Springer-Verlag, New York, 1995, pp. 41-60.
7. J. Lee et al., "UM-PRS: An Implementation of the Procedural Reasoning System for Multirobot Applications," *Proc. AIAA/NASA Conf. Intelligent Robotics in Field, Factory, Service, and Space*, NASA Center for Aerospace Information, Linthicum Heights, Md., 1994, pp. 842-849.
8. M.P. Wellman, "A Market-Oriented Programming Environment and Its Application to Distributed Multicommodity Flow Problems," *J. Artificial Intelligence Research*, Vol. 1, No. 1, Aug. 1993, pp. 1-23.
9. T. Mullen and M.P. Wellman, "A Simple Computational Market for Network Information Services," *Proc. First Int'l Conf. Multiagent Systems*, Amer. Assn. Artificial Intelligence Press, Menlo Park, Calif., 1995, pp. 283-289.
10. E. Soloway, "Beware, Techies Bearing Gifts," *Comm. ACM*, Vol. 38, No. 1, Jan. 1995, pp. 17-24.
11. A.L. Brown and J.C. Campione, "Psychological Theory and the Design of Innovative Learning Environments: On Procedures, Principles, and Systems," in *Contributions of Instructional Innovation to Understanding*

Learning, L. Schauble and R. Glaser, eds., Erlbaum, Hillsdale, N.J., 1996 (in press).

Daniel E. Atkins is dean and professor at the School of Information and professor of electrical engineering and computer science at the University of Michigan. He is the director of the NSF-ARPA-NASA UM Digital Library (UMDL) Project, the NSF Upper Atmospheric Research Collaboratory (UARC), and a Kellogg Foundation grant to restructure graduate education for information systems professionals. His research focuses on the design and evaluation of network-based knowledge work environments. He received a PhD in computer science at the University of Illinois in 1970.

William P. Birmingham is an associate professor in the Electrical Engineering and Computer Science Department at the University of Michigan, with a joint appointment in the School of Information. His research interests include large, distributed information systems in areas such as distributed optimization and design, concurrent engineering, and digital libraries. He received a PhD from Carnegie Mellon University in 1988 for his dissertation on developing and maintaining large knowledge bases for design applications. Birmingham was named an NSF Presidential Young Investigator and is a member of Sigma Xi, AAAI, ACM, and IEEE.

Edmund H. Durfee is an associate professor in the Department of Electrical Engineering and Computer Science at the University of Michigan, where he conducts research in multiagent systems, real-time intelligent control, and cooperative problem-solving for applications ranging from interacting unmanned vehicles to supporting human collaboration. He received a PhD in computer science from the University of Massachusetts in 1987 and was named an NSF Presidential Young Investigator in 1991.

Eric J. Glover is a graduate student in the Department of Electrical Engineering and Computer Science at the University of Michigan, pursuing degrees in VLSI and computer science. He received a magna cum laude BSE in electrical engineering in 1990 from the University of Michigan.

Tracy Mullen is a PhD student in the Department of Electrical Engineering and Computer Science at the University of Michigan. Her research interests include the design of distributed information service environments based on computational market technology. She previously worked at Lockheed Software Technology Center in Palo Alto, California, and received a BS and an MS from Rutgers University.

Elke A. Rundensteiner is an assistant professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. Her current research interests include object-oriented database technology for nontraditional applications, view and schema evolution tools, database support for digital libraries, and multimedia information systems. She received a PhD in computer science from the University of California, Irvine. Rundensteiner has received a Fulbright Scholarship, an IBM Scholarship, an NSF National Young

Investigator Award, and an Intel Young Investigator Engineering Award from the Engineering Foundation.

Elliot Soloway is a professor in the Department of Electrical Engineering and Computer Science and in the School of Education at the University of Michigan. His current research interests lie in exploring the roles that computational media can play in self-expression, communication, and learning and teaching. Soloway is editor of *Interactive Learning Environments*, a journal devoted to exploring next-generation computational and communications technologies for learning and teaching. He received a PhD from the University of Massachusetts, Amherst, in 1978.

José M. Vidal is a PhD student in the Department of Electrical Engineering and Computer Science at the University of Michigan. His research interests are in agent modeling, software agents for multiagent systems, and distributed AI. He received an SB from the Massachusetts Institute of Technology and an MS from Rensselaer Polytechnic Institute, both in computer science.

Raven Wallace is a PhD student in educational technology at the University of Michigan. Since receiving MS degrees in mathematics and civil engineering, she has taught at the college, secondary, and elementary school levels. Her current research addresses cognitive implications of digital libraries in secondary schools.

Michael P. Wellman is an assistant professor in the Department of Electrical Engineering and Computer Science at the University of Michigan. His current research focuses on computational market mechanisms for distributed decision making. He received a PhD in computer science from the Massachusetts Institute of Technology in 1988 for work in qualitative probabilistic reasoning and decision-theoretic planning. He received an NSF National Young Investigator Award in 1994.

For more information about this article, contact Wellman at the Department of EECS, University of Michigan, Ann Arbor, MI 48109, wellman@umich.edu.

Sidebar:

The remora agent

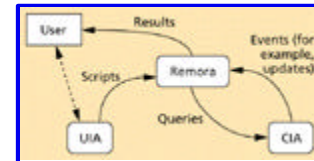
[Return to the main text](#)

The remora is one example of the value-added services the UMDL accommodates. A mediator agent, the remora offers event-driven notification services for a variety of library resources. Users specify events of interest and receive notifications when such events, like new items appearing in a collection, occur.

We got the name "remora" from a kind of fish that attaches itself to sharks and other large oceanic creatures. In the UMDL, remoras attach themselves to CIAs

for the purpose of detecting events. On behalf of other UMDL agents, the remora accepts scripts that specify events of interest and the actions they trigger. For example, one script might ask for e-mail notification whenever a collection adds a new Hubble Space Telescope image. Another script might define filters to extract articles matching current curricular items from a Web page, and the script might include processing instructions to add the articles to a particular portfolio document in a specified way. Figure A depicts the interaction of the remora with other UMDL agents.

Figure A. *The remora agent provides event-driven notification services by querying collections according to user scripts.*



The remora participates in the UMDL information economy through several markets. Remoras compete with each other, and perhaps with other subscription agents, to supply the service of running scripts. They must also bid to receive events—that is, attach to CIAs—and to acquire the necessary computational resources.

[Return to the main text](#)

Guest Editors' Introduction

University of Illinois

University of California at Berkeley

Carnegie Mellon University

University of California at Santa Barbara

Stanford University

University of Michigan

[Computer](#) | [Computer Society home page](#)

Send comments and questions about this page to Christine Miller, cmiller@computer.org

Send general comments and questions about the IEEE Computer Society's Web site to



THE NSF/DARPA/NASA SPONSORED UNIVERSITY OF MICHIGAN PROJECT **DIGITAL LIBRARY**

UMDL TECHNOLOGIES

ARCHITECTURE: AGENTS AND ONTOLOGIES

ACCESS: ARTEMIS INTERFACE

CONTENT: COLLECTIONS

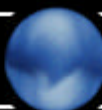
ECONOMY: COMPUTATIONAL MARKETS

ADVANCED USER INTERFACE

CONSPECTUS & IR

PRODUCTION SYSTEM

MISSION • ACCOMPLISHMENTS • IN ACTION



UMDL TECHNOLOGIES • IMPACT • TEAM

Strategies for Digital Library Development: The University of Michigan Case Study

Wendy P. Lougee
University of Michigan

Prepared for TICER, Second International Summer School, Tilburg, The Netherlands, August 1997

Abstract

Beginning in 1993, the [University of Michigan](#) launched a partnership between the [Information Technology Division](#), [University Library](#), and [School of Information](#) to develop the campus environment for networked information resources. The collaborative effort focuses on building critical infrastructure components as well as useful information resources and services. A strategy has been employed which develops the capability, within projects, to support a variety of information formats, search engines, and applications. As the program matured, a virtual organization has been created which provides production support for digital collections and manages programs of access for University users and remote customers. Critical to the success of the joint program has been the integration of expertise and resources of the partner organizations and a commitment to address campuswide practices and policies.

Introduction

There is no single strategy to employ in building a digital library. Strategies, like libraries themselves, are born out of the time, the culture, and opportunities. I've been asked to share the [University of Michigan](#) (UM) experience as a case study and hope to convey the context and opportunity-based factors which have shaped our endeavor. Our story is really a story about relationships and the critical interweaving of expertise and resources which, I believe, has been the hallmark of our success.

This overview is structured to provide: the history of our program, some perspective on the forces shaping digital libraries, a description of the [University of Michigan Digital Library](#), and finally some sense of how our organization has matured from a series of projects to a more mature production-focused service.

History of the Digital Library Program

The roots of our campuswide digital library program can be found in a year-long committee

process in 1991. Three faculty committees were charged to look at issues surrounding electronic information: collection needs, funding models, and user support. The process was co-sponsored by the [University Library](#), [the Information Technology Division](#) (ITD) (the campus computer services organization), and the then [School of Information and Library Studies](#) (SILS). Subsequent to the reports of these groups, a smaller executive team was convened (chaired by the Dean of SILS and a former University President) and issued a brief, but strategic report. Their recommendations focused on the emergence of the distributed computing environment and the impending chaos that could occur on campus if we failed to build coordinating structures and practices. Specifically, the group recommended that:

- the campus should harness the complementary expertise of the library and technology sectors,
- a project-based approach should be undertaken which would create visible, useful information products and services for faculty and students, and
- attention to campus information policies would be critical to ensure the development of an "information community" rather than a highly distributed and fragmented information environment.

These recommendations formed the basis for a formal collaborative partnership between the three organizations. Launched in 1993, the partnership is supported through contributed funds from each organization. From the beginning, it was thought imperative that there be focused leadership and a position was created to oversee the joint efforts. Though administratively associated with the Library, this position reports jointly to the Director of the [University Library](#), the Dean of the newly renamed [School of Information](#) and the Executive Director of the [Information Technology Division](#) and has responsibility for the near-term development of the digital information environment. More recently, a fourth organizational partner has emerged as the University has created an [Academic Outreach Program](#) (AOP) which facilitates distance-independent programs and services.

A central theme of the partnership is the synergy found in the integration of expertise of the partner organizations as depicted in Figure 1.

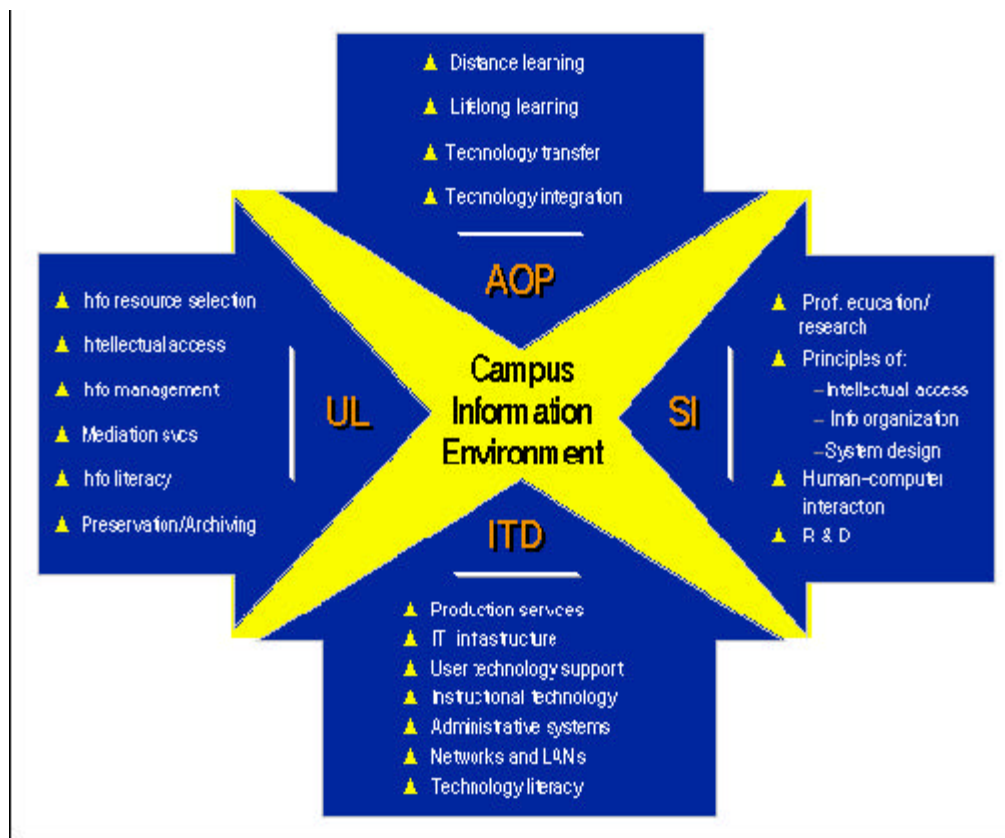


Figure 1: Core Competencies/Expertise of the Partner Organizations

An important opportunity was added to the arena in 1994 with the receipt of a large-scale federal grant (from the joint Digital Library Initiative of the National Science Foundation/Advanced Research Projects Agency/National Aeronautics and Space Administration). This multi-disciplinary research project is developing an agent architecture for the organization of and access to distributed digital libraries. (The [UMDL research project](#) will be described in a later TICER session.)

Another more recent organizational dimension has been the further evolution of the School of Information and Library Studies into the newly named [School of Information](#) (SI). The new school brings together faculty from library and information science, computer science, economics, psychology, education, and public policy and will help shape the information professional of the future. In addition, the research base within SI and the many related programmatic initiatives underway provide fertile ground for digital library development.

The Changing Landscape

We are witnessing a change in our shared vision of a "library." One could argue that the technology forces have prompted this new vision, but there are in fact many more factors which will shape the evolution of libraries. Changes within disciplines are effecting changes in the academy, and higher education in general is beset by new stresses of accountability. The publishing sector is weathering the dual pressures of new technologies and questions about its added value role. Many institutions are pursuing new markets and new time and location-independent strategies for delivery of courses.

The impact of these forces on libraries is simplistically forecasted in the contrasting descriptions of Figure 2. As the potential grows for any individual or any unit to act as author, library, or publisher, it is clear that the information environment will become more distributed and the role of librarians may increasingly stress the unique expertise that the profession brings to the environment--i.e., enabling the selection, organization, access, and long-term preservation of knowledge resources.

"Traditional" Library	21st Century Library
<ul style="list-style-type: none"> ● facility-centric ● centralized collections ● hierarchical organization ● common description & access protocols ● value = collection size ● just-in-case collection strategy ● central funding model; "free goods" ● campus focus 	<ul style="list-style-type: none"> ● user-centric ● distributed collections ● collaborative teams ● heterogeneous access protocols ● value = access & services ● just-in-time information delivery ● mixed funding models ● global reach

Figure 2: The "Traditional Library" vs. the Future Library

We can envision that in a mature distributed information environment, libraries will no longer be facility-centric, but what role should they play? Are libraries destined for a limited, archival role? How should the expertise of librarians be transformed in this new digital era? The rise of digital technologies challenges us to look beyond mere extrapolation from existing functions and services and seek out innovative models which extend library and librarian functions more broadly. In addition, there is enormous potential to harness new collaboration technologies and create a broader vision of libraries themselves--"collaboratories" or "knowledge unions" as they are sometimes called.

The 21st century library offers enormous potential to build highly integrated library/collaborative work environments which meet the needs of tomorrow's scholars. Michigan's strategy for digital library development is firmly based in this notion of broadening the influence of libraries and the sharing of librarian expertise outside traditional organizational bounds. The strategy has also promoted the imperative of collaborative teaming between librarians and technologists. Finally, the close relationship between our near term initiatives and the research environment hopefully provides opportunities to explore this broader collaborative environment in addition to building the necessary networked information infrastructure.

The UMDL Strategy

The initial "blueprint" for building the [University of Michigan Digital Library](#) recognized that technology development was only one piece, albeit an important piece. Equally important are the issues which involve organizational development for the partners and also policy development for the campus and the broader scholarly community. The technology and organizational elements have been addressed largely by *doing*--i.e., actual projects which have put in place real collections and services for real users. Increasingly, however, projects have been approached which address a critical policy domain--e.g., institutional publishing on the Internet, economic and pricing issues of scholarly information, rights management, and archiving of digital information.

Initially, a two-pronged approach toward the digital library arena was undertaken. First, an overarching framework for digital information resources on campus was developed--beginning to create the notion of "campus as library" and bringing library access principles to bear on the campus environment. Second, a series of format-based projects were developed, each addressing critical infrastructural elements for our digital library.

A Framework for Campus Digital Resources

An overall framework for campus networked information resources has been created through development of the UM [Information Gateway](#) project. As individuals and units have expanded their roles as information providers and the use of the World Wide Web has proliferated on campus, it has become increasingly clear that the institution needs a strategy for managing its information assets. The Gateway provides a mechanism for federating the information resources available to the campus community, creating a more coherent "whole" out of the the rich array of distributed resources.

At the simplest level, the Information Gateway is an institutional homepage. The policy structures supporting the homepage development define the appropriate content linkages (administrative and unit-sponsored resources) and provide guidance to information providers about publishing on the Internet. (The University's policy document for Internet Publishing can be found at <http://www.umich.edu/~gateway/policies/>). A joint team with representatives from the Information Technology Division, the University Library, and the University Relations office oversee the development of the Gateway, including its functional capabilities. In addition to the typical full-text search capability for the sites linked to the Gateway, a mechanism has been developed to provide more structured search capability and, in addition, a strategy to assist the campus in managing its distributed information resources.

Not surprisingly, the University's Internet environment is populated with a wide range of different types of electronic resources: descriptive information about University units, course catalogs, research data sets, publications, licensed publisher content, video, real time data, etc. In order to "federate" the very diverse types of resources and facilitate searching across resources, a minimal set of metadata has been developed and deployed in an [SGML registry database](#). The registry has been designed which will describe key attributes about each digital collection accessible via the Gateway (e.g., resource manager, topical coverage, rights ownership, deliverable formats, etc.). These metadata will allow users to search the registry

and dynamically create views of the resources in response to an information need. As the size of the campus digital environment grows, this functionality for filtering, searching, and navigation will become critical.

In addition to campus resources being registered within the Gateway, a companion [digital library registry](#) is being created by subject specialist librarians. This service will include not only campus resources of scholarly interest, but will also provide a vehicle to select and register Internet sites relevant to specific disciplines.

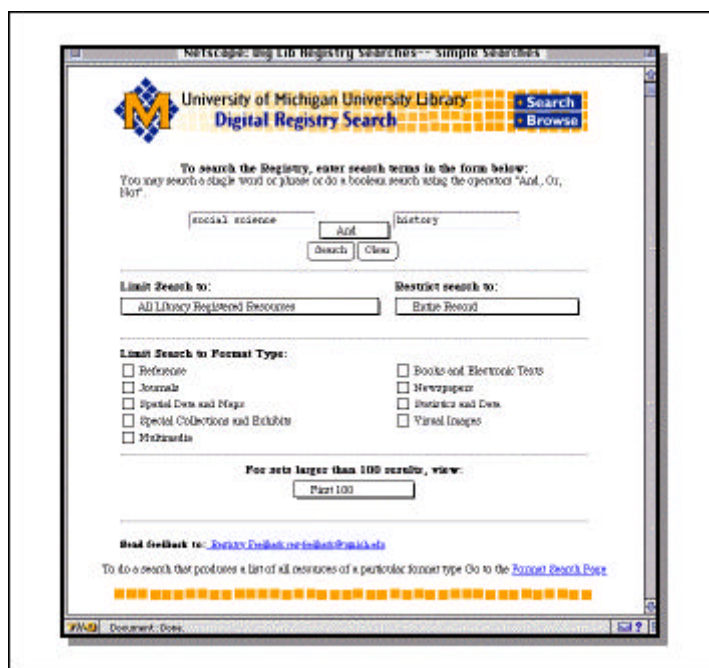


Figure 3: Digital Library Registry

Format-Based Projects

Early on we recognized that the central challenge for digital libraries was gaining expertise and developing infrastructure to handle diversity--e.g., diversity of formats, diversity of underlying search engines and retrieval systems, diversity of users, diversity of pricing and funding systems. Our approach has addressed this diversity through the implementation of format-specific projects (e.g., encoded text, visual images, spatial data, etc.), with the assumption that each project will establish a model for dealing with additional content in these format categories and simultaneously build key pieces of the information infrastructure (e.g., security, accounting/billing, distributed printing, etc.). In addition, elements which will help federate these individual collections have been pursued. Projects are also embedded within programs of outreach to particular discipline domains as well, ensuring that services developed are sensitive to the unique research methods and vocabularies of those subject areas. A conscious decision was made to use readily available web browsers wherever possible as the primary (though not exclusive) framework for user access. The types of resources pursued are represented in Figure 4.

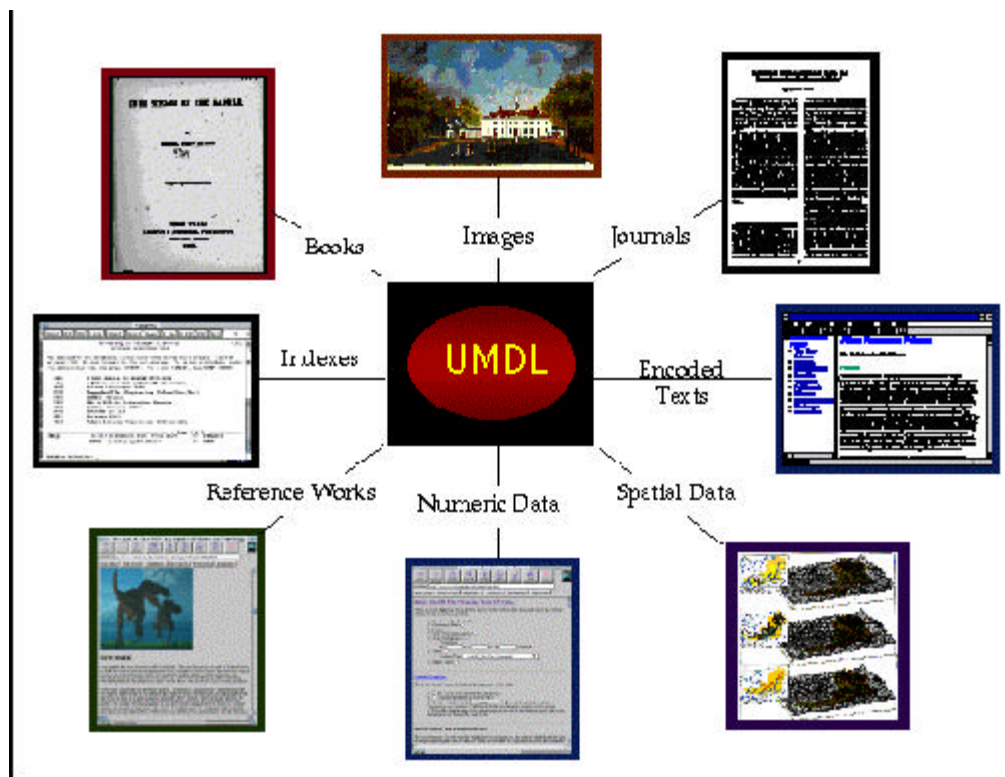


Figure 4: Collections Components for Federation in the UMDL

The sections which follow will describe the individual projects of UMDL, highlighting the infrastructure components addressed and some of the lessons learned. Throughout, I hope to also convey the program's strategic goal to stretch each project beyond the actual deployment of content, and to leverage the activity toward the exploration of broader digital information issues.

Journal Projects

Beginning with the Elsevier Science [TULIP project](#) in 1991, the University of Michigan has been involved in several projects focused on the delivery of journal content. TULIP addressed issues of delivery, use, and economics of electronic journals through implementation of 43 journals in materials science. Michigan was fortunate to be able to draw on a locally developed search engine in its implementation as well as a multi-organizational design team. TULIP gave us important early experience with an array of issues, including: storage requirements, client development, the importance of critical mass of content, the key role of promotion of new services, and costs. We also began the development process for distributed printing and accounting /billing functions. TULIP paved the way for a number of other page-image-based text projects, including an expanded partnership with Elsevier Science.

A second project undertaken in partnership with [UMI](#), was launched in 1994 and targeted two specific requirements for the digital library. First, we wanted to experiment with migrating proprietary and stand-alone systems to a more open implementation. Second, we were interested in exploring the linking of full text products with abstracting and indexing tools.

The Core Journals Project has created a substantial networked collection of journals, by networking UMI's Powerpages product. Over 600 journal titles have been made available by linking the images of journal articles (managed on a local CD jukebox system) to existing journal indexes in our online catalog (through matching algorithms for the citation data). Users are able to search the index, identify that a scanned article image exists, and request a print be delivered to selected campus printers. Printing at campus sites (including libraries) can be charged to individuals' computing accounts.

A third journal undertaking, the [JSTOR](#) (Journal Storage) project, was conceived and initially funded by the [Andrew W. Mellon Foundation](#). JSTOR, now a not-for-profit organization based in New York, is a large-scale undertaking to scan, OCR, and index the backfiles (up through approximately 1990) of selected journals. Michigan is serving as a service agency and host for JSTOR, overseeing the conversion and quality control processes and providing user support to the over 200 institutional sites.

Through JSTOR, we added to our repertoire important hands-on experience with large-scale conversion and the necessary added value aspects of indexing and quality control. For example, as the project grew in scope and volume, strategies for sampling during the quality control process became essential. In addition, the JSTOR team has grappled with enhancements to indexing specifications in order to reflect the unique structure of each journal--e.g., thematic issues, supplementary material, etc.

Perhaps the most ambitious local journal project to date is our expanded collaboration with Elsevier Science. The [PEAK \(Pricing Electronic Access to Knowledge\)](#) Project will make available all 1100 journals of Elsevier, testing our capability for production and host service support and also providing an environment in which we can explore issues surrounding pricing of electronic journals. Professor [Jeffrey MacKie-Mason](#) of Michigan's Department of Economics and School of Information, working with a team of librarians and technologists, has developed a methodology to explore pricing models which range from traditional subscriptions to article transactions to the creation of "virtual journals" through user-selected bundles of articles. In addition, we hope to explore non-linear approaches to pricing.

Several critical lessons have emerged from our journal project experience. First, a critical mass of useful content is important in securing user attention and investment. This suggests that libraries may be more likely to embrace collection development techniques which build volume more quickly than might be the case in traditional selection and acquisition practices. While users often develop habits of targeting inquiries to a particular digital collection, traditional techniques of identifying relevant articles via journal indexes remain important. This raises the second lesson of our experiences, namely the importance of linkage to "legacy" systems and services such as bibliographic indexes and catalogs.

Early experiences in journal delivery provided only mechanisms to deliver images of journal pages--e.g., our UMI project. However, increasingly users demand the functionality which comes with full text, hyperlinks, and more fully developed content indexing schema. Therefore, the format of the delivered journal content--particularly more robust and re-purposeable formats such as SGML--will be central to future activity.

Finally, a more mundane lesson: easy, flexible, and high quality printing capability is absolutely essential to the success of journal delivery.

Encoded Texts/Text Analysis

The University of Michigan has had a long term commitment to building encoded text collections in the humanities to support text analysis. Local capability has now been extended through a program of host services and middleware provision to other institutions (the [SGML Server Program](#)). In addition, the initiative has been rounded out with a program, the [Collaboratory for the Humanities](#), to support the development of new scholarly editions and new research foci for faculty.

[The Humanities Text Initiative](#) (HTI) has developed a substantial production capability for mark-up and delivery in SGML. HTI has its roots in work begun in the 1980's to build a substantial archive of literary and historical resources in SGML. That archive has evolved and expanded through conversion, commercial sources, and cooperative projects. Implementations of these SGML collections allow the user to navigate easily within large, complex bodies of data through efficient displays and representation of the layers and structure of the collections.

HTI has been a beta site for Soft Quad's Panorama, the first SGML "helper app" for the Web. Using Panorama, users can view the rich SGML which is especially important for representing certain information (e.g., scientific notation and formulae). In addition, SGML resources are dynamically rendered on the fly to HTML for access through web browsers. This "just-in-time" delivery strategy is described in [Just-In-Time Conversion, Just-In-Case Collections](#).

Some of the notable projects supported by HTI include:

- the [American Verse Project](#) which is converting American verse published before 1922,
- the [Middled English Compendium](#) which is creating an electronic version of the Middle English Dictionary and a hyperbibliography of Middle English text resources, and
- the [Advanced Papyrological Image System](#) which is converting papyri with transliterations/translations and links to published works.

The text projects have provided a fruitful venue to impact scholarship through functionality for text analysis and also new forms of authorship. The active engagement of scholars in this enterprise has been fundamental to its success.

Electronic Reference Shelf

Our SGML capabilities within HTI have been exploited in the deployment of several heavily used reference sources, primarily through partnerships with publishers to co-develop the digital products. Since many of these titles are widely held on campus, these projects have provided an opportunity to begin to raise issues surrounding the funding of licensed, core

digital tools. Figure 5 depicts the on-the-fly created table of contents information from long articles within the *Encyclopedia Americana*. The capability for using structure to present the information in ways to aid navigation and fully validated hyperlinks have added significant value to the functionality of these products.

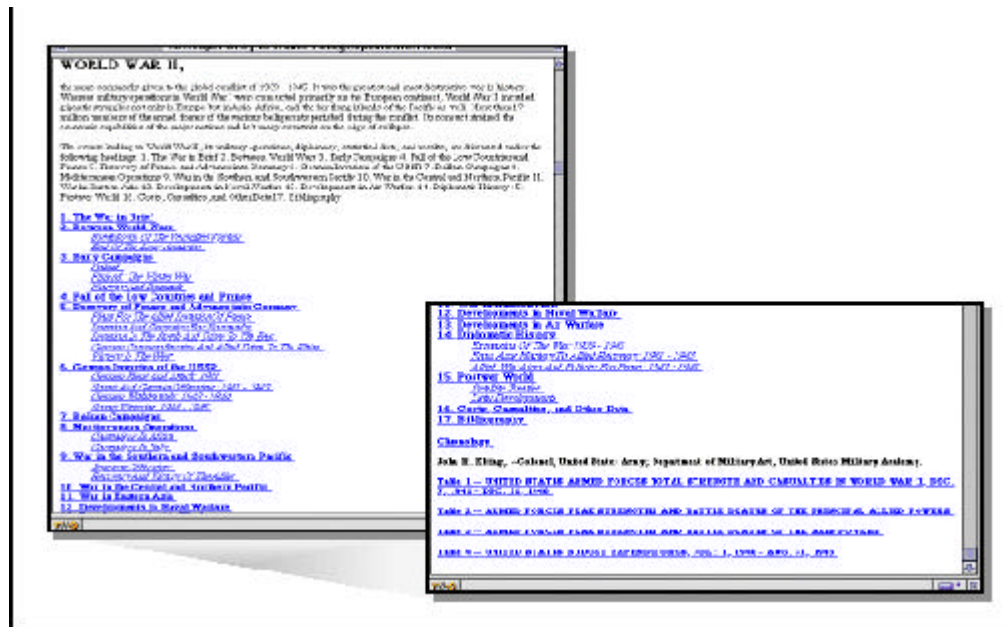


Figure 5: World War II Article Table of Contents from the Encyclopedia Americana

Conversion & Preservation

Digital libraries, particularly converted historical material, cannot live in isolation and strategies to develop conversion standards and link related collections are necessary. A project funded by the Andrew W. Mellon Foundation to convert American historical materials (initially focusing on 1850-1876) has been developed in partnership with [Cornell University](http://www.cornell.edu). As part of this undertaking we are exploring those elements of the digital library which must be done collaboratively to build a coherent library out of separate collections. [The Making of America Project](http://www.makingofamerica.org) has scanned and preserved 5000 volumes in the initial phase and a second phase with many more institutional participants is under development.

Michigan's approach toward conversion of book-length texts within Making of America has moved the conversion process further to include adding full text and basic structural elements. Figure 6 depicts the conversion steps which have been developed, with a vendor providing the basic indexing and scanning. Automated routines have been locally developed to auto-generate OCR and low-level (page-level) SGML mark up which aids search and navigation.

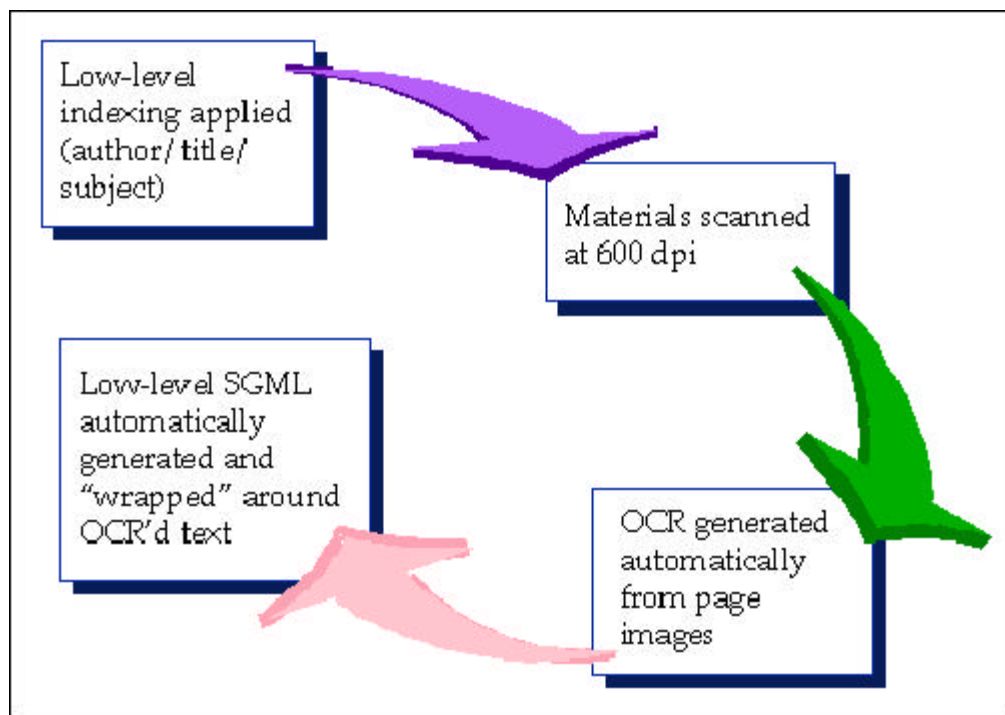


Figure 6: MOA Conversion Process

This process has resulted in a very cost-effective process of adding structure and search functionality to long documents.

Visual Images

The projects described thus far have been primarily text-based. Clearly, strategies for describing and deploying visual materials are an important capability for a digital library. Michigan's participation in the [Museum Educational Site License](#) (MESL) project has provided an opportunity to explore the necessary infrastructure for visual content. MESL is sponsored by the [Getty Art History Information Program](#). Six museums and the [Library of Congress](#) are providing content for MESL; in addition to the art images, documentation for each item is provided and described through over 30 attributes (e.g., creator, nationality, medium, etc.).

MESL has challenged us to more fully explore the requirements of deploying a digital collection which will have high instructional impact. We've learned a great deal about teaching styles with visual materials and also about the less-than-optimal state of most classrooms to handle use of digital media. Central to MESL's mission has also been the exploration of educational site licensing models for museum data. To build on this experience, additional commercial content has been added, and a Visual Image Service has been created to assist campus units in converting and managing visual image collections.

Numeric & Spatial Data

Numeric and spatial information poses a number of unique challenges. The research community knows well the past problems of handling tapes, computational resource needs, and complex analysis software. A joint development team with faculty input developed a

project which addressed the issues of complexity and access as well as the integration of related materials. Several core data sets (including some produced at Michigan such as the American [National Election Studies](#)) were moved to an online environment, codebook documentation was also converted to digital form, and a Web interface developed.

Our campus [Geographic Information Systems](#) group is also a collaborative effort. The group has launched workshops, fostered course development, and has also addressed the infrastructure requirements for spatial data. A distributed environment of data providers is assumed, but metadata and data management processes are being pursued to ensure coordination of these resources.

Digital Library Production

Michigan's digital library initiatives developed significantly during the first two years of the partnership. Projects were launched and there was general recognition that the project-based approach served us well. However, it became increasingly clear that we needed to move prototypes and projects to production and build a more clearly articulated organizational framework for production support and future project activity.

The framework for structuring near-term digital library activity--i.e., production operations and management-- builds on our previously articulated notions of core competencies for the partner organizations. In addition, several assumptions were identified about the scope of the evolving production capability, specifically:

- The environment should support multiple information types: image, audio, multimedia, structured text, visual images, numeric/spatial data, etc.
- Production services should be available (by contract) for campus unit information providers and off campus communities.
- Staff for production operations should be drawn from the partner organizations, yet retain current affiliations.
- As a virtual organization which relates, by design, to multiple organizational units, it is appropriate that its strategic directions will be set by a jointly populated advisory board.

In May 1996, the [Digital Library Production Service](#) was launched through joint sponsorship and funding. These core responsibilities were assigned:

- Near-term architecture development
- Document/data structure assessment
- Application development (i.e., new project support)
- Application maintenance, (i.e., existing project maintenance and enhancements)
- Server management
- Operations: data loading, indexing, system management/maintenance.

We also have defined the critical relationships which must support or draw on production services. These related organizational and functional components are illustrated in Figure 7.

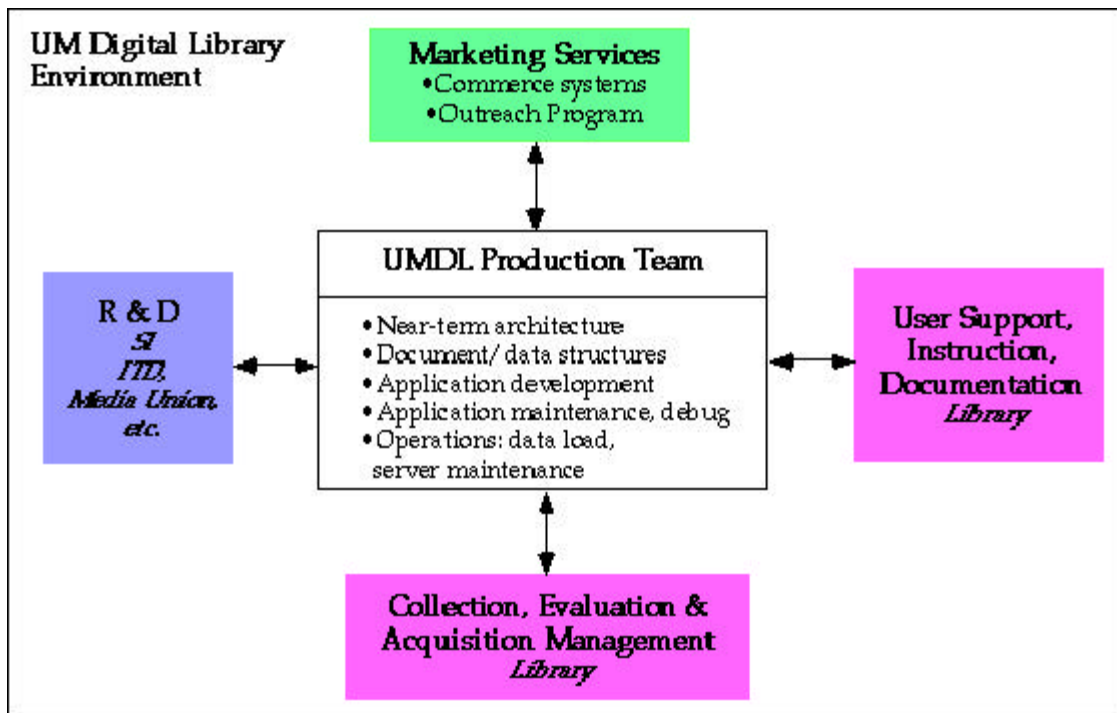


Figure 7: UMDL Production Organization and Functional Relationships

Federating Digital Libraries

The development of a production organization is evidence of the maturation of our digital library activity. Similarly, other research libraries are grappling with these same issues--i.e., how do we move beyond individual projects and address the broader issues of architectures for digital libraries? In response to these emerging issues, a group of 15 research libraries (including the Library of Congress) have come together in the [National Digital Library Federation](#) to address the implementation of a distributed, open digital library.

This group has focused on three themes in its two-year history. First, it has addressed issues of *discovery and retrieval* in a distributed environment (i.e., how do users find digital resources and how do we enable cross-collection search?). Second, the institutions are interested in new *economic models* for supporting the creation and distribution of digital information. Finally, in response to the Archiving of Digital Information Task Force Report, we are interested in addressing issues of *long-term access* to digital resources. These three themes have been addressed through projects sponsored by the NDLF participants.

Projects underway include multi-institutional efforts for papyri, numeric data, medieval manuscripts, and the aforementioned expanded Making of America project. In each of these projects, attention toward a distributed architecture (e.g., administrative and structural metadata, encoding standards for interoperability, etc.) is a central issue, along with the obvious development of digital content of broad scholarly use.

Concluding Remarks

The UMDL strategy has been successful in tapping necessary expertise and resources. Each project undertaken has added significant functionality and experience as well. If one looks broadly at the array of near-term projects and research underway at Michigan, a picture emerges of significant exploration and development in nearly every functional component essential to digital library development (or at least those that we can envision). Figure 8 is suggestive of the elements that we are exploring as they relate to the traditional library functions of collection development, access, mediation, and archiving.

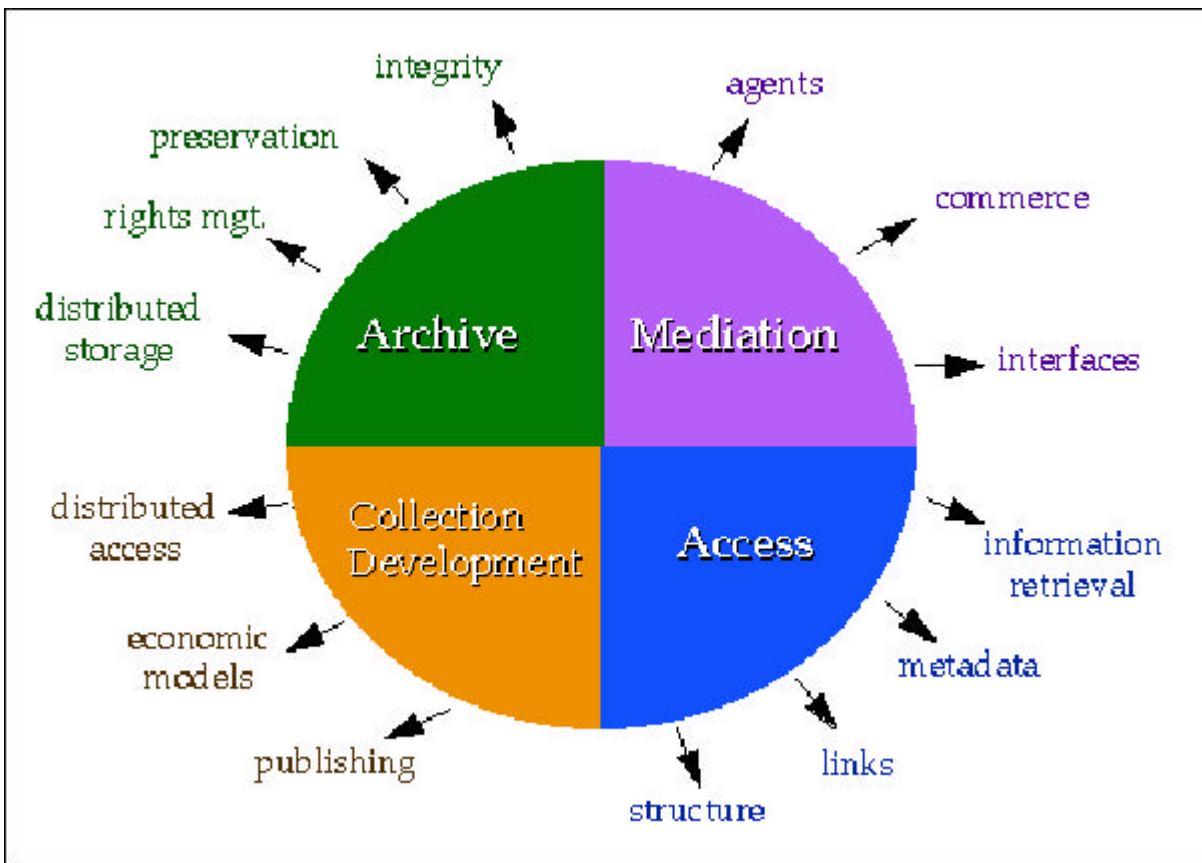




Figure 8: UMDL: Areas of Functional Exploration


Obviously, the development of University of Michigan Digital Library has been an evolutionary and iterative process. What began with a fairly straightforward approach of launching projects has now been more fully articulated as a broad agenda which builds content and infrastructure and enables the exploration of issues central to the future of scholarship. We have witnessed, too, an evolution of the notion of "library"--from a digital replication of traditional print resources to more robust environments for inquiry. Finally, we're beginning to see the emergence of "knowledge communities" or "collaboratories" which bring together digital content, analytic tools, and integral communication and collaboration capabilities. This suggests a maturation of the role of digital libraries which has been enabled by technology and collaboration with the scholarly community.


JOURNAL
STORAGE

REDEFINING ACCESS TO SCHOLARLY LITERATURE

ENTER JSTOR 

 ABOUT JSTOR

NEWS & NOTES 

 DEMO

[Enter JSTOR](#) | [About JSTOR](#) | [News & Notes](#) | [Demonstration Database](#)

©1998 JSTOR
JSTOR® and the JSTOR logo are Reg. U.S. Pat. & Tm. Off.
All Rights Reserved
[Contact JSTOR](#)

[BROWSE](#)[SEARCH](#)[RANDOM](#)[HELP](#)[ABOUT](#)

The original MESL cooperative agreement, concludes **June 1, 1998.**

Good News! Several of the museums have agreed to extend Michigan's use of their images. The images and searchable text will be migrated to a new system in [Digital Library Production Services](#), [Image Services](#) by the end of June 1998. In the meantime, please continue to use this site, and also come and see the new system, including the [History of Art Department Visual Resources Collections](#).

Extended use has been granted by: George Eastman House, Museum of Fine Arts: Houston, Fowler Museum of Cultural History: UCLA, and National Museum of American Art.
Still pending are: Library of Congress, and National Gallery of Art.

Concluded: Harvard University Art Museums.

Please see [Background and Purpose](#) for more info.

Users of these images and texts agree to adhere to the [Conditions of Use](#).

The **Museum Educational Site Licensing Project** is a pilot that exists to make museum information more accessible through electronic technology. Please explore the thousands of images of museum objects that are provided for educational use at the **University of Michigan**. This material is restricted to local users under a license agreement executed by all participating institutions.

Please send questions and comments to the MESL Web Team at... mesl.web@umich.edu



Fifteen of the nation's largest research libraries and archives have agreed to cooperate on defining what must be done to bring together--from across the nation and beyond--digitized materials that will be made accessible to students, scholars, and citizens everywhere, and that document the building and dynamics of United States heritage and cultures.

What's New NEW!

- *Name Change:* National Digital Library Federation becomes **Digital Library Federation**
-

About the Digital Library Federation (DLF)

- [America's Heritage: Mission and Goals](#) for a Digital Library Federation
 - [DLF Constituted as Charter Organization: Adopts Three-Point Agenda](#)
 - [DLF Policy Committee Members](#)
 - [DLF Planning Task Force Members](#)
-

DLF Policy Committee

Summary of Meetings:

- [June 19, 1996](#)
 - [September 18-19, 1996](#)
 - [November 22, 1996](#)
-

DLF Planning Task Force

- [DLF Planning Task Force: Final Report](#)

Summary of Meetings:

- [June 13, 1995](#)
- [November 14, 1995](#)
- [January 29-30, 1996](#)
- [March 18-19, 1996](#)

- [September 18-19, 1996](#)
-

Federation Member Web Sites and Digital Library Projects

- [Commission on Preservation and Access](#)
 - [Columbia University](#)
 - [New York State Museum Bulletins Project](#)
 - [Cornell University](#)
 - [Networked Computer Science Technical Reports Library](#)
 - [Prototype Cornell Digital Library \(includes Making of American Project\)](#)
 - [Emory University](#)
 - [Emory University Virtual Library](#)
 - [Harvard University](#)
 - [Information Infrastructure Project](#)
 - [Library of Congress](#)
 - [American Memory](#)
 - [National Archives and Records Administration](#)
 - [New York Public Library](#)
 - [Pennsylvania State University](#)
 - [Princeton University](#)
 - [Stanford University](#)
 - [Stanford Digital Libraries Project](#)
 - [Stanford University Computer Science Electronic and Technical Reports Library](#)
 - [University of California, Berkeley](#)
 - [UC Berkeley Digital Library Initiatives](#)
 - [NSF Digital Library Project](#)
 - [Computer Sciences Technical Reports Project \(NCSTRL\)](#)
 - [University of Michigan](#)
 - [Digital Library Initiatives](#)
 - [NSF Digital Library Project](#)
 - [Making of America \(MOA\)](#)
 - [University of Southern California](#)
 - [University of Tennessee](#)
 - [Yale University](#)
 - [Open Book Project](#)
-

Federation-sponsored Meetings and Conferences

- [Organizing the Global Digital Library II and Naming Conventions \(May 21-22, 1996\)](#)
-

Other Internet Resources on Digital Libraries

- [Digital Library Resources and Projects](#): A Library of Congress Internet Resource Page
-

Go to:

- [Library of Congress Greetings Page](#)
 - [Library of Congress Special Programs and Services Page](#)
 - [Library of Congress Home Page](#)
-

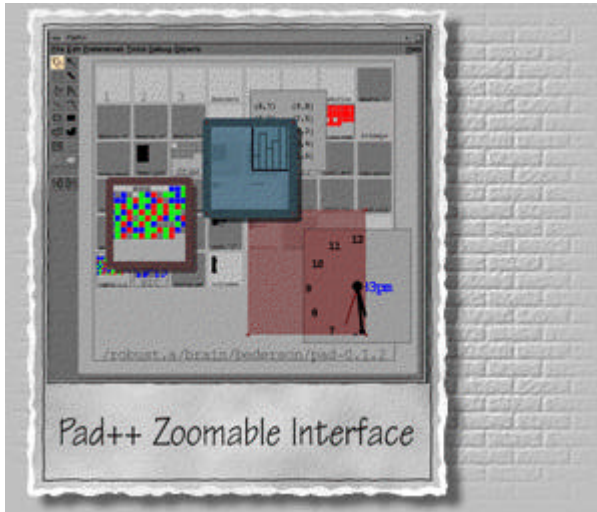


Library of Congress

Comments: lcweb@loc.gov (11/06/97)

Pad++: Zooming User Interfaces (ZUIs)

University of Maryland, College Park
Human-Computer Interaction Lab



We are exploring Zooming User Interfaces (ZUIs) where zooming is a fundamental part of the user's interaction with the computer (also known as *multiscale interfaces*).

in collaboration with [UCSD](#), [NYU](#), and [UNM](#)

You can find out about Pad++ through

- [A guided tour](#)
- [Frequently asked questions](#)
- [On-line papers](#)
- [Documentation](#)
- [Some places we've been in the press](#)
- [Download it now](#)

Zooming [Site Map](#)

Check out Ben Shneiderman's [use of Pad++](#) in his plenary talk at CHI98.

Pad++ is supported in part by DARPA grant #N660011-94-C-6039.



Just-in-time Conversion, Just-in-case Collections

Effectively leveraging rich document formats for the WWW

John Price-Wilkin
Head, Digital Library Production Service
University of Michigan
Ann Arbor, Michigan
jpwilkin@umich.edu

D-Lib Magazine, May 1997

ISSN 1082-9873

-
1. [Introduction](#)
 2. [Why on-the-fly?](#)
 - 2.1 [Converting for actual use, not potential use](#)
 - 2.1.1 [Case 1: Grolier's *Encyclopedia Americana*](#)
 - 2.1.2 [Case 2: Making of America](#)
 - 2.2 ["Re-purposing" and managing data](#)
 - 2.2.1 [Flexible delivery](#)
 - 2.2.2 [Collection management](#)
 3. [Conclusion](#)
-

1. Introduction

The University of Michigan's Digital Library Production Service (DLPS) has developed substantial experience with dynamic generation of Web-specific derivatives from non-HTML sources based on several key projects and consideration of how users work with key resources. This article is based on DLPS's experience and resultant policies and practices that guide present and future projects. In a rapidly changing world where the implications of information technologies for broad yet differentiated clienteles are mysterious, we hope that our experience will contribute to a better understanding of practical strategies.

The WWW has long included the ability to offer access to documents stored in formats other than HTML. Beginning with NCSA's "htbin" mechanisms, and soon after using the now widely embraced Common Gateway Interface (CGI), managers of large document collections have been able to store materials in a variety of formats, while offering these documents to a wide consumer base. This was the model used by the author in 1993 (when NCSA introduced Mosaic) to build access to large collections of documents stored in a variety of forms of SGML.¹ Once CGI was

introduced, in fact, the Internet quickly saw the introduction of gateways to popular enterprise database systems such as Oracle, and a variety of other data types such as numeric data files (e.g., [American National Election Studies](#)).

A fundamental part of this strategy is the real-time creation of Web-friendly versions of material in formats not natively supported by Web browsers. Frequently, this strategy is taken as a matter of course: after all, relational databases and numeric data files are not "documents", and choices have always been made about real-time derivatives for display. In the world of documents, however, especially those encoded as SGML or as high-resolution page images, document managers have needed to choose between a strategy that **pre-computes** and **stores** derivatives for the WWW, or generating the Web-specific version **on-the-fly** for the end-user.

The choice between creating derivatives dynamically and storing static Web-accessible collections is, fundamentally, a management strategy. Neither strategy is intrinsically "good" or intrinsically "bad", though either can be adopted for ill-conceived reasons and implemented in shortsighted or wasteful ways. The choice between pre-computed derivatives and dynamically-derived documents is sometimes seen as a simple dichotomy. For example, in *comp.text.sgml*, one writer described the choice as a basic opposition of computing cycles and disk space: if computing cycles are more expensive than disk space, then we should conclude that creating and storing HTML derivatives of non-HTML sources is more cost effective because we can use cheap disk to reduce the CPU load caused by repetitive conversions. This reductionist view is frequently wrong, and more importantly it obscures the importance of administrative decisions involved in each choice.²

2. Why on-the-fly?

The DLPS currently offers dozens of collections, including more than 2,000,000 pages of SGML-encoded text and more than 2,000,000 pages of material using TIFF page images.³ All of the material in these collections is offered through the WWW, and nearly all of it is presented in Web-accessible formats through real-time transformations of the source material. Two primary considerations go into our decision to make the material available through dynamic rather than pre-computed and stored transformations. First, we assume that the patterns of use in our collections mirror those of traditional libraries, where many of the materials go unused for significant periods of time, and where many resources are used only once in an extended period of time. Consequently, creating derivatives for *potential* use will result in most derivatives being unused and both computational and human resources being wasted. Second, by maintaining the data in the richest possible format, we are able to use the same source in a variety of different ways as circumstances and tools allow. This allows for forward migration as the technologies for access and manipulation of information on the desktop mature.

2.1 Converting for actual use, not potential use

As digital libraries increase in size, they begin to exhibit many of the same patterns of use that the traditional library has seen: many resources are infrequently used. As materials are added to our currently relatively small digital libraries, increasingly smaller segments receive the sort of focused attention that we associate with actual use such as reading or printing. The extraordinary boon of the electronic format includes the ability to search across large bodies of material, "touching" many documents along the way. Still, relatively few of the documents will be read or printed.⁴

At the same time that we continue to build these massive collections of digital documents, HTML and online delivery mechanisms are changing frequently. When the Humanities Text Initiative (HTI), a DLPS unit that supports SGML collections, first mounted the *Information Please Almanac*, Web browsers did not yet support tabular presentation of data through the use of the TABLE element in HTML. The strategy employed for the *Almanac* was one that used HTML's <PRE> to force a fixed-pitch representation of the extensive tabular data. When Netscape began to support tables, the HTI staff made a small change to the transformation routines, thus enabling the display of the tabular data through Netscape's new functionality.

Each passing month sees the introduction of significant new functionality--recently frames, multiple fonts in a single document (and thus multilingual documents), Unicode support, and soon math. This trend should continue indefinitely, and we should expect to see a continually richer capability for presentation of information through the Web. Historical deficiencies such as a lack of tabular support, the inability to support superscript and subscript, or multilingual document support were remedied by interim measures such as inline, transparent GIFs, but history has shown that mechanisms will eventually be introduced to support all commonly used document layout features. By continuing to store *only* the richest version of these documents (rather than the derivatives), and by computing new versions only when needed, we are able to continue to tap the best features of the Web without either compromising the document or needing to manage a parallel collection.

For the DLPS, an approach that involved caching the majority of our data for prospective use--use that will not take place before we need to re-convert the material--would result in "wasted" computing cycles and wasted management effort. If we were to continue to convert and store versions of our document stores with each added capability in HTML, we would inevitably devote attention to detail for a variety of documents that would not be used during an "era" of HTML. Moreover, except for the strongest willed among us, automatic conversion from one format to another will almost always lead to some degree of dissatisfaction with the result. A project that generates static derivatives will be tempted to manually introduce "improvements" to the cached material. The simple argument that precomputed derivatives save computing cycles may actually prove false as collections grow in size, as smaller percentages of materials are used, and as we increase functionality.

2.1.1 Case 1: Grolier's [Encyclopedia Americana](#).

It may be helpful to examine a case where DLPS scrutinized these decisions. Grolier's *Encyclopedia Americana* is one of the more heavily used reference collections provided to the University of Michigan user community by DLPS. The *Encyclopedia Americana* is provided to the DLPS in a very rich SGML, directly from the publisher's editorial process. The same SGML used to create the printed encyclopedia is used to build the online system. The *Encyclopedia Americana* is brought online by indexing the SGML and writing programs that allow users to navigate content and structure, for example by presenting the user with a hierarchical representation of the organization of an entry. These "versions" of the encyclopedia--both HTML articles and browsable intermediates--are generated for the user in real-time, creating HTML from the single, large online SGML collection. Users are provided a more easily navigated resource (i.e., one whose navigational information can be made to reflect an arbitrarily arrived at context), as well as one that makes the most of current HTML display capabilities (e.g., by introducing font-based support for non-Roman alphabets).

Transaction-level use data were analyzed for the period beginning in August 1996, and ending in early April 1997, for a total of eight months. Users conducted a total of 43,959 searches in the *Encyclopedia Americana* during this period (see [Table 1](#)). CGI transactions that retrieved articles were extracted to help determine the unique set of articles viewed by users (and thus, the articles that would have benefited by being converted to HTML).⁵ Despite being heavily searched, only 9.3% of the total number of available articles were viewed during the eight month period, and of these, 65% were used only once.

What can we conclude from these patterns of use? The *Encyclopedia Americana* is certainly heavily used, but despite this, only a small percentage of the articles in the encyclopedia are read or printed. Fewer still (35%) were displayed more than once. Investing significant resources in converting all 42,882 articles in the *Encyclopedia Americana* to HTML would result in the majority of these resources expended for no real purpose. Further, the likelihood that new features of HTML browsers would make it possible to improve navigation and display of results would mean that these same articles would need to be re-converted when the new features were available. At the same time, continuing to hone the capabilities of the real-time conversion capabilities means that the display of results can be improved with relatively little trouble.

2.1.2 Case 2: [Making of America](#)

The Making of America project will be treated at greater length in a forthcoming issue of *D-Lib*, but some brief comments on its deployment may be helpful in understanding another approach to just-in-time conversion. The Making of America (MOA) project is a collaborative collection building enterprise between Cornell University and the University of Michigan. The MOA collection at the University of Michigan consists of some 2,500 volumes of nineteenth-century publications (primarily monographs, but eight journals as well) published in the United States between 1850 and 1877, and focusing on American social history from the antebellum period through reconstruction. The materials are stored in TEI-conformant SGML that points to 600dpi TIFF images of pages. Materials are prepared for conversion at the University of Michigan, and are scanned as TIFF images by a service bureau. Significant local processing takes place to create OCR and the encoded text over which searches are conducted.

Few of the volumes in the online collection are corrected, fully-encoded SGML. The rough OCR that sits behind the scenes provides an effective mechanism for users to navigate the vast quantities of material (650,000 pages), but users are presented only with brief navigational information in HTML. The book itself is viewed on screen or is printed by using the images of the pages. The 600dpi images are an excellent source for creating printed volumes, and because of the relatively high resolution, a number of different resolutions can be easily derived. Using [tif2gif](#), a program developed by Doug Orr for University of Michigan Digital Library projects, GIF images suitable for several different display resolutions (reflecting the user's preference) are created in real time. The same single source can thus be used to create printed editions, and to serve a variety of different user needs. The choice to do this in real time, rather than pre-computing and storing these derivatives, does not adversely affect performance for users (i.e., it happens quickly enough that users are unaware that the version is being created), and helps us avoid the need to convert, store, and manage pages that are seldom or never used.

Data similar to those collected for the *Encyclopedia Americana* are not yet available for the Making of America project. However, because the printed materials were held in the Library's storage facility and were used only infrequently (e.g., once every twenty years), we assume that

they will be used less frequently than the online reference sources. Historical resources such as these will probably exhibit overall less use and less repeated use. Regardless which type of collection, however, large numbers of books or articles go unused, except to report to the user of the full-text search that they do not contain results, or to present a "pick list" choice for a user that is subsequently unused.

2.2 "Re-purposing" and managing data

Maintaining the data in its original structured or high-resolution format rather than HTML or derived images allows DLPS to re-purpose the data in a variety of flexible ways. Texts, and especially portions of texts, are more easily re-purposed. Portions can be displayed for different uses. For example, section heads can be drawn from documents to create context-aware tables of contents. Moreover, the awareness (by the system) of the arrangement of documents--of parts to the whole--provides a powerful tool for collection management.

2.2.1 Flexible delivery

When the system "knows" about the structure of the document, different portions can be delivered to the user for different needs and in different contexts. A volume/issue browse can be generated from the same set of data as a search resulting in links from authors and titles. A system that provides relevance at the level of the subdivision can show the relationship, dynamically derived, of the part to the whole (see [Figure 1](#)). A small document subset, such as quotations in the OED, can be delivered as ends in themselves (see [Figure 2](#)) instead of links to dictionary entries that frequently exceed ten printed pages. Synthetic documents, assembled by pulling together parts from many different wholes, can be created for both real and whimsical purposes (see the poem in [Figure 3](#), where each line was automatically extracted from a collection of several thousand poems in real time; cf. <http://www.hti.umich.edu/bin/amv-idx.pl?type=random>)

2.2.2 Collection management

For large document stores--especially large collections of large documents--a cached HTML approach results in version control issues of tremendous proportions. Consider, for example, the HTI's American Verse Project collection, where the current collection of seventy-four volumes is divided into 2,308 major sections (most of these too large for web-based delivery), 2,222 secondary sections, 629 tertiary sections, and so forth. A pure HTML approach will result in approximately 5,000 separately named objects for this relatively small collection and more than 40,000 objects by the time the project is complete. The need for frequent revision of the objects, if not accomplished through entirely automatic mechanisms, will present problems of collection management not present when the parts of the whole are represented through markup.

3. Conclusion

While different project imperatives should necessarily lead to different approaches, the relatively large collections of richly represented documents in the University of Michigan Digital Library has led to an approach that favors "just-in-time" rather than "just-in-case" conversion. Real-time transformation of high-resolution page images and richly encoded documents has proven possible without noticeably diminishing performance for end-users. This strategy has allowed us to store materials in forms that pre-date the web and continue to exceed the Web's capabilities for display, and yet to make them available in Web-capable formats, in increasingly attractive or informative

ways. The WWW will continue to change with remarkable speed, and choosing to store these large collections in their richest form while dynamically converting to today's HTML makes it possible for us to continue to keep these rich libraries alive while avoiding the creation and maintenance of a series of interim collections.

The speed and direction of the Web's improvements makes it seem likely that this dichotomy of dynamic and static conversion will become increasingly moot. At the same time that we see continuing enhancements being made to HTML, we are also witnessing an increased interest in native support for richer formats. The recent creation of [XML](#), sponsored by the W3C, may make it possible to send natively encoded SGML documents to Web browsers. For our page image systems, the native TIFF compression scheme by [Cartesian Inc.](#), CPC, should begin to make it possible to send the TIFFs themselves, rather than derivatives. At the same time, by continuing to rely on the richer encoding, especially of SGML documents, we are better able to support navigation and the display of partial documents as needed.

The strategy of "just-in-time conversion" paired with "just-in-case storage" goes to the heart of digital libraries. We cannot reliably predict which materials will be used or relevant for research. *Effective* digital libraries will be those that make their resources available in ways that do not influence research by using predictive methods that penalize the user who steps outside the mainstream. Relying on dynamic transformation methods for large digital collections positions the digital library in ways that allow us to take advantage of future capabilities without losing access to historical collections.

Table 1: *Encyclopedia Americana* use from August 1996 to March 1997

	Aug-Dec '96	Jan-Mar '97	Total	Percent of total arti
Total EA transactions:	24,073	19,886	43,959 ^[6]	
Total article retrievals:	2,767	1,237	4,004	9.3%
Total unique retrievals:	1,962	870	2,625 ^[7]	6.1%
Total articles in EA:			42,882	

Table 2: Representative DLPS collections

Collection	Type/Domain	Format
Encyclopedia of Science and Technology ; Physician's GenRx	Document retrieval, sciences	SGML using unique DTDs from different publishers
Encyclopedia Americana ; Oxford English Dictionary ; American Heritage Dictionary	Document retrieval, general reference	SGML using unique DTDs from different publishers
<i>Bryn Mawr Classical Review</i> ; <i>Bryn Mawr Medieval Review</i>	Document retrieval, humanities	journal articles, using TEI encoding
American Verse Project ; Corpus of Middle English Prose and Verse	Document retrieval and analysis, humanities	SGML locally-encoded using TEI
Elsevier Science journals	Document retrieval, sciences etc.	rough OCR behind scenes (no encoding--simple use of file system for management); 300 dpi bitonal TIFF page images for viewing
Making of America	Document retrieval and document analysis, humanities	loosely-encoded TEI of OCR behind scenes, 600 dpi bitonal TIFF page images for viewing

Notes

[1] Price-Wilkin, John. "A Gateway between the World Wide Web and PAT: Exploiting SGML Through the Web." *The Public-Access Computer Systems Review* 5, no. 7 (1994): 5-27.

[2] As one might assume from this position, not all DLPS collections rely on real-time derivation of Web-friendly sources. In our high resolution imaging projects in support of papyrology and manuscript study, we store a variety of reduced resolution versions for different types of applications. Of course this practice may change with available tools. A brief synopsis of some of the more notable DLPS projects and their formats is appended in [Table 2](#).

[3] A fuller discussion of two DLPS initiatives and their collections will appear in a forthcoming issue of *D-Lib*. For a brief discussion of some DLPS collections, see also "Just-in-time Publishing; Just-in-case Libraries: Cooperation at the University of Michigan." Presented at Rice University, April 10, 1997. [Available on the WWW](#).

[4] We might hope or expect that *more* and *different* documents will be read or printed by virtue of the increased level of access available in full-text search. The digital library should indeed affect the way that we conduct research and the results that come out of that research, but researchers will not necessarily use significantly larger amounts of material.

[5] Because not all article retrievals take place in the same way through the CGI mechanisms, this process excluded 361 "cross-reference" article retrieval transactions in the '96 (11% of all article retrievals in the period) and 199 "cross-reference" article retrieval transactions in '97 (14% of all article retrievals in the period).

[6] The difference between total transactions and article retrievals is accounted for by a higher number of searches than actual article use, and by browse transactions for parts of the articles represented in article retrievals.

[7] This figure is not a sum of the figures to the left. Instead, duplication across the entire period was taken into account.

Copyright © 1997 John Price-Wilkin



hdl:cnri.dlib/may97-pricewilkin

[Library of Congress:](#)

- American Memory <http://lcweb2.loc.gov/>
 - Call about American Memory <http://lcweb2.loc.gov/ammem/award/>
 - Sponsors and Contributors to the National Digital Library Program
<http://lcweb2.loc.gov/ammem/sponsors.html>
-

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Centers\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

[\[text-only\]](#)

June 22, 1998

Today in History

The LIBRARY of CONGRESS



32 Collections Now Available

SEARCH[Options](#) | Simple search: **BROWSE**[Titles](#) | [Topics](#)
[Library Divisions](#)

----- Collection Types -----

[Photos & Prints](#)[Documents](#)[Motion Pictures](#)[Maps](#)[Sound Recordings](#)**LEARN**Organized help for using the
collections

Show case

Sample Collection

[Washington](#)**New Offerings**[Lincoln](#)[A Century
of Lawmaking](#)[Northern
Great
Plains](#)[Map Collections](#)

[A Unique Public-Private Partnership](#)
[Supporting the National Digital Library](#)

See who is helping to bring a virtual library to all Americans for the 21st Century

[National Science Foundation Digital Libraries Initiative](#)
Co-sponsored by the Library of Congress

[Previews and Future Collections](#)**[How to View These Collections](#)**[LC/Ameritech Competition](#)[Copyright and other Restrictions](#)[Background Papers and Technical
Information](#)[About the Special Collections of
the Library of Congress](#)Send comments about these historical collections to ndlpcoll@loc.gov

am 4-28-98

Library of Congress
General Comments:
lcweb@loc.gov

NOTICE



National
DIGITAL
Library



A Unique Public-Private Partnership

Supporting the National Digital Library



The Library of Congress, with the bipartisan support of the United States Congress, the Executive Branch, and America's entrepreneurial and philanthropic leadership, is bringing the National Digital Library to the nation.

The Library has proposed a public-private partnership for fiscal years 1996 - 2000 that would include a minimum of \$3 million in annual appropriations from Congress. To date, the Congress has appropriated more than \$9 million toward a total appropriation of

\$15 million by the year 2000.

To achieve the program's fundraising goal of \$60 million, the Library has embarked on a major campaign to raise the remaining \$45 million from private funds.

Sponsors and Contributors to the National Digital Library Program

The Library gratefully acknowledges the generosity of the following sponsors and contributors whose support is instrumental to the success of the National Digital Library.

The United States Congress

Founding Sponsors*

Mr. John W. Kluge
The David and Lucile Packard Foundation

*Indicates contributions and pledges of \$5 million or more

Charter Sponsors**

Ameritech
Anonymous
AT&T Foundation
Bell Atlantic Corporation
Citicorp Foundation
Discovery Communications, Inc.
Donaldson, Lufkin & Jenrette
Eastman Kodak Company
H.F. Lenfest
Jones Family Foundation
Glenn R. Jones (Jones Intercable)
Federal Express Corporation
W.K. Kellogg Foundation
Laurance S. and Mary French Rockefeller
McCormick Tribune Foundation
Pew Charitable Trusts
Occidental Petroleum Corporation
Reuters

**Indicates contributions and pledges of \$1 million to \$5 million

Contributors

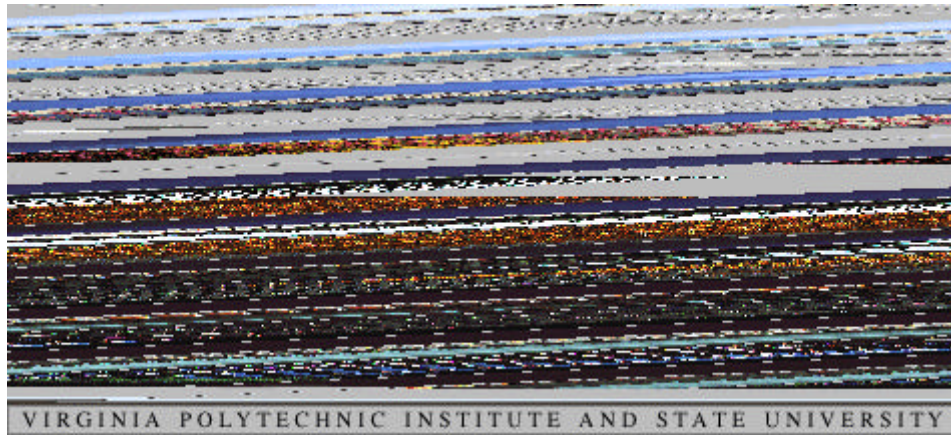
Bankers Trust
COMPAQ Computer Corporation
R.R. Donnelly & Sons
The Ford Foundation
The Hearst Foundation, Inc.
David H. Koch Charitable Foundation
Mr. Carl H. Lindner
Lucent Technologies
NYNEX Foundation
Shell Foundation
Texaco Foundation

In-kind Contributors

Hewlett-Packard Company
International Business Machines Corporation
LizardTech

UK Electronic Libraries Programme ([eLib](#))

- There are online [working papers](#)
- It is funded by the Joint Information Systems Committee (JISC).
- It is based on the Libraries Review by the UK Higher Education Funding Councils, chaired by Professor Sir Brian Follett in 1993. They prepared the [Joint Funding Council's Libraries Review Group Report, Prof. Sir Brian Follett, HEFCE, 1993](#).
- As a response, JISC started eLib with 15 million pounds over 3 years, to engage the Higher Education community in developing and shaping the implementation of the electronic library.
- There have been 2 separate calls and over 60 projects in areas:
 - access to network resources
 - digitisation
 - document delivery
 - electronic journals
 - electronic short loan collections
 - images
 - on demand publishing
 - pre-prints and grey literature
 - quality assurance
 - supporting studies
 - training and awareness
- One part is the [Arts and Humanities Data Service](#) and its service for the Visual Arts.
- On preservation, a workshop was held, with report Long Term Preservation of Electronic Materials: a JISC/British Library workshop as part of the Electronic Libraries Programme, Organised by UKOLN, 27-28 November 1995, U. of Warwick, prepared by the Mark Fresko Consultancy, 1995, also available [online](#).
- [Ariadne](#) is an eLib ejournal providing current information on eLib and digital libraries in general.



CS Dept. NSF-Supported Education Infrastructure Project / ei.cs.vt.edu

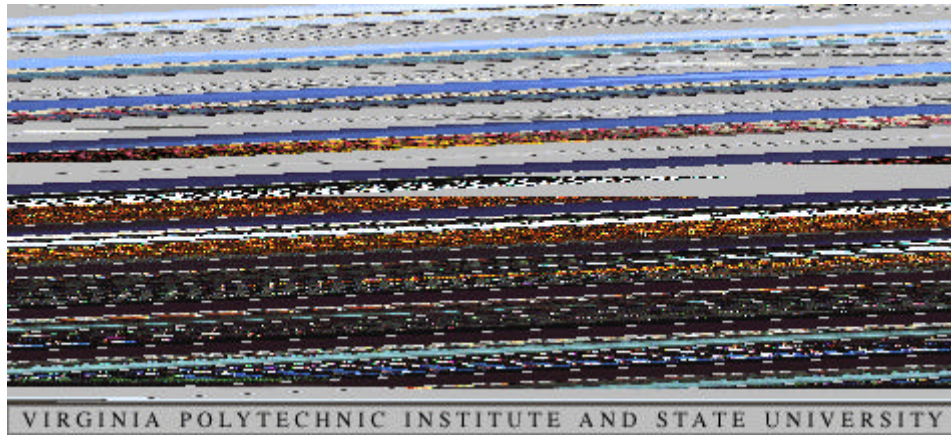
Welcome to the home page for the [NSF-supported](#) project "Interactive Learning with a Digital Library in Computer Science". We hope you find some of the results of our project useful! Please [send me comments and suggestions](#)! Thanks, Prof. E. A. Fox.

- [Courses](#) (over 40, with over 10K files)
- [Search ei.cs.vt.edu etc. \(e.g., all courses\)](#)
- [National EI Projects Home Page](#)
- [Computer Science Teaching Center](#)
- [Curriculum Resources in Interactive Multimedia](#)
- [Computer Science Education Innovation Workshop](#) June 15-21, 1997 **NEW!**
- [QUIZIT Software and thesis](#) **NEW!**;

Papers:

- [QUIZIT: An interactive online quiz system for WWW-based instruction](#) - Tinoco, L. C., Fox, E. A., Ehrich, R. W, Fuks, H. In Proceedings of the VII Brazilian Symposium on Educational Technology. Belo Horizonte, Brazil, Nov. 1996.
- Online Evaluation in WWW-based Courseware - Tinoco, L. C., Fox, E. A., Barnette, N. D. In Proceedings ACM SIGCSE'97, San Jose, Feb. 1997:
[paper in PDF](#), [presentation in PowerPoint](#)
- [Audio and Video Tutorials on Popular Tools and Systems](#)
- [Project Overview](#)
- [ENVISION Project \(that led to development of the digital library\) Final Report](#)
- [SWAN \(algorithm vizualization system\)](#)
- [Electronic Submissions of Student Programming Assignments used in CS3204](#)
- [Faculty Development Institute](#)
- [References](#)
 - [Project Overview \(for FIE'96, in PDF\)](#)
 - [Project Interim Report, Oct. 1996](#)
 - [Project Report for NSF EI PI Meeting, Nov. 1996](#)

This DEC Alpha, ei.cs.vt.edu, supports the Virginia Tech/Norfolk State University project "Interactive



CS Courses

Welcome to one of the largest (over 40 courses, over 10K files) repositories of Computer Science courseware! I hope you benefit and [send me comments and suggestions!](#)

Regards, Prof. E. A. Fox for

[Virginia Tech CS Dept.'s NSF Education Infrastructure Project](#)

[Search ei.cs.vt.edu etc. \(e.g., all courses\)](#)

- [CS1004: Computer Literacy](#)
- [CS1014: Numerical Computational Techniques](#)
- [CS1024: Computing For Business](#)
- [MaSc1044: Computer Science: A Liberal Arts Approach](#)
- [CS1044: Programming in C](#)
- [CS1104: Introduction to Computer Science](#)
- [CS1205: Operating System Tools I](#)
- [CS1344: Introduction to C Programming](#)
- [CS1206: Operating System Tools II](#)
- [CS1604: Computers and Networked Information](#)
- [CS1704: Introduction to Data Structures & Software Engineering](#)
- [CS2304: Self Study in a Programming System \(Java\)](#)
- [CS2304: Self Study in a Programming System \(UNIX\)](#)
- [CS2604: Data Structures and File Processing](#)
- [CS2704: Object-Oriented Software Design and Construction](#)
- [CS2964: Field Studies](#)
- [Honors 3004: Digital Libraries](#)
- [Honors 3004: Multimedia Technology and Projects \(1996\)](#)
- [UH3004: High-Performance Scientific Computing](#)
- [CS3204: Operating Systems](#)
- [CS3304: Comparative Languages](#)
- [CS3304sm: Comparative Languages \(Summer, 1997, offering\)](#)
- [CS/Math 3414: Numerical Methods](#)
- [CS3604: Professionalism in Computing](#)
- [CS3724: Introduction to Human-Computer Interaction](#)
- [CS4104: Data and Algorithm Analysis](#)

- [CS4114: Formal Languages](#)
 - [CS4124: Theory of Computation](#)
 - [CS4204: Computer Graphics](#)
 - [CS4214: Simulation and Modeling](#)
 - [CS4234: Parallel and Distributed Computing](#)
 - [CS4414: Issues in Scientific Computing](#)
 - [CS4504: Computer Organization](#)
 - [CS4624: Multimedia, Hypertext and Information Access](#)
 - [CS4964: Field Studies](#)
 - [CS4984: WWW - The Underlying Technology](#)
 - [CS5014: Research Methods in Computer Science](#)
 - [CS5024: Models and Analysis](#)
 - [CS5034: Models of Computation](#)
 - [CS5104: Computability and Formal Languages](#)
 - [CS5114: Theory of Algorithms](#)
 - [CS5204: Operating Systems](#)
 - [CS5224: Systems Simulation](#)
 - [CS5314: Concepts of Programming Languages](#)
 - [CS/ECE5515: Computer Architecture](#)
 - [CS/ECE5516: Communication Networks](#)
 - [CS5604: Information Storage and Retrieval](#)
 - [CS5614: Database Management Systems](#)
 - [CS5724: Models and Theories of HCI](#)
 - [CS5734: Computer-Supported Cooperative Work](#)
 - [CS5814: Digital Picture Processing](#)
 - [CS6104: Algorithmic Number Theory](#)
 - [CS6204: The World-Wide Web: Beyond the Basics](#)
 - [CS6204: Java and the WWW](#)
 - [CS6404: Advanced Topics in Mathematical Software](#)
 - [CS6604: Digital Libraries](#)
 - [CS6604: Interactive Accessibility \(1995\)](#)
 - [CS6724: HCI of Collaborative Systems](#)
 - [Digital Libraries - self study](#)
-

Catalog Pages

- [Ugrad](#)
- [Grad](#)

Class Data Archives

- [EI Archives](#)
- [CS Department Archives](#)

Usage Statistics

All materials prepared for these [Dept. of Computer Science](#) courses are

Copyright 1995, 1996 [Virginia Tech](#)

Linking to or using these works for educational use is encouraged.

Commercial use of these works is strictly prohibited.

See also

- [CS listing for World Lecture Hall](#)
 - [NSF Computer Science Courseware Repository \(NSFCSCR\)](#)
 - [Computational Science Education Project](#)
-

Author: [Edward A. Fox](#)

Email: fox@vt.edu

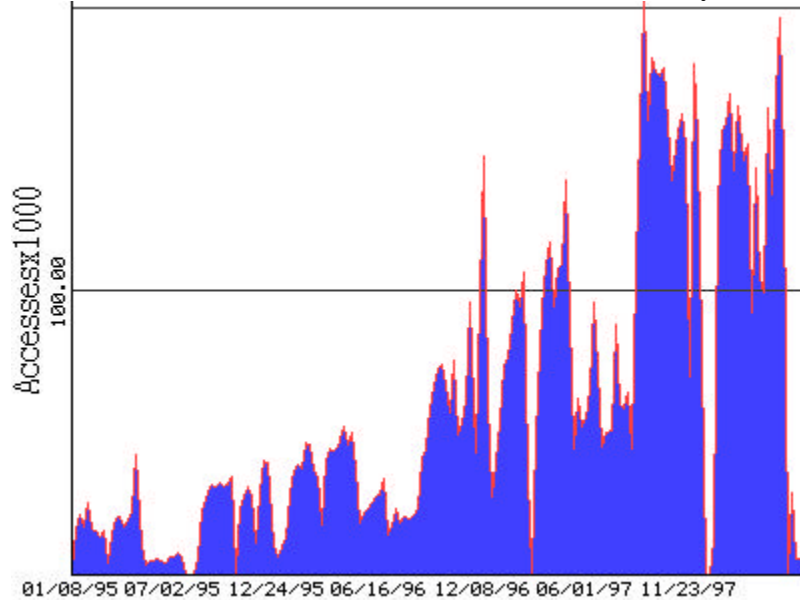
EI Statistics

History from 01/08/95 to 06/20/98

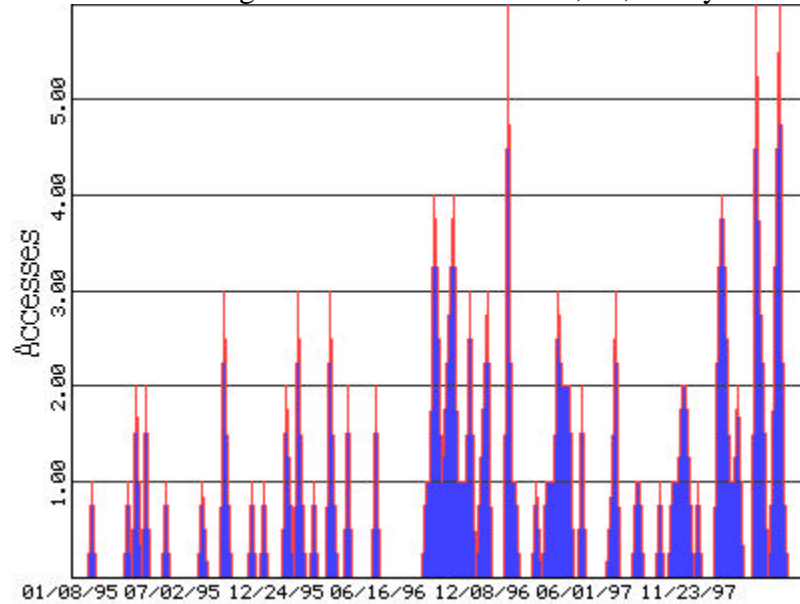
[Skip to Weekly Reports](#)

Totals

- Overall Hits: 10,706,544 accesses, 125,705,431,180 bytes



- Courses Home Page Accesses: 136 accesses, 11,940 bytes



Reports

- [Reports for the Year 1998](#)

D-Lib Magazine, December 1995

Project Briefings and Updates

Making a Digital Library

The Chemistry Online Retrieval Experiment A Summary of the CORE Project (1991-1995) December 1995

Contributed by:

Richard Entlich, Cornell University
Lorin Garson, American Chemical Society <http://pubs.acs.org>
Michael Lesk, Bellcore <http://community.bellcore.com/lesk/home-page.html>
Lorraine Normore, Chemical Abstracts Service
Jan Olsen, Cornell University
Stuart Weibel, OCLC <http://www.oclc.org:5046/~weibel>

The CORE project was an electronic library prototype of primary journal articles in chemistry, containing about four years of twenty primary journals published by the American Chemical Society (about 400,000 pages). CORE included both scanned images and an SGML (Standard Generalized Markup Language) marked-up version for on-the-fly rendering for screen display. Each page was scanned and segmented, with graphical units isolated and linked to figure references in the articles. The original machine-readable typography was converted to SGML format and the results were used to build databases with indexes for full-text Boolean searching.

Each page image was stored as a 300 dpi bitonal image for printing, and 100 dpi greyscale for screen display. All text data and the most recent page images were available on Unix-based magnetic storage at any given time, with additional (older) page images stored on a WORM (Write Once, Read Many) jukebox.

Complex scientific material (superscripts, tables, equations, special fonts and symbols, etc.) presents substantial problems for representation and display, especially when the material is being converted from previously published information, as were these journals.

The tasks of building and maintaining electronic journal databases remains formidable (especially if conversion from older formats is involved). However, experiences with chemists in this project suggest that electronic publishing will be popular with scholars, even though there remain significant disadvantages and impediments to adoption.

Analysis of user studies and transaction logs is ongoing and will be submitted for publication in the near future.

Further information on the CORE Project can be found at:

<http://www.oclc.org:5047/oclc/research/projects/core>

Acknowledgments: The CORE project thanks Sony of America, Digital Equipment Corporation, Sun Microsystems, and the Cornell Theory Center (which receives major funding from the National Science Foundation, and New York State, and additional funding from ARPA, the National Institutes of Health, and IBM Corporation).

[d-lib home](#)[d-lib magazine](#)[d-lib next](#)[comments](#)

hdl://cnri.dlib/december95-briefings.2

TULIP - The University Licensing Program

When you scroll further down this page you'll find

- [Introduction](#)
 - [The TULIP Final report](#)
 - [TULIP Newsletters](#)
 - The [Journal Titles](#) in TULIP
 - The [Universities](#) involved in TULIP
 - [Contact information](#)
-

Introduction

TULIP is a cooperative research project testing system for networked delivery and use of journals, performed by Elsevier Science and [nine Universities](#) in the USA. The participants set three objectives at the outset:

Technical

To determine the technical feasibility of networked distribution to and across institutions with varying levels of sophistication in their technical infrastructure. "Networked distribution" means sending the information both across the national Internet and over campus networks to the desktops of students and faculty. Elsevier will deliver the journal information to participating universities in standard formats. The universities will incorporate the information in local prototype or operational systems. A wide variety of delivery alternatives, search and retrieval systems and print-on-demand options will be compared.

Organizational and economic

To understand, through the implementation of prototypes, alternative costing, pricing, subscription and market models that may be "viable" in electronic distribution scenarios; comparing such models with existing print-then- distribute models; and understanding the role of campus organizational units under such scenarios. The overall goal is to reduce the unit cost of information delivery and retrieval. "Viable" means economically and functionally acceptable to all parties.

User behaviour

To study reader usage patterns under different distribution (technical, organizational and economic) situations. Improvement in the functionality of the information, whether as to article structure or retrieval tools, will also be considered. Certain data will be collected uniformly at all sites for analysis in the aggregate and for comparison among different systems.

[Return](#) to top of this page

The TULIP Final report

The [final report](#) for the TULIP project is currently available, both in a Web and in a printed version.

The Web version

For easy printing of the entire report by your Webbrowser, it is divided in eight files:

1. The [top level document](#) including the [Table of Contents](#), the [Executive Summary](#) and the [Introduction](#) (23 Kilobytes)
2. [Chapter I. Description of the project and participants](#) (40 Kilobytes)
3. [Chapter II. Technical aspects](#) (75 Kilobytes)
4. [Chapter III. Promotion](#) (9 Kilobytes)
5. [Chapter IV. User behavior](#) (45 Kilobytes + artwork 380 Kilobytes)
6. [Chapter V. Organizational and economic issues](#) (32 Kilobytes)
7. [Chapter VI. Implications of the TULIP project for the future of the development of digital libraries](#) (19 Kilobytes)
8. The HTML document with appendices I through XIV (35 Kilobytes). Please note that some appendices are not available in the Web version.



The printed version

We have a limited supply of the paper version of the TULIP Final Report available. Please fill in the [order form](#) to receive a free copy of the TULIP Final Report.

[Return](#) to top of this page

TULIP Newsletters

The following TULIP Newsletters are available for browsing

- [TULIP Newsletter no. 7 - July 1996](#)
- [TULIP Newsletter no. 6 - May 1995](#)
- [TULIP Newsletter no. 5 - September 1994](#)
- [TULIP Newsletter no. 4 - April 1994](#)
- [TULIP Newsletter no. 3 - January 1994](#)
- [TULIP Newsletter no. 2 - Augustus 1993](#)
- [TULIP Newsletter no. 1 - November 1992](#)

[Return](#) to top of this page

Journals in TULIP in the field of Material Science

The participating universities have in common strength in the physical and engineering sciences. In looking within these disciplines for a target area, we wanted a field in which the researchers were comfortable with computer applications and had a higher than average installed base of workstations. An

obvious choice might have been computer science itself, but we felt these users would be so atypical in their computer facility as to make it hard to generalize results to other disciplines. Materials science provided a field in which there was both a sufficiently large corpus of frequently-cited material within one publishing company and interested faculties. Therefore [83 journal titles](#) were chosen from the collection of Elsevier Science journal titles.

[Return](#) to top of this page

Universities participating in TULIP

- **University of California** (all campuses)
 - Berkeley
 - Davis
 - Irvine
 - Los Angeles
 - Riverside
 - Santa Barbara
 - Santa Cruz
 - San Diego
 - San Francisco
- **Carnegie Mellon University** (Pittsburgh, PA)
- **Cornell University** (Ithaca, NY)
- **Georgia Institute of Technology** (Atlanta, GA)
- **University of Michigan** (Ann Arbor, MI)
 - A [demo version](#) of the prototype of Michigan's library system based on World Wide Web
- **Massachusetts Institute of Technology** (Cambridge, MA)
- **University of Tennessee** (Knoxville, TN)
- **Virginia Polytechnic Institute and State University** (Blacksburg, VA)
- **University of Washington** (Seattle, WA)

[Return](#) to top of this page

For general information on the TULIP Program

Project Manager

Jaco Zijlstra (The Netherlands), voice +31 20-485 3722, fax +31 20-485 3354, e-mail J.Zijlstra@Elsevier.nl

Project Leader

Karen Hunter (USA), voice +1 212-633 3787, fax +1 212-633 3764, e-mail K.Hunter@Elsevier.com

Technical Coordinator

Paul Mostert (The Netherlands), voice +31 20-485 3574, fax +31 20-485 2734, e-mail P.Mostert@Elsevier.nl

Mail Address

Elsevier Science



IBM Digital Library Version 2

An end-to-end solution for managing multimedia content.



- Reach new markets and establish new sources of revenue through improved management and reuse of media assets
- Preserve your assets from physical deterioration
- Protect your assets with advanced rights management
- Consolidate management of text, images, audio and video with easier, faster access
- Save money with electronic delivery
- Be ready for Year 2000

Developed with a variety of key customers and business partners, IBM Digital Library has helped businesses and institutions in the areas of higher education, media and publishing, entertainment, culture, health, and commerce provide greater access to their digital assets, while enhancing their growth and new revenue opportunities. IBM Digital Library Version 2, building on the strength of these technologies, enables literally petabytes of text, images, audio and video, to be created or transformed into digital form and distributed over any network, with security, to users around the world.

New features in Version 2 include:

- Enhanced platform support now includes Windows NT and Macintosh
- Multi-language development tools
- Enhanced rights management
- Integrated multi-search capability
- Integrated support for IBM media servers
- Java-based system administration interface
- Easier installation with graphical user interface guides
- IBM DB2 Universal Database components

[Key Features of IBM Digital Library Version 2](#)

[IBM Digital Library Architecture](#)

[IBM Digital Library at a Glance](#)



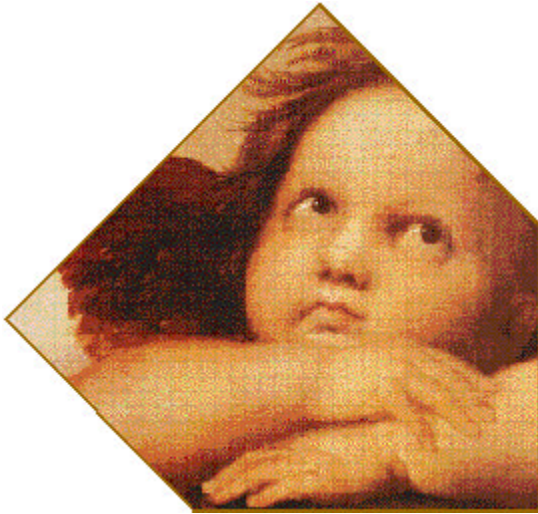
©1998 IBM Corporation



IBM Digital Library Collection Treasury

Extending worldwide access to special collections

Museums and libraries holding special, unique collections have two distinct missions. On the one hand, these institutions want to share their collections with as many people as possible. On the other hand, as copyright holders or caretakers for irreplaceable cultural artifacts, they want to minimize exposure to their holdings, in order to ensure their preservation.



Working with museums and libraries, IBM has developed IBM Digital Library(TM) Collection Treasury--a solution that will enable these institutions to maximize their ability to share their holdings, while substantially reducing risks to preservation. The solution, based on IBM Digital Library technology, enables institutions to provide worldwide access to their holdings via the Internet and vastly increase the potential "visitor traffic" to the institution. At the same time, none of these visitors ever touch the holdings, reducing risk of damage, loss and theft.

In effect, IBM Digital Library Collection Treasury extends the walls of the traditional library or museum, making possible "virtual libraries" and "virtual museums" that can be explored at any time, from anywhere in the world.

Sophisticated search technology enables the "virtual visitor" to sort through vast collections with ease, finding a specific document or image among thousands. This powerful search and access capability also enables the visitor to conduct lengthy ongoing research conveniently from a distance, integrating this research into a normal daily schedule.

By contrast, consider the traditional situation. Many university researchers are a plane ride away from any specific museum or library. If they wish to conduct research involving documents or images at a given institution, finding travel funds is often difficult. If they find the funds, they need to set aside at least two days to travel to and from the site, and then need to compress their research time into an arbitrary block of time reserved for the research. If they have follow-up questions, another trip must be scheduled--or the research questions go unanswered.

Now, IBM Digital Library Collection Treasury offers the potential to transform museums and libraries for the digital age, making their unique collections more relevant to more people than ever before. Thanks to IBM's rights management and watermarking technologies, institutions will continue to exercise tremendous control over their collections--minimizing illegal copying, and restricting access to any specific audience. What has changed is the breadth of access: with the tyranny of distance removed, libraries and museums can truly "serve the world."

IBM Digital Library at work

IBM Digital Library Collection Treasury shows the power of IBM Digital Library at work. The same core IBM Digital Library technology is being used by international historical archives to digitize and manage their ancient manuscripts, and by major Hollywood studios to manage brand new production content.

With IBM Digital Library's special rights management and watermarking capabilities, valuable original holdings can be managed online with appropriate restrictions on audience access and reuse of the images. Future capabilities might include electronic commerce extensions that can provide revenue to content owners through fee-based access and reuse of content.

Putting it to work for you

IBM Digital Library and IBM Digital Library Collection Treasury are highly scalable, enabling institutions to start by digitizing small collections, and then expand to encompass millions of holdings.

This scalability makes IBM Digital Library Collection Treasury ideal for pilot projects, which provide institutions with an affordable, low-risk approach to test the feasibility, process and benefits of digitizing their collections. Once benefits have been demonstrated, the institution can commit to digitizing much larger collections, building on the same technology base.

While IBM Digital Library Collection Treasury can connect to the Internet to extend worldwide access to collections, it can also be used to provide much more selective access -- within a single campus or company, for example, or to a set of specifically authorized individuals.

Toward virtual libraries

IBM Digital Library Collection Treasury has three goals. First, the preservation of cultural heritage by creating digital records of books, manuscripts, photographs, artwork and other cultural artifacts. Second, managing those records efficiently by storing them with appropriate descriptive information. Third, allowing more open access to those records for researchers, students and public users, while protecting the assets from unauthorized use.

Think of IBM Digital Library Collection Treasury as the electronic bricks and mortar supporting future virtual museums and libraries. When that future arrives--when a majority of the world's important collections have been digitized and put online--the bricks and mortar will be all around us. We will, from that point on, be *within* a virtual library of libraries, a museum of museums--and "research" will no longer be a question of privileged access, but more a matter of simple human curiosity and desire.

For more information

For more information on IBM Digital Library Collection Treasury, contact [Jackie Mahoney](#), Segment Manager, IBM Corporation, at 941-646-1083 or learn more about IBM Digital Library by calling your IBM representative or IBM authorized software reseller, or by visiting our Web site at www.software.ibm.com/is/dig-lib.

© International Business Machines Corporation 1997



International Business Machines Corporation
Old Orchard Road
Armonk, NY 10504

IBM is a registered trademark of IBM Corporation.
IBM Digital Library is a trademark of IBM Corporation.
GC26-9112-00

IBM Home	Support	Contact IBM	Employment	Privacy	Legal
--------------------------	-------------------------	-----------------------------	----------------------------	-------------------------	-----------------------

©1998 IBM Corporation



This site received a
4 star rating from McKinley Group's editorial team
and
"Best of the Web!" by Snap! Online




QBICTM -- IBM's Query By Image Content

On-line collections of images are growing larger and more common, and tools are needed to efficiently manage, organize, and navigate through them. We have developed the QBIC system which lets you make queries of large image databases based on visual image content -- properties such as color percentages, color layout, and textures occurring in the images. Such queries use the visual properties of images, so you can match colors, textures and their positions without describing them in words. Content based queries are often combined with text and keyword predicates to get powerful retrieval methods for image and multimedia databases.


QBIC is available for download with a free 90 day trial license. The download package includes the image indexing and search engine (for AIX, Linux, Windows NT/Windows95, and OS/2), a Web front end, APIs for imbedding QBIC in other applications or extending QBIC with new query functions, and even a sample image collection. You can download it from [IBM software download site](http://www.ibm.com/software/awdtools/qbic/).

News Bulletins

- 
 IBM AND [MAGNIFI](#) ANNOUNCE LICENSING AGREEMENT -- IBM Research Technology gives Magnifi the cutting edge in Visual Searching Capabilities. Click [here](#) for more information.

- IBM, VIRAGE [ANNOUNCE](#) BROAD CROSS-LICENSING AGREEMENT
- Stay tuned here for announcements about a coming new release of QBIC.

QBIC demos on this site:

-  [A collection of all U.S. stamps before 1995, searchable by QBIC and DB2 with a Java GUI.](#)
- [A prototype trademark browsing and retrieval site.](#)
- [Our old Stock Photo demo spruced up with our Java GUI.](#)

Other sites showcasing QBIC technology:

- [The Art and Art History QBIC project at UC Davis.](#)
 - [Imagebase](#) at the Fine Arts Museums of San Francisco.
 - [Image collections from the French Ministry of Culture.](#)
 - [Querying the "Electronic Visualization Library" by images.](#)
-

To send comments on this on-line demo, or to contact the QBIC group, write us at:

qbicwww@almaden.ibm.com.

To get information on obtaining a full use licenses, contact tedl@almaden.ibm.com.

To be added to our mailing list, enter your name in the following box and press Enter:

Your e-mail address:

Check out QBIC's availability in the [DB2 Image Extenders](#), which are components of IBM's scalable, multimedia, Web-enabled [DB2 Universal Database](#). Other related sites include

- [IBM Digital Library - Related technologies for information management.](#)
 - [Technical paper requests on QBIC \(please provide surface mailing address in your request.\)](#)
-

[[IBM home page](#) | [Order](#) | [Search](#) | [Contact IBM](#) | [Help](#) | [\(C\)](#) | [\(TM\)](#)]



IBM Digital Library Media & Entertainment Solutions

News Archive Solution

***A solution for news asset management
--spanning text, photos, graphics, transcripts, film, video and audio***

Imagine if retail stores took their cash income at the end of each day, stuffed it in a mattress, and forgot about it forever.

That may sound ludicrous. But without good management, it's essentially what happens in broadcast news when video assets aren't managed as valuable cash assets.



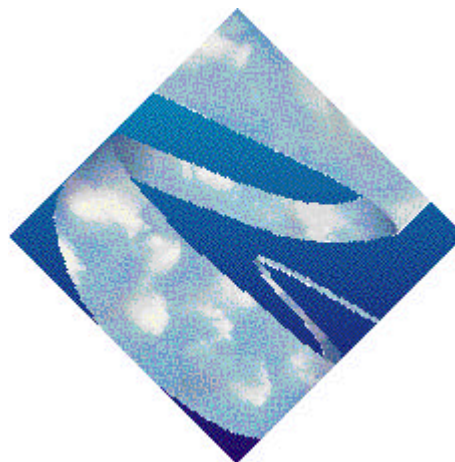
Broadcast news operations work hard every day, and spend a great deal of money, to capture the best video content possible for current news stories. Much less attention is spent on managing and reusing this content after the fact. Since video content is essentially the cash commodity of the business, that's like stuffing wads of hundred dollar bills in a mattress.

Do the math. The average finished minute of broadcast news costs the industry thousands of dollars to produce. However, once produced, that finished minute can be reused and resold again and again. It's like selling lemonade, and then putting the lemonade back in the bottle to sell again.

Often, it's more than just an economic issue. News, by definition, involves the reporting of "new" events, requiring an instant and professional response to an unpredictable world. When a major leader is overthrown in a coup, you want footage of that leader now. When two sports teams engage in a blockbuster trade, you want highlight footage of the traded players now.

The IBM News Archive Solution makes such responsiveness easily possible, helping you to break news faster, beat the competition with higher quality news stories, and end the day with more cash in your pocket than in the proverbial mattress.

Designed specifically to serve the needs of a major TV network, the IBM News Archive Solution is robust enough to support any broadcast news operation in existence today. A complete solution, it packages application software, systems software, implementation services and support. The packaged solution means that you don't have to be a multimedia database expert yourself to integrate powerful database management into your news operation. With IBM's help, you can be up and running quickly.



The News Archive Solution helps you manage all the elements that make up your news archives--including text, photos, graphics, transcripts, film, video and audio. A simple graphical user interface enables easy storing, searching, retrieval, viewing, and online ordering of your assets. This graphical user interface runs on any Intel-based PC client using the Netscape browser and/or

Windows® 95. The application can run on multiple client PCs, all accessing a single Windows NT®-based server.

The News Archive Solution offers the following business case benefits:

- **Greater reuse, saving re-creation and purchase dollars**--News-editors and producers gain easy access to archived video content through advanced preview to assure high reuse of existing film content
- **Additional sources of revenue**--Full-function search allows access to content for internal use and outside remarketing
- **Responsive archive based on accurate logging information**--Powerful logging and indexing features with full VTR transport control and capture capabilities allow for faster and more complete logging with time codes
- **Avoid cost of extinct technology**--Easy migration from existing Library Management System to a more state-of-the-art platform
- **Cost saving on inventory management** -- Physical videotape storage management is enhanced through tighter inventory controls via barcoding
- **Controlled migration to digital**--Ability to perform low-resolution multimedia capture with logging
- **Digital content storage savings**--Ability to archive digital content to less expensive storage medium automatically.

The IBM News Archive Solution is built on robust, proven IBM database technology. That same "digital library" technology is being used by international historical archives to digitize and manage their ancient manuscripts, and by major Hollywood studios to manage brand new production content. Quite simply, if you've got valuable multimedia content to manage, IBM has the technology you need to do it right--today, and into the future.

The solution is highly expandable and will grow with you as your archive grows. On the client side, the solution enables multiple users to store and view text, images and video right at the workstation. Assets can be both physical media (such as videotapes and film) or digital media.

Powerful and intuitive search capabilities allow users to search via parametric, free text, and IBM's "Query by Image Content." Parametric searches enable queries by such precoded parameters as clip date, subject name, producer, location, etc. Free text searching enables queries across text notes, scripts and other non-coded text fields. Query by Image Content enables you to search for clips and images based on what they *look* like--a "fuzzy" characterization based on subject shape, color, etc. Looking for clips of a bright red fire truck? For fashion designs using lime green? Query by Image Content will help you find them.

The application intuitively allows you to search your News Archive whether the content is on the shelf or in the computer. Through a folder-based workflow, field-value prompting and full-function help facility, your work gets done easier and quicker.

Don't sleep on your hard-won broadcast news assets. With the IBM News Archive Solution, you can cash in today on the true power of your news operation.

For more information on the IBM News Archive Solution, contact Tom Davis, IBM's Solution Manager for News Archive, at (203) 431-5891 or thd@us1.ibm.com. Learn more about IBM Digital Library by calling your IBM representative or IBM authorized reseller, or by visiting our Web site at www.software.ibm.com/is/dig-lib.



© International Business Machines Corporation 1997
International Business Machines Corporation
Old Orchard Road
Armonk, NY 10504

IBM is a registered trademark of IBM Corporation.
IBM Digital Library is a trademark of IBM Corporation.



☑ Hyperwave Information Server 4.0

Hyperwave is proud to announce the availability of Information Server 4.0 - the most powerful and scalable enterprise Information Management platform available in the market today. By combining ease-of-use with an enhanced open architecture...

▶ Hyperwave forges Knowledge Management Alliances

© Copyright 1997, 1998 Hyperwave GmbH. All Rights reserved.



HYPERWAVE AUTHOR

OVERVIEW

HyperWave Author is integrated hypermedia authoring software, specially designed to help you create and edit documents on HyperWave servers.

Special navigation tools such as collection browsers and hyperlink maps make the creation of high-quality webs easier than ever before. Integrated support for the HyperWave Server's search engines helps the user to find existing documents and reuse them. HyperWave Author will prevent you from getting "lost in hyperspace".

Distributed data management is the keyword for working with HyperWave software within large companies. The combination of HyperWave's access control features and HyperWave Author lets each department keep track of its information without having to set up its own server. This leads to a sharp decrease in software and administration costs.

- [KEY FEATURES](#)
- [TECHNICAL SPECIFICATIONS](#)
- Online documentation for [HyperWave Author for Windows \("Amadeus"\)](#)
- Online documentation for [HyperWave Author for UNIX \("Harmony"\)](#)

User: Guest • Owner: gmesaric • Last modified: 04/25/1996 15:19:31



HYPERWAVE AUTHOR

KEY FEATURES

HYPERMEDIA AUTHORIZING

HyperWave Author is the ultimate interactive authoring tool for HyperWave Servers. Users can author remotely over network boundaries: the Internet or any other TCP/IP based network can be used. HyperWave author provides full support for the HG-CSP network protocol, special HyperWave Server features such as database and search facilities are seamlessly integrated into the interface.

OBJECT DATABASE

Object orientation is one of the key concepts of HyperWave. HyperWave Author provides full support for HyperWave Server's object-oriented database system, allowing easy insertion and editing of server-side objects. HyperWave Author for Windows additionally provides a local version of the database, letting users author web applications offline which they can later easily upload to a HyperWave Server.

VRML AND POSTSCRIPT

HyperWave Author software includes IICM's free VRML scene viewer and a viewer for PostScript documents. VRML is the standard 3D data format in the WWW. PostScript is the industry standard for electronic publishing. HyperWave Author provides integrated PostScript viewer software, including the facility for inserting hyperlinks in PostScript documents: annotations to non-HTML documents are possible because of HyperWave's link database approach.

ADVANCED NAVIGATION

Critics of the WWW often mention the so-called "lost in hyperspace" syndrome. HyperWave Author provides advanced navigation concepts and demonstrates that there are solutions to this problem: tree-view collection browsers let you navigate easily through big web servers and dynamically generated hyperlink maps help you keep masses of interlinked information up to date.

DISTRIBUTED INFORMATION MANAGEMENT

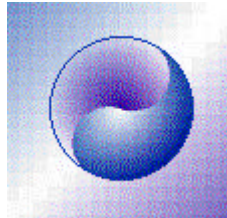
HyperWave's authoring software provides the facility of distributed information management: every logical part of a company can have a virtual web server without having the overhead of setting up its own real web server. A company can have one corporate identity on the web, running a WWW service where every department of the corporation is responsible for its own part.

MULTILINGUAL DOCUMENTS

HyperWave Author supports easy creation and editing of multilingual web applications. HyperWave's support for multilingual document clusters is especially interesting if your company is located for example in Europe or Asia, or any other part of the world where more than one



Welcome to Harmony: The Hyperwave Administrator for Unix/X11



Harmony is the Unix/X11 client for [Hyperwave](#), the first second-generation, publicly available, networked hypermedia information system running over the Internet. Hyperwave integrates hyperlinking, hierarchical structuring, sophisticated search, and access control facilities into one single system, and is interoperable with other network information tools like Gopher and WWW.















Here you can find information about

- [where to get Harmony by anonymous FTP](#)
- Harmony's special features
 - [information structuring facilities](#)
 - [orientational aids](#)
 - [multilinguality](#)
 - Harmony's [document viewers](#)
- Installation guide
 - [installation](#)
 - [user configuration](#)
- User support
 - the [Hyperwave mailing list](#)
 - the [Harmony FAQ](#)
 - further information

Please direct any feedback (comments, suggestions, bug reports, etc.) concerning Harmony by electronic mail to:

harmony@iicm.tu-graz.ac.at

We are very interested in your feedback, even though we may not be able to respond personally to every piece of mail.

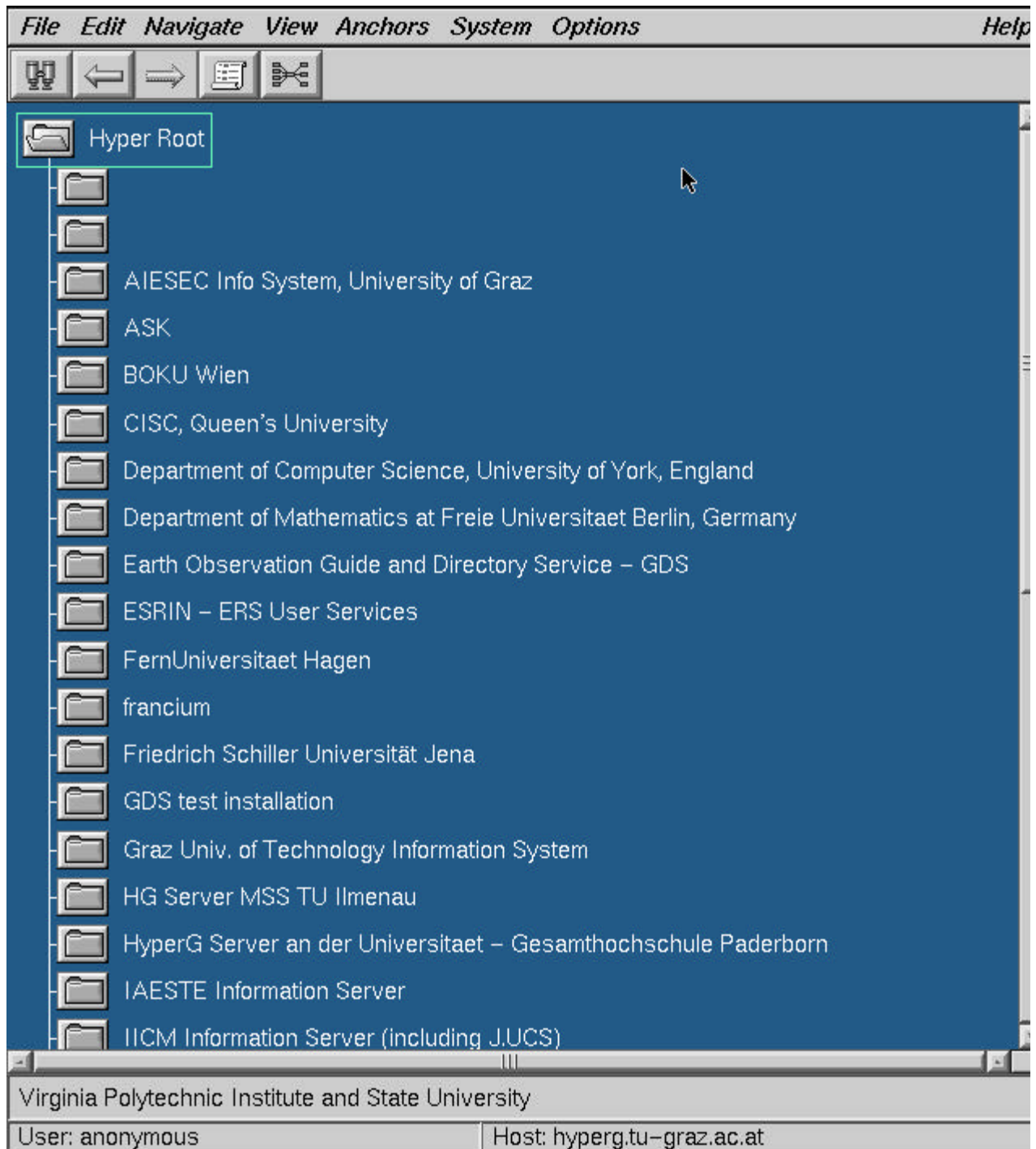
-   [Harmony Release Notes & Where to get Harmony by FTP](#)
-   [General Information About Harmony](#)
-   [Installation and user configuration of Harmony](#)
-   [Harmony User Support](#)
-   [Harmony- Frequently Asked Questions](#)
-   [The Harmony User Guide](#)
-   [Harmony Quick Reference](#)

User: **Guest** • Owner: **kandrews** • Last modified: **04/22/1996 15:21:29**

Hyper-G --- Harmony Illustrations

Illustrations of the use of Hyper-G with the Harmony (UNIX) client include:

- connection to the global root



- expansion of the collection of nodes accessible from the root to those at the Virginia Tech server



File
Navigate
Anchors
View
Options
Help

Search
Anchors

hginscoll (1)

Name

hginscoll – insert a new collection

Synopsis


```
hginscoll [-h] [-i FCollId | -n FCollName] [-N CollName] [-c] [-A Author]
[-C CDate] [-E EDate] [-O ODate] [-F][-T Title] [-R Rights] [-D
Description] [-S SortOrder] [-L Language] [-r hghost] [-d hgport]
```

Description

hginscoll builds a collection or cluster object and insert it into the Hyper-G database.

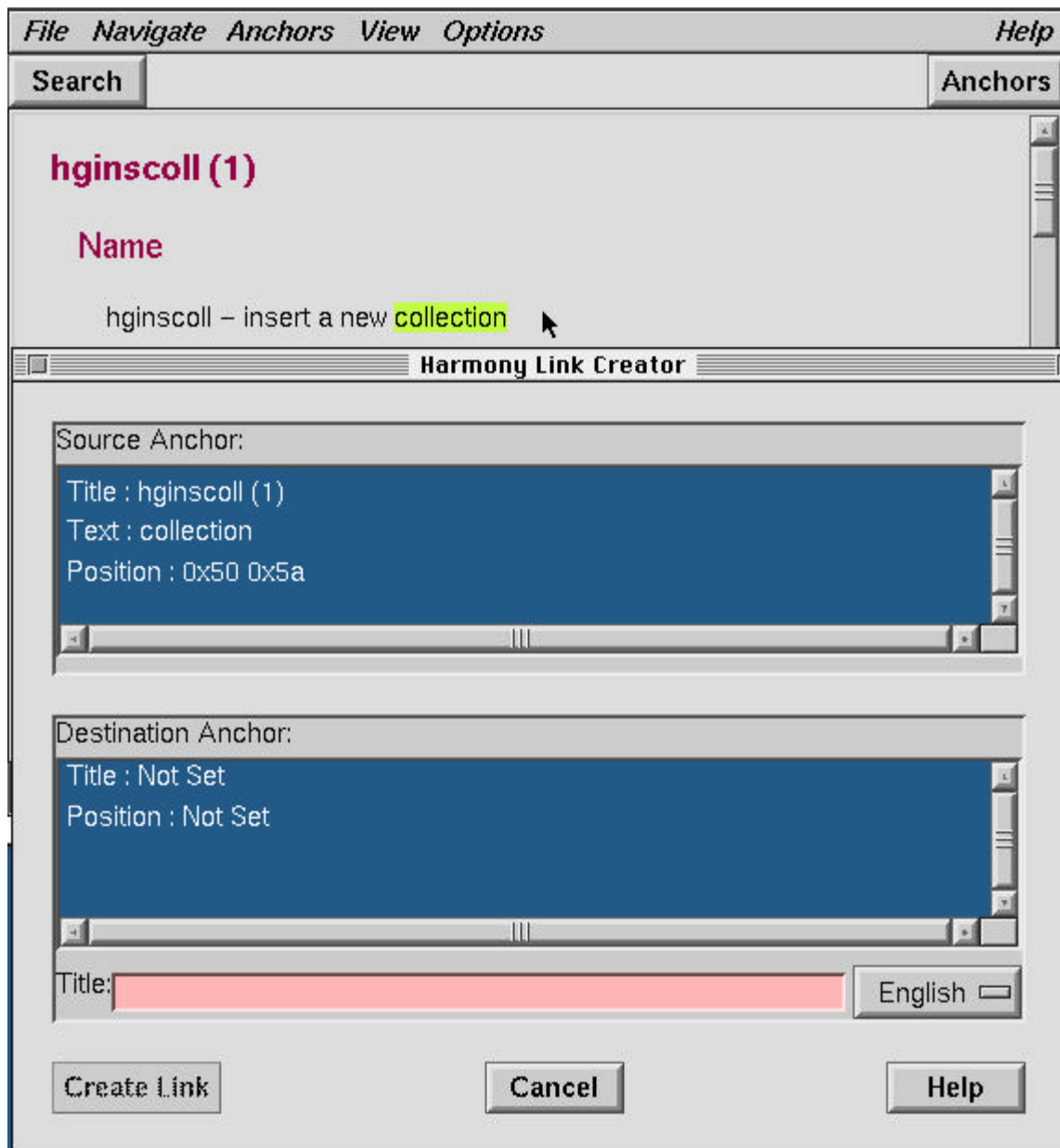
hginscoll (1)

Environment

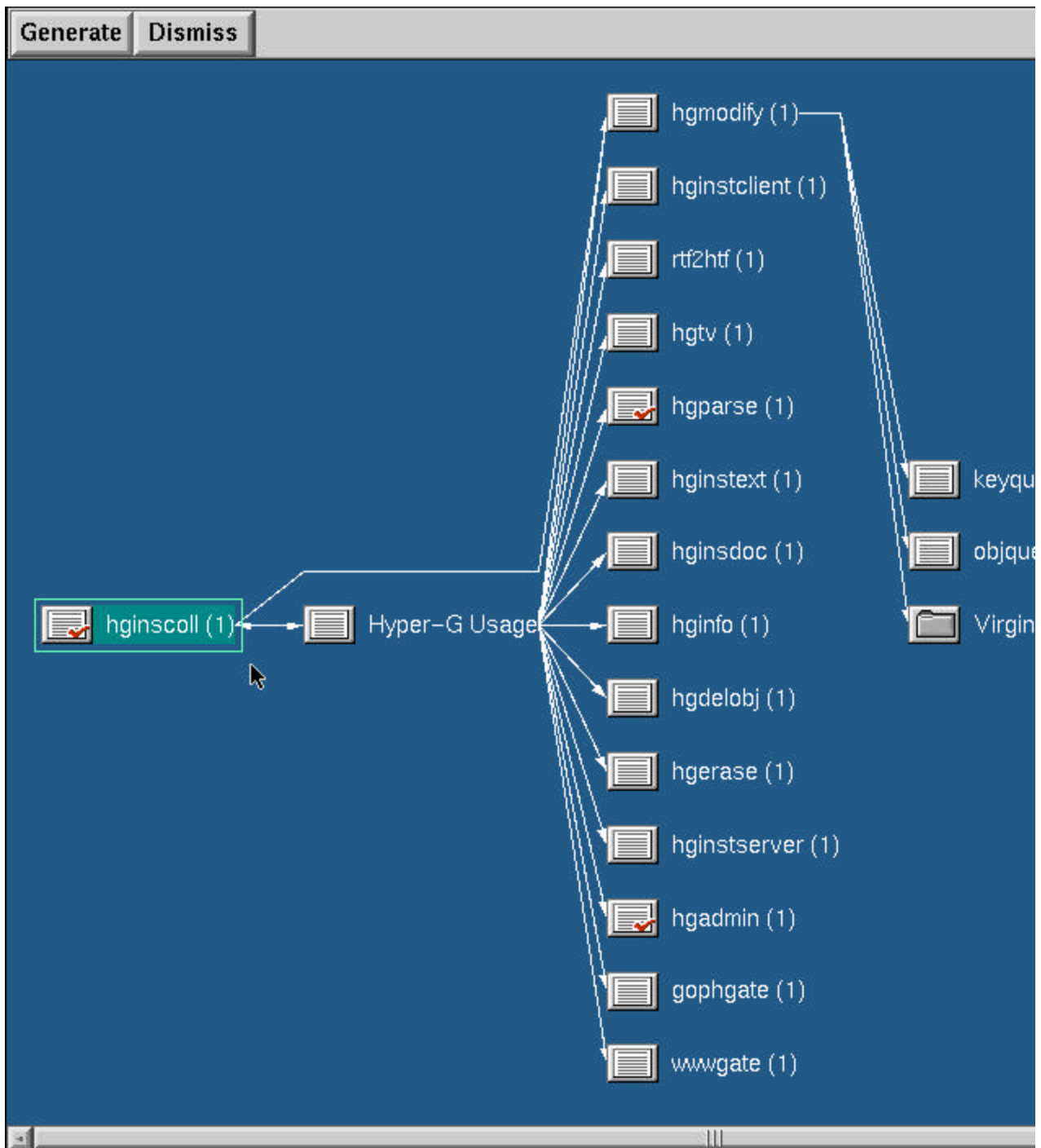
HGAUTHOR:	 Author
HGRIGHTS:	Rights
HGDESCRIPTION:	Description
HGSORTORDER:	SortOrder
HGFATHERCOLL:	FCollName
HGLANGUAGE:	Language

hginscoll (1)

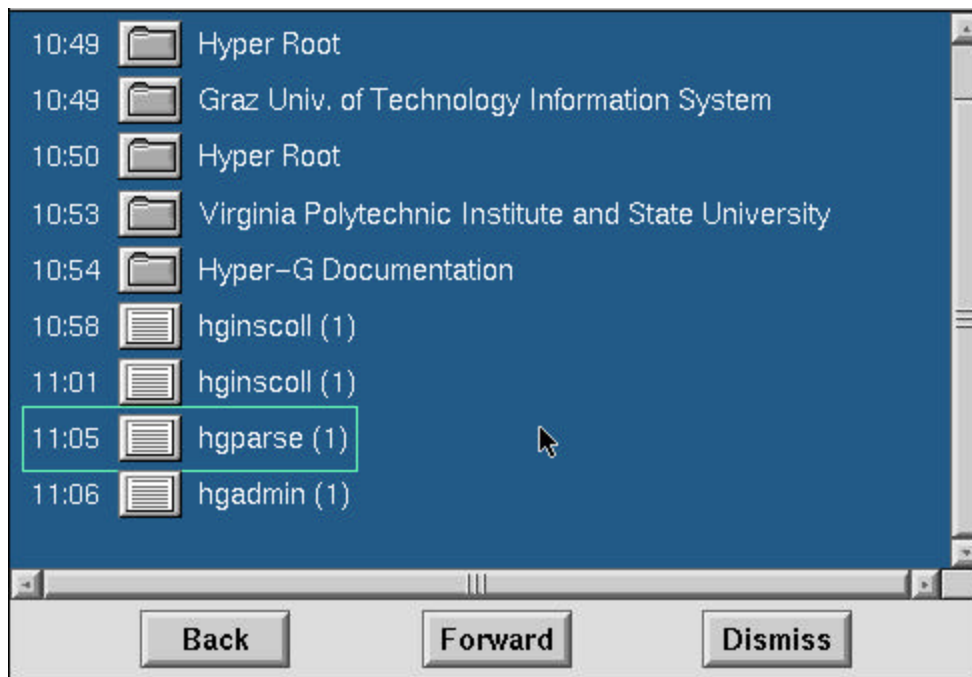
- making a link



- viewing a local map



- reviewing the history of accesses





BLACKSBURG

electronic village



Community
Arts
Organizations
Religion
Sports
Education
Library
Museums
Schools
People
Discussion
Seniors
Government
Health
Village Mall
Visitor's Center

Search Tools

Help Desk

▼ Full Index

News

[BEV User Survey](#)

Please complete this local Internet survey to provide feedback to help shape the future of the BEV.

Live in an apartment with Ethernet?

The BEV/BA apartment Ethernet system will be transferred to the apartment owners beginning July 1st. The transition should be minor for most users, but [read the FAQ](#) to see if you will be affected.

Whitewater Kayaking Classes

Wednesday evenings we offer roll classes in the county pool, Sundays we offer river classes on the New River. Classes are instructed by Back Country Ski & Sport's ACA certified instructors, with a low student-instructor ratio. Roll class or previous experience required before you register for river class. \$15/roll class + \$5/boat rental. \$50.00/river class + \$15/boat rental. Pre-registration required - classes fill quickly! Call 382-6980 if you need more information.

[SEEDS Summer Field Camps](#)

Seek Education, Explore, DiDiscover - SEEDS has a few openings left for the 1998 full-day Summer field camps. Coed sessions for ages 7-9 and 10-12. Go exploring with SEEDS this Summer!

[Sign the Bicentennial Guestbook!](#)

Here's your chance to be a part of history.

Summer Playcamp Registration

We are a licensed day care facility offering care with a summer camp atmosphere. Arts & crafts, nature, games, pool time, free lunch and field trips are just a few of the exciting activities we offer from 7:30 - 5:30, Mon-Fri. June 15 - August 7 for \$55.00 a week. Pre-registration required, deadline is prior Wednesday.

[\[Old Messages || POST a Message\]](#)



[Weather Underground Forecast](#)
[Blacksburg National Weather Service](#)

[About the BEV](#) | [Services](#) | [Training](#) | [Research](#) | [Starting a Village](#)

[Community](#) | [Education](#) | [Library](#) | [Museums](#) | [People](#) | [Discussion](#) | [Seniors](#)
[Government](#) | [Health](#) | [Mall](#) | [Visitors](#) | [Search](#) | [Help](#) | [Index](#) | [Home](#)



© 1998 Virginia Polytechnic Institute and State University

comments to webmaker@bev.net

Visit the [Town of Blacksburg](#) home page.

updated Friday, 19-Jun-98 23:27:00 EDT

<http://www.bev.net/index.html>



Welcome to the **BEV HistoryBase**, a WWW History Page for the [Blacksburg Electronic Village](http://history.bev.net/bevhist/)! Try out the [BEV History Timeline](#) to learn more about the history of our electronic community. For a non-graphical alternative, check out the [Textual BEV History Timeline](#). Both contain the same information so feel free to browse either.



[[Main Timeline](#) | [Contribute](#) | [What's New?](#) | [Search](#)]

[Message of the Day Listings](#)

[Blacksburg Telecommunications Advisory Committee Meeting Minutes](#)

[BEV Media Coverage Archive](#)

[BEV Group Home Pages](#)

This project is supported by NSF Grant CDA-9424506. A [copy of the grant proposal is online](#).

Last updated 27 October 1995 / schmidt@cs.vt.edu

[HistoryBase
Main Page](#)

[Contribute](#)

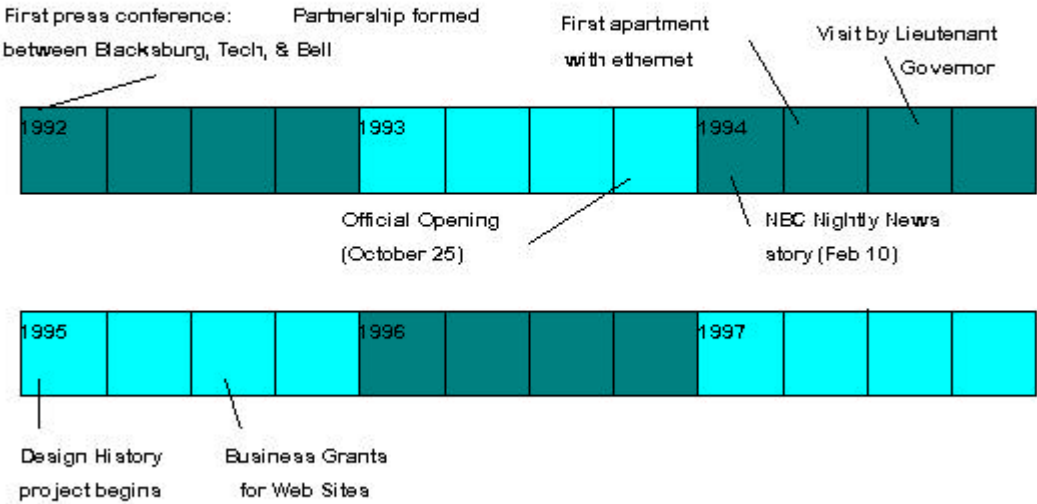
[??? What's
New?](#)

[Search](#)

[BEV
Homepage](#)

BEV HistoryBase: Main Timeline

Click in a box to see a more detailed history for that quarter



Click in a box to see a more detailed history for that quarter

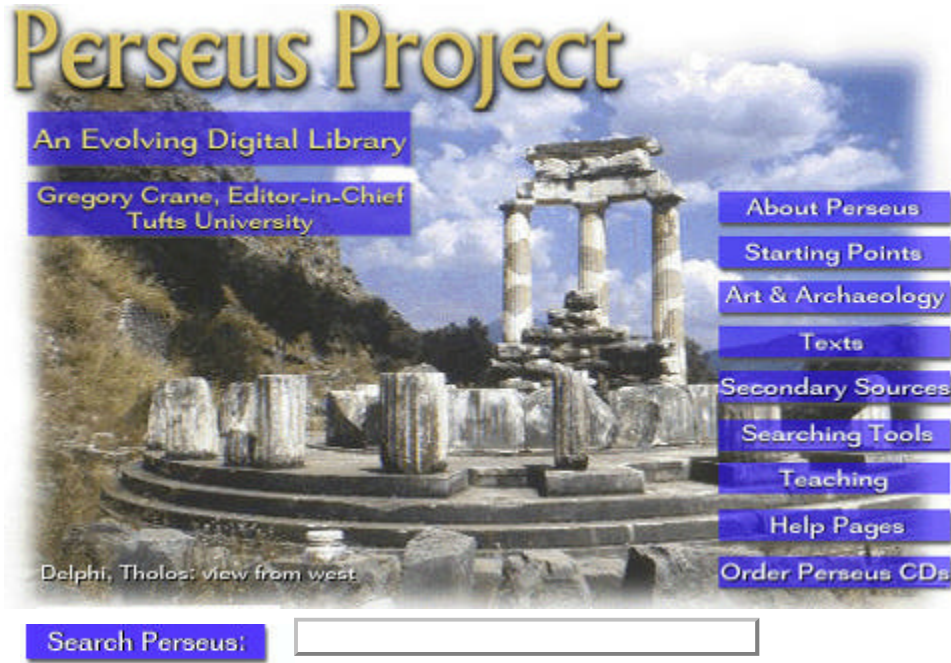
[HistoryBase
Main Page](#)

[Contribute](#)

[??? What's
New?](#)

[Search](#)

[BEV
Homepage](#)



Perseus Project

An Evolving Digital Library

Gregory Crane, Editor-in-Chief
Tufts University

- About Perseus
- Starting Points
- Art & Archaeology
- Texts
- Secondary Sources
- Searching Tools
- Teaching
- Help Pages
- Order Perseus CDs

Delphi, Tholos: view from west

Search Perseus:

[Copyright](#) | [General Policies of this Site](#) | [FAQ](#)

Read the important announcement on [planned changes for access](#) to the Perseus web site.



- New [interactive web atlas](#).
- Revised Catalogue of [Greek sculpture](#), including images from the [Museum of Fine Arts, Boston](#)
- Newly accessible [ancient art resources](#): over 670 vase pictures, and the landmark Caskey and Beazley vase catalog, from the Museum of Fine Arts, Boston
- Announcing our new site on [Julius Caesar](#)

Congratulations to the city of Athens,
site of the 2004 Summer Olympic Games!

Many are the sights to be seen in Greece, and many are the
wonders to be heard; but on nothing does Heaven bestow more
care than on the Eleusinian rites and the Olympic games.

Pausanias, [Description of Greece 5.10.1](#)



[Awards and reviews](#) | [Text-only home page](#) | [Related sites](#)

The Perseus Project is supported by the [Annenberg/CPB Project](#), the [National Science Foundation](#), [Apple Computer](#) and the [National Endowment for the Arts](#), the [National Endowment for the Humanities](#), the Packard Humanities Institute, the [Getty Grant program](#), [Xerox Corporation](#), [Boston University](#), [Harvard University](#), and the [Fund to Improve Post-Secondary Education](#).

Perseus is a non-profit enterprise, located in the [Classics Department](#), [Tufts University](#).

Mail problems and suggestions to:
webmaster@perseus.tufts.edu

Search the NCSTRL Collection



Search *ALL* bibliographic fields ...

Search for:

Sort results by:

rank



Search *SPECIFIC* bibliographic fields ...

Author:

Title:

Abstract:

(Combine fields with ☒ **AND** ☐ **OR**)

Sort results by:

rank



Search

Clear fields

If you would like to view the NCSTRL collection by *year* or by *institution*, use the [browse form](#).

Digital Libraries for CS

Here are some pointers to Digital Libraries / bibliography servers related to CS.

[ACM Digital Library Collection at Virginia Tech](#)

Small test collection of CACM articles from those scanned in as part of the NSF-supported Envision project.

[ACM Graphics Bib. DB](#)

SIGGRAPH Online Bibliography Database

[ACM Computer Graphics Courseware Repository](#)

SIGGRAPH Computer Graphics Courseware Repository (ftp)

[ACM HCI Bib. DB](#)

interactions Bibliographies on Human-Computer Interaction

[BibNet Project](#) and [TeX Users Group](#) FTP bibliographies

bibliography collections from Nelson Beebe including HTML with extensive internal and external hypertext links. See examples: [IBM Systems Journal](#), [DEC Technical Journal](#). See [program to build these from BibTeX](#).

[CACM Collection \(1959-1979\) using Inquiry](#)

U. Mass. CIIR demo of Inquiry with CACM test collection

[Collection of Computer Science Bibliographies](#)

from Alf-Christian Achilles; updated monthly; 790 locally stored bibliographies; more than 530,000 references; 20,000 references contain URLs to an online version of the paper; more than 1600 links to other sites carrying bibliographic information; uses Glimpse

[Databases and Logic Programming \(mirror\)](#)

bibliography server by Michael Ley

[Hypertext Bibliography Project](#)

Hypertext Bibliography Project (Glimpse search of many publications)

[NCSTRL](#)

Networked Computer Science Technical Report Library

[Table of Contents re LIS](#)

Table of Contents for JASIS, IPM, etc. - may be slow

[Univ. of Wales Cardiff CS Courseware](#)

Courseware on Algorithms, AI, C, Graphics, Image Processing, Parallel Processing, Vision, X

ETD Electronic Thesis and Dissertation Initiative

Welcome to the Virginia Tech Electronic Thesis and Dissertation home page!

- Browse the [ETD Library](#). Hundreds of titles!
 - Learn about our parent project, the [Networked Digital Library of Theses and Dissertations](#)
 - Learn what the Graduate School expects when you [submit your ETD](#) **HOT!**
 - Attend an [ETD Workshop](#) **NEW!**
 - Learn [how to create an ETD](#) step by step **HOT!**
 - Please complete a [survey](#)
 - What is [PDF](#) anyway? **NEW!**
 - Learn about [LaTeX](#) and [ETD-ML](#) submissions
 - Learn about [publishers and copyright](#)
 - Need help? We have installed Adobe PDF software in [computer laboratories](#) all over campus
 - Still puzzled? Try our list of [frequently asked questions!](#) **HOT!**
 - All else has failed? Contact us: etd@vt.edu
-

[Campus Labs](#) | [ETD Library](#) | [ETD-ML](#) | [FAQ](#) | [How-to](#) | [LaTeX](#) | [NDLTD](#) | [PDF](#) | [Submission Guidelines](#)

etd

Revised: Wed Feb 18 11:12:14 EST 1998



[index.sl](#)

*scholarly communications project*

Virginia Tech <ETD> Submission Form

Instructions:

Please fill out the form completely. Cut and paste, from your document and into the form, as necessary. Read the [help file](#) for help on cutting and pasting your abstract and for selecting keywords. Once you are done filling out the form read the [copyright statement](#) at the bottom of the page and if you agree to it click "Preview".

 **WARNING: There have been problems reported when using these forms with Netscape version 3.02. If you see an error message of the type "Document Contains No Data", please make sure that you are not using Netscape Navigator 3.02.** 

Document Type:

Please select the type of document you are submitting.

Master's Thesis 

Name:

Please enter your name just as it appears on the title page.

First and Middle name (if desired):

Last Name:

Suffix (Jr., Sr., III, etc.):

Email:

Please enter your email address.

Title:

Please enter the title just as it appears on the title page.

Degree:

--

Department:

--

Committee Chair:

--

Chair

Committee Members:

Please enter the names (and optionally, positions) of the remainder of your committee.

[illegible]

Keywords:

Please enter [keywords or phrases](#) separated by commas
(ex.: vortex field, art history, equinology).

Date of defense:

Please enter the date of your defense (ex.: April 10, 1997).

Abstract:

Enter your **abstract** below. If you are cutting and pasting **be sure to read the [help file section on abstracts](#) first!** If you do not it is possible that your abstract will not look the way you want it to when it is viewed by the Graduate School or when it is publicly available on the WWW.

Availability:

Select when you want your work to be available to the public.

- ☒ Release the entire work immediately worldwide.
- ☐ Release the entire work for Virginia Tech access only. After one year release worldwide only with written permission of the student and the advisory committee chair.
- ☐ Secure the entire work for patent or proprietary purposes. After one year release worldwide only with written permission of the student and the advisory committee chair.

Copyright Statement:

I hereby grant to Virginia Tech or its agents the right to archive and to make available my thesis or dissertation in whole or in part in the University Libraries in all forms of media, now or hereafter known. I retain all proprietary rights, such as patent rights. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

☐ **I Agree**

Continue ETD Submission

Virginia Tech Graduate School

Electronic Thesis and Dissertation (ETD) Submission

Approval Form

Student Name: _____

ID#: _____

Department: _____

Degree: ☐ Bachelor's ☐ Master's ☐ Doctoral degree

Document Type: ☐ Project Report ☐ Thesis ☐ Dissertation

Document Title: _____

Student Agreement:

I hereby certify that, if appropriate, I have obtained and attached hereto a written permission statement from the owner(s) of each third party copyrighted matter to be included in my thesis, dissertation, or project report, allowing distribution as specified below. I certify that the version I submitted is the same as that approved by my advisory committee.

I hereby grant to Virginia Tech and its agents the non-exclusive license to archive and make accessible, under the conditions specified below, my thesis, dissertation, or project report in whole or in part in all forms of media, now or hereafter known. I retain all other ownership rights to the copyright of the thesis, dissertation, or project report. I also retain the right to use in future works (such as articles or books) all or part of this thesis, dissertation, or project report.

Student and Committee Agreement:

Part A. We agree that the above mentioned document be placed in the ETD archive with the following status: (*choose one of 1, 2, 3, or 4*)

- ☐ 1. Release the entire work immediately for access worldwide.
- ☐ 2. Release the entire work for Virginia Tech access only.
- ☐ 3. Secure the entire work for patent and/or proprietary purposes for a period of one year. During this period the copyright owner also agrees not to exercise her/his ownership rights, including public use in works, without prior authorization from Virginia Tech. At the end of the one year period, either we or Virginia Tech may request an automatic extension for one additional year. At the end of the one year secure period (or its extension, if such is requested), the work will be handled under option 1 above, unless we request option 2 or 4 in writing.
- ☐ 4. Release the entire work for Virginia Tech access only, while at the same time releasing the following parts of the work only (e.g., because other parts relate to publications) for worldwide access (check all that apply or provide an attached list):
- ☐ Abstract and key bibliographic data (i.e., from submission form)
- ☐ Files named as follows (i.e., separate PDF or multimedia files):
- _____
- _____
- _____

Part B. (use only if you checked 2 or 4 above). Our preference regarding being contacted to see if we will give written approval to expand the access to the above mentioned document is: (*choose one*)

- ☐ in 1 year
- ☐ in 3 years
- ☐ probably never (e.g., since a publisher will release a book version soon)

Part C (optional proxy). To cover cases such as when one or more of the student and committee signing this form becomes inaccessible, each of the following people (indicated by their names printed)

Printed name of proxy: _____

Printed name of proxy: _____

Printed name of proxy: _____

is authorized to serve as a proxy in submitting future versions of this form, so submissions with any of these proxies signing are officially recognized just as if the student and full committee signed. For example, it is suggested that the committee chair be a proxy.

Review and Acceptance:

The above mentioned document has been reviewed and accepted by the student's advisory committee. The undersigned agree to abide by the statements above, and agree that this Approval Form updates any and all previous Approval Forms submitted heretofore.

Signed:	<div></div>	<div></div>
	(student)	(date)
Committee:	<div></div>	<div></div>
	<div></div>	<div></div>
	<div></div>	<div></div>
	<div></div>	<div></div>
	(committee member)	(date)

19971105

ETD Multimedia File Formats

General Comments

Including complex multimedia objects in an ETD is a relatively new possibility. Those attempting this are pioneers. You are encouraged to work with those on your committee interested in this to gain their approval and assistance. Ultimately they should check your final submission, and should be prepared and agree to do so with the multimedia part, else you may think about putting your multimedia work into some other document (e.g., report, WWW site).

There are locations on campus to help with multimedia work. One is the New Media Center in Newman Library, which supports the campus and local community. Hancock Hall houses a multimedia laboratory for Engineering and Architecture. The Center for Digital Music in Squires supports work with audio. The Information Access Laboratory in McBryde 110, supports scanning, digital audio, and digital video. Experts in digital library technology are available in the Digital Library Research Laboratory.

- [New Media Center](#)
- [Center for Digital Music](#)
- [Information Access Laboratory](#)
- [Digital Library Research Laboratory](#)

It is likely that complex multimedia objects will each reside in a different file, located in the same directory as the rest of your ETD. You may wish some icon or thumbnail or other small form of the complex multimedia object in the body of your ETD, and to have that linked to the complex multimedia object.

Archiving

Be careful to consider issues of long-term archiving.

- Always include the highest resolution version of your object, not just a version suitable for today's devices, since technology may improve. You can include several versions, to help those with a variety of devices, particularly if the media itself is not scalable. For example, scan a slide at 2700dpi, but have 640x480 and 320x240 versions as well.
 - If you can, include a version using a well-accepted international standard. Thus, for video, MPEG is encouraged. If you start with QuickTime, include that, but also include MPEG if possible.
 - If you use some proprietary software, include a viewer if that is allowed by the vendor. That way people can view your object without buying that software. Realize, however, that in a few years this object may not be readily usable due to changes in versions and technology.
-

To Learn More

To learn more about multimedia, you may want to take a course like:

- [CS 4624: Multimedia, Hypertext and Information Access](#)
-

Acceptable File Formats

- Thesis Body
 - [PDF](#) Portable Document Format
 - [ETD-ML](#) Electronic Thesis and Dissertation Markup Language
- Text
 - ASCII (.txt)
- Images
 - PDF (.pdf) use Type I PostScript fonts
 - JPEG (.jpg)
 - CompuServe GIF (.gif)
 - TIFF following version 6.0 or later, including CCITT G4 (.tif)
 - CGM Computer Graphics Metafile (.cgm)
 - PhotoCD

Note: We recommend a minimum of 600 dpi resolution for images of pages with text.

- Video
 - MPEG (i.e., MPEG-1, MPEG-2) (.mpg)
 - QuickTime - Apple (.mov)
 - Audio Video Interleaved - Microsoft (.avi)
- Audio
 - MPEG-2
 - CD-DA
 - CD-ROM/XA (A or B or C)
 - AIF (.aif)
 - SND (.snd)
 - WAV (.wav)
 - MIDI (with timing information) (.midi)
- Authoring
 - Authorware
 - Director (MMM, PICS)
- Special
 - Spreadsheet - Excel (.xcl)
 - AutoCAD (.dxf)

ETD Digital Library

Networked Digital Library of Theses and Dissertations

Digital Library of ETDs

Official Nodes in the NDLTD

- [North Carolina State University](#)
- [University of Virginia](#)
- [Virginia Tech](#)
- [West Virginia University](#)
- [University of Waterloo](#)

Other Sites with ETDs

- [University of Michigan](#)
 - [Independent ETDs](#)
-

Federated Search (Demo) for NDLTD

Please try out the following demonstration of how federated search of NDLTD may occur. Report suggestions to James Powell at jpowell@vt.edu

- [Federated Search Demonstration](#)
-

Collection Highlights - Notable ETDs

- [Notable ETDs](#)
-

[NDTLD](#)

etd

Revised: Mon Jun 15 16:50:00 1998

[index.sl](#)

NDLTD Networked Digital Library of Theses and Dissertations

Universities, students, publishers, other interested parties, Welcome!

- Researchers, see <http://www.theses.org/> to **search** and **browse** our library of electronic theses and dissertations (ETDs).
- Students, see <http://etd.vt.edu/> for help creating and submitting ETDs.

What We Are

- An [initiative](#) to improve graduate education, increase sharing of knowledge, help universities build their information infrastructure, and extend the value of digital libraries
- A federation of [member universities](#)
- A project [supported by FIPSE and SURA](#)
- A [project team](#) based at [Virginia Tech](#)
- A recent topic in the [news](#)
- Led by [steering committee](#) and a [technical committee](#)

What We Do at Virginia Tech

- Require students to develop and submit Electronic Thesis or Dissertations (ETDs)
- Provide a [web site](#) to help students
- Support a [digital library](#) of ETDs
- Develop a [workflow model](#) for submitting ETDs
- Give [talks](#)
- Write [papers](#)

How YOU Can Participate

- Come to [organizational meetings](#)
- [Join us](#) and develop your own NDLTD member site with our help!
- Contribute to our [e-mail list\(s\)](#)

Our Objectives

- **To improve graduate education** by allowing students to produce electronic documents, use digital libraries, and understand issues in publishing
- **To increase the availability of student research** for scholars and to preserve it electronically
- **To lower the cost** of submitting and handling theses and dissertations
- **To empower students** to convey a richer message through the use of multimedia and hypermedia technologies
- **To empower universities** to unlock their information resources
- **To advance digital library technology**

Further Information

- Statistics on [usage](#) of Virginia Tech collection
- General and historical [information](#)
- Information for [publishers](#)
- Issues in [copyright](#)
- Doctoral students can win an [Innovation Grant](#)
- Links to [related projects](#)
- Links to [related \(meta-\)initiatives](#)

Questions? Comments? etd@ndltd.org

NDLTD History, Description, and Scope

Early History

The concept of electronic theses and dissertations (ETDs) was first openly discussed at a 1987 meeting in Ann Arbor arranged by UMI, and attended by representatives of Virginia Tech (Ed Fox from Computer Science and Susan Bright from the Computing Center), University of Michigan, SoftQuad, and ArborText. As followup, Virginia Tech funded development of the first SGML Document Type Definition (DTD) for this purpose, by Yuri Rubinski of SoftQuad.

Virginia Tech's Dean Gary Hooper agreed to finance further development in 1991. Ed Fox and John Eaton (Dean of the Graduate School) have collaborated on this project since that time, investigating problems associated with production, archiving and access, initially with a local faculty committee. Since 1992 they have worked with the Coalition for Networked Information (CNI), the Council of Graduate Schools (CGS), UMI and other interested organizations, helping run a series of design and discussion meetings. Additionally, the University Library's Scholarly Communications Project developed the procedures and systems for processing, archiving, and providing public access to Virginia Tech's graduate research works.

SURA Support

In 1993, at the inception of the Monticello Electronic Library Project, supported by SURA and SOLINET, Professor Edward Fox of Virginia Tech became Co-Chair of its Working Group on Theses, Technical Reports and Dissertations. In 1994 SURA funded a workshop at Virginia Tech to develop plans for electronic theses and dissertations (ETDs), selecting Adobe's Portable Document Format (PDF) and the Standard Generalized Markup Language (SGML) for representation and archiving. To help implement these plans, SURA has funded a [research, development, and dissemination effort](#) based at Virginia Tech for 1996.

Goals

The main goals of the ETD initiative are:

- for graduate students to learn about electronic publishing and digital libraries, applying that knowledge as they engage in their research and build and submit their own ETD,
- for universities to learn about digital libraries, as they collect, catalog, archive, and make ETDs accessible to scholars worldwide,
- for universities in the Southeast and beyond to learn how to unlock the potential of their intellectual property and productions,

- for graduate education to improve through more effective sharing, and
 - for technology and knowledge sharing to speed up, as graduate research results become more readily and more completely available.
-

Recent Virginia Tech Activities

Since 1994, the short term solution at Virginia Tech has been for students to submit their documents as Portable Document Format (PDF) files. Students create PDF files using software running on Windows, Macintosh, or UNIX systems. These PDF files may be moved across computer platforms and operating systems and still retain all their formatting (the electronic documents look just like the paper copy---indeed a paper copy can be printed from the PDF file!). Use of PDF costs the students nothing: the Adobe Acrobat Reader software that is necessary to read the document is free and may be downloaded from the World Wide Web. The student submits his/her ETD via a WWW submission page, by file transfer protocol (FTP), or by submitting a floppy disk.

When the Graduate School receives the PDF file, it is reviewed for errors in formatting. If the ETD passes published quality requirements, the library catalogs the ETD and places it on the electronic bookshelf for ETDs, which supports flexible browsing. A simple search engine facilitates access too, and will be replaced by a more powerful system when the number of documents warrants.

Library patrons can use the online catalog to locate ETDs and use the given Internet addresses (URL) to go to the Web ETD resource. Patrons can then down-load them to their own computers or to library workstations and view them or print them, as desired.

1996 Pilot Project

Virginia Tech is developing tools for students to submit ETDs both as SGML and PDF documents. For the SGML version, SGML constructs can refer to non-text objects, and those objects would be stored in widely accepted standard representations (e.g., JPEG for color images, MPEG for video). SGML documents are more easily archived, more easily searchable, more reusable (e.g., to copy an entry in a bibliography, or to test a new hypothesis using the data and model in a spreadsheet), and therefore are more valuable to scholars.

As the software is developed, other southeastern universities (e.g., Auburn, Clemson, Delaware, Georgia, Georgia Tech, Oklahoma State, Mississippi State, NCSU, and West Virginia) will help test the ETD software. When the software is released, it will be available to other institutions for local use as a part of the Monticello Electronic Library project.

Virginia Tech also will coordinate development and implementation of a distributed digital library system, so that ETDs from all participating institutions can be easily accessed. This will allow browsing and searching (based on institution, date, author, title, keywords, and full-text), as well as downloading for local reading or (selective) printing.

MAGAZINE

National Digital Library of Theses and Dissertations

A Scalable and Sustainable Approach to Unlock University Resources

Edward A. Fox, John L. Eaton, Gail McMillan
Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer
Virginia Tech
Blacksburg, Virginia
<http://etd.vt.edu/etd/>
etd@vt.edu

Project Director: Edward A. Fox

D-Lib Magazine, September 1996

ISSN 1082-9873

Table of Contents

1. [Introduction](#)
2. [Expected Benefits](#)
3. [How You Can Help](#)
4. [Since 1987](#)
5. [Pilot Efforts at Virginia Tech](#)
6. [Related Work](#)
7. [Sustainability](#)
8. [Scalability](#)
9. [Plans](#)
10. [Acknowledgments](#)

1. Introduction

As of September 1, 1996, the U.S. Department of Education provided grant support for a three-year, Virginia Tech-led project to *Improve Graduate Education with a National Digital Library of Theses and Dissertations* (NDLTD), adding to 1996 funding from the Southeastern Universities Research Association (SURA) for *Development and Beta Testing of the Monticello Electronic Library Thesis and Dissertation Program*. True success in these projects will potentially mean a permanent change in graduate education and scholarly publishing, with digital libraries playing a more dominant role in supporting and disseminating research.

This article serves as an overview of the project, indicating what benefits are likely, what roles various partners (including, we hope, you, the reader) may play, and what related work has occurred in the past. It is also an invitation to universities to unlock their resources in connection with this collaborative project.

If many in the international community join in, the project could lead to a multilingual corpus of vast proportion and significance. Our collection focus is on doctoral dissertations and masters theses, so we will repeatedly refer to TDs (theses and dissertations) or ETDs (electronic theses and dissertations). However, we also welcome special reports (especially those prepared by graduate students) and bachelor theses. Since there are over 40,000 doctoral and 360,000 masters degrees awarded in the U.S. alone each year, and since our aim is for all graduate students to learn how to publish electronically, the annual growth rate of the collection could exceed 100,000 new works per year by the turn of the century. If there is a fair amount of multimedia content included, as we expect will be the case, the collection might increase in size at the rate of about a terabyte each year.

2. Expected Benefits

The NDLTD should help almost everyone, and so, through broad cooperation and participation, should be a sustainable effort. Let's take a moment to consider its likely effects on the key parties involved: students, universities, the research community, and the publishing world.

Students

Our project is primarily an effort to improve graduate education. We will work so that graduate students become *information literate*, learning how to become electronic publishers and knowing how to use digital libraries in their research. The overall process is shown in our diagram of the [Life Cycle of an ETD](#). Toward this end we continue to develop written documents, extensive WWW materials, and a distributed education and evaluation program in which universities accept responsibility for local support.

With access to the NDLTD, graduate students will be able to find the full texts of related works easily, to read literature reviews prepared by their peers, and to follow hypertext links to relevant data and findings. Their professors will be able to point to the best examples of research in their area, even to the level of an interesting table, an illustrative figure, or an enlightening visualization. Also, students can benefit by learning how to become electronic publishers, preparing them for their future work. Since this educational initiative targets all graduate students, it is unique in its potential to train future generations of scholars, researchers, and professors. If they can publish electronically and add to digital libraries, future works they write will not have to be scanned or re-keyed.

Graduate students also may be empowered to be more expressive as they prepare their submissions for the NDLTD, if such is allowed by their committee, department, and university. Some students have already prepared hypertexts as literature, included color images or graphics, illustrated concepts with animations, explained processes with video, or used audio when dealing with musical studies. One masters project about training students to use *Authorware* included an Authorware tutorial in the appendix. This has already helped people in South America learn more about multimedia technology.

Access begets access, so having more graduate works in the NDLTD is likely to simulate greater interest in theses and dissertations (TDs). Studies at Virginia Tech of the average number of times a paper TD circulates per year indicate a steady growth from 0.55 to 0.85 circulations between years 2 and 4 for

dissertations, with a roughly parallel line for theses reflecting an increase from 0.4 to 0.68 circulations over the same period. Based on the increases we have seen in numbers of accesses to electronic journals as they became available on WWW, we expect that there will be a dramatic increase in the average number of accesses to TDs when they shift from paper to NDLTD availability. Most students are eager for their works to be cited, and we plan for our monitoring and evaluation system to record such accesses. With bibliographies on-line too, a citation index among NDLTD entries will be possible as well, helping students keep track of new studies related to their investigations.

Finally, students are likely to benefit financially from the NDLTD. Publishing electronically should save them the costs of preparing at least some of the paper copies now required. There also may be lower fees from their university and other parties for filing their final copy.

Universities

Few universities have a university press, and many of those are not profitable. Yet, through the NDLTD, every university can publish the results of its graduate students with a minimal investment. This should increase university prestige, and interest outsiders in the research work undertaken.

University libraries can save shelf space that would otherwise be taken up by TDs, and the costly handling of paper TDs by personnel in the graduate school and library can be reduced or eliminated. At Virginia Tech, for example, the catalogers decided to have students assist with cataloging by adding keywords to the cover page, thus reducing processing in the library. It seems likely that at least one person in each large university can be freed to work on other tasks if proper automation takes place, resulting in simplification of the work flow related to TDs. In addition, library on-line catalogs can provide fuller information by including the abstract from the electronic text.

Research and Publishing

Student research should be aided by the NDLTD since graduate students will have a single repository for the work of their peers, supported by full-text search. Other researchers, including people in companies interested in opportunities for technology transfer, can look to the NDLTD as a way to quickly learn of new findings.

Suppliers of electronic publishing software already have found it valuable to participate in the NDLTD. Adobe is making generous donations of their Acrobat software, in part because they realize that having all graduate students exposed to the Acrobat line of tools will ensure a large base of future users. Associations like ACM (the *First Society in Computing*) are supportive, in part because having their members learn to publish electronically in graduate school can help reduce the anticipated cost of shifting to electronic publishing, when authors will be expected to submit final copies of acceptable articles in proper forms (e.g., using SGML) for publication and electronic archiving.

Indeed, it is likely that future shifts in publication practices will be facilitated by the effort to build the NDLTD. This is of particular interest to universities, which now cannot control what happens to the research publications they support, and later spend large sums to buy back research publications from commercial publishers. Through the NDLTD, universities can control one important class of the intellectual property they produce, and can share it freely with other universities to reduce overall costs.

3. How You Can Help

Since almost everyone stands to benefit from the development of the NDLTD, we encourage you to help in this process in a way that fits the mission of your institution. For example, if you are engaged in the development of software or systems for digital libraries, or helping with standards efforts, you can help directly with building the NDLTD. If you are at a university, you can help build local consensus and devise a supporting infrastructure so the NDLTD is a key part of graduate education.

Software, Standards, Systems

The NDLTD presents unique challenges on many fronts, and help is needed in various technical areas. On the one hand, it is desirable for graduate students to be expressive, using multimedia representations, but this can lead to very large works, even when compression is required. While we observe many ETDs only require on the order of a megabyte, we expect that with images and other media forms, the average size will approach 5-10 megabytes. A single video file can consume one or two orders of magnitude more space; it is fortunate that a computer system with two terabytes of hierarchical storage is available at Virginia Tech to support this project! But even more important will be good software to undertake content analysis of multimedia information. Other software is needed to help with electronic publishing, and other aspects of digital library operations.

Standards also are essential for the success of the NDLTD. If the archive will last for decades, hopefully centuries, its content must be usable many years after publication. If authors work with standard representations, those are more likely to be understood than are representations that are unpublished and proprietary in nature. If the number of standards supported is kept to a minimum, there will be less work in *refreshing* the archive as technologies and standards change, calling for conversion to more modern storage and representation schemes.

The NDLTD must operate as a production service if it is to replace current library approaches to handling TDs. Thus, reliable, commercially supported digital library systems are needed for long term success. Companies like University Microfilms International (UMI), IBM, and Online Computer Library Center (OCLC) are participating in the unfolding of the NDLTD. Thus, IBM donated a large SMP computer that will serve as the central host for this effort, and which can run IBM digital library software. Various IBM products for handling databases, image collections, searching on image content, and rights management have great potential for helping with the NDLTD.

Building Local Consensus

At universities, while moving toward the NDLTD is clearly advantageous, such a shift requires many changes in policies and practices. The best way to accomplish this seems to be to develop a local plan, with guidance from key staff in the graduate school(s), the library, and the computing or information technology operations, as is illustrated in our diagram of [ETD Site Implementation](#). Then, workflow changes and automation opportunities are likely to be grasped and become practice. With leadership from these three groups, students and faculty can be consulted and involved in detailed planning. It appears likely that a transition period of about a year is required to effect the change from introduction of concept to widespread acceptance and participation in the NDLTD. Note that real benefits of workflow improvement and universal access to online graduate research require a nearly complete shift to electronic submissions of all TDs.

Supporting Infrastructure

If **every** graduate student is to submit an ETD, enhancement to the campus infrastructure is required in

most institutions. Usually, this is more a matter of will and coordination than large expense, and most would agree that the end result is highly desirable. Let us consider several possible scenarios.

First, if Adobe's Portable Document Format (PDF) is the target representation, most PC, Mac and Unix systems can run the software required to prepare PDF files. Though there are minor complexities related to fonts and special formatters like LaTeX, these can be worked out, and the investment by students or labs in Adobe software is not high (e.g., about \$40 per copy of *Exchange*).

Second, if SGML is the target representation, there are various solutions. One is to use a standard editor, inserting tags, much like what is done by many HTML authors. While possible, the number of tags (see our illustration of the [Parts of an ETD](#)) needed makes this cumbersome. Thus, it is better to use an SGML editor, but those are expensive. Microsoft is assisting with the investigation of the *SGML Author* extension to *Word* as an appropriate tool, which could be made available in small numbers in campus labs. Virginia Tech is working on conversion software and templates to enable students to use preferred environments like LaTeX, and to automatically make a 100% accurate conversion to SGML.

Third, there is the question of images. Since many TDs have some type of artwork, color scanners with high quality capture capability at 600dpi must be available, along with computers, adequate RAM and disk storage, software (e.g., Adobe Photoshop), technical assistance, and network transfer capabilities to move the results to locations students can more easily access.

Finally, there is the higher end requirement of supporting special multimedia forms. Special systems for audio and video capture and compression are required for these media types. Note, however, that if there are no special multimedia laboratories available on campus, students can pay for such services themselves.

4. Since 1987

Though the ND LTD is new to many readers, work on it actually began late in 1987! A brief history is in order.

UMI and Electronic Manuscripts

Nick Altair, then at UMI, who had recently worked on the Electronic Manuscript Project, convened a meeting in 1987 in Ann Arbor, Michigan. Representatives from University of Michigan, ArborText, SoftQuad, and Virginia Tech participated.

Soon after, Yuri Rubinsky of SoftQuad worked with Virginia Tech to develop the first SGML Document Type Definition (DTD) for TDs. (This was only revised in 1996, in connection with recent efforts supported by SURF - see below.) Virginia Tech continued work in 1989 and 1990, experimenting with conversion of TDs that were obtained on diskette from student volunteers.

CNI and Project Discovery

In 1992, the Coalition for Networked Information sponsored a project discovery workshop with 11 invited universities, each of which had documented the interest of their graduate school, library and computing/information technology groups. This meeting was planned by representatives of UMI, Council of Graduate Schools, and Virginia Tech. Subsequently, a number of further discussions were held at CNI meetings. In connection with one of these, representatives from UMI and Virginia Tech

visited Adobe, to learn about plans for the Adobe Acrobat family of tools.

SURA/SOLINET and Unlocking University Resources

In 1993, SURA and SOLINET (Southeastern Libraries Network) joined forces to work toward the Monticello Electronic Library. At the first open meeting, Edward Fox of Virginia Tech was invited to give a presentation, re-introduced the idea of the ETDs, and subsequently became co-chair of the working group on theses, dissertations, and technical reports. There was widespread interest in this and subsequent meetings, and University presidents saw the potential benefits as well at various SURA discussions. Consequently, a group of interested universities sent representatives to a workshop at Virginia Tech in August 1994, hoping to develop specific plans for ETDs. One key decision from that meeting was to work toward a dual representation scheme. Thus, two copies of each TD would be archived, one using Adobe PDF and the other using SGML. Virginia Tech and UMI agreed to explore the SGML conversion problem in more detail. Virginia Tech began to convert some of the TDs it received to PDF.

SURA and Beta Testing

Late in 1995, Virginia Tech prepared a pre-proposal to the U.S. Department of Education regarding a three year effort to build the NDLTD, and also requested that SURA fund initial work on establishing a part of the Monticello Electronic Library for ETDs for the Southeast. The first of these led to funding September 1, 1996 and the latter covered calendar year 1996 pilot efforts in the Southeast. North Carolina State University was the first institution seeking to join the initiative, and initial electronic submissions are expected there in October. The first regional workshop for Southeastern universities was held August 1-2, 1996, hosted by University of North Carolina, Charlotte. Many discussions have been held, and presentations given, in the region, nation, and even internationally. There appears to be interest in such institutions as: Auburn, Clemson, Georgia Tech, Michigan State, Mississippi State, MIT, Oklahoma State, University of Georgia, University of Utah, University of Virginia, and Vanderbilt.

5. Pilot Efforts at Virginia Tech

Interest in ETDs has continued and spread since 1987.

DTD Development

While SGML has always seemed the logical choice for an archive of TDs, there have been serious technical and economic problems that have delayed its usage. First, few graduate students had heard about SGML, and it seemed unlikely that we could educate them about it. However, with the growth in interest in HTML, this problem has been largely eliminated. Second, there are few freely available editors for SGML. While this continues to be the case, discussions are underway with a number of vendors to work out economically feasible solutions in the context of the NDLTD. Third, there has been the problem of how to find an acceptable DTD that would be suitable for authors, technically sound, and could be adopted nationwide. We believe we have solved this problem - see the DTD and related documentation at our WWW site (<http://etd.vt.edu/etd/>). While it may evolve as comments are received, we hope some version of it will be universally adopted so that TDs are tagged to facilitate searching and formatting. In particular, we have developed software to convert from documents prepared according to the DTD to HTML (for WWW delivery - see our illustration of the [ETD Hyper-Text Structure](#)) or LaTeX (for rendering to paper or page images).

Finally, there is the outstanding problem of conversion from word processors and formatters to SGML. We are developing a set of LaTeX macros to ensure reliable conversion from LaTeX to SGML. Similar efforts are planned for Microsoft Word, but may be simplified if SGML Author for Word will fit into the plans.

Capture with PDF

In the last several years, PDF has matured and been more widely supported, with freeware tools like *xpdf* aiming to round out the ability to read such documents on UNIX systems. Any computer with Adobe Exchange can write to the PDFwriter instead of a laser printer, and create a PDF file. PostScript files can be converted to PDF using Adobe Distiller. Since almost every tool used in document creation can either work with the PDFwriter or yield a PostScript file, electronic publishing to PDF is relatively straightforward and can be taught during a 1-2 hour training session.

One technical problem with PDF that troubled our early efforts has been solved in 1996. That is, there are publicly available *outline* fonts that allow authors who work with LaTeX to prepare PDF files without including bitmap fonts (which increase file size, make display and reading on screen impossible, and restrict text searching options). We are developing automated services on Sun systems to allow authors to prepare PDF files with the proper outline fonts included.

Workflow Automation

Automation is the key to increasing efficiency in handling TDs. The Library and Graduate School have completely redone the flow of work at Virginia Tech so as to eliminate steps and carryovers from the world of paper. For example, authors now are encouraged to submit single spaced documents, which are easier to read on the screen than double-spaced documents. Authors assign keywords to their documents, since catalogers have trouble assigning categories to new works like TDs. Authors directly upload their submissions to a library server, where the graduate school can check for proper form; there is no longer a need to *deliver* to the graduate school and have them move completed works to the library.

Central to our automation is a WWW submission page, which is filled in by the author, and leads to uploading and archiving of the TD. This operation includes students authorizing the university to handle access (non-exclusively) to their works, classifying the work (e.g., thesis or dissertation), and providing email information about them and their chair (so completion of processing can lead to automatic notification).

When SGML submissions are easily accomplished, they will be the basis for a variety of derivatives. One is the HTML version already mentioned. Another is the MARC record needed for cataloging. Third is the entry for *Dissertation Abstracts*. Once these can be produced, the submission process will be simplified even further.

Workshops

Since Spring 1996, there have been a variety of workshops to train students regarding electronic publishing (using PDF, tools like Word and Exchange, LaTeX) and the automated submission effort. By holding events every few months, handling email questions, making special visits to interested groups, and providing on-line FAQ files, the needs of graduate students are being addressed.

The Faculty Development Initiative at Virginia Tech involves training the entire faculty over a four year

period about electronic publishing, workstations, networked computing, and educational technologies. A regular part of the FDI is for faculty to learn about Adobe Acrobat and the handling of ETDs - thus over 600 have been trained about this initiative. Others have been exposed in College meetings, through newspaper explanations, or through the open workshops.

Requirement

In Spring 1996, the Commission on Graduate Affairs agreed to require ETDs at the start of 1997. Thus, all students will prepare an electronic submission, and the Library and Graduate school will not accept or receive paper submissions. This is a serious plan! We hope that after months or perhaps a year of working with the NDLTD, other universities will follow this scheme, so that students really will learn how to publish electronically and use digital libraries.

6. Related Work

Development of the NDLTD fits in with other digital library and other electronic publishing efforts. Some of the most closely related ones are as follows.

NCSTRL

Beginning in 1992, with the Wide Area Technical Report Service (WATERS), Virginia Tech has been involved in digital library efforts related to computer science technical reports. In 1995, the WATERS group joined the CSTR group to form the Networked Computer Science Technical Report Library (NCSTRL). Virginia Tech is a regular member. Fox is a member of the NCSTRL Working Group, and the NCSTRL backup server runs in the Virginia Tech Computing Center.

The software used with NCSTRL is available for use with NDLTD and can support a distributed system including situations in which UMI and Virginia Tech, for example, serve sites that do not wish to maintain their own servers.

NCSTRL is one of the early adopters of the CNRI handle system. Virginia Tech has obtained permission for a top-level naming authority for *theses* and will run a local handle server for TDs so that persistent names can be guaranteed.

Envision

Since 1991, Virginia Tech has worked with ACM and others to develop a prototype digital library for computer science and to apply it to improve related educational efforts. Some of the software developed may be of use for NDLTD. The methods and tools used for monitoring WWW use and analyzing that data will be a part of the evaluation component for NDLTD.

IBM Digital Libraries

IBM has collaborated with Virginia Tech in several ways regarding digital libraries. The central server for the NDLTD will run IBM digital libraries software. Where possible, local development will be reduced when commercially available solutions apply.

7. Sustainability

For digital libraries to be successful, they must be sustainable, scalable and usable. With a world-class Center for Human-Computer Interaction at Virginia Tech, and with a Department of Computer Science whose main focus is HCI, working toward a usable system will be an ongoing and central concern for our efforts. Usability labs and research in remote usability evaluation should help our efforts, as will related projects for WWW monitoring and analysis. So, we turn our attention to the other two legs of successful digital libraries, starting with sustainability.

Mission

Every university with a graduate program is obliged to deal with TDs and to ensure that graduate students are properly educated. As argued above, the NDLTD is in the best interest of students and universities. Thus, to carry out the mission of educating graduate students and handling their TDs, universities should ensure that they know how to publish electronically and how to use digital libraries, which can be accomplished most efficiently by joining the NDLTD effort.

Similarly, many university libraries and/or archives have assumed the responsibility of having copies of works written by local faculty, staff and students. This has been a particularly strong tradition in the arena of theses and dissertations. On many campuses the library is committed to maintaining such works indefinitely, which fits into the long term goals of the NDLTD.

Infrastructure

Universities support students in their roles of publishers and researchers. Having the right infrastructure to support local involvement in the NDLTD fits in with the general type of support that universities need to provide.

Economics

Because of saving the costs of copying and submitting paper versions of their TD, we believe students have an economic incentive to participate in the NDLTD. Similar savings are expected for universities, in particular their graduate school and library. Since students still will provide some payment to their university when submitting the TD, there is an economic foundation for continuing the effort as a self-sustaining enterprise.

8. Scalability

The NDLTD effort is scalable by its very nature. First, it builds upon a system of higher education (in the United States) that has demonstrated its ability to scale to meet needs throughout the twentieth century. Second, it makes use of technology that is modular and distributed, and which is addressing needs of a growing number of computer science departments. Further, this effort piggybacks upon other normal activities of universities, relating to education, scholarly communication, and libraries - each of which demonstrates a fair degree of scalability.

University

Each university has responsibility for its own TD collection, but can handle that locally or assign it to others. At the level of a university the problems are not terribly large - even if a thousand ETDs are submitted in a year, the disk space required to store them probably would cost less than \$3,000.

State

In some cases, there are statewide consortia for library information sharing. Thus, the Virtual library of Virginia (VIVA) initiative could allow for a statewide coordination of part of the NDLTD, supporting the needs of small universities where running suitable servers is not warranted.

Region

As in the case of the Monticello Electronic Library, having a regional consortium for NDLTD is quite sensible and feasible. There are parallel groups to SURA, SOLINET, and the Conference of Southern Graduate Schools in other regions of the U.S.

Nation

In the U.S., the NDLTD represents the national effort. Researchers in other countries like Korea are looking into a similar effort in their country that would connect with NDLTD.

Repositories

For NDLTD to be successful, there must be long term support. UMI already has an archive of 1.3 million TDs, in microform, and is willing and able to provide long term electronic archive services. Other groups also are interested in this opportunity. Negotiations between universities and UMI are needed to work out the proper arrangement for all parties, in the context of the growth of NDLTD.

9. Plans

Future work on the NDLTD is laid out in the proposal to the U.S. Department of Education, which is included in PDF form on the WWW pages for the project. Collaboration with UMI is expected on all fronts. Some of the other highlights are as follows.

Technical Development

The NDLTD effort will involve collaboration with the Cornell Digital Library Research Group, which has developed the software used with NCSTRL, and with CNRI, which has developed the handle system and other digital library services. There also is collaboration with IBM regarding their digital library systems and software. OCLC has promised support from its Office of Research, especially regarding useful tools. Other collaboration will take place in the context of electronic publishing work, such as with Adobe.

Administrative Collaboration

The NDLTD has support from many groups interested in universities, graduate education, libraries and networked information. There will be close coordination with the national and regional graduate school groups, presentations supported by CNI, and of course ongoing work with SURA and SOLINET, as well as similar associations in other regions.

Education and Evaluation

Since we aim to improve graduate education, we must afford equal access and undertake a careful

evaluation. A detailed evaluation plan is given in the proposal, to include surveys, logging, focus groups, and other efforts. Usability studies will help with detailed analysis and improvements of interfaces. It is expected that about one-third of the project will relate to evaluation issues, both formative and summative.

Thus, we hope to not only develop a large and valuable digital library to support graduate education and research, but also to show that it has proved to be of benefit, and that graduate students indeed know how to publish electronically and how to use digital libraries.

10. Acknowledgments

The U.S. Department of Education's Fund for the Improvement of Post Secondary Education supports NDLTD. Authorized funding for the first year is in the amount \$69,762. Anticipated future funding for years 2 and 3 are: \$69,337 and \$68,941. If all federal funding is received as planned, the total would be \$208,040. Virginia Tech will provide institutional support which gives federal/nongovernmental percentages 53.3/46.7. Additional in-kind support for the FIPSE proposal has been promised by: ACM, Adobe, Council of Graduate Schools, Coalition for Networked Information, Cornell Digital Library Research Group, Council of Southern Graduate Schools, IBM, OCLC, State Council of Higher Education for Virginia, SOLINET, SURF, UMI, and University of Utah Graduate School.

Copyright © 1996 Edward A. Fox, John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer



[hdl://cnri.dlib/september96-fox](http://cnri.dlib/september96-fox)

M A G A Z I N E

Networked Digital Library of Theses and Dissertations

An International Effort Unlocking University Resources

Edward A. Fox, John L. Eaton, Gail McMillan
Neill A. Kipp, Paul Mather, Tim McGonigle, William Schweiker, and Brian DeVane
Virginia Tech
Blacksburg, Virginia 24061-0106
<http://www.ndltd.org>
etd@ndltd.org

Project Director: Edward A. Fox

D-Lib Magazine, September 1997

ISSN 1082-9873

Table of Contents

- [1. Introduction](#)
- [2. Progress](#)
 - [2.1. Collaboration](#)
 - [2.2. Infrastructure](#)
 - [2.3. Tools](#)
 - [2.4. Collection](#)
- [3. Controversy](#)
 - [3.1. Social Issues](#)
 - [3.2. Time, Effort, Impact, Reward, and Quality](#)
 - [3.3. Loyalties](#)
 - [3.4. Economics](#)
- [4. Future Growth](#)
 - [4.1. Digital Libraries and Education](#)
 - [4.2. Universities Working Together](#)
 - [4.3. Evolution](#)
- [5. References](#)
- [6. Acknowledgments](#)

1. Introduction

On the first anniversary of funding by the U.S. Department of Education (FIPSE) for a National Digital

Library of Theses and Dissertations, we review its origins (see [\[FOX96a\]](#) for an overview of the project), describe progress-to-date that warrants its now being called the Networked Digital Library of Theses and Dissertations (NDLTD), explain some of the controversy that has led to widespread publicity and dissemination, and explore future growth possibilities.

The first workshop about electronic theses and dissertations (ETDs) took place in 1987 with a technical focus on standards, namely applying SGML to the description of research. Ten years later, we realize that the proper aim should be improving graduate education by having students enter ETDs into a digital library which facilitates much broader access. Achieving that goal calls for a sustainable, worldwide, collaborative, educational initiative of universities committed to encouraging students to prepare electronic documents and to use digital libraries - NDLTD. Since students often learn best by doing, this competency-oriented program should ensure that the next generation of scholars is prepared more completely for the Information Age, in which they can apply and pass on their skills in academia or other research situations.

With funding in 1996 from the Southeastern Universities Research Association (SURA), our Virginia Tech team was able to build upon local efforts, including a solid foundation of library-developed processes, to facilitate a beta program in the Southeast. Additional support from FIPSE, and in-kind contributions from a number of sources, especially Adobe, IBM, and Microsoft, have enabled expansion to the national and international levels. Public forums afforded by the Coalition for Networked Information (CNI), the Council of Graduate Schools (CGS), and many other groups, have made the idea of an ETD initiative a familiar topic to hundreds of leaders at diverse universities. Much larger numbers have heard about the topic through newspaper, radio, and TV coverage [\[NDLTDa\]](#). Yet, because news coverage often focuses on controversy, the discussion below attempts to concentrate on progress made and to dispel some misconceptions that may have arisen.

2. Progress

As NDLTD has expanded, we have seen progress in many places. For example, prompted in part by the NDLTD, UMI, which has the world's largest microform archive of theses and dissertations, has launched its ProQuest Direct service of scanning (at 300 dpi) and using optical character recognition software to convert the scanned documents into text files (into PDF, so text is recognized as accurately as current tools allow) works it receives after 1996. Many groups, including CGS, have established committees to explore the concept of ETDs. Although these deserve concentrated attention, we will focus mainly on collaborative efforts and work that is specifically oriented toward building the NDLTD.

2.1. Collaboration

Local progress toward NDLTD has been made possible as a result of efforts by the Library, Graduate School, funded project team, and other parts of Virginia Tech. Several student project teams in courses in computer science (CS4624, CS5604) have made important contributions, assisting in the preparation of multimedia training materials and prototype digital library implementations. Students studying digital libraries (in University Honors 3004 and CS6604) in Fall 1997 have already started to select term projects to assist our efforts. Professor Jong-Min Bae has come from Korea to spend a sabbatical year starting August 1997, providing further aid at Virginia Tech. In addition, there are students, faculty, and staff at other universities and organizations providing assistance by testing, adapting, and extending Virginia Tech's programs and processes.

At the University of Waterloo, a team has been studying about ETDs, and prepared a survey of

worldwide activities [WATE97]. North Carolina State University recently established an ETD Web site, and the University of Virginia makes available on WWW a student-run ETD resource directory, plus pointers to publications showing student interest in the initiative [KIRS97]. From these sites one can learn about investigations and pilot efforts in Australia, Aalborg University, The University of Texas at Austin, and University of South Florida. We invite others involved in related efforts to provide pointers so that we may cite their work.

Though it may take 12-18 months for a university to investigate the idea of ETDs, develop suitable policies, reach consensus, launch a pilot effort, begin to train students, and enhance local infrastructure to facilitate network submission by students, some institutions have moved more rapidly. Among those institutions joining NDLTD most quickly are those in Dagstuhl, Germany and Monterey, California. At the Darmstadt University of Technology, it is likely that interest was stimulated because of local expertise in multimedia information and systems. ETDs allow students to apply those technologies directly and go beyond the limits of paper theses or dissertations by including audio, image and video illustrations and by adding hypertext links. In the case of the Naval Postgraduate School (NPS), building upon prior digital library activities [NORR97], a team of Navy reserve officers studied the matter and reviewed documentation provided by Virginia Tech. Very shortly after a telephone conference that obtained additional information, university officials decided to join NDLTD. NPS is obligated to provide access to its theses and dissertations to the Navy worldwide; this is much less expensive if electronic distribution methods can be employed. This shows a clear economic benefit.

During its first year, NDLTD has grown to 20 members, with scores of other institutions interested and in a number of cases, visited or briefed on the initiative. An online status file is maintained to document the current situation [NDLTD6]. At present, Florida is the state with the largest number of members in NDLTD. A team at University of South Florida is helping prepare an edited sourcebook on electronic theses and dissertations, having produced a call for contributions, with publication planned in 1998.

The University of Virginia has taken the initiative on adapting the Dienst system (developed at Cornell, and used in the Networked Computer Science Technical Report Library, <<http://www.ncstrl.org>>) to use for ETDs [MOOR97]. Interoperability tests with Virginia Tech are planned for Fall 1997. Access using Dienst will mean that end-users will have a single view of the distributed set of ETDs. They can use the WWW to browse among the dispersed collections at NDLTD sites, by author, topical area (i.e., department), or year. Alternatively, they can search the full-text of metadata (including abstract) for the full collection or parts thereof, i.e., issue one query to search all sites in parallel. Furthermore, the NDLTD as a whole could have archival and search engines flexibly structured and located to suit economic, political, and social preferences. Universities could keep their own archive or have it managed by an archival service, and search engines could be at each university or run by state, regional, national, or other services. For performance reasons, backup and regional replication systems can be included in the overall architecture. Further work with Dienst should afford other user services, especially if Dienst is used to handle a large portion of computer science preprints, and is extended to manage user profiles and selective dissemination of information.

Steering Committee. Guidance for NDLTD is provided by an international steering committee. The committee meets in the middle of March and September each year in Washington, D.C., and has email discussion during the intervening months. Members represent Canada, UK, World Bank (African Virtual University), universities and libraries in the Southeast (SURA, SOLINET), Western Area Graduate Schools, the National Science Foundation, Adobe, CNI, CGS, IBM, CIC (Big 10), Job Accommodations Network, NSF, OCLC, U.S. Department of Education, and other constituencies.

After hearing reports from UMI and OCLC about archival and access services, members at the March 1997 meeting decided to encourage maximizing access, allowing as many "players" as become interested to provide various services for those interested in ETDs. This will be feasible if all member institutions freely share among themselves and their agents MARC (library catalog) records describing their ETDs, and if each record contains one or more URNs pointing to authentic full copies, such as might reside in a university archive. Thus, the NDLTD support team at Virginia Tech is working to arrange interoperability tests, building upon existing library and digital library infrastructure.

2.2. Infrastructure

To support digital library activities at Virginia Tech, IBM has donated a variety of hardware. One server, acquired to run IBM Digital Library software, and to serve multimedia files associated with ETDs, has four terabytes of hierarchical storage, roughly 40,000 gigabytes---enough for about 40 million average-sized ETDs. Virginia Tech will be hosting a user group meeting (October 20-22, 1997) for groups employing IBM Digital Library systems; the focus will be on human-computer interaction issues. It is hoped that the IBM system will be extended to support gateway and federated search capabilities that will allow interoperability tests among NDLTD institutions.

One of the IBM systems runs OCLC's SiteSearch, thanks to a license donation by OCLC. OCLC is providing over a million MARC records that refer to theses and dissertations from its WorldCat database. This will provide information previously not readily available since few masters theses are included in the UMI database. SiteSearch supports Z39.50, which enables access through a variety of clients [[LYNC97](#)]. It also can be adapted to provide similar functionality to Dienst, so that "federated search" is afforded, with client or gateway merging of results from remote sites [[PAYE97](#)].

In addition to hardware and software to support NDLTD, Virginia Tech also has a rich network infrastructure, including a vBNS (high speed Internet research and education backbone) connection through "Network Virginia," the statewide ATM network that it runs and which includes educational institutions all over the Commonwealth. Local access, so that students can submit their works electronically, is provided by the campus network as well as through the town (Blacksburg Electronic Village, <http://www.bev.net/>). While such an infrastructure is not necessary to participate in the NDLTD, it does improve online processing time and it enhances user access.

2.3. Tools

Students at Virginia Tech use a variety of tools developed to help them prepare ETDs, thanks in part to support from SURF aimed to support growth of NDLTD in the Southeast. We have adopted a scenario-based design approach, and in addition to assembling commonly available inexpensive software packages, have been constructing other files and tools to support low-cost document manipulation as well as efficient workflow processes [[KIPP97](#)]. These include:

- a document type definition for ETD-ML, the markup scheme which has been repeatedly refined to be easy to use and yet powerful enough to capture the important metadata and structure of ETDs;
- a template for *Microsoft Word* to use with *Microsoft SGML Author for Word* (copies of which have been donated by Microsoft) to prepare an SGML version of ETDs;
- Word Perfect and AuthorEditor templates for SGML;
- a template for LaTeX that will allow production of page-based versions through *LaTeX2e* as well as conversion to an SGML version;
- Panorama style sheets for viewing SGML ETDs over WWW;

- a converter for SGML ETDs to prepare a network of HTML pages, to ease dissemination of information and to simplify navigation (i.e., which automatically produces links between chapters, and between the table of contents and parts of the ETD); and
- multimedia training materials explaining how to use PDF tools, how to produce a PDF document, how to add links to PDF documents, how to scan images for inclusion into ETDs, etc.

Collaboration with staff at the University of Virginia who are involved in the Text Encoding Initiative (TEI) includes demonstrating interoperability between documents marked up with ETD-ML and those marked up according to the *TEI Guidelines*. Collaboration with a student at Rhodes University in South Africa deals with testing many of the tools discussed above, and complements efforts underway with various institutions in the Southeastern United States.

2.4. Collection

Institutions involved in NDLTD are all working toward having students prepare ETDs so they can learn from that experience and at the same time help build a large and smoothly functioning digital library. The contents of the Virginia Tech WWW site are distributed on CD-ROM to institutions that join the NDLTD. This is intended to help other institutions provide information and access to their community. Therefore, Virginia Tech's material has been reorganized into three parts, to discriminate clearly among the following:

1. NDLTD (the project and its members);
2. student submission (including policies, checklists, training materials, and automated scripts);
3. the searchable collection.

Virginia Tech students have submitted over 500 works that are included in the Library's online catalog. A variety of additional services are provided on an interim basis until the collection gets larger and until distributed digital library software is tested in conjunction with other NDLTD members. Thus, browsing is supported, with a separate list for recent works, as many local students are eager to have their work immediately accessible, and many outsiders look for the latest findings. The OpenText system indexes each text or PDF part of each ETD, and handles full-text searching by all interested parties. Other software will index image files to support searching on image content.

A number of other NDLTD members already have online documents, including: Naval Postgraduate School, NC State University, and University of Virginia. In addition, searches over the Internet and discussions with personnel from a variety of universities have turned up small collections of works, made available by individual departments or centers. We have contacted each new interested party to see if they would join NDLTD.

3. Controversy

Since the inception of FIPSE support for NDLTD, we have addressed several controversies, working with students, faculty, and publishers. While we anticipated concerns of publishers, and released in 1996 a *Statement about Publications* specifically targeted to assuage concerns of that audience [\[FOX96b\]](#), it appears that few have read that document or been assuaged by it. A variety of efforts are now underway to prepare paper booklets and eventually books to explore matters more fully, document the various perspectives, and explain many of the legal and technical complexities. We hope that such efforts will broaden the discussion and understanding, facilitate cooperative agreements between all parties involved, and further our aim of having students and universities understand more about preparing

electronic documents and using digital libraries.

Meanwhile, there has been extensive news coverage related to NDLTD, e.g., an *NPR Morning Edition* story, an article in the *NY Times* that was later picked up by a number of regional newspapers, and an interview on a Singapore TV morning show [NDLTDa]. Much of that coverage concerns Virginia Tech's making ETDs freely available in connection with the NDLTD, and statements by publishers that they would not accept submissions that appear on WWW.

We believe that worries of publishers in this regard can be resolved by some variant of the Approval Form [NDLTDc], which is explained in an open letter to students [FOX97a]. In particular, this form requires students and their faculty committee members to sign an agreement in which they:

- indicate having permission for including in their ETD any items with third-party copyright;
- give the University non-exclusive license to archive the ETD and make it accessible;
- select one of several options regarding allowing access to their ETD:
 1. release worldwide;
 2. prohibit access for 1 year, renewable for an additional year, to allow patent application or to satisfy proprietary restrictions related to the research reported;
 3. restrict access to the campus wherein the work was prepared;
 4. allow access from the local campus to the entire ETD, and allow worldwide access to selected parts (explicitly listed);
- indicate when the author can be contacted to see if access can be extended (i.e., after a year, after three years, or probably never);
- optionally identify those who might serve as proxy to change access restrictions in the future (e.g., if some of those signing will not be available).

We discuss further aspects of the controversy related to ETDs in the subsections below.

3.1. Social Issues

As has been discussed at events such as the 1996 Allerton conference (<<http://edfu.lis.uiuc.edu/allerton/96/>>) about social and user aspects of digital libraries, the success of a digital library project depends strongly upon how it relates to the activities of individuals, groups, organizations, and institutions, as well as the broader social context. Additional research on these matters should be given high priority [BORG96].

While this philosophy has been adopted since the early 1990s in connection with developing a program for ETDs, the ramifications and practical impact of various concerns of students, faculty, and publishers have not yet been summarized for the digital library community. Fundamentally, those concerns fall into three main categories, covered in the next three subsections, namely those relating to: time, effort, impact, reward, and quality; loyalties; and economics.

3.2. Time, Effort, Impact, Reward, and Quality

Theses and dissertations are written as part of the requirements for graduate studies. While there have always been particular quality constraints enforced on those works by faculty and officials handling graduate affairs, and while few have been willing to complain about those rules, changing those rules in a significant manner has caused a number to complain vociferously.

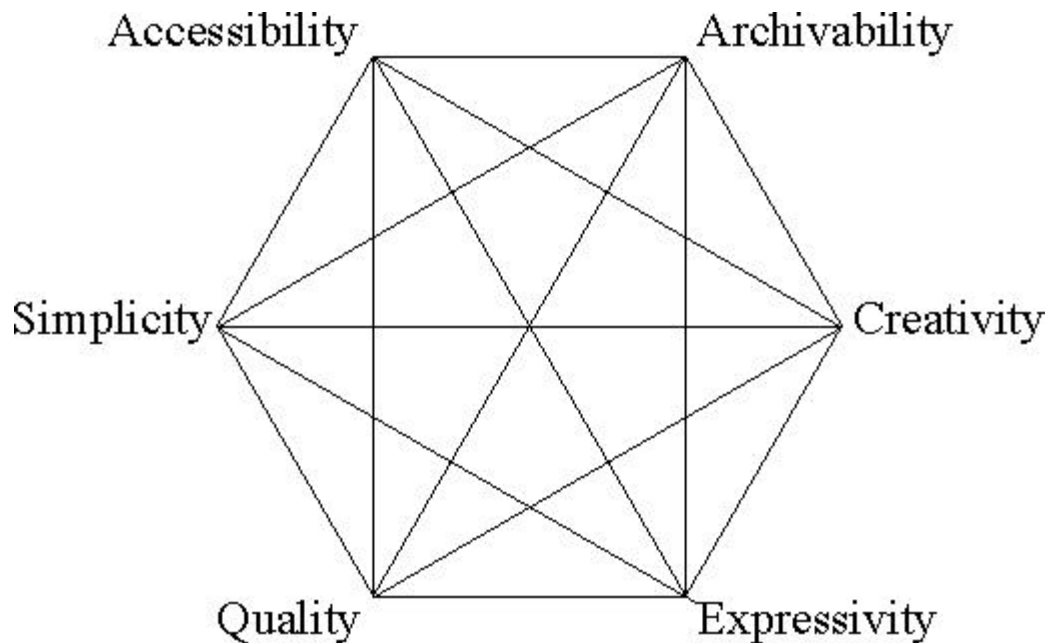
In 1996, even with an economic incentive (waiving \$20 archiving fee), only a fraction of those turning in theses or dissertations elected to do so electronically. Since project objectives are for students to learn, Virginia Tech officials agreed during the spring of 1996 to make submission a requirement, starting in 1997, in effect forcing students to learn what was deemed beneficial for them. Though diverse publicity and training efforts took place on campus to alert students to the policy and to help them prepare their ETDs, when deadlines for spring graduation came near in April 1997, we received many vocal complaints.

There appear to be several explanations for such concerns, e.g.,:

1. These changes clearly are requirements. The academic culture is such that new requirements are frequently subject to at times heated discussion before they are accepted.
2. If students had not planned to prepare an ETD, or did not get immediate assistance when faced with word processing problems, a requirement for electronic submission could lead to delays in completing their degree requirements, at a time when there are intense pressures (e.g., facing hard deadlines, planning to move, preparing for a job change).
3. Though Virginia Tech has a high degree of computer literacy among its student body and faculty, some are reluctant to learn new systems.
4. Since most students understand little about scholarly publishing and the complex laws related to copyright and publishers handling works submitted for publication, and since publisher and editor policies vary widely -- all reasons why NDLTD aims to increase knowledge in this arena -- students feel uneasy about any change that brings these matters to their attention.
5. Though most students had experience with WWW, many lacked experience with making their own works accessible, and some were concerned about potential negative effects of unknown readers studying their ETD.

These all arise naturally, because of the unique context, wherein NDLTD was established to change graduate education significantly and address many of these concerns. The timing is crucial, since nowhere else in a student's career can a requirement to achieve competency in preparing electronic documents be easily enforced.

Underlying these concerns are key issues regarding preparing theses and dissertations. First, writing a thesis or dissertation takes time and effort, usually more than was expected. Hence, anything that might increase the time required is very likely to be resisted. Second, many students are unsure about the impact of their works, or about what rewards they can expect from their effort, due to the complex system of credit given to people engaged in publishing. Finally, students are uncertain about the many tradeoffs and interconnections between aspects of electronic publishing, as shown in the following figure.



Quality results from time and effort that usually is prompted by hoped-for reward, such as impact on ones' scholarly community. That impact depends on a work being accessible, which is much more likely with ETDs than previously. Similarly, impact may increase if a student can more directly and simply express, using multimedia technology for example, the key ideas and message of their research. Creative expression thus may be facilitated through an electronic document. However, that may make it more difficult to archive the document, and extensive use of diverse multimedia representations may also reduce accessibility. Balancing these six aspects calls for more thought than most students, faculty, and librarians may have given to electronic publishing, but is a key to building digital libraries and is an important goal of the NDLTD.

Having students prepare electronic documents, even though based on sound pedagogical and career growth principles, also brings up a key issue which is at the heart of distinguishing digital libraries from the WWW. In the culture of the Internet, many vehemently argue for **free** information, regardless of the quality that results from such a policy. For example, in the arena of computer science technical reports, on the basis of experience with the WATERS and NCSTRL initiatives, few authors or departments are concerned about the correctness of bibliographic data that facilitates access, or the reliability of servers supporting searching. Some argue that fully automated systems, that gather data for searching without requiring work by authors or departments [WITT96], are adequate. Given such attitudes in the WWW culture, it is not surprising that students are unclear regarding how much time and effort they should invest.

3.3. Loyalties

Another underlying issue relating to ETDs is the diversity of opinion among students and faculty regarding loyalties. Why should a student support NDLTD, which aims to promote knowledge sharing and scholarship, and is endorsed by ones' university, when there are competing influences from ones' advisors, research group, discipline, and associations? Why should a student give copyright to a publisher and not retain rights to their intellectual property, e.g., that allow inclusion in their own thesis or dissertation as well as distribution of those important documents to interested scholars?

In some disciplines, students are further from center stage in research groups than others, and efforts to

give their work more attention as opposed to that of their advisors may meet with some resistance. That attitude may be reflected in the amount of time spent by advisors in reading, editing, and helping refine a thesis. It also may be reflected in differences between disciplines regarding if a thesis should be made largely of chapters very similar to published works, or if dissertations should be more book-like, telling an in-depth story of the research undertaken. In the humanities and social sciences, dissertations often are more like a book. In science and engineering there are closer ties to conference proceedings and journal articles.

The complex mix of loyalties relating to publishing of student works seems to be at the heart of concerns raised by faculty regarding NDLTD. While a reasonable solution to these concerns appears to be allowing students and their committee to discuss and agree upon access to each ETD, in the long term it is likely that the answer will depend upon:

- keeping citation counts to ETDs, so impact can be measured;
- evolution of the genre of ETDs, so their place in the publication mix is better understood relative to conference papers, journal articles, and books;
- discussions in various disciplines, and at numerous universities, of the credit that should be afforded to students and their works relative to that of their advisors and research group.

3.4. Economics

NDLTD relates to many issues with an economic basis. For example, during the first six months of the initiative, considerable attention was given to relationships with commercial efforts such as that of UMI (recall [Section 2.1](#) above). Only at the March 1997 NDLTD Steering Committee meeting was it decided that such matters were beyond the purview of the initiative, and that the focus should be education and on maximizing access.

Another basically economic issue is the relationship of ETDs to other forms of publication. If one assumes a zero sum game (which in the context of access to information through the Internet is probably not appropriate), giving more prominence to theses and dissertations might be viewed as threatening to other publication enterprises. On the other hand, theses and dissertations have been produced for over a hundred years, and have supplemented other types of publications without conflict, through a variety of changes in technology. The number who will read hundreds of pages about a topic as opposed to a short summary article is likely to be quite small. **It seems unlikely that NDLTD will have a negative financial impact on publishers.**

The approval form allows students and faculty to establish restrictions on access imposed by publishers, and those restrictions can be implemented using digital library technology [\[GLAD97\]](#). It would be beneficial to those in the scholarly community interested in ETDs to reduce such restrictions, however.

Compromises have been agreed upon, so that financial risk to publishers is minimized. In cases where an ETD or part thereof relates closely to an article, delaying worldwide access to the ETD for three months or even a year after the journal article is published is adequate protection. Similarly, in the case of a book that is published which is closely related to a dissertation, blocking outside access to the ETD from the time the book appears, till two years later, is more than adequate protection for publishers, but denies traditional access through interlibrary lending. In short, concerns of publishers, and related concerns of students and faculty regarding economic issues associated with access to ETDs on the Internet, can be resolved. While current solutions maintain an uneasy peace, they must be further negotiated so that economic concerns are addressed in coordination with access concerns for students, educators, and

researchers. Ultimately, digital libraries may need to evolve past their binary basis, where access is not either completely free to the world or severely restricted, where charges are not either zero or a very large sum, and where access to student research is not either solely through a publisher or solely through a university.

4. Future Growth

The future of the NDLTD is continued growth, as concerns are addressed, and benefits increase. We consider three key aspects.

4.1. Digital Libraries and Education

Fundamentally, NDLTD is an effort to improve education while building a digital library and expanding current library services and resources. While many education efforts have focused on how students can learn through accessing a digital library, NDLTD does not solely concentrate on that important issue. It also deals with how students learn by preparing an electronic document and submitting it to a digital library. Further, and key to solving various concerns raised, NDLTD strives to ensure that students are prepared to work with the world of publication.

As the collection of works related to NDLTD increases, log analysis and surveys will be used to determine how ETDs are used in graduate education. During the first year of widespread access to the Virginia Tech collection, the number of downloads per work appeared to be almost two orders of magnitude more than the number of circulations of the library copy. Additional factors to be analyzed include institution, topic, length, and use of multimedia relate to learning, and what measures prove informative: numbers of accesses, professions of those downloading copies, or types of use of ETDs.

4.2. Universities Working Together

True success of the NDLTD depends upon growth of a collection to the scale of hundreds of thousands of works, its ease of access, and the amount of use it gets. Widespread involvement of universities and their students will be determining factors.

In an era where there are increasing political and social pressures on universities to increase efficiency, be more open about their research findings, and share more with similar institutions, such collaboration seems appropriate. As universities see more need to archive their electronic works, and realize the economies of scale that result from cooperative ventures in the electronic publishing and archiving arenas, the type of initiative exemplified by NDLTD may become more commonplace. As has occurred in the context of state and regional library consortia (e.g., VIVA, OhioLink, CICNET), having a large market block to deal with publishers [\[NORR97\]](#) seems likely to motivate agreements, such as those over access policies. In the context of NDLTD, efforts in this direction are likely to expand, based on good experience in the Southeast, especially as interoperability tests proceed.

4.3. Evolution

Access to university information has evolved through various stages, leading to sophisticated programs for interlibrary loan and universally accessible library catalogs. As more use of the WWW takes place in colleges and universities, and as technology advances to better support URNs and electronic archives, it will be easier to move into the realm of fully functional digital libraries. Challenges still remain, regarding federated search, and multilingual access [\[BORG97\]](#). Efforts like NDLTD are likely to evolve

along with the technology, as universities aim to improve education and learn the benefits of collaborative initiatives.

5. References

[BORG96] Borgman, C.L.; Bates, M.J.; Cloonan, M.V.; Efthimiadis, E.N.; Gilliland-Swetland, A.; Kafai, Y.; Leazer, G.L.; Maddox, A. (1996). "Social Aspects Of Digital Libraries." *Final Report to the National Science Foundation*; Computer, Information Science, and Engineering Directorate; Division of Information, Robotics, and Intelligent Systems; Information Technology and Organizations Program. Award number 95-28808. <<http://www.gslis.ucla.edu/DL/>>

[BORG97] Christine L. Borgman (1997). "Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries: Or How Do We Exchange Data In 400 Languages?" *D-Lib Magazine*, June 1997. <<http://www.dlib.org/dlib/june97/06borgman.html>>

[FOX96a] Edward A. Fox, John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, and Scott Guyer (1996). "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources." *D-Lib Magazine*, September 1996. <<http://www.dlib.org/dlib/september96/theses/09fox.html>>

[FOX96b] Edward A. Fox (1996). "Statement About Publications." <<http://www.ndltd.org/info/pubs.htm>>

[FOX97a] Edward A. Fox (1997). "Letter to Virginia Tech Students Preparing an ETD." <<http://etd.vt.edu/submit/letter.htm>>

[GLAD97] Henry M. Gladney (1997). "Safeguarding Digital Library Contents and Users: Document Access Control." *D-Lib Magazine*, June 1997. <<http://www.dlib.org/dlib/june97/ibm/06gladney.html>>

[KIPP97] Neill A. Kipp (1997). "A scenario from the Networked Digital Library of Theses and Dissertations: The life of an ETD from creation to dissemination." <<http://www.ndltd.org/howto/etdlife.htm>>

[KIRS97] Matthew G. Kirschenbaum (1997). "Electronic theses and dissertations in the humanities: A directory of on-line references and resources." <<http://etext.lib.virginia.edu/ETD/ETD.html>>

[LYNC97] Clifford A. Lynch (1997). "The Z39.50 Information Retrieval Standard: Part I: A Strategic View of Its Past, Present and Future." *D-Lib Magazine*, April 1997. <<http://www.dlib.org/dlib/april97/04lynch.html>>

[MOOR97] Mariahna Moore (1997). "UVA SEAS Electronic Undergraduate Thesis Pilot." <http://univac.cs.virginia.edu:3066/SEAS_ETD.html>

[NDLTDa] NDLTD Team (1997). "NDLTD in the News." <<http://www.ndltd.org/news/>>

[NDLTDb] NDLTD Team (1997). "NDLTD Status of Universities." <<http://www.ndltd.org/join/status.htm>>

[NDLTDe] NDLTD Team (1997). "NDLTD Related Projects."

<<http://www.ndltd.org/projects/index.htm>>

[NDLTDc] NDLTD Team (1997). "Virginia Tech Graduate School Electronic Submission Approval Form." <<http://etd.vt.edu/submit/approval.htm>>

[NORR97] Bob Norris and Denise Duncan (1997). "Sink or Swim? The U.S. Navy Virtual Library (NVL)." *D-Lib Magazine*, March 1997. <<http://www.dlib.org/dlib/march97/navy/03norris.html>>

[PAYE97] Sandra D. Payette and Oya Y. Rieger (1997). "Z39.50: The User's Perspective." *D-Lib Magazine*, April 1997. <<http://www.dlib.org/dlib/april97/cornell/04payette.html>>

[WATE97] University of Waterloo Electronic Thesis Project Team (1997). "Terms of Reference and Team Members." <<http://www.lib.uwaterloo.ca/~uw-etpt/>>

[WITT96] Ian H. Witten, Sally Jo Cunningham, and Mark D. Apperley (1996). "The New Zealand Digital Library Project." *D-Lib Magazine*, November 1996.
<<http://www.dlib.org/dlib/november96/newzealand/11witten.html>>

6. Acknowledgments

The U.S. Department of Education's Fund for the Improvement of Post Secondary Education supports NDLTD. Additional in-kind has been provided by: Adobe, Arbortext, Council of Graduate Schools, Coalition for Networked Information, IBM, OCLC, SOLINET, and SURA.

Copyright © 1997 Edward A. Fox, John L. Eaton, Gail McMillan, Neill A. Kipp, Paul Mather, Tim McGonigle, William Schweiker, and Brian DeVane



hdl:cnri.dlib/september97-fox

People:

[Dan Atkins](#) University of Michigan Digital Library Project Director.

[Edward A. Fox](#) Director of the [Digital Libraries Research Group](#) at Virginia Tech.

[Hector Garcia-Molina](#)

- [Papers](#)

[Henry Gladney](#) [Peter Graham](#)

[Michael Lesk](#)

- [Images: Quantity is not always Quality - U. KY talk](#)
- [digital libraries](#)
- [library preservation](#)
- [information retrieval](#)
- [networking, etc.](#)
- [Projections for Making Money on the Web](#)

[Gary Marchionini](#)

- [U. Md. DL Home Page](#)
- [Encyclopedia article draft](#)
- [CACM April 1995 article](#) in [that year's volume online in ACM DL](#)

[Michael Mauldin](#) (Lycos, CMU)

[Bruce Schatz](#) University of Illinois at Urbana-Champaign, DLI Principal Investigator

[Marvin Sirbu](#)

- [publications available online](#)

[Terry Smith](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Digital Library Research Laboratory

840 University Boulevard, Suite 8
Mail Stop 0368
Virginia Tech
Blacksburg VA 24061

Table of Contents

- [1](#) Projects
 - [2](#) Prototypes
 - [3](#) Research Staff
 - [4](#) Books, Papers, Publications, Presentations
-

1. Projects

- [NDLTD](#) Networked Digital Library of Theses and Dissertations
 - [ETD](#) Electronic Thesis and Dissertation Initiative, Virginia Tech
 - [4S](#) Sets, Streams, Structures, and Scenarios (4S): Towards a Formal Model of Digital Libraries
-

2. Prototypes

- [SAUCER: Virginia Tech Speculative Fiction \(prototype\)](#)
-

3. Research Staff

- [Edward A. Fox, Director](#)
 - [Ghaleb Abdulla](#)
 - [Robert France](#)
 - [Tom Johnson](#)
 - [Neill A. Kipp](#)
 - [Binzhang Liu](#)
 - Paul Mather
-

4. Books, Papers, Publications, Presentations

- [Fox, et al](#)

Henry Gladney:

- Access Control Articles in D-Lib Magazine:
Gladney et al., Safeguarding Digital Library Contents and Users:
 - [Assuring Convenient Security and Data Quality](#),
 - [Document Access Control](#)
 - [Digital Images of Treasured Antiquities](#)
 - [A Note on Universal Unique Identifiers](#)
 - [Storing, Sending, Showing, and Honoring Usage Terms and Conditions](#)
- [Gladney et al. report on DL requirements and architecture \(PostScript\)](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[People\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

STORIES

D-Lib Magazine
May 1998

ISSN 1082-9873

Safeguarding Digital Library Contents and Users

Storing, Sending, Showing, and Honoring Usage Terms and Conditions

Henry M. Gladney and Jeff B. Lotspiech
IBM Almaden Research Center
San Jose, California 95120-6099
[\(gladney, lotspeich\)@almaden.ibm.com](mailto:(gladney, lotspeich)@almaden.ibm.com)

Abstract

This article knits together ideas and technologies discussed in several prior articles in the *Safeguarding ...* series in **D-Lib Magazine**.

We discuss languages for representing intellectual property usage terms and conditions in databases, for network transmission, and for presentation to and editing by human beings. Prototypes, one in each domain, can be knit together as a component of digital library services. We favor an approach based on cryptographic envelopment of document packets because this provides end-to-end protection and requires less network infrastructure and administration than alternatives. It needs protocols for enforcing information owners' rules -- protocols which govern how a user might select, request, possibly pay for, and eventually gain access to what she wants.

A deployed permission management and revenue collection mechanism will implement at least three system roles: a source **S** which encrypts and bundles valuable objects, an end user system **U** which manages requesting and receiving protected information, and a clearance center **C** which checks users' commitments to observe owners' conditions. We consider three alternative configurations.

We remind the readers why incomplete schemes based on much-ballyhooed "trusted systems" are fundamentally flawed, and suggest why it is unlikely that these notions will evolve to practical personal computer services. People might be less likely to be misled if this elusive objective were called "**trustworthy**"

systems".

Introduction

The *Safeguarding ...* series in **D-Lib Magazine** is intended to explore and illustrate technical contributions to mitigating intellectual property exposures which digital representations have raised. That technology can contribute only in a complex of administrative, legal, contractual, and social practices is well known; the current article is driven, more than any previous article in the series, to considering relationships among technical and other measures.

Until now, each article in our *Safeguarding ...* series has presented some narrow theme without connecting it carefully to other elements needed to realize "complete" digital library services. Articles in the series and elsewhere have discussed identifying *what* is to be protected [[Gladney 1998](#)]; how users might inspect and edit protection rules [[Walker 1998](#)]; how protection rules can record for future years the terms and conditions for each property [[Alrashid 1998](#)]; transmitting rule sets from where they are generated to where they are needed [[Ciccione 1996](#)]; efficient payment mechanisms [[Herzberg 1998](#)]; trustworthy identification of *who* is generating a rule set (authentication), providing a document, or requesting one; how properties can be bundled for distribution with first-rate protection against many different kinds of misuse [[Lotspiech 1997](#)]; and so on. The current article makes a start towards describing how these proposals could be combined.

Until now, each of our *Safeguarding ...* articles has also described work complete to at least a prototype and pilot implementation. The current article shifts from this retrospective approach to a prospective one, considering modest and feasible next steps.

We are forced to reconsider what might be a controversial issue, personal computers as so-called "trusted systems". We believe that what has been widely publicized under this rubric is not only a repetition of old work, but also impractical today for reasons similar to those which caused it to be abandoned 15 years ago.

In fact, one value of technological aids is to provide *mitigations for mistrust*, augmenting legal, contractual, and social pressures by making cheating difficult and forcing cheaters to take overt steps which remind them of property rights and create evidence of violations. There is another practical value to the work we are about to describe: although the terms and conditions for each work might be simple, the aggregated terms and conditions of millions of works held for decades -- often beyond the job tenures of the individuals who negotiated for each work -- constitute an administrative nightmare that digital storage, communication, and analysis go a long way towards relieving. What we describe are essential elements of larger complete solutions.

Languages Expressing Usage Terms and Conditions for Intellectual Property

We need to represent terms and conditions in at least three domains: on screens in a style that administrators and end users can edit, understand, and analyze with a minimum of prior training or "help" text; in databases made reliably durable for survival over decades and longer; and for transmission among heterogeneous computing systems, i.e., supporting "open" systems so that software consumers have the benefit of multiple technology sources.

In what follows, we use *language* somewhat more broadly than some readers may be accustomed to. *Information representation on a screen together with patterns of interaction, considered together with the interpretation of their meanings, is **language**. Database tables together with programs to interpret them and to map to/from other representations are **language expressions**.*

These languages could be different -- in fact, it is best to make them so. For example, the best storage representation is one that allows the administrative data to be reliably preserved for many years; this can be done at low cost only by holding the information in a database used by other applications -- we strongly favor relational database technology. In contrast, the transmission format must be linear, which could be simply a linearization of the database format. And finally, the external language should be whatever is best for human comprehension and convenience, with minimal compromise for easy programming. We have three candidates from three sources that began their work independently of each other.

The first comes from James Barker and his colleagues at Case Western Reserve University [[Alrashid](#)]; this work includes a database representation [[Barker1995](#)], defined as a set of relational database tables and their interpretation. The basic tables and their columns have names as shown in the following table. Although space does not permit a careful description of the language rules and interpretation, it is, in fact, simple enough for the reader to infer what can be expressed and for the developers to extend to anything needed for environments beyond those of the CWRU prototypes.

Schema of CWRU Permission Manager Base Tables

Table	Column Names
Billing	Licensor ID, User Email ID, Holding ID, Date of Use, Time of Use, Rule Identification Number Category of Work Code, Title of Work, Use Description, Charges Incurred, User ID, Last Name First Name, Middle Initial, Name Prefix, Name Suffix, Street Addr - Line1, Street Addr - Line2, City, State or Province Code, Zip Code, Country Code, Phone Number, FAX Number, Language Code
Element Permissions	Holding ID, Element ID, User Category Code, Use Type, Rule Type, Element Rate Type, User ID, Rule ID, Organization Category, Element Major Type, Element Minor Type, Internet Address Profile Code, Transmission Profile Code, Protection Profile Code, Processing Profile Code, Percent Excerpt Limit/Year, Percent Excerpt Limit/Term, Rule Begin Date, Rule End Date, Element Rate, Maximum Concurrent Users, Language Code
Elements	Holding ID, Element ID, Element Description, Element Major Type Code, Element Minor Type Code, First page # in element, Number of pages in elem, Disc # w/in disc set, Track # on the disc, Length of performance, Language Code
Holding Permissions	Holding ID, User Category Code, Use Type, Rule Type, Holding Rate Type, Rule ID, User ID, Organization Category, Element Major Type, Element Minor Type, Internet Address Profile Code, Transmission Profile Code, Protection Profile Code, Processing Profile Code, Percent Excerpt Limit/Year, Percent Excerpt Limit/Term, Rule Begin Date, Rule End Date, Holding Rate, Maximum Concurrent Users, Language Code
Holdings	Holding ID, License Agreement ID, Copyright Effective Date, Copyright Expiration Date, Title of Work, Category of Work Code, Number of Elements, Creator ID, Work Order ID, Language Code
Licence Agreement Permissions	License Agreement ID, User Category Code, Use Type, Rule Type, License Rate Type, Rule ID User ID, Organization Category, System Major Type, System Minor Type, Internet Address Profile Code, Transmission Profile Code, Protection Profile Code, Processing Profile Code, Percent Excerpt Limit/Year, Percent Excerpt Limit/Term, Rule Begin Date, Rule End Date, License Rate, Maximum Concurrent Users, Language Code
Licence Agreements	License Agreement ID, License Agreement Type, Licensor ID, License Description, Effective Date of License, Expiration Date of License, Copyright Notice, Language Code
Licensor Permissions	Licensor ID, User Category Code, Use Type, Rule Type, Licensor Rate Type, Rule ID, User ID, Organization Category, Licensor Major Type, Licensor Minor Type, Internet Address Profile Code, Transmission Profile Code, Protection Profile Code, Processing Profile Code, Rule Begin Date, Rule End Date, Licensor Rate, Language Code
Licensor Profiles	Licensor ID, Licensor User ID, Licensor Email ID, Licensor Organization Name, Licensor - Las Name, Licensor - First Name, Licensor - Mid Initial, Licensor Name Prefix, Licensor Name Suffix, Licensor Address - Street1, Licensor Address - Street2, Licensor Address - City, Licensor State/Province, Licensor Address - Zip Code, Licensor Country Code, Licensor Phone Number, Licensor FAX Number, Contact User ID, Contact Email ID, Contact Last Name, Contact First Name, Contact Mid Init, Contact Name Prefix, Contact Name Suffix, Contact Address - Street1, Contact Address - Street2, Contact Address - City, Contact State/Prov, Contact Address - Zip Code, Contact Country Code, Contact Phone Number, Contact FAX Number, Agent User ID, Agent Email ID, Agent Last Name, Agent First Name, Agent Mid Init, Contact Name Prefix, Contact Name Suffix, Agent Address - Street1, Agent Address - Street2, Agent Address - City, Agent State/Prov, Agent Address - Zip Code, Agent Country Code, Agent Phone Number, Agen FAX Number, Language Code
System Permissions	User Category Code, Use Type, Rule Type, System Rate Type, Holding ID, Element ID, Rule ID, User ID, Organization Category, Element Major Type, Element Minor Type, Internet Address Profile Code, Transmission Profile Code, Protection Profile Code, Processing Profile Code, Rule Begin Date, Rule End Date, System Rate, Language Code
User Profiles	User ID, User Email ID, User Last Name, User First Name, User Middle Initial, User Name Prefix, User Name Suffix, User Organization ID, User Address - Street1, User Address - Street2, User Address - City, User State/Province, User Address - Zip Code, User Country Code, User Phone, User FAX, Language Code

Not shown are a larger number of administrative, logging, and support tables. The support tables are key to what the CWRU RightsManager System allows; the values in columns of the basic tables shown are not restricted by software, but rather by administrators' entries in support tables; this permits tailoring to any installation's needs together with validity checking of permission table entries.

A second language, for transmission of the same rights management information, has been outlined by a team at the Xerox Corporation [[Stefik 1997a](#), [Stefik 1997b](#)]. This is Xerox's DPRL (Digital Property Rights Language) [[Ciccione 1996](#)]; the example immediately below hints at its origin in artificial intelligence work. We could use this linear language to carry terms and conditions from content repositories **S** to content users **U** as needed by the [network configurations discussed below](#).

(Work: (Description: "Title:'Fanciful' Author:'I.A. Fancy' Copyright:'I.A. Fancy'")

(Owner: "J Books, Inc.")

(Rights-Group: "Distributor" (Comment: "Rights limited to licensed distributors")

(Bundle:(Access:(Security-Class:5) (User-Authorization: "IDG Books Worldwide"))))

(Copy: (Access (Fee: (Ticket: "IDG Inventory 12345"))))

(Play:))

(Rights-Group: "Consumer" (Comment: "Rights for any purchaser")

(Bundle:(Access: (Security-Class:3)))

(Copy:(Next-Copy-Rights: (Delete:"Distributor)

(Fee:(Per-Use:10)(To:"Account IDG35"))))

(Play:(Fee:(Metered:(Rate: .09)(Per: 1:0:0)(To:"Account IDG36"))))

(Delete:(Comment:"This right is unrestricted"))

(Transfer:))))

A more recent alternative linear language is XML; we need to consider this because it seems about to become the popular choice for Web documents, and also because there is a [proposed W3C XML standard](#).

A third language provides our preferred human interface; also coming from an artificial intelligence tradition; it is Adrian Walker's Internet Knowledge Manager (IKM) [[Walker 1998](#)]. It is the best candidate we know that:

1. allows human beings to understand and write rules
2. that can interface transparently with relational databases; and
3. permits people to ask not only what the permissions and prices of access to a holding are, but also to inquire what instances of general policy rules were used to find the answer. For example, this is of interest when pricing is dependent on factors such as organizational affiliation and prior

purchases.

A [Web-friendly IKM implementation](#) is freely available for readers' inspection and experiments.

In a session with the IKM, one uses an ordinary Web browser to write agents, and also to run them. In doing this, one can make use of a library of agents that have already been written, for business subjects such as insurance, international transfer pricing, and so on. Here is an outline of an example in which a distributor gets a discount from a publisher, based on the volume of sales of multilayer documents. We first write a table saying how a document is made up of other documents.

I	Concept:	a document has an immediate component document
	Items:	document component
		Web Encyclopedia All About Ants
		Web Encyclopedia Bugs You Should Know
		All About Ants Ant Farm Video

Figure 1: A Simple IKM Table of Documents and Their Components

In the table, the Web Encyclopedia has a component that in turn has a subcomponent. To collect all its components we write a general rule like this.

I	IF
	Concept: a document has an immediate component document
	Items: document component
	Concept: a document has a sub component document
	Items: component sub component
	THEN
	Concept: a document has a sub component document
	Items: document sub component

Figure 2: An IKM Rule about the Subcomponents of a Document

After writing some more tables and rules, we can ask what discount a trader called NetVidStore got in 1997. Answer is a table like this.

	Concept:	a trader rates a maximum % discount from a supplier in a year			
	Items:	trader	maximum % discount	supplier	year
		NetVidStore	20	Random House	1997

Figure 3: An IKM Answer Table from the Discount Agent

Even in a simple example like this, it's good to be able to see the reasons for an answer. The

IKM provides an overview like this.

[SINCE			
Concept:	a trader has bought documents worth a grand total amount from a supplier		
Items:	NetVidStore	188980	Random Hou
Concept:	a supplier gives a % discount for more than an amount in a year		
Items:	Random House	20	100000
Concept:	an item is greater than another item		
Items:	188980	100000	
WE HAVE			
Concept:	a trader rates (at least) a % discount from a supplier in a year		
Items:	NetVidStore	20	Random House 1997

Figure 4: The Main Reasons for an IKM Answer

and we can drill down into more detailed reasons if we so wish.

You are invited to [look at the full example, called Market-1, and also to run it](#)

The authors of these prototypes have examined each other's work sufficiently to be confident that the needed translators will be easy to build. Why have we not already built them? Although some big publishers have vigorously urged the need for tools to store, audit, and manage their contracts with authors, photographers, and other original sources and similar relationships with their customers, none of these publishers has yet been ready to deploy a pilot to scale. We are leery of building something without a committed user community, because software built on speculation so often misses the mark.

Trust Management Involves at Least Three Administrative Domains

We know of three practical network configurations for automated distribution of valuable intellectual property: (1) publishers' repositories delivering under contract to libraries which provide access to limited communities; (2) publishers delivering massive encrypted content to potential end users who negotiate with clearance centers for access for selected small subsets of the content; and (3) a variant of the second scenario in which the user's workstation can view, at most, incomplete works locally, but render full works on protected terminals or printers.

The first arrangement was explored by IBM and ISI (Institute for Scientific Information, Philadelphia) in a dozen customer pilot installations [[Choy 1996](#)]. Scientific, engineering, and medical periodicals are mostly sold in subscriptions to libraries, which provide access primarily to limited communities (e.g., the members of a university), but extend limited access to larger communities (e.g., anyone who goes to the library building). We discovered that publishers cautiously accept the digital distribution layout depicted in Figure 5, with the library delivering limited and tracked amounts of in-the-clear content to end user workstations. Its attractions include immense performance improvements for large user groups distant from their libraries, easy protection of the

anonymity of individual readers, single points of authentication for the thousands of users of each of many (university) libraries, and emulation of the common practice of institutional subscriptions to periodicals.

In this network layout, the manager of the publisher's or distributor's repository, Simon Supplier, delivers to Linda Librarian, or makes accessible for rapid download, all volumes of each subscribed periodical. Linda publishes her library catalog and enables every Ulrich User in her community for download of individual pages or individual articles. Simon, Linda, and Ulrich each accept this scenario, without necessarily being delighted, because it enables workable compromises: Simon and Linda each want the budgeting predictability of annual subscriptions; Simon can shift the responsibility of limiting access to a contractually-defined community to Linda, who is incented to comply both because universities intend to be honorable and because she does not want to risk loss of license to the materials; Linda further gets the ability to protect her readers' privacy; and Ulrich gets access to what he needs. Because of the nature of the material (each individual article is of interest only to a small number of scholars), and because violations can readily be detected and traced, Simon is not greatly worried by the possibility that Ulrich will violate "fair use" by wide distribution of licensed content.

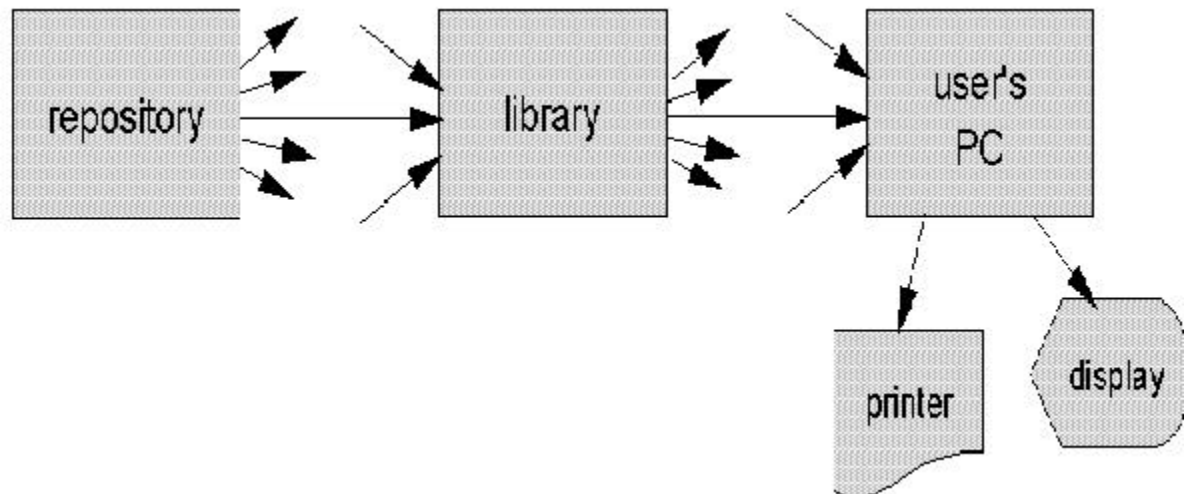


Figure 5: Library redistribution of publisher's content

Of course publishers also want to distribute both subscriptions and individual content elements directly to end users for money or other considerations. Figure 6 suggests how this is enabled by our previously described Cryptolope (TM) technology [Lotspiech 1997]. Simon packages each attractive set of materials as a set of files, each encrypted under a different key; he further includes descriptive and promotional material in the clear, a statement of terms and conditions both in the clear and encrypted, a bill of materials, and an encrypted file of the prior encryption keys. The master key for this data set of individual document keys is either the public encryption key of a clearance center (Simon would need to provide such a key file for each potential clearance center) or a secret shared with clearance centers by an independent channel.

Ulrich decides from the promotional material and the clear-text terms and conditions which portions of a package he wants to buy, and sends to a clearance center this information together with the encrypted terms and conditions, the encryption key files, and whatever information about himself will be needed to check authorization, doing so under the public key of the clearance center. The clearance center checks whether what the information owner demands is satisfied, forwards bookkeeping entries as needed, and returns to Ulrich the encryption keys of the sections he has purchased, doing so under Ulrich's public key.

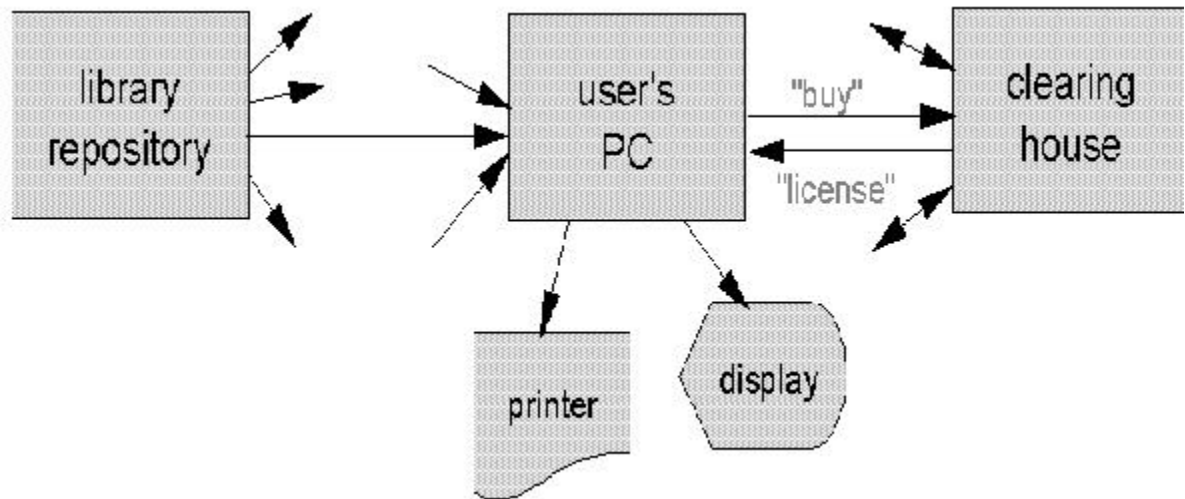


Figure 6: Cryptolope delivery with delayed purchase by information consumer

A third layout, made feasible by the advent of printers with sophisticated embedded computers (and in the future, other presentation devices), might be attractive to large libraries (Figure 7). It is made practical by the willingness of content providers to enter trust relationships with university and public libraries, which would manage the printers in controlled environments (e.g., behind the counters of a circulation desk).

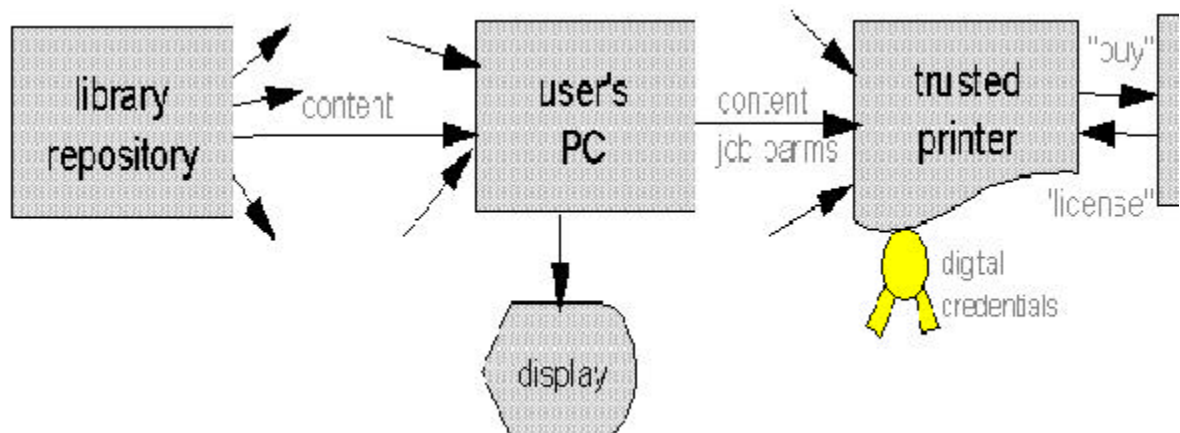


Figure 7: Copying limited by a Trustworthy Intelligent Print Server
(in this layout, the clearing house could be packaged as part of the printer)

It will often be inconvenient to package very large materials (e.g., feature-length movies, large scientific files) or sessions of indeterminate length (e.g., interactive consulting services mediated by digital communications) within Cryptolopes. It is, in fact, sufficient to send the administrative information in a Cryptolope, together with whatever addressing and other information Ulrich would need to access the large objects directly. After the administrative checks are made for such a session, the networked systems can choose and set up the most efficient channel between the repository and the printing or display device. This would work in any of the layouts shown.

The above schemes require each clearinghouse to receive secret information from each repository whose offerings it will mediate. No confidential delivery channels are shown because public key cryptography can hide secrets in the depicted channels.

What makes each layout shown attractive is the willingness of content providers (publishers, movie studios, various kinds of distributors) and institutional libraries to enter into predictably-priced agreements to supply content from each information resource to large numbers of end users. Such agreements are usually explicitly or implicitly contractual, with defined penalties for failures to perform. Similar agreements directly between each of thousands of providers with each of millions of recipients would be impractical and often inconvenient. Note that each network configuration requires at least three processes: an information source, a personal environment, and a clearing house.

Convergence of Access Control and Permission Management

The reader who compares the CWRU permission management database schema above schema in access control subsystems [[Gladney 1997](#)] will see similarities. These suggest that access control and permissions management might be made to grow together. As access control is enriched by finer granularity for object-oriented programming, delegation of privileges for office environments, more flexible grouping of users and privileges based on organizational affiliation rather than directly on user identification, and privileges sensitive to environmental circumstance (e.g., time of day, funds availability), the similarity of the supporting databases will increase. We intend to investigate whether this convergence is as attractive as it superficially seems.

Trusted Systems? Trustworthy Services Belong in Glass Houses

We encounter colleagues who are not directly involved in this kind of research who project unrealistic expectations onto the technology. At least some of the publicly available literature (see, for example, [[Stefik 1997a](#), [Stefik 1997b](#)]) seems to fuel these fancies -- although the unpublished white papers and

technical discussions may have moved beyond the information contained in the publicly accessible record. Nevertheless, the published material is unfortunately incomplete on critical points: what trust is to be held by whom in whom else, what attributes a system must have to be trustworthy, and what technical means can realize such attributes reliably. Since the published writings are incomplete on these points, we must infer what we can from the available articles in [Scientific American](#) and the [Berkeley Law Review](#) and then relate that to the best available prior work, [\[Weingart 1987\]](#) and [\[White 1987\]](#), as well as some 1998 IBM work which we learned about just as the final drafts of this article were being edited.

The simple inferences to be drawn from this research is that notion of trust is the same as what we understand from common usage in natural language. Specifically, what is called for is:

- that what is to be trusted is that valuable documents are printed only if specifically stated conditions, including payments required, have been satisfied
- that copying is limited to bounds which may demand destruction of the copy instances in one system when copies are forwarded to another system; and
- that technical means to enforce such compliance are known.

Further, that personal computers and workstations will achieve such means of enforcement is implied by examples in pictures and text in the **Scientific American** article [\[Stefik 1997b\]](#) from which much of our understanding of the "trusted systems" approach is based (absent other sources of information). We argue below that little, if any, of this is practical, because:

- trust is limited to human beings, but may be extended to include corporate entities;
- trust can be extended to inanimate systems only indirectly and cautiously; and
- such systems do not include personal computers or workstations today and will not, we believe, soon do so.

The notion of trust antedates any digital system. Is there is a reasonable extension to inanimate systems? Human trust relationships come into existence only when one individual knows another sufficiently to be confident that some limited responsibility will be faithfully discharged (e.g., that my neighbor's daughter, whom I have known for 5 years, is trustworthy as a baby sitter) or that some defined risks will be avoided (e.g., that my son is a skilled and careful driver who will avoid damage to my new Volkswagen). In most cases of trust conferred, there are adverse consequences to a breach of trust; these may be explicit but are often implicit (if the babysitter ignores my wailing child, I will

probably not employ her again, and I tell my son that if he misuses the automobile, I will not lend it to him again). Trust relationships are extended from individuals to corporate entities by agreements made with human agents of those corporations and tend to be more explicit both in their scopes and their damage commitments than are those between individuals, i.e., explicit or implicit contracts are frequent when the trusted entity is a corporation.

We know no sensible and economic way to extend such notions to computing or communications systems, unless "system" is construed to include the human beings who manage the machinery -- a construction which is neither conventional nor voiced by the "trusted system" articles already cited. Attempts were made 15 years ago to enforce outside rules with workstation components [White 1987]. Their authors decided such efforts impractical then, and continue today to stand by their conclusions; we'll summarize the specifics below, and direct the reader to some continued consideration of the conundrum.

One problem is that we must make it possible, in advance, to know that a remote target machine managed by someone else truly satisfies certain attributes, e.g., that it contains a certain kind of security coprocessor with appropriate installed software, and that these have not been modified or bypassed by known or unknown people. Parts of this problem have only recently been addressed by Smith and colleagues [Smith 1998], who discuss the following scenario:

Suppose Sam develops and sells some rights management software for our secure platform, and Alice and Bob are (distributed) participants. If Alice trusts:

- that public key crypto works
- that IBM builds and certifies only bona fide devices
- that the certification Alice has in hand truly comes from IBM
- that Sam's software behaves as Sam alleges

then she can always distinguish between

- a message from Sam's program, running on an untampered device at Bob's site
- and a message from a clever adversary

even if

- Alice, Bob, and Sam have never met
- the adversary might be using Sam's software on a tampered device, or other software on an untampered device
- there are no "trusted couriers" or "trusted security officers" anywhere.

The Cryptolope-exploiting network layouts we suggest above are made practical by agreements between content providers and enterprises which run library centers, protected printers, or clearance centers. The number of such enterprises will be 100- to 10,000-fold smaller than the number of individual users; each enterprise will be motivated strongly to honor commitments; and each also has the ability to purchase and manage relatively sophisticated machines in "glass house" environments. Such characteristics are unlikely for individuals, as is evidenced by the very large numbers of software copies installed out of licence. Like the software industry, intellectual content providers will make some offerings available in the clear to end users on their own workstations, but will do so not out of trust but rather based on market estimates which take into account massive unlicensed use.

Returning to the possibility of improving the situation by some device built into personal computers to enforce the conditions specified by content providers, we note that a device would have to include a hardware component, because software is readily bypassed or substituted. A practical fact is that this particular horse has long ago escaped the barn. 100,000,000 personal computer owners will neither pay for a hardware addition that inhibits freedoms they currently have nor permit enforced installation of new devices. They also will not give up print redirection which currently permits them to capture the unencrypted form of any file at all for any use they subsequently choose.

Even if this horse had not already escaped, it would have been impractical to build a strong barn. This is what Weingart and White attempted; they found that the PC builders were unwilling to install any device which increased their manufacturing cost if the device did not benefit the immediate customer. This is because Ulrich User will refuse, as a matter of principle, any unlegislated taxation intended to benefit Simon Source. The only possibility for "trust technology" is that it is legislated, as has been attempted from time to time for music reproduction devices, or agreed to privately by commercial enterprises. For example, as part of deploying DVD technology, a consortium of movie studios, consumer electronic companies, and computer hardware companies are trying to enforce constraints by licensing device producers.

Specifically, about 15 years ago, an IBM research team [[White 1987](#), [Weingart 1987](#)] designed a personal computer security coprocessor, called ABYSS, and an ABYSS operating system with security kernel primitive operations which could enforce constraints called for by permissions languages. They packaged it to resist code substitution and other tampering. Although the 1985 projected incremental cost of ABYSS enablement was under \$10, the IBM Personal Computer product groups refused to include the technology because it would increase prices for PC purchasers without increasing their direct benefits.

The idea of coprocessors whose owners were constrained in clearly articulated and certified ways from certain explicitly defined changes [[Yee 1994](#)] has not been abandoned in the IBM T.J. Watson Research Center. Some ABYSS design ideas have evolved into more expensive security processors for "glass house" systems, such as the IBM 4758 Crypto Controller(TM), an application

which makes sense because managers of corporate computing servers have economic incentives (contract obligations, loss of essential licensing, reputation for integrity) to enforce access control policies. The embedded processor is in fact a much more capable engine than ABYSS, being an Intel 486 (TM) with 2 Mbyte of storage, of which a small portion is protected against unprotected change. Just in case a way can be found to motivate personal computer users to purchase and install security coprocessors (e.g., by persuasively cost-saving applications made available only by way of PC's with such hardware installed), work continues to define and harden such machinery. Being pursued are engineering modifications and repackaging of evolutions of what went into the IBM 4758; even assuming that persuasive applications are found, the cost seems to us more than an order of magnitude too high in the current embodiments whose protection is good and whose power is sufficient for the kinds of applications conjectured by Smith and colleagues [Smith 1998] (commercial applications, rather than information distribution applications).

We re-emphasize that we cannot make a contract with a machine -- a contract in which the machine undertakes to execute or to avoid executing each of a list of carefully described actions.

The questions of trust debated above were the topic of a panel and public discussion in the May 1998 IEEE Symposium on Security and Privacy. Our perception is that both the panelists and the 300 people in the room agreed that trust is a human attribute without any direct machine analogue. Even if we construe "system" to include the individual or organization that ensures that the machinery complies with agreed-upon rules, the notion of "trusted system" is weak because the phrase implies something about the attitude of the people giving trust. For these reasons, we would prefer at most to consider the feasibility of practical "trustworthy services" in which legal or contractual constraints figure as part of enforcement of explicitly stated rules and limitations.

Returning to the DVD case, the above chicken-and-egg deadlock can be addressed if the market is being created from scratch. DVD players are a case in point. Each player (or "movie-compatible" PC) contains some logic to make it *difficult, but not impossible*, for the casual consumer to copy a DVD movie. The manufacturers are not legally compelled to put this logic in their boxes. Instead, the movies are scrambled, and to learn the scrambling secret, the manufacturers sign a licence by which they are contractually bound to certain restrictions. (Of course, they sign the licence, because a DVD player that cannot play Hollywood movies would have a negligible market.) Although the situation is still somewhat murky, it is possible that the DVD descrambling licence may become the initial domino by which other copy protection schemes become deployed: for example, protection on the "Firewire" (consumer digital video connection) and copy watermark detectors in recorders and players. The dominos may fall as follows:

1. DVD players are required by descrambling licence to put copy protection on the Firewire bus.

2. Digital TVs and VCRs will want to connect to DVD players and will need to get the licence for the Firewire copy protection scheme.
3. All licences will require that watermark detection logic exist to help find and block illegal copies.

The concern above is for situations in which content recipients have little motivation for observing constraints wanted by content owners. For this situation, the DVD example illustrates what might be possible when there is the luxury to design the system from scratch -- as might happen if the PC technology ground rules suddenly change. However, in today's PC arena, we view the idea of "trusted systems" with hearty skepticism. For situations in which content recipients share objectives with content owners, as is discussed for limiting children's access to pornographic materials, PC-enforcement is plausible. Blaze [[Blaze 1997](#)] discusses mechanisms, trust models, and an implementation for such cases.

Sandbox Protection for End Users

The sections above are mostly concerned with protecting the interests of copyright holders. What about the interests of personal computer users? Solutions based on "trusted systems" create risks for them also, if such solutions require their machines to execute programs written by content providers or their agents. It is not easy to protect end users from such risks. This is because a marketplace with thousands of providers and millions of consumers would probably be administratively efficient only if each producer loaded into each *actual* consumer's machine the software required to interpret and enforce the kinds of terms and conditions suggested by the language examples above. If this is done, for user safety, the personal computers must somehow fence in the execution of the imported code so that it cannot capture control of the entire machine. Such a fencing is sometimes called a "sandbox" [[Anderson 1972](#)].

We'll discuss the feasibility of sandbox architecture after disposing of the only alternative we know of, trustworthy security kernels. The hope is similar to that suggested by "trusted systems", except that the sought-for protection would be for the personal computer user rather than for the content provider. Presuming that a technical solution could be devised, this would have to be tested and demonstrated in a sufficiently public way or with sufficient promises of indemnification to consumers suffering invasions with breach of the solution. Without such measures, rational consumers would not buy and install the technology. Assuming that all this could be accomplished, the technology would have to be deployed into an economically significant fraction of the installed or newly installing personal computer population. We understand the infrastructure and delays such certification would require; the model is the testing required for various levels of computer security certification by the U.S. Department of Defense; it is expensive and introduces a delay of several years. Such challenges are so high that no manufacturer is following this course for

commercial personal computers, or even considering it as far as we know. We believe combined hardware/software protection can make economic sense only for entirely new business segments, as illustrated by the above discussion of DVD.

Since such trustworthy security kernels seem impractical, significant effort is currently being expended on various software-only "sandbox" possibilities. Currently, attention is focused on Java(TM) as a cross-platform program transmission vehicle. That accepting programs from unknown and unpredictable sources is very risky is illustrated by work which has recently identified a Java security exposure -- a way to use the subroutine return protocol [Malkhi 1998] -- and is proposing a change to Java virtual machines to close this loophole. (It is in the nature of such loopholes that a malicious application program can capture control of the computer and can do whatever damage it's author wants, without the machine owner being aware that it is happening until it is too late.) I.e., sandbox protection for personal computer users is being looked into, but it is a difficult challenge. Of course, it can be solved with computer operating systems similar to those on "big iron", but this is not a current prospect.

Questions of Public Policy and Law

The technical topics that are the focus of this paper lead directly to open questions of legal interpretation and policy -- questions that are being carefully considered in public discussions, in other articles, by legal, political, and economics scholars, and in some cases by legislative committees. We feel impelled to mention some of these topics, but will limit this to suggesting issues in which technical considerations intersect broader domains.

One such is the meaning of "copy" in interpreting current copyright law and suggestions how current law might need to be refined or extended to cope with the pliability of digital representations and derivative works. Stefik [Stefik] touches on some aspects, calling for enforced rules about how many copies of a work a licensee can create, but does not settle a more fundamental concern. It is not clear whether or not the copyright law sees as copies the several instances computers make today in order to make any work accessible to a single reader. (Current U.S. copyright law defines a "copy" to be a physical or material object; this definition essentially appears in the Berne convention; some people believe that this definition has been superceded by practice.) A possible remedy is to define different forms of "copy" and to formulate rules distinguishing among these kinds, e.g., a cached copy would be different from a screen image copy, and both would be different from a print copy. (This distinction was suggested to us by Professor Pamela Samuelson of University of California, Berkeley, but has surely occurred to many people.)

A controversial topic is inherent in packaging intellectual property cryptographically. In the context of a recent public panel debate, one participant took strong exception to the notion, on the grounds that it will deny "fair use rights of access" to scholars. Although we are sympathetic to his motivation,

the political value of open information, we are also skeptical of his case in current law. As we understand the U.S. copyright law, "fair use" is an effective defense against an action claiming copyright violation. However, "fair use" in no way compels any owner of content to make it available to anyone, or to make it as available to one person as he has made it to another. We hasten to say that our point here is not to argue one side or another of this important question, but rather to illustrate how intimately the technology is intertwined with difficult questions of law and public policy.

This last question may lure some civil liberties extremists into an absurdity. We would not be surprised to see some lobby simultaneously insisting that governments should not limit private use of cryptography, as some police and defense lobbies propose, and also that intellectual property owners should not be permitted to use encryption to deny access to their holdings. Again, our purpose here is not to suggest what makes sense, but to illustrate that we are faced with policy choices -- choices that will probably be settled differently in different jurisdictions.

Another controversial topic is the doctrine of "first sale", which holds that once a publisher has sold a copy of a work, the current owner of that copy can lend or give it to anyone else without permission or further payment to the publisher. Although DPRL provides language to express transfer, no reliable implementation has been built. Publishers are understandably reluctant to agree that the notion of "first sale" of physical copies has a digital equivalent.

Such questions are important enough, urgent enough, and difficult enough that the U.S. National Science Foundation has commissioned a U.S. National Research Council-managed [Study Committee for Intellectual Property Rights in the Emerging Information Infrastructure](#). [Individual members of this committee](#) would like to hear carefully considered opinion on any topic within the committee scope.

Conclusions

Our objective has been to show likely direction in which previously discussed intellectual property protection technologies will be knit into complete solutions. Among other things, languages and their interpreters are needed to express, record, and administer whatever rules are chosen. For situations in which information providers and information users have conflicting economic motivations, **practical enforcement scenarios require three (or more) processing environments for every transaction**: a content originator's, an end user's, and a clearance center which could be folded back into the content originator's environment in some situations.

In contrast, we argue that it is **not reasonable to expect to use personal computers to enforce content providers' interests as so-called "trusted systems"**. Further, we believe it misleads the public to refer even to clearance centers as "trusted systems"; to convey what useful function such machines and their human managers can provide, it would be better to call them

"trustworthy services".

Processes and databases to record the rules for managing intellectual property and access control databases can be made to be similar. We believe these similarities will offer simplifications for both users and software providers. IBM work on a human-intelligible rules language with easy Web and database interfaces [Walker 1998] seems to us ready to deploy. What the transport language for rules should be is an open question; the momentum in the near future favors DTDs and interpreters for XML rule expressions.

We find it impossible to discuss rights management technology without encountering unsettled questions of policy. We have identified a few of these that are intertwined with network delivery of protected information; it will be possible to continue the technical development to accommodate some likely policy choices, but others will be beyond practical technical measures. It will be **important for the technical and legal community to communicate to policy makers which policies can and cannot be effectively administered by digital computers**, and which seemingly distinct policy objectives are, in fact, incompatible.

Acknowledgements

This article was made possible by conversations with many colleagues -- Jim Barker, John Hurley, Paul Karger, Steven Newell, Sean Smith, Adrian Walker, Steve White, and others -- who shared their deep understanding of the field, pointed us at the seminal works, and critiqued drafts of the article. We are also indebted for access to unpublished materials to Jim Barker and his CWRU team, to Mark Stefik and his Xerox colleagues, and to Adrian Walker.

Bibliography

[Alrashid 1998] Tareq M. Alrashid, James A. Barker, Brian S. Christian, Steven C. Cox, Michael W. Rabne, Elizabeth A. Slotta, and Luella R. Upthegrove, [Safeguarding Copyrighted Contents: Digital Libraries and Intellectual Property Management](#), D-Lib Magazine, (April 1998).

[Anderson 1972] Investigating the sandbox approach was first suggested in James P. Anderson, [Computer Security Technology](#), ESD-TR-73-51, Vol. II, pp. 58-69, (Oct. 1972) (HQ Electronic Systems Division, Hanscom Field, Bedford, MA).

[Barker 1995] J. Barker et al., [RightsManager System: Permissions Manager Subsystem \(Version 2 draft\)](#), from Library Collections Services, Case Western Reserve University, (July 1995).

[Blaze 1997] M. Blaze, J. Feigenbaum, P. Resnick, and M. Straus, [Managing Trust in an Information-Labeling System](#), European Transactions on Telecommunications 8(5), 491-501, (September 1997).

[Choy 1996] D.M. Choy, J.B. Lotspiech, L.C. Anderson, S.K. Boyer, R. Dievendorff, C. Dwork, T.D. Griffin, B.A. Hoenig, M.K. Jackson, W. Kaka, J.M. McCrossin, A.M. Miller, R.J.T. Morris, and N.J. Pass, [A Digital Library System for Periodicals Distribution](#), in Proceedings of ADL96 - A Forum on Research & Technology, Advances in Digital Libraries, IEEE Computer Society Press, Los Alamitos, CA, pp. 95-103, (1996).

[Ciccione 1996] B. Ciccione, K. Duong, S. Okamoto, P. Ram, X. Riley, and M. Stefik, The Digital Property Rights Language, private communication, (1996). [Stefik includes a language sample.](#)

[Gladney 1997] H.M. Gladney, [Safeguarding Digital Library Contents and Users: Document Access Control](#), D-Lib Magazine, (June 1997). This is a synopsis of [Access Control for Large Collections](#) (ACM Trans. Info. Sys. 15(2), 154-194, (1997)), which shows the database schema alluded to.

[Gladney 1998] H.M. Gladney, [Safeguarding Digital Library Contents and Users: a Note on Universal Unique Identifiers](#), D-Lib Magazine, (April 1998).

[Herzberg 1998] A. Herzberg, [Charging for Online Content](#), D-Lib Magazine, (January 1998).

[Lotspiech 1997] J.B. Lotspiech, U. Kohl, and M.A. Kaplan, [Safeguarding Digital Library Contents: and Users: Protecting Documents Rather Than Channels](#), D-Lib Magazine, (September 1997).

[Malkhi 1998] D. Malkhi, M.K. Reiter, and A.D. Rubin, [Secure Execution of Java Applets using a Remote Playground](#), 1998 IEEE Symposium on Security and Privacy, 40-51, (May 1998).

[Smith 1998] S.W. Smith, E.R. Palmer, S.H. Weingart, [Using a High-Performance, Programmable Secure Coprocessor](#), FC98: Proceedings of the Second International Conference on Financial Cryptography. Anguilla, BWI, Springer-Verlag LNCS, 1998 (to appear). S.W. Smith, S.H. Weingart, [Building a High-Performance, Programmable Secure Coprocessor](#), IBM Research Report RC21102, (1997).

[Stefik 1997a] M. Stefik, [Shifting the Possible: How digital property rights challenge us to rethink digital publishing](#), Berkeley Technology Law Journal, 12(1), 137-159, (1997).

[Stefik 1997b] M. Stefik, [Trusted Systems](#), Scientific American 276(3), 78-81, (1997).

[Walker 1998] A. Walker, [The Internet Knowledge Manager: Dynamic Digital Libraries, and Agents You Can Understand](#), D-Lib Magazine, (March 1998). The Internet Knowledge Manager, and its Use for Rights and Billing in Digital Libraries. Proc First International [Conference on the Practical Applications of Knowledge Management](#), March 1998.

[Weingart 1987] S.H. Weingart, [Physical Security for the microABYSS System](#), Proceedings of the 1987 IEEE Symposium on Security and Privacy, Oakland, CA, pp. 52-58, (April 1987).

[White 1987] S.R. White and L. Comerford, [ABYSS: A Trusted Architecture for Software Protection](#), Proceedings of the 1987 IEEE Symposium on Security and Privacy, Oakland, CA, pp. 38-51, (April 1987).

[Yee 1994] B.S. Yee, [Using Secure Coprocessors](#). Ph.D. dissertation, Carnegie Mellon University, Department of Computer Science (1994).

Copyright and Disclaimer Notice

Copyright IBM Corp. 1998. All Rights Reserved. Copies may be printed and distributed, provided that no changes are made to the content, that the entire document including the attribution header and this copyright notice is printed or distributed, and that this is done free of charge. We have written for the usual reasons of scholarly communication. Wherever this report alludes to technologies in early phases of definition and development, the information it provides is strictly on an as-is basis, without express or implied warranty of any kind, and without express or implied commitment to implement anything described or alluded to or provide any product or service. Use of the information in this report is at the reader's own risk. Intellectual property management is fraught with policy, legal, and economic issues. Nothing in this report should be construed as an adoption by IBM of any policy position or recommendation.

The opinions expressed are those of the authors, and should not be construed to represent or predict any IBM position or commitment.

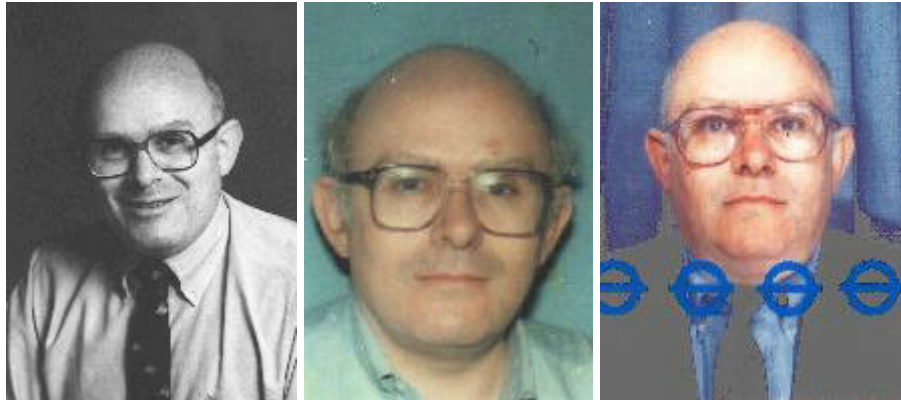
[Top](#) | [Magazine](#)
[Search](#) | [Author Index](#) | [Title Index](#) | [Monthly Issues](#)
[Previous Story](#) | [Next Story](#)
[Comments](#) | [E-mail the Editor](#)

hdl:cnri.dlib/may98-gladney

Michael Lesk's Grade Crossing on the Information Superhighway

Please change any address/link to this page to <http://www.purl.net/NET/lesk>. The address 'purl.net' refers to 'permanent URL' and this address should survive local administrative changes. Thank you.

This page is also available from a [site in the United Kingdom](#).



Professional

Amateur

Coin-operated

Now out: my new book *Practical Digital Libraries: Books, Bytes and Bucks*, [Morgan Kaufmann](#), July 1997.

Position: Division Director, Information and Intelligent Systems, National Science Foundation, <http://www.cise.nsf.gov/iis>.

Also: Visiting Professor, University College London, Department of Computer Science.

Biography

In the 1960's I worked for the SMART project, wrote much of their retrieval code and did many of the retrieval experiments, as well as obtaining a PhD in Chemical Physics. In the 1970's I worked in the group that built Unix and I wrote Unix tools for word processing (*tbl*, *refer*), compiling (*lex*), and networking (*uucp*). In the 1980's I worked on specific information systems applications, mostly with geography (a system for driving directions) and dictionaries (a system for disambiguating words in context), as well as running a research group at Bellcore. And in the 1990s I have worked on a large chemical information system, the CORE project, with Cornell, OCLC, ACS and CAS.

I am also Visiting Professor in computer science at University College London; I'm on the Visiting Committee for the Harvard University Library; and I've worked with the Commission on Preservation and Access addressing digital preservation issues. I received the "Flame" award for lifetime achievement from Usenix in 1994, and I am a Fellow of the ACM. You can read my [publication list](#) if you wish. The previous paragraph is available in [Japanese](#).

Where?

Michael Lesk
[National Science Foundation](#)
4201 Wilson Boulevard, Room 1115
Arlington, Virginia 22230
703 306-1930 [Voice]
703 306-0599 [Fax]
lesk@acm.org

Interests

[Digital Libraries](#)



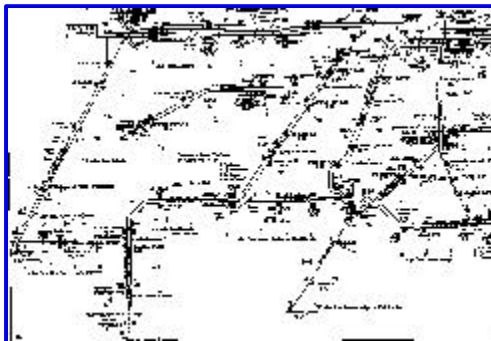
[Library preservation](#)



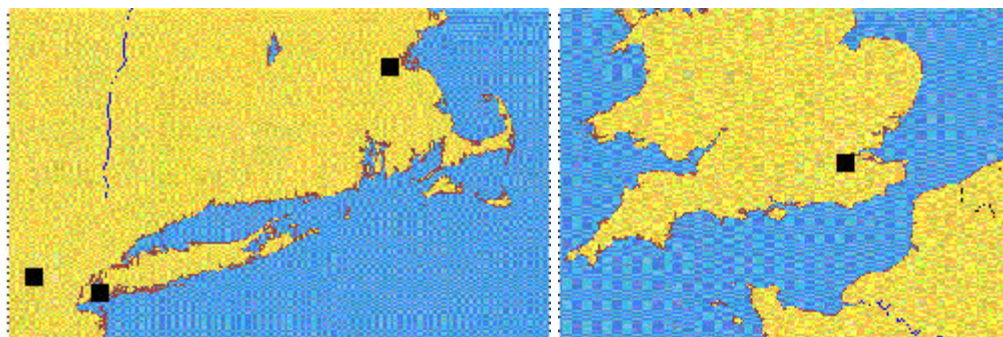
[Information Retrieval](#)



[Networks & Misc.](#)



Places I have lived



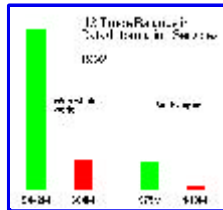
[New Jersey](#) . . [Brooklyn](#) [Cambridge, Mass.](#) [London](#)
..... [transport](#)

[Serving Human Needs Through Human Centered Systems](#). Draft of NSF subgroup report from workshop held February 1997. Contributors to text include Gio Wiederhold, Ben Shneiderman, and Jim Hollan.

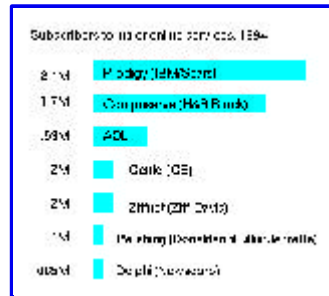
lesk@bellcore.com Michael Lesk
Last changed: 3 June 1998

Factoids

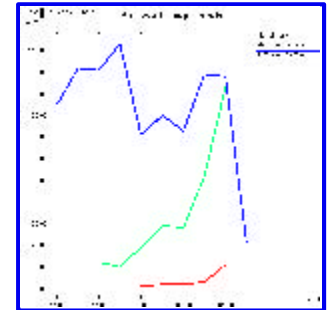
United States balance of trade in information services



Number of customers of online services



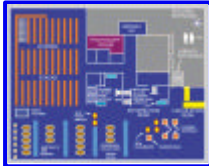
Trends in buzzwords



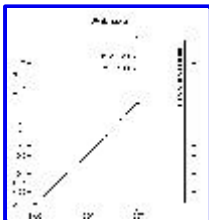
Material on digital libraries



[US Digital Library Programs: What Goals?](#)



[The Organization of Digital Libraries](#)



[How Much Information Is There in the World?](#)



[Digital Libraries: A Unifying or Distributing Force?](#), to be presented at *Scholarly Communication and Technology*, a conference sponsored by the Andrew W. Mellon Foundation, Atlanta, Georgia (April 24, 1997).



[Mad Library Disease: Holes in the Stacks](#), Lazerow Lecture, given at University of California Los Angeles, 18 April 1996, to appear in print later in 1996



[Libraries and the Web](#), to appear *Libraries and Information World Wide*, 1996



[Economics of Digital Libraries](#), course outline for lectures given Jan-Apr 1996,



Columbia University, New York.



[*Why Digital Libraries*](#), Follett lecture on electronic libraries, given 19 June 1995, BBC Conference Center, London, England.



[*The Future Value of Digital Information and Digital Libraries*](#); lecture given 9 November 1995 at the Kanazawa Institute of Technology Roundtable on Libraries and Information Systems, Kanazawa, Japan.



[*Making a Digital Library: The Contents of the CORE Project*](#); draft paper, October 1994; to appear, ACM TOIS

The roles of digital libraries in teaching and learning

**Gary Marchionini (University of Maryland, College Park, MD/USA,
email: march@umdd.umd.edu)**

**Hermann Maurer (Graz University of Technology, Graz/Austria,
email: hmaurer@iicm.tu-graz.ac.at)**

CACM April 95-Volume 38, Number 4 pg 67-75

Introduction

Libraries have long served crucial roles in learning. The first great library, in Alexandria two thousand years ago was really the first university. It consisted of a zoo and various cultural artifacts in addition to much of the ancient world's written knowledge and attracted scholars from around the Mediterranean who lived and worked in a scholarly community for years at a time. Today, the rhetoric associated with the National/Global Information Infrastructure (N/GII) always includes examples of how the vast quantities of information that global networks provide (i.e., digital libraries) will be used in educational settings [16].

This paper describes how digital libraries are evolving to meet the needs of teaching and learning and identifies issues for continued development. We distinguish formal, informal, and professional learning and argue that digital libraries will allow teachers and students to use information resources and tools that have traditionally been physically and conceptually inaccessible. We illustrate the types of information resources that digital libraries offer to teachers and learners and discuss some of the issues and challenges that digital libraries present for teaching and learning.

How do libraries support teaching and learning?

A library is fundamentally an organized set of resources, which include human services as well as the entire spectrum of media (e.g., text, video, hypermedia). Libraries have physical components such as space, equipment, and storage media; intellectual components such as collection policies that determine what materials will be included and organizational schemes that determine how the collection is accessed; and people who manage the physical and intellectual components and interact with users to solve information problems.

Libraries serve at least three roles in learning. First, they serve a practical role in sharing expensive resources. Physical resources such as books and periodicals, films and videos, software and electronic databases, and specialized tools such as projectors, graphics equipment and cameras are shared by a community of users. Human resources--librarians (also called media specialists or information specialists) support instructional programs by responding to the requests of teachers and students (responsive service) and by initiating activities for teachers and students (proactive services). Responsive services include maintaining reserve materials, answering reference questions, providing bibliographic instruction, developing media packages, recommending books or films, and teaching users how to use materials. Proactive services include selective dissemination of information to faculty and students, initiating thematic events, collaborating with instructors to plan instruction, and introducing new instructional methods and tools. In these ways, libraries serve to allow instructors and students to share expensive materials and expertise.

Second, libraries serve a cultural role in preserving and organizing artifacts and ideas. Great works of literature, art, and science must be preserved and made accessible to future learners. Although libraries have traditionally been viewed as facilities for printed artifacts, primary and secondary school libraries often also serve as museums and laboratories. Libraries preserve objects through careful storage procedures, policies of borrowing and use, and repair and maintenance as needed. In addition to preservation, libraries ensure access to materials through indexes, catalogs, and other finding aids that allow learners to locate items appropriate to their needs.

Third, libraries serve social and intellectual roles in bringing together people and ideas. This is distinct from the practical role of sharing resources in that libraries provide a physical place for teachers and learners to meet outside the structure of the classroom, thus allowing people with different perspectives to interact in a knowledge space that is both larger and more general than that shared by any single discipline or affinity group. Browsing a catalog in a library provides a global view for people engaged in specialized study and offers opportunities for serendipitous insights or alternative views. In many respects, libraries serve as centers of interdisciplinarity--places shared by learners from all disciplines. Digital libraries extend such interdisciplinarity by making diverse information resources available beyond the physical space shared by groups of learners. One of the greatest benefits of digital libraries is bringing together people with formal, informal, and professional learning missions.

Formal learning is systematic and guided by instruction. Formal learning takes place in courses offered at schools of various kinds and in training courses or programs on the job. The important roles that libraries serve in formal learning are illustrated by their physical prominence on university campuses and the number of courses that make direct use of library services and materials. Most of the information resources in schools are tied directly to the instructional mission. Students or teachers who wish to find information outside this mission have in the past had to travel to other libraries. By making the broad range of information resources discussed below available to students and teachers in schools, digital libraries open new learning opportunities for global rather than strictly local communities.

Much learning in life is informal--opportunistic and strictly under the control of the learner. Learners take advantage of other people, mass media, and the immediate environment during informal learning. The public library system that developed in the U.S. in the late nineteenth century has been called the "free university", since public libraries were created to provide free access to the world's knowledge. Public libraries provide classic nonfiction books, a wide range of periodicals, reference sources, and audio and video tapes so that patrons can learn about topics of their own choosing at their own pace and style. Just as computing technology and world-wide telecommunications networks are beginning to change what is possible in formal classrooms, they are changing how individuals pursue personal learning missions.

Professional learning refers to the on going learning adults engage in to do their work and to improve their work-related knowledge and skills. In fact, for many professionals, learning is the central aspect of their work. Like informal learning, it is mainly self-directed, but unlike formal or informal learning, it is focused on a specific field closely linked to job performance, aims to be comprehensive, and is acquired and applied longitudinally. Since professional learning affects job performance, corporations and government agencies support libraries (often called information centers) with information resources specific to the goals of the organization. The main information resources for professional learning, however, are personal collections of books, reports, and files; subscriptions to journals; and the human networks of colleagues nurtured through professional meetings and various communications. Many of the data sets and computational tools of digital libraries were originally developed to enhance professional learning.

The information resources--both physical and human--that support these types of learning are customized for specific missions and have traditionally been physically separated, although common technologies such as printing, photography, and computing are found across all settings. This situation, is depicted in Figure 1.

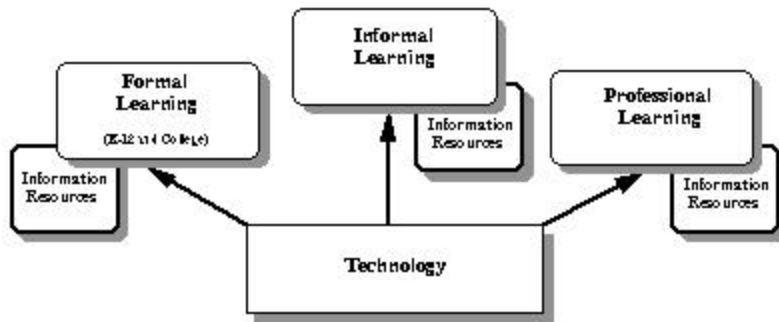


Figure 1. Current model of technological support for types of learning

Digital libraries combine technology and information resources to allow remote access, breaking down the physical barriers between resources. Although these resources will remain specialized to meet the needs of specific communities of learners, digital libraries will allow teachers and students to take advantage of wider ranges of materials and communicate with people outside the formal learning environment. This will allow more integration of the different types of learning, as depicted in Figure 2.

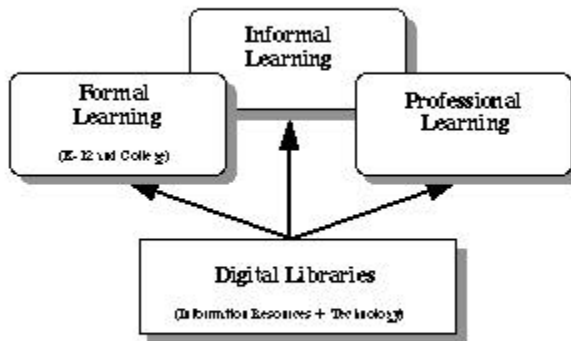


Figure 2. Digital libraries lead to integrated resources and type of learning

Although not all students or teachers in formal learning settings will use information resources beyond their circumscribed curriculum and not all professionals will want to interact even occasionally with novices, digital libraries will allow learners of all types to share resources, time and energy, and expertise to their mutual benefits. The following sections illustrate some of the types of information resources that are defining digital libraries.

Scientific data sets.

An enormous amount of attention is being given to making data sets collected by scientific projects available to broader communities of users. International efforts such as the Earth Observing System (EOS) and the human genome project demand large investments of public resources and create huge volumes of data. Multiple forces act to cause the development of digital libraries of scientific data from these projects. First, the tools used to collect, transmit, and analyze data generate or require digital signals, thus the information materials are in digital form rather than paper form. Second, the data must be made available to scientists worldwide on a timely basis and digital electronic networks make this possible. Third, the huge public investments encourage scientists to disseminate data as widely as possible to maintain public support and further educational and social progress. Providing access to these data sets through electronic libraries is a important challenge, especially in the U.S. where law mandates that publicly supported scientific data be made freely available to citizens (see the sidebar by Gey).

One example of how primary data sets are used in education is the Earth System Science Community Curriculum Testbed project that links students and teachers in high schools and universities in an effort to build an earth system science (ESS) community (<http://www.circles.org/ESSCC/ESSCC.GIF>). The project aims to build a curriculum for the interdisciplinary field of ESS by linking teachers of physics, chemistry, biology and other sciences to ESS scientists and NASA data sets. Topics such as acid rain and global warming are explored by teams of students in each classroom by taking advantage of a growing electronic community of students, teachers, and researchers. Using tools such as Mosaic, FTP, and Stella, teachers and students in schools in North America access data sets at different levels of representation, analyze the data, simulate scenarios, collaborate with scientists and students at remote sites, and publish reports. This project has been funded as part of the NASA digital library initiative and illustrates how electronic technology can support collaboration among scientists and students and how digital libraries of data, messages, and student reports are grown and managed. The ESS community is thus manifested as an organic, evolving digital library that includes primary data sets, conversations about them, and the results of using them.

Other Data Sets.

Textual databases of classic works (out of copyright) and image collections for important artistic exhibits or museums have been assembled by scholars and made available through the Internet. (See [13] for a collection of arts and humanities electronic resources and projects). As more schools and individuals acquire access tools and funds, it is likely that private digital libraries will move out of specialized markets to provide access to primary information for a fee. For-profit companies such as publishers of print, music, and film products and radio and television broadcasters own enormous volumes of information, and international information infrastructures will create new markets for that information. Teachers and learners will likely not be heavy on-demand users for this information but rather want to use it as the raw material for study and for integration into instructional presentations. How these materials are made available and what "fair use" policies evolve are yet to be determined.

Electronic journals.

Although electronic journals are becoming more common, they have not achieved as much penetration as many expected [23, 24]. As electronic journals develop, they will certainly improve informal and professional learning and will likely become useful resources in the K-12 arena which has traditionally

maintained only modest journal collections in schools. Two common approaches to electronic journals are to: (i) store files in LaTeX, PostScript or ASCII form in a fileserver and email the files, or allow FTP access to them ("generic approach"); and (ii) store documents in hypertext/hypermedia systems and allow online browsing and perusal ("hypertext approach"). Table 1 gives a sample of electronic journals that use the generic approach and Table 2 gives a sample of those using the hypertext approach.

The main problems that these publications solve to different degrees are related to information retrieval support, display of complex graphics and formulas, and distribution speed and reliability. A recent journal using the "hypertext approach" is J.UCS, the Journal of Universal Computer Science (see http://www.iicm.tu-graz.ac.at/Cjucs_root or send an email for general information to jucs@iicm.tu-graz.ac.at with subject [info]). It addresses these three problems by using a range of searching techniques including scoped searches; using HFT, RTF and particularly LaTeX and PostScript as file formats to provide high quality display; and using a world-wide network of initially 65 "foundation servers" to remove much of the access-time problems associated with earlier attempts.

Newsgroups, listservs and mail archives.

Perhaps the first examples of digital libraries in networked environments were the archives produced by the many USENET newsgroups and listservs available through global networks. News reading and filtering programs [21, 25] and search tools such as Archie and Veronica [8] provide rudimentary aids for locating information in these electronic discussions. Listservs are used for specialized projects (e.g., the ESS project above and the Perseus project both have listservs) and for distance education courses. In a cable television course taught by Marchionini, a listserv was used by students to present "one-minute papers" at the conclusion of each session. This provided continuity between sessions and personalized the interactions between the instructor and students, who would otherwise have only remote telephone access during live sessions. In another semester, students in graduate seminars in human-computer interaction taught by Marchionini at the University of Maryland and Christine Borgman at UCLA collaborated on term projects through email and FTP services. Students gained broader perspectives by virtue of the diversity in backgrounds that students from the different schools brought to the courses, and both positive collaborations and "techno-bullying" were observed. See [15] for a set of experiences in the virtual classroom.

In another setting, Maurer used Hyper-G [17] both as electronic library and discussion forum. In a 200-student class on "Societal Aspects of Computer Science" some 50 high-quality papers from specialists were made available to students via Hyper-G as the basis of a wide ranging electronic discussion. Students were able to comment on papers and earlier comments, the structure of the discussion being visualized using the X-Windows client Harmony [11]. The experiment created a network of over 4000 hyperlinked documents. Students remained "semi-anonymous" to encourage free discussion: i.e. students were allowed to choose arbitrary pen-names known only to each individual and to the instructor, the latter since student evaluation was based on the quality of contributions of the students. The experiment exemplifies blurring of the borderlines between electronic libraries and CSCW [7]--the semi-structured threads of conversation that make up news archives and lists provide another type of digital library product that will find increasing use in both formal and informal learning.

Specialized hypermedia corpuses.

A variety of hypermedia materials are becoming available and these collections are often served from a library rather than dedicated machines in classrooms. The Perseus hypermedia corpus (2.0) includes about 200 plays, books, poems, and text fragments in Greek and English translation; almost 25,000

24-bit color images of vases, sculpture, coins, and sites; maps; site plans; and a variety of search, navigation, and display tools. [6, 20]. Hundreds of colleges and scores of high schools are currently using Perseus to support instruction in Greek language, ancient history, Greek literature, religion, archaeology, and art history. In many sites, Perseus is delivered through a campus network. In some sites, Perseus is provided on a stand-alone machine in a library. The many CD-ROM corpuses now available for specialized topics challenge schools and individuals to be judicious in acquisition and use of these materials, thus increasing the need for resource sharing functions of libraries.

Another instance of an emerging corpus of material entering digital libraries is the PC-library [19] a product developed by a publishing consortium. Originally designed for stand-alone PC applications it has now migrated to client/server architecture. At the time of writing some 40 substantial reference volumes including a 10-volume encyclopedia ("Meyer A-Z"), dictionaries for most European languages ("Langenscheidt dictionaries"), the famous German-English "Oxford Duden," and standard scientific reference books on medicine, computer science and CAD are either available or in preparation, some of them containing high quality diagrams and pictures. There are a number of aspects of the PC library particularly worth mentioning: First, an arbitrary subset of the books in the library can be "activated" at any time, and all searches (including fuzzy full text) are carried out only within the books activated. Second, the PC library is not just a set of static books but can be used in a variety of not-only-read mode: persons can leave comments (for themselves or for others); searches can be activated from other applications and the results used in such other applications; books can be augmented by additional (personal) entries, including multimedia material (e.g. personal pictures or video clips); and material is automatically hyperlinked using a keyword based technique.

As vendors develop new products and as these specialized corpuses become available through global networks, libraries should take responsibility for ensuring secure and legal usage by students and teachers. One example of how libraries and computing centers cooperate today is in negotiating site licenses for these products and maintaining firewall services to ensure that licensing agreements are met. These collaborations can only grow as acquisition, organization, and dissemination of specialized software and hypermedia corpuses increase.

Table 1. Selected Electronic Journals Providing Generic Access

Numerische Mathematik Electronic Edition

Sponsor: Journal of the same name

Topics: Mathematics

Format: TeX and LaTeX

Features: every electronic issue some 2 weeks before the printed issue.

Access: EM-Helpdesk@springer.de.

Electronic Publication

Sponsor: MIT

Topics: Theoretical computer science

Format: LaTeX or PostScript

Features: Subscribers receive a notice each time an article is published; available for FTP

Access: Fisher@mitvma.mit.edu.

EJournal

Sponsor: University of Albany

Topics: Theory and practice surrounding electronic "text" and also social psychological, literary, economic and pedagogical implications of computer-mediated networks.

Format: Plain ASCII

Features: Listserved

Access: EJOURNAL@ALBANY.bitnet.

Asia-Pacific Journal (APEX-J)

Sponsor: University of Hawaii

Topics: Education in multicultural, international campuses

Format: Plain ASCII

Features: quarterly

Access: JamesS@UHunix.UHcc.Hawaii.edu.

Digest of Physics News Items

Sponsor: American Institute of Physics, by Phillip F. Schewe

Topics: physics

Format: Plain ASCII

Features: Posted in the Internet newsgroup sci.research and back issues can be downloaded by FTP from NIC.HEP.NET.

Access: physnews@aip.org

Table 2. Selected Electronic Journals Providing Hypertext Access

MUSE

Sponsor: Johns Hopkins University Library and Homewood Academic Computing

Topics: JHU Press journals

Access: telnet://jhuniverse.hcf.jhu.edu:20001/

Journal of Computer-Mediated Communication (JCMC)

Sponsor: Annenberg School of Communication, University of Southern California

Topics: Interpersonal and social aspects in communication networks.

Access: http://www.huji.ac.il/www_jcmc/jcmc.html

Electronic Journal of Combinatorics

Sponsor: Georgia Institute of Technology and the American Mathematical Society.

Topics: Combinatorics, graph theory and discrete algorithms.

Access: <http://ejc.math.gatech.edu:8080/Journal/journalhome.html>

Newsletter of the National Research Center on Student Learning (NRCSL).

Sponsor: Learning Research and Development Center

Topics: education

Access: <gopher://gopher.pitt.edu/11/news/lrdc>

Journal of Universal Computer Science (JUCS)

Sponsor: Springer Pub.Co. and Graz University of Technology

Topics: All areas of Computer Science

Access: http://www.iicm.tu-graz.ac.at/Cjucs_root

Indexes and directories.

There are a host of bibliographic and catalog databases that may be included in digital libraries. These range from the more than 20 million record database of bibliographic citations in OCLC and the millions of citations in online databases for specialized literatures such as medicine (e.g., MEDLINE) and engineering (e.g., NTIS). Tertiary databases such as citation indexes and databases of directories make information seeking more effective but require specific skill and effort on the part of information seekers. Many of the thesauri for specialized literatures are available in electronic form (e.g., Medical Subject Headings, ACM Computing Reviews Classification System) and techniques for merging and filtering these languages to allow users to search across multiple databases are emerging. Although most indexes to image and sound collections currently use words from captions or titles, new pattern-matching techniques are emerging to categorize and classify multimedia objects [10]. In the past, bibliographic instruction has been provided by librarians as a supplement to "regular" courses, but widespread availability of digital libraries will require remote instruction and support related to information-seeking skills and knowledge.

Electronic search and display tools.

It often has been said that the Internet is starting to provide the largest library humankind has ever had. As true as this may be, it is also the messiest library that ever has existed. Navigation and display tools such as Mosaic allow users to browse the World Wide Web and display text and multimedia objects. Search tools such as the Wide Area Information Server (WAIS) and Archie and Veronica allow people to search specific directories or list archives (see [22] for an overview of tools). However, in addition to index and directory services or navigation tools, it has become apparent that such 'a posteriori' tools to organize the unstructured Internet universe are not sufficient. Rather, some 'a priori' structuring is necessary. This was first done quite successfully with Gopher [3] and later with WWW [5]. However, "first generation hypermedia techniques" do not seem to be sufficient for large amounts of data: "second generation techniques" [4] involving distributed database mechanisms, scope definition facilities for searches, bidirectional link databases [14] for automatic link maintenance, and other advanced techniques are emerging. For example, Harmony [11], the X-Windows client for Hyper-G, WWW, Gopher and WAIS provides sophisticated navigational facilities when used in conjunction with a Hyper-G database: the facilities include visual "local maps" of all in- and outgoing hyperlinks, a 3-D landscape generator, a history and hierarchy browser, Boolean searches on attributes, full text searches including approximate matches in user-defined scopes that may arbitrarily cross even the physical boundaries of servers. Such features will make working with large electronic libraries less frustrating than it is sometimes now, and will certainly assure that the use of electronic libraries is more efficient than using large amounts of printed material. As these tools evolve, better integration of search and display will be necessary. One approach is dynamic queries [2] that provide graphical representations for

database elements and sliders for adjusting parameters on those elements. As parameters are changed, the graphical display is immediately updated, providing immediate visual answer sets.

Digital libraries in education: Promises, Challenges and Issues

The examples above illustrate that digital libraries have obvious roles to play in formal learning settings by providing teachers and learners with knowledge bases in a variety of media. In addition to expanding the format of information (e.g., multimedia, simulations), digital libraries offer more information than most individuals or schools have been able to acquire and maintain. Digital libraries are accessible in classrooms and from homes as well as in central library facilities where specialized access, display, and use tools may be shared. Remote access allows possibilities for vicarious field trips, virtual guest speakers, and access to rare and unique materials in classrooms and at home. The promise is one of better learning through broader, faster, and better information and communication services. These physical advantages promise several advantages to teachers and learners by extending the classroom, however, as with all technologies, there are costs and tradeoffs to these advantages.

One clear difference between traditional libraries and digital libraries is that digital libraries offer greater opportunity for users to deposit information as well as use information. Thus, students and teachers can easily be publishers as well as readers in digital libraries. The number of student-produced "Mosaic home pages" and gopher sites continues to grow as teachers and students not only bring digital library information into the classroom but move the products of the classroom out into the digital libraries. Just as distinctions between publishers and readers are becoming less clear in networked environments, Internet access in classrooms blurs distinctions between teaching and learning. Students bring interesting and important information to class discussions and in many cases lead teachers and classmates to new electronic resources and tools. Teachers' increasingly will find themselves in the important roles of moderator and critic, modeling for students how to examine and compare points of view and look critically at information. Teachers who have begun using networked materials in their classes are early adopters of new ideas and technologies and are comfortable sharing power with students. Just as "authority of information" has become an issue in professional communities that leverage networks, the authority of information in classrooms that has traditionally rested solely with teachers will increasingly be challenged by students locally and remotely.

Digital libraries will support communities of interest and allow more specialized courses to be offered. For example, students at different high schools in the CoVis project collaborated by sharing a digital library of weather data [12] and students in the Earth System Science Community Project described above share a variety of NASA data in classes in Washington D.C., St. Louis, Los Angeles, and New Mexico. Telecourses have already allowed rural schools to offer advanced placement courses to a few students by sharing teachers across geographical distances. As network access improves in schools, highly specialized courses offered on a distributed basis will become common, and it is likely that some of these will be offered by students. Internet-based courses have already been offered successfully, although mainly on the topic of the Internet itself, and network based electronic conferences have proven effective (e.g., University of Maryland Professor, Thomas O'Haver recently ran a chemistry conference that involved 450 participants from 33 countries).

The most important changes that digital libraries bring may be in advancing informal learning. The same advantages that accrue to classroom learning also accrue to individuals pursuing their own learning. In many ways, the development of Freenets are extensions of the public library system. Digital libraries are digital schools that offer formal packaging for specific skills and topics as well as general browsing for creative discovery and self-guided, informal learning. The design community has already begun to

consider ways to support learning on demand in electronic environments [9] to address problems of coverage (since no learning system can cover all things learners may need) and obsolescence (systems and knowledge changes).

For the promises to obtain, issues of access and intellectual property must be addressed. Although the U.S. Library of Congress has committed to becoming a digital library, it can make available only documents or finding aides created within the Library or government agencies, items out of copyright, and representations from exhibits or events sponsored by the Library. Although these represent enormous quantities of information, the core holdings of the Library--the books, films, and recording--cannot be made available electronically under current copyright law. Whether the copyright law will change to allow materials to be accessed electronically under some educational fair use arrangement remains to be settled. Curators, theater owners and publishers are loathe to give up restricted access due to understandable self-preservation concerns. Some of these fears may be unfounded. For example, in the 1930's owners of professional baseball clubs allowed only World Series games to be broadcast on the radio because they feared that attendance at regular games would go down if all games were broadcast. When Lawrence McFale in Cincinnati began to broadcast the Reds' games in 1938, entire new markets opened up beyond the traditional male attendees--women and men who previously did not know much about baseball became interested and attendance went up (Ken Burns' PBS series, *Baseball*). Additionally, entire new revenue streams from advertising became available, which today eclipse attendance profits. However, historical examples are not likely to be enough to convince publishers and other information industry entities to make their "property" available electronically without secure mechanisms for profit.

Even more challenging, however, is building intellectual infrastructures for digital libraries. These include techniques for using electronic information in teaching and learning [18]. Teachers must learn how to teach with multimedia resources and to share informational authority with students. Designing activities that take advantage of digital library resources requires time and effort to examine what is available and integrate information into modules and sequences appropriate to the students and curriculum. Furthermore, modeling the research process for students requires teachers to grapple with problems on-the-fly, make mistakes, recover, react to dead ends, and demonstrate all the other uncomfortable and frustrating aspects of problem solving. Like Euclid, who presented the products of geometric research in the form of neat, polished deductive proofs (rather than the empirical and intuitive thought that led to the theorems), teachers are more comfortable providing polished packages/modules rather than the messy details of discovery and problem solving. Applying digital libraries in classrooms requires different attitudes and tolerances for such learning conditions.

Just as teachers must learn new strategies for using electronic tools in teaching, students must learn how to learn with multimedia (both actively and passively) and how to take increased responsibility for directing their own learning. In our observations of students in classrooms where Perseus was used, students expressed concerns about taking notes--because a screen of text, a screen of vases, and the instructor's verbal comments were concurrently available, they did not know what to write down! Although better technological tools such as networked laptop computers may solve the technical problem, the issues of what to attend to and how multiple streams of information should be integrated require new combinations of perceptual, cognitive, and physical skills for learning. In short, building intellectual infrastructures requires intellectual, emotional, and social breakthroughs for teaching and learning.

At the nexus of physical and intellectual infrastructure is the interface to the digital library. Tools for finding, managing, using, and publishing electronic information must be both powerful and easy to use.

Digital libraries must provide a mix of software and people to provide reference assistance and question answering services (e.g., Ackerman's Answer Garden system for handling X-Windows questions, [1]). The people in the digital library will go beyond reference to serve as teachers on demand. These humans must be aided by software that shunt "typical" questions toward pathfinders or frequently-asked-question services. Thus, digital libraries will extend what has been the most beneficial feature of electronic networks--communication---to teaching and learning settings. Good interfaces will allow learners to take advantage of digital resources equally well in classrooms, homes, and offices.

Clearly, digital libraries have important roles to play in teaching and learning. Existing physical schools and libraries will continue to exist since they serve cultural and social roles as well as informational roles. There will always be a need for physical objects and social settings in learning; the vicarious is not enough. Parents will continue to demand child care, assurances of organized and shared culture beyond television, and human direction and guidance in learning at all levels. These demands will also be augmented by digital environments. Digital libraries will allow parents, teachers, and students to share common information resources and communicate easily as needed. In special cases, work, school, and play may become one--novice and professional learners collaborating with common information resources to solve real problems. In many respects, digital libraries will become digital schools. This represents a return to Alexandria, where learners of all types come together to share and explore information and expertise.

References

1. Ackerman, M.S. Answer Garden: A tool for growing organizational memory. Ph.D. Thesis, MIT, 1993.
2. Ahlberg, C., Williamson, C., & Shneiderman, B. Dynamic queries for information exploration: An implementation and evaluation. In B. Shneiderman (Ed.). *The sparks of innovation in human-computer interaction*. Ablex, Norwood, NJ. 1994, pp. 281-294.
3. Alberti, B., Anklesaria, F., Lindner, P., McCahill, M., & Torrey, D. Internet Gopher Protocol: A Distributed Document Search and Retrieval Protocol; FTP from boombox.micro.umn.edu, directory pub/gopher/gopher_protocol. 1994.
4. Andrews, K., Kappe, F., Maurer, H., Schmaranz, K. On Second Generation Hypermedia Systems, IIG Report, Graz (1994); FTP server iicm.tu-graz.ac.at directory pub/Hyper-G/papers.
5. Berners-Lee, T., Cailliau, R., Groff, J. WorldWideWeb: The Information Universe, *Electronic Networking: Research, Applications and Policy* 1,2 (1992), 52-58.
6. Crane, G. *Perseus 2.0*. New Haven, CT: Yale University Press. 1994.
7. Devan, P. A Survey of Applications of CSCW Including Some in Educational Settings, *Proceedings of ED-MEDIA'93*, (June 23-26, Orlando, FL). AACE, Charlottesville, VA, 1993, pp. 147-152.
8. Deutsch, P. Resource Discovery in an Internet Environment-the Archie Approach, *Electronic Networking: Research, Applications and Policy* 1,2 (1992), 45-51.
9. Eisenberg, M. & Fischer, G. Symposium: Learning on demand. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, (June 18-20, Boulder, CO), Erlbaum, Hillsdale, NJ, 1993, pp. 180- 186.
10. Faloutsos, C., Equitz, W., Flickner, M., Niblack, W., Petkovic, D., & Barber, R. Efficient and effective querying by image content, *Journal of Intelligent Information Systems*, in press.
11. Fenn, B., Maurer, H.: Harmony on an Expanding Net; *ACM Interactions*. 1,3, (October 1994), 26-38
12. Fishman, B.J. & D'Amico, L.M. Which way will the wind blow? Networked computer tools for studying the weather. *Proceedings of ED-MEDIA 94* (June 25-30, Vancouver, BC). AACE,

- Charlottesville, VA, 1994, pp. 209-216.
13. Getty Art History Information Program, The American Council of Learned Societies, & The Coalition for Networked Information. Humanities and arts on the information highways: A profile. Santa Monica, CA: Getty Art History Information Program. (1994).
 14. Haan,B.J., Kahn,P., Riley,V.A., Coombs,J.H., Meyrowitz,N.K. IRIS Hypermedia Services, Communications of the ACM 35,1 (1992), 36-51.
 15. Hiltz, S.R., & Turoff, M.. Virtual classroom plus video: Technology for educational excellence. Proceedings of ED-MEDIA 94 (June 25-30,Vancouver, BC). AACE, Charlottesville, VA, 1994, pp. 26-31.
 16. Information Infrastructure Task Force. The national information infrastructure: Agenda for action. Washington, DC: NTIA. 1993, Also available via anonymous ftp from ftp.ntia.doc.gov.
 17. Kappe,F., Maurer,H., Scherbakov,N. Hyper-G-a Universal Hypermedia System, Journal of Educational Multimedia and Hypermedia 2,1 (1993), 39-66.
 18. Marchionini, G. & Crane, G. Evaluating hypermedia and learning: Methods and results from the Perseus Project. ACM Transactions on Information Systems, 12(1), (1994), 5-34.
 19. Maurer, H., Muelner,H., Schneider, A: The PC Library and Applications in an Educational Setting, Symposium Didaktik der Mathematik, Klagenfurt (1994)
 20. Mylonas, E. An interface to classical Greek civilization. Journal of the American Society for Information Science, 43(2), (1992), 192---201.
 21. Oard, D.W. The Information Filtering Laboratory. 1994. WWW site:
<http://www.umiacs.umd.edu/labs/CLIP/filter.html>.
 22. Obraczka, K., & Danzig, P., & Li, S. Internet resource discovery services. IEEE Computer, 26, (1993), 8-22.
 23. Odlyzko, A., M. Tragic Loss or Good Riddance? The impending demise of traditional scholarly journal, To appear in: Notices of the AMS (1994).
 24. Schaffner, A. The future of scientific journals: Lessons from the past. Information Technology and Libraries, 13(4), 1994, 239-247.
 25. Stevens, C. Knowledge-Based Assistance for Handling Large, Poorly Structured Information Spaces. Ph.D. Dissertation, University of Colorado at Boulder Technical Report Number CU-CS-640-93, January 1993.

[Back one topic.](#) [Back to Outline.](#) [Back to Digital Library Homepage.](#)

Countries & Regions:

(Chapter 11, page 245, "Books, Bytes and Bucks", Michael Lesk)

- **United States of America:** In the US, NSF, NASA and ARPA have funded six important Digital Library efforts, called the DLI (Digital Libraries Initiative). These programs each involve a large consortium of cooperating institutions but the six main ones are : University of California at Berkeley, University of Santa Barbara, University of Michigan, Carnegie Mellon University, Stanford University, and the University of Illinois.
 - University of California at Berkeley: Image content queries along with Xerox PARC, database extraction from documents, multivalent documents, NLP. Headed by Robert Wilensky.
 - University of Michigan: Scalability and Education. They are also investigating the use of agent architectures for Digital Libraries and trying to merge DLI with their other digital library efforts such as JSTOR and TULIP. Headed by Dan Atkins.
 - University of Illinois: Concentrating on using scientific journals as their base collection with diversity in both documents as well as publishers, making the transition process from SGML to HTML smoother, defining semantic spaces. Headed by Bruce Schatz.
 - Stanford University: concentration is on the infrastructure development such as base networking and databases to support digital libraries. Also concerned with interoperability between different digital library projects. Headed by Hector Garcia-Molina.
 - University of California at Santa Barbara: spatial indexing and retrieval , image processing. Headed by Terry Smith.
 - Carnegie Mellon University: digital video, image analysis, speech recognition, face recognition, natural language understanding. Headed by Michael Mauldin and Marvin Sirbu.

Other than DLI, many research projects are underway at some other universities such as Virginia Tech and Texas A&M. In the near future, extensive funds are expected to be allocated for Digital Libraries.

The Library of Congress, under James Billington is digitizing 5 million of its items in a massive \$60 million effort. Other universities involved in related projects are Georgia Tech, Cornell, MIT, University of Tennessee, Washington and California and Virginia Tech (known for the Envision system of Ed Fox). Other limited efforts include University of Virginia, University of Georgia and Columbia University.

- **United Kingdom:** Though efforts are still limited to penny-pockets, 20 million pounds have been set aside for digital library projects. The program originally called FIGIT, now known as E-LIB funded 35 projects. Work includes cataloguing of archives, digitization of documents and data sharing. Some of the more notable efforts are : Digitizing the Burney collection of pre-1800 newspapers and scanning of Batley News, the CANTERSBURY TALES project that involves scanning all pre-1500 manuscripts and some other similar projects. However, the most notable is the Electronic Beowulf project which is a US/UK collaboration between Kevin Kiernan (University of Kentucky), Paul Szarmach (Western Michigan University) and the British Library.
- **France:** Work includes some scanning of old manuscripts with the most notable being the Tresor de la Langue Francaise project at the University of Nancy. The French, along with the Japanese

are also leaders in the Group 7 project which is a museum project. Other efforts are INIST and FOUDRE (1989 to 1992) followed by EDIL and ELITE.

- **The EU:** The European Union funds a large number of international efforts in digital libraries. (Please see page 255 of Michal Lesk's book for details)
- **Japan:** Japan is involved in some digitization and cataloguing efforts and has a \$50M project on. They are also working on modern document delivery and OCR.
- **Australia:** Australia has recently made a modest effort to enter into digital library research. They are planning some digitization projects with a \$10M (Australian) digitization project on the anvil. They are also interested in digitizing Aborigine scriptures and paintings.
- **Elsewhere:** Many other countries are involved in digital library research on much smaller scales. Notable amongst them are Canada, Singapore, Korea and China.

NOTE 1: For detailed information on any of the above please refer to Dr. Lesk's book (recommended as supplement text for this course).

NOTE 2: See also the table pointing to various national digital libraries from April 1998 CACM [online pages](#)

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

URLs from Comm. ACM, 41(4), April 1998

This page supplements the Guest Editors' page <http://purl.lib.vt.edu/dlib/pubs/CACM199804> related to the April 1998 CACM Special Section on Digital Libraries: Global Scope, Unlimited Access. Below are all the URLs from that special section, for easy navigation. Please feel free to report any problems or changes. - eaf

List of articles in the Special Section, with URLs

- Legally Speaking: Encoding the Law into Digital Libraries, by Pamela Samuelson
 - (p. 17) [Digital Future Coalition Web site](#)
- Toward a Worldwide Digital Library, by Edward A. Fox and Gary Marchionini
 - (p. 30) Table 1. National libraries of countries mentioned in this section
 - [Australia - National Library of Australia](#)
 - [Brazil - University of Sao Paulo Library](#)
 - [Canada - National Library of Canada / Bibliotheque Nationale du Canada](#)
 - [Denmark - Det Kongelige Bibliotek](#)
 - [Finland - Helsinki University Library / National Library of Finland](#)
 - [France - Bibliotheque Nationale de France](#)
 - [Germany - Die Deutsche Bibliothek](#)
 - [Hungary - Hungarian Electronic Library](#)
 - [Japan - National Center for Science Information Systems - may not respond](#)
 - [Korea - Five Library Consortium](#)
 - [Netherlands - Koninklijke Bibliotheek](#)
 - [New Zealand - National Library of New Zealand](#)
 - [Singapore - Multiple Agencies](#)
 - [United Kingdom - British Library](#)
 - [United States - Library of Congress](#)
 - p.32 References
 - 1. Borgman, C. [Social Aspects of Digital Libraries](#)
 - 3. Paepcke, A. [Digital libraries: Searching is not enough: What we learned on-site](#)
 - 4. Scherlis, W. [Repository Interoperability Workshop: Towards a repository reference model](#)
- Interoperability for Digital Libraries Worldwide, by Andreas Paepcke et al.
- Accessing Distributed Cultural Heritage Information, by William E. Moen
 - [Aquarelle](#)
 - [Dublin Core Metadata Element Set](#)
 - [Z39.50 Maintenance Agency](#)
- Digital Access to Antiquities, by Henry M. Gladney et al.
 - [Supporting Web pages](#)
- FedStats Promotes Statistical Literacy, by Cathryn S. Dippo
 - (p. 59) [Current Population Survey Data](#)
 - (p. 60) Reference 1. Hert and Marchionini, [Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations](#)
 - [U.S. Department of Labor: BLS Handbook of Methods, Ch. 1, Labor Force Data Derived from the Current Population Survey](#)
 - U.S. Bureau of Labor Statistics: [Labor Force Statistics from the Current Population Survey, Bureau of Labor Statistics Data, Employment Situation](#)
 - U.S. Office of Management and Budget: [Statistical Policy Working Paper 22 - Report on](#)

[Statistical Disclosure Limitation Methodology](#)

- New Role for Community Networks, by D.D. Cowan et al.
 - [CTT Community Network \(CTTnet\)](#)
 - (p. 63) Reference 3. Hecker, [Advice for Community Network System Designers](#)
 - [Lead author's home page](#)
- Viewing Multilingual Documents on Your Local Web Browser, by A. Maeda et al.
 - [gateway server and multilingual e-text collection](#)
 - [multilingual folk tales](#)
 - (p. 65) Reference 1. Chase et al., [Web Fonts - WC3 Working Draft](#)
 - (p. 65) Reference 3. Dartois et al., [A multilingual electronic text collection of folk tales for casual users using off-the-shelf browsers](#)
- Distributed Chinese Bibliographic Searching, by M.K. Leong, L. Cao, Y. Lu
- NSF-EU Multilingual Information Access, by Judith L. Klavans and Peter Schauble
 - [Digital Library Collaboratory Working Groups information](#)
 - [Multilingual Information Access working group information](#)
- A Public Library Based on Full-text Retrieval, by Ian H. Witten et al.
 - (p. 75) Reference 7. Witten et al., [The New Zealand digital library project](#)
- Students Access Books and Journals through MeDoc, by Albert Endres and Norbert Fuhr
 - [Project home page](#)
 - (p. 77) Reference 5. Fuhr, [Optimum database selection in networked IR, in Proc. SIGIR'96 Workshop on Networked Information Retrieval](#)
- Discovery of Resources within a Distributed Library System, by Laszlo Kovacs
 - [Computer Science Technical Reports](#)
 - [DELOS Working Group](#)
 - [Digital Library Collaboratory Working Groups](#)
 - [ERCIM Digital Library Initiative](#)
 - [European Research Consortium for Informatics and Mathematics](#)
 - [MTA SZTAKI, Department of Distributed Systems](#)
 - [Networked Computer Science Technical Reports Library](#)
 - [NCSTRL overview](#)
 - [Resource Indexing and Discovery in a Globally Distributed Digital Library Workshop](#)
 - (p. 77) Reference 2. Kahn and Wilensky, [A Framework for Distributed Object Services](#)
- Initiatives That Center on Scientific Dissemination, by Marcos Andre Goncalves and Claudia Bauzer Medeiros
 - [Digital Agro-Library - EMBRAPA](#)
 - [UNICAMP Database Group page - see pointers for Geo-Library-Framework \(GeoLib\)](#)
 - [Scientific Electronic Library Online - SciELO](#)
 - [Tropical Database](#)
- R&D for a Nationwide General-Purpose System, by Sung Hyon Myaeng
 - (p. 85) Reference 1. Feedman, [WILLOW: Technical overview](#)
 - (p. 85) Reference 4. NIST, [Guide to Z39.50/PRISE 1.0: Installation, Use, and Modification](#)
- Many Projects That Depend on Collaboration, by Cliff McKnight
 - [Electronic Libraries Programme, eLib](#)
 - [Follett Report, 1993](#)
 - (p. 87) Reference 1. McKnight and Dillon, [User-centred design of library information systems: HyperLib - may not respond](#)
- Libraries' New Role in Electronic Scholarly Publishing, by Andrew E. Treloar
 - [Project home page links](#)
- Semantic Information Retrieval, by Annelise Mark Pejtersen

- Discourse Analysis of User Requests, by Sanna Talja et al.
 - Sorting Out Searching: A User-Interface Framework for Text Searches, by Ben Shneiderman et al.
-
-

Author: [Edward A. Fox \(CV, directions, hours, photo\)](#)

Curator: [Virginia Tech](#); [Dept. of Computer Science](#)

Last Updated: 98/3/27

Email: fox@vt.edu

(c) Copyright 1998 Edward A. Fox

Centers, sites and organisations:

Some major Digital Library centers and research programs, separately described:

- [Carnegie Mellon University](#)
 - [CNRI](#)
 - [Library of Congress](#)
 - [Stanford University](#)
 - [University of California at Berkeley](#)
 - [University of California at Santa Barbara](#)
 - [University of Illinois](#)
 - [University of Michigan](#)
 - [Texas A&M](#)
 - [Virginia Tech](#)
-

Selected other sites:

ACM DL : Tap into the ACM Digital Library, a vast resource of bibliographic information, citations, and full-text articles.

IEEE

- [IEEE Computer Society Digital Library News \(DLN\) Archive](#)
- [IEEE Digital Library Task Force](#)

IBM

- [IBM DL Home page](#)
- [IBM Renaissance Consortium Panel](#) and [workshop](#)
- [images - QBIC](#)

OCLC (OCLC is a nonprofit, membership, library computer service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs.

- Research <http://www.oclc.org/oclc/menu/research.htm>
SiteSearch <http://www.oclc.org/oclc/menu/site.htm>

Xerox

- [Home Page](#)
- [Scientific American article](#)
- [Scatter/Gather examples](#)
- Questions:
 - Compare
 - What are the various interfaces built? How do they compare? What is the best use of

each?

- Scatter/gather
 - Explain clustering, relate it to scatter/gather.
 - What are special problems with large category systems and how can they be solved?

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

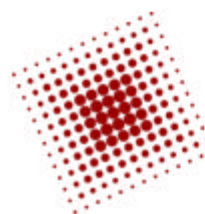
CNRI:

- home page (site map) http://www.cnri.reston.va.us/site_map.html/site_map.html
 - Architecture
 - Kahn-Wilensky Framework for Distributed Digital Object Services
<http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html>
 - key architectural issues <http://WWW.CNRI.Reston.VA.US/home/cstr/arch/slides.html>
 - architecture for information in digital libraries
<http://www.dlib.org/dlib/february97/cnri/02arm s1.html>
 - Digital Object Architecture Project <http://www.cnri.reston.va.us/doa.html>
 - CS-TR Computer Science Technical Reports <http://www.cnri.reston.va.us/cstr.html>
-

[\[Main\]](#) [\[Contents\]](#) [\[Resources\]](#) [\[Centers\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



The Corporation for
National Research Initiatives

Site Map



web-curator@cnri.reston.va.us

About CNRI

- [CNRI Mission](#)
- [Directions to CNRI](#)
- [Employment Opportunities](#)
- [Information Technology Infrastructure](#)
- [Officers and Directors](#)

Publications

- [XIWT White Papers](#)
- [D-Lib Magazine](#)
- [IETF Proceedings](#)
- [Infrastructure History Series](#)
- [Recent CNRI Publications](#)
- [CNRI Publications Archive](#)

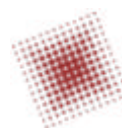
Programs and Activities

- [Application Gateway System \(AGS\)](#)
- [Cross-Industry Working Team \(XIWT\)](#)
- [D-Lib and D-Lib Magazine](#)
- [Defense Virtual Library](#)
- [Digital Object Architecture](#)
- [Digital Object Identifier System](#)
- [Electronic Payments Forum](#)
- [Grail](#) and [Python](#) and [JPython](#)
- [The Handle System](#)
- [Infrastructure History Series](#)
- [IETF Secretariat](#)
- [IOPS.ORG](#)
- [Knowbot Programs](#)
- [MAGIC](#)
- [National Digital Library Program](#)
- [The Registry](#)
- [Repository Architecture](#)
- [US Copyright Office](#)

Recent Activities

- [Computer Science Technical Reports](#)
- [Gigabit Testbed Initiative](#)

Last Updated: March 26, 1998
Corporation for National Research Initiatives



A Framework for Distributed Digital Object Services

Robert Kahn
Corporation for National Research Initiatives

Robert Wilensky
University of California at Berkeley

May 13, 1995
cnri.dlib/tn95-01

1. Introduction

This document describes fundamental aspects of an infrastructure that is open in its architecture and which supports a large and extensible class of distributed digital information services. Digital libraries are one example of such services; numerous other examples of such services may be found in emerging electronic commerce applications. Here we define basic entities to be found in such a system, in which information in the form of **digital objects** is stored, accessed, disseminated and managed. We provide naming conventions for identifying and locating digital objects, describe a service for using object names to locate and disseminate objects, and provide elements of an access protocol.

We use the term **digital object** here in a technical sense, to be defined precisely below. Files, databases and so forth that one may ordinarily think of as objects with a digital existence are not digital objects in the sense used here, at least not until they are made into an appropriate data structure, etc., as we will describe shortly.

Only the most basic elements of the infrastructure are described herein. These elements are intended to constitute a minimal set of requirements and services that must be in place to effect the infrastructure of a universal, open, wide-area digital information infrastructure system ("the System"). We anticipate that many other services and elaborations will come into existence as the System is further developed, either building upon or otherwise added to these elements.

This paper focuses on the network-based aspects of the infrastructure, namely those for which knowledge of the contents of digital objects is not required. Definition of the content-based aspects of the infrastructure is purposely not addressed in this paper. An important goal in limiting the description of the infrastructure in this way is not to constrain the higher level user and service level choices that, for many reasons, might be inappropriate to fix upon at this point in time. With only the most basic elements of the infrastructure in place, technological evolution would not be overly constrained. Further, the likelihood of achieving widespread interoperability of services at some early point in the future will be preserved. No doubt the resulting capability will have a greater potential for enhancement and evolution through the participation of many others in helping to define it.

2. Overview and Definitions

In this section, we first present an informal overview of the elements of the System, sketching its

elements and how they are supposed to function together. These elements include the notions of **digital objects**, **handles**, **metadata** and **key metadata**, **repositories**, **handle generators**, **originators**, **users**, **global naming authorities** and **local naming authorities**, and a **repository access protocol**. Then we provide more formal definitions of these entities, and explicate their details.

2.1 Informal Overview

Conceptually, the System works as follows: An **originator**, i.e., a user with digital material to be made available in the System, makes the material into a **digital object**. A digital object is a data structure whose principal components are digital material, or **data**, plus a unique identifier for this material, called a **handle** (and, perhaps, other material). To get a handle, the user requests one from an authorized **handle generator**. A user may then deposit the digital object in one or more **repositories**, from which it may be made available to others (subject, of course, to the particular item's terms and conditions, etc.). Upon depositing a digital object in a repository, its handle and the repository name or IP address is registered with a globally available system of **handle servers**. Users may subsequently present a handle to a handle server to learn the network names or addresses of repositories in which the corresponding digital object is stored.

Interactions such as depositing digital objects or accessing digital objects in repositories is accomplished using a **repository access protocol (RAP)**, which all repositories must support.

A digital object stored in a repository, and whose handle has been registered with the handle server system, is called a **registered digital object**. Registered digital objects are of primary concern to us here, as they are explicitly constructed to be known about by others, presumably for widespread availability. However, we do not constrain repositories to contain only registered digital objects. Nor are repositories constrained to operate only via the repository access protocol, although they must all support it.

Handles are the primary global identifiers for digital objects. However, we do not anticipate that users will necessarily manipulate handles directly; nor is the system of handle servers intended as the only means by which users will locate objects. More likely, location services will be accomplished by various value-added providers not defined as part of the infrastructure. Rather, the handle server system provides a kind of public safety net which facilitates the location of a digital object given only its handle.

We emphasize that the term **digital object** is used here in a technical sense of a particular sort of data structure, and not in the general sense of any object that may have digital form. Perhaps a term such as **digital infrastructure object** would better capture this intention. However, we have found this alternative terminology to be somewhat cumbersome in practice, and have therefore chosen to retain the simpler term digital object instead.

2.2 Definitions

We now define our terminology more formally, and describe the operation of the various components of the System in some detail.

Formally, a digital object is an instance of an abstract data type that has two components, **data** and **key-metadata**. The data is typed, as is described below. The key-metadata includes a **handle**, i.e., an identifier globally unique to the digital object; it may also include other metadata, to be specified. Possible primitive and composite data types for digital object data are discussed below.

A **repository** is a network-accessible storage system in which digital objects may be stored for possible subsequent access or retrieval. The repository has mechanisms for adding new digital objects to its collection (**depositing**) and for making them available (**accessing**), using, at a minimum, the **repository access protocol**. The repository may contain other related information, services and management systems.

Repositories have official, unique names, assigned or approved to assure uniqueness by a **global naming authority**. In general, the global naming authority will assign a name to a local naming authority. The local naming authority may use this name as the name of a repository. In addition, it may extend this name to create new names by suffixing the name with a ".", followed by a new (relatively) unique name component. Each such name represents a naming authority and potential associated repository. (I.e., in general, repositories will have unique names of the form "X.Y.Z".)

Note that a repository name is not necessarily the name of a particular host. For example, it may correspond to a set of hosts at different physical locations.

A **stored digital object** is a digital object stored in a repository. In addition, handles are expected to be made known to a system of **handle servers**, as described below. Such a handle is a **registered handle**. A **registered digital object** is a stored digital object whose handle has been registered. (Note that a handle cannot be registered until its corresponding digital object is stored) Repositories provide users access to stored objects under terms and conditions that may be set by the depositor and/or a given repository.

Registered digital objects are the entities of primary concern to the infrastructure, since they are stored in a repository and made known via the registration of their handles. Intermediate entities, such as stored digital objects, are defined only because they may arise in implementations of repositories that provide access to registered digital objects. However, their existence is not strictly necessary. For example, a repository may offer a service in which it deposits a digital object and registers the handle simultaneously, therefore creating a registered digital object without creating a prior stored, but not registered, digital object. (It is possible, of course, to create other useful classes of digital objects. For example, we may define a **proposed digital object** as a digital object whose handle field contains a string that has not yet been registered and whose uniqueness may not yet be known.)

Each repository contains a **properties record** for each of its stored digital objects. The properties record comprises all metadata for a digital object, including its key-metadata, but also, other metadata the repository may maintain for that digital object. Notionally, the key-metadata component is a subset of metadata which is invariant for a digital object over repositories. No attempt is made in this paper to delineate how much of the metadata should be included in the key-metadata, other than requiring that it include the mandatory handle. Possible examples of repository-dependent metadata are the general terms and conditions for access and usage of the digital object, and the date and time of deposit.

A simple **repository access protocol (RAP)** is supported by each repository (and defined in section 3.1). Only the minimal necessary aspects of the RAP are specified here. We anticipate that these aspects of the RAP, or the RAP itself, will be a subset of the interface protocol used by repositories, and require only the functions or operation of the RAP not be affected by any implemented supersets of the protocol. In particular, the RAP allows for accessing a stored digital object or its metadata by specifying its handle, a service request type and additional parameters. If this request is complied with, the output of the service request is termed a **dissemination**. A dissemination is the result of an access service request, along with additional data affixed to it, to be specified below.

An **originator** is an entity that authorizes or validates a set of digital objects; it is responsible for each such digital object including making it available in the System and defining terms and conditions for its use. Every digital object has an originator, which may be an individual or an organization (there may be a number of kinds of originators worth distinguishing, but we do not differentiate them here). Originators may deposit and access the digital objects they authorize or validate and may authorize others to do so (this also includes the right to withdraw or modify the objects), subject to the procedures established by individual repositories. Naming authorities have the right to insert handle entries for handles they generate into the handle server system and to authorize others to do so. The relationship of the originator to the naming authority is left unspecified here. An originator and/or a naming authority may also delegate this authorization ability to others (typically this would be to one or more repositories). Such delegation includes at least the right to authorize the further deposit of digital objects on behalf of the originator and insertion of designated groups of handles on behalf of the naming authority. Repositories may establish additional requirements of various kinds. The process by which an originator or a naming authority informs a repository of any such authorization is left unspecified here.

The initial repository used to deposit a registered digital object is designated the **repository of record (ROR)**. The ROR is responsible for authorizing additional instances of the digital object at other repositories, and for making changes or withdrawals of such additional instances of the digital objects, usually upon the direction of the originator. Once designated, the ROR may subsequently be changed by an authorized party to another repository, but the method for achieving this is not specified here. The notion of ROR is not defined for stored digital objects that are not registered.

A handle is a globally unique string, produced by an authorized **handle generator**. It consists of two logical parts, concatenated with an intervening separator character. The two logical parts are: 1) name of a **local naming authority**, which controls the handle generation process, and 2) a locally unique string, which is assigned by (one of) its handle generator(s). An originator may ask a handle generator for a handle, or it may propose a local string to be used. The local handle generation process should insure that local strings are unique. Handles have no prescribed maximum length in principle, but there will be a default length in existence at any time which can be adjusted upwards if necessary.

For handles to be unique, the names of local naming authorities are controlled by the global naming authority for the System. The global naming authority generates names for local naming authorities, and assigns these to local naming authorities for use by the handle generators they authorize. A prospective local naming authority may propose a name for itself to the global naming authority for validation and registration. A local naming authority, named, say, "X", may create additional, derived naming authorities of the name "X.Y", etc., each authorizing its own handle generator. (At this point, it is left unspecified whether the naming authority name spaces for repositories and for handle generators are distinct.)

In addition to the first globally assigned component (e.g. "X"), each subsequent component field of a naming authority name (e.g. "Y", or "Z") must be non-null and not contain the character ".". There may be other restrictions on the non-alphanumeric characters to be used in naming authority names. In particular, the default separator character is "/" (so, e.g., "X.Y/local-string" is a typical handle from the naming authority "X.Y") Other separator characters, and a syntax for defining another separator characters, (from a restricted class of non-alphanumeric characters) may be defined, and may entail other restrictions on the possible characters used in naming authority names. e.g., a conceivable syntax is to specify a non-default separator by an initial non-alphanumeric character, so that "%X.Y%local-string" is a valid handle. We leave unspecified at this point how this might be accomplished, whether otherwise

identical handles with different separators are identical or distinct, whether an **escape character** for restricted characters exists, and whether the separator characters are restricted (e.g., whether "a/b" is a possible naming authority name that can only be used with a non-default separator). Initially, naming authority names will be issued conservatively, being restricted to alphanumeric characters.

The handle generator may be a person, an organization, or a fully-automated process running on some machine or a set of machines. An originator may control a naming authority, but there may be naming authorities that are not controlled by originators. The details of interaction with handle generators are left unspecified.

It is also unspecified what an originator must supply to a handle generator in order to receive a handle. An originator may propose handles to be assigned to its digital objects. Moreover, the handle generator need not assume any responsibility for insuring that a handle which it generates is associated with any particular digital object; that correspondence may be left to the originator.

A stored digital object may have associated with it in a repository a **transaction record**, which records transactions of that repository involving the digital object. The transaction record may contain entries such as the time and date of deposit of the object, the time and date of each request for retrieval of the object, the identity of the requesting party, the handle and service request for the object, and the applicable terms and conditions including amount and method of payment. Transaction records will only be made available to authorized parties. Repositories are not required to have transaction records persist for any period of time and it may store transaction records at various times and places as deemed necessary subject to administrative controls.

The data of each digital object is typed. Data types assumed to be in the System include **bit-sequence**, **digital-object**, and **handle**, and also **set-of-bit-sequences**, **set-of-digital-objects** and **set-of-handles**. Other data types can be defined and made available to the System via the type construction operators **set-of** and **compose**; these types are then registered in a global type registry. The mechanism for this registration is currently unspecified. Note also that there is, at present, no (defined) registration of methods associated with types.

In contrast, one can create subtypes of digital objects by introducing new fields of metadata; these may be arranged hierarchically. For example, one might create a subtype of digital object called **computer-science-technical-report** which has metadata for **author**, **institution**, **series**, and so forth.

We shall informally refer to digital objects whose data is a set, one of whose elements is of type **digital-object**, as **composite digital objects**. A digital object that is not composite is said to be **elemental**. (Note that this definition explicitly excludes the application of the adjective **composite** to a digital object whose data is another digital object, i.e., whose data is of type **digital-object**, as distinguished from a singleton set of this type. Nothing precludes the existence of such objects, however.)

The terms and conditions of a composite object may implicitly or explicitly be unioned with those of its constituent objects to arrive at the terms and conditions for those constituent objects. Terms and conditions may be explicitly imposed only on the composite object, in which case they would apply to each constituent object; or each constituent may have its own separate terms and conditions in addition. (Of course, creating composite digital objects may be subject to copyright and any other legal restrictions pertaining to its constituent objects.)

A digital object's data may incorporate information or material in which copyright, design patent or other rights or interests are claimed. There may also be rights associated with the digital object itself. An author may have submitted a digital object for purposes of registering a claim to copyright in a work that may be incorporated in the object. Since the copyright pertains to the underlying work fixed in the form of the particular submitted representation, the rights would normally pertain to all representations of the work, including, but not limited to, those representations of the work that are contained in other digital objects.

While we intentionally avoid issues of content in the infrastructure, we note that the entities provided thus far give users a number of means to include digital objects that contain or may be interpreted to manifest the same or similar information or material. As an example, a literary work may be fixed in a number of different formats, e.g., LaTeX, PostScript and GIF page images. Each fixation may correspond to a distinct (elemental) digital object, each with its own unique handle, and other metadata). A composite digital object may then be created whose data is the set of these digital objects. Similarly, one could create a composite digital object whose constituent objects were the fixations of the literary works of Shakespeare in PostScript. The handle of this composite digital object, in effect, names the PostScript collection of Shakespeare's literary works.

Note that it is possible to construct objects with similar effects without using composite digital objects. For example, the single digital object intended to correspond to a work could have data of type **set-of-bit-sequences**, rather than of type **set-of-digital-objects**, and contain each of the forms of fixation therein. In this case, digital objects may not exist corresponding to the individual fixations. Another possibility is to have a digital object whose data is of type **set-of-handles**. In this case, the handles would name the individual fixations (which may not even be available from the same repository). Such a digital object may contain other data fields that further describe (or annotate) the handles. Yet another possibility is to create a markup language which admits handles, plus other conventions for expressing how they relate to each other (for example, whether the individual handles are meant to be interpreted as different fixations of the same work, or a list of bibliographic citations, etc.) A digital object whose data comprise sentences in this markup language could serve to represent the same entities as do composite digital objects.

We use the informal term **meta-object** to refer to a digital object whose primary purpose is to provide references to other digital objects. Both digital objects whose data are of type **set-of-handles** and digital objects in a markup language that admits handles, would be instances of meta-objects.

A digital object may be **mutable** in that it may be changed after it is placed in a repository. Although none of the key-metadata may be changed, nor may any known digital object that it contains be changed (unless the original digital object is also changed), most other changes are permissible. Minor changes might be made to correct a misspelling or other such error; changes to the title of a mutable digital object may be permissible. A mutable composite digital object could be modified to add the representation of an underlying work in a new format. Mutability would also be a useful way to allow digital objects that are designed to change with time or are dynamically computed.

A digital object that cannot be changed is said to be **immutable**. If an object is immutable, then, once it is placed in a repository, the result of all subsequent requests to that repository that are functionally dependent on the data of the object must be identical. (However, it may be possible to remove an immutable object from a repository, or deny access to it at different points in time.) That a digital object is immutable may be reflected in its key-metadata. It is also possible that a given repository may preclude changing a stored object by an indication in its non-key-metadata.

Once set, the mutability or immutability of a digital object cannot itself be changed. Users who wish to achieve a comparable effect would have to create a new digital object with similar data and altered metadata. The original digital object may then be withdrawn or not, as desired.

There is no requirement that a digital object be stored in a repository in any particular manner. Conceptually, the description of a digital object is strictly a logical one and is not intended to describe any particular implementation. In particular, it is possible that, in response to a request to access a particular digital object, a server runs a program that computes the digital object on the fly. It is possible for multiple digital objects to be embedded in a program (e.g., a data base manager or knowledge based system) that emits them upon request. The program may itself be a digital object. Thus, accessing and depositing are virtual processes, and may or may not involve the actual depositing and retrieval of actual objects per se, although such actual storage and retrieval is likely to be prevalent.

3. Accessing Digital objects

3.1. Repository Access Protocol (RAP)

Each repository must support a simple protocol to allow deposit and access of digital objects or information about digital objects from that repository. This is called **Repository Access Protocol**. RAP is meant to provide only the most basic capabilities and may evolve over time. Repositories may support other more powerful query languages that allow users to access objects that meet meaningful criteria. At present, the RAP includes deposit of digital objects, access to digital objects by handle, and related repository services. Each of these capabilities will produce different results, depending on the specific nature of the service request.

(i) Access to a digital object (ACCESS_DO)

Access to a digital object will generally invoke a service program that performs stated operations on the digital object or its metadata depending on the parameters supplied with the service request. Defined service requests include **metadata**, **key-metadata** and **digital object**; the first requests only the metadata, the second only the key-metadata, and the latter, the entire digital object (i.e., the key-metadata and the data). Other systems-level services may be defined. Possible examples of such additional services might be **encrypt**, i.e., return the digital object in some encrypted form, or **compress**, i.e. store a fewer set of bits than supplied with the property that the original bits can be regenerated, perhaps exactly. However, we do not define such additional requests, here.

In addition, it is possible that data-type-dependent service requests will be introduced. Possible examples of such data-type-dependent services requests might be **execute** (for digital objects a portion or all of whose data component is of type **program**), or **subpart** (which requests only a component of the data or metadata of the digital object, further specified by some parameter). We emphasize that such data-type-dependent service requests are not defined as part of the System infrastructure.

When a digital object is accessed via **ACCESS_DO**, the recipient receives a **dissemination**, that is, the result of the service request, along with information such as the key-metadata of the digital object, the identity of the repository, the service request that produced the result, the method of communication (if appropriate) and a transaction string corresponding to an entry in the transaction record. The transaction string is unique to the repository. In addition, the dissemination may contain an appropriately authenticated version of some portion of the properties record for that object, including the specific

terms and conditions that apply to this use of the digital object and the materials contained therein.

As noted above, depending on the nature of the **ACCESS_DO** service request, the dissemination may not be stored as a digital object per se. It might instead include data that is not contained in any registered digital object, such as a portion of a digital object's data, the digital object data in a compressed format, or the result of executing the data of the digital object. In all cases, however, the key-metadata (including, of course, the handle) of the digital object is included.

From a copyright perspective, if the service request produced a dissemination that was derived from a particular digital object, the digital object may be **contained** in the dissemination, in the sense that the dissemination may be encumbered by the rights associated with the digital object. For example, if the data of a stored digital object represents an episode of a television program, and the dissemination contains the data corresponding only to the first two minutes of this television program, the dissemination may be said to contain the digital object in a legal sense, even if it does not properly contain all of its data.

(ii) Deposit of a digital object (DEPOSIT_DO)

Several forms of DEPOSIT_DO are possible. For example, one form may take data, a handle, and perhaps other metadata as arguments, and produce a stored digital object and properties record from these arguments. Another possible form may take a digital object as argument, perhaps with additional metadata, and simply deposit it. Yet another form may take only data and certain non-key-metadata, and automatically request a handle from a handle server, and then simultaneously store the object and register the handle.

The DEPOSIT_DO command may be used to replicate an existing digital object at additional repositories. The exact method of controlling such replication, if any, is unspecified here. A DEPOSIT_DO command may also be used to directly modify an existing mutable digital object. Alternatively, a modified version of an existing digital object may be stored as a new digital object rather than by modifying the existing one.

(iii) Access to reference services (ACCESS_REF)

This command provides a uniform and understood way to identify alternate means of accessing a specified repository and/or information about objects in that repository. Two possible responses are (i) **No information**, and (ii) a list of **servers, protocol-name** pairs, with the interpretation that each server, speaking the named protocol, will provide information about the contents of the repository. (That is, we provide a means of allowing a repository to have its contents indexed, queried, or otherwise described. It is possible, for example, that a repository will be its own provider of information about its contents, and list only itself, and some protocol, as the information provider about its contents. However, it is not required that any accounting of the contents of a repository be available, or that it be available from any one service. This is because we do not require that repositories per se correspond to coherent collections, which may be distributed across independently operated repositories.)

The initial RAP has been purposely kept simple, and all the more complex transactions are assumed to be handled by other protocols, or by subsequent extensions of the RAP. In the first case, a primary use of the RAP for more sophisticated repositories is to have it present the other protocols that it supports (e.g., Z39.50, SQL3, ZQL, Dienst) as alternative access methods.

It may be desirable to extend the RAP in any number of ways, for example, to explicitly include, for example, a payment mechanism or a negotiation mechanism or a more sophisticated interactive model-based interaction mechanism.

Above we described the possibility that a user may construct a single digital object whose data is the set of all fixations (i.e., known formats) of a given work. If so, then there is as yet no formally defined method within the RAP to determine what formats are available, and then, to extract one of them. We expect a set of mechanisms to be developed which expand upon the internal structure of the objects in the infrastructure, but this level of description has intentionally been omitted here.

3.2. The Handle Server Infrastructure

A highly reliable distributed system of **handle servers** is maintained as part of the infrastructure. These servers map handles to network resources at which the corresponding digital objects are available. Handle directory servers are also stipulated; these will be located at certain well known locations and will maintain a table of network addresses of handle servers (generally, each handle server will contain such a directory). This table will generally be downloaded by each participating site frequently enough to be "acceptably " up-to-date at all times. Local handle servers may also exist. A local handle server could be run by an organization if it wishes to keep a store of pertinent handles locally. These local servers may access the global system of handle servers, but are not themselves necessarily accessible from the global system. Caching handle servers also may be run at local workstations on behalf of individual users to store location information for frequently used handles.

The handle server system is intended to be a means of universal basic access to registered digital objects. In the worst case, a user can present a handle to a handle server and be advised of some repository which an authorized party has asserted contains the digital object designated by the handle. The handle server is not meant to be the only, or even primary, means, to locate repositories. Primary access may be provided locally and also by value-added service providers, likely in a variety of different and possibly incompatible ways. Users interacting with such services may not encounter handles; and such services may interact with repositories via RAP or via protocols that do not involve handles.

Handle servers provide a number of services, three of which are RESOLVE, INSERT, and DELETE. A party that is authorized to insert, delete and otherwise change handle entries for a particular naming authority is called a handle administrator. A naming authority will generally designate one or more repositories to act as handle administrators on its behalf. This designation will be made known by the naming authority to the handle server system.

(i) RESOLVE: A handle is sent to a handle server to locate network addresses of repositories containing that object. The handle is first mapped to locate the handle server from the handle directory server table but is not otherwise interpreted. One can also supply a handle to a separate system, which invokes the above procedures to find the stated object. Local handle servers may use any technique to do the mapping. The handle servers maintained as part of the infrastructure map the handles by hashing them.

No guarantee is made that the identified repositories will provide the designated object. Rather, the user is assured only that the specified repositories are where authorized maintainers of repository services have indicated particular digital objects reside.

Since a handle is just a unique string, it can be mapped to an actual repository by any of several mechanisms, including a mechanism that attempts to interpret the string. Repository names are not

actual network addresses; they must first be mapped to network locations. The method for accomplishing these mappings is not specified. The handle service is one available means for both kinds of mappings; it would specify at least the location of the interface that supports the RAP protocol for a given repository. There may also be a need to explicitly provide a country identifier for repositories, naming authorities and/or originators. For the present, however, country identifiers are omitted.

When a repository is identified by a handle server, it will be most efficient to map the handle directly into the network address (or addresses) of the repository. This mapping avoids having to do a double lookup from repository name to repository location. However, if the location of the repository were to change, the handle server would have to be notified so it could make the corresponding changes. It is possible that certain repository names may resolve to broadcast addresses to locate specific machines. This might be the case where a single repository consists of multiple machines on a local area network at a given site. The handle administrator may determine whether to store IP addresses or domain names or other information in the handle server. The entries are typed and therefore one or more of the above information types may be provided by the administrator for retention in the handle server.

(ii) INSERT (DELETE): Information associating handles with network services are inserted into (deleted from) the handle server system by the handle administrator or other parties authorized by it. Such authorized parties include repositories of record. The repository of record is presumed to make known to the handle server system that it contains (or no longer contains) a particular digital object some reasonable time after the digital object is deposited in (withdrawn from) it. Similarly, the repository of record would make known to the handle server system the identity of other repositories which it authorizes to store a given digital object. The handle server system may perform certain administrative functions upon receipt of unauthorized requests. In addition, some form of reporting may be desirable to insure that entities that misbehave can be detected.

3.3 Value-added Reference Services

The handle server system is intended as a **safety net** of information about where digital objects reside. There will no doubt be other, valuable services that provide information to users about the location of digital objects in repositories. However, we do not consider these services per se to be part of the infrastructure of the System. Instead, they comprise value-added services whose nature we do not see as appropriate to constrain.

In addition, as mentioned above, we do not require repositories to provide a description of their contents. Repositories may not house coherent collections, and hence, querying or searching a repository may be a service appropriate only to the repository administrator, not to a user. Presumably, such capabilities will exist in the form of value-added services. It is such services, rather than repositories per se, that users would interrogate to identify digital objects of a certain nature. Such services may, of course, be offered by repositories themselves, especially in the case when one is intended to house a coherent collection. However, such a server is not a requirement of a well-behaved repository.

4. Imposing Semantics on Handles

As discussed above, a handle is presumed to have two logical components, a local naming authority name, and an identifier unique to that naming authority. These naming authorities will be assigned in a manner. For example, there may be a "naming authority" named "berkeley", which will authorize other naming authorities within the "berkeley" domain. Within the "berkeley" domain, names are locally assigned to other naming authorities. Thus, the name "berkeley.cs" might be assigned to the authority

responsible for naming the UCB Computer Science technical report series (or to several such series). Note that this particular naming authority will not generally correspond to a valid Internet address, even though it may follow similar syntactic conventions.

Particular naming authorities may follow their own conventions for assigning semantic or non-semantic strings to their objects. For example, "berkeley.cs" may follow a proposed convention for its technical reports, and give each of the corresponding digital objects (whether composite objects or meta-objects) a local handle, e.g., "csd-93-712". (The "csd" -- for "Computer Science Division" is perhaps redundant; however, we use it here to indicate the possibility of a single naming authority issuing several distinct series.)

The full unique handle for this digital object would be

```
berkeley.cs/csd-93-712
```

where the "/" separates the naming authority name from the string unique to that authority.

In addition, digital objects may exist for this work in each of a number of fixations (formats). The handles for these fixations may also be semantically interpretable, e.g., the string "csd-93-712/all.ps" might be the unique local part of the handle for the digital object corresponding to the PostScript version of this work; "csd-93-712/all.tif" the handle for the tiff representation. (Note that the character "/" is allowed in the local name. It may also be desirable to distinguish other characters, but this is not discussed further in this paper.)

Other schemes may be used to generate handles in other ways. For example, the local portion of a handle might correspond to a date- time format, so that the digital object above might instead have the handle

```
berkeley.cs/1994.12.05.23.42.12;7
```

These handle forms can be embedded within various syntactic wrappers to distinguish them in various contexts from other notations. For example, the handle might be expressed in URN syntax as follows:

```
<URN:ASCII:ELIB-v.2.0:berkeley.cs/csd-93-712>
```

Here "ELIB-v.2.0" is supposed to suggest (via "ELIB") that this is a URN for electronic library material, and also, (via "-v.2.0") that some particular naming convention is used by the naming authority. Another possibility is the notation used by Grass and Arms (GA1994), which resembles that for URLs, and proceeds that handle with the prefix "hdl://" (to denote that a handle follows), or just "/" (if it is important to distinguish a global root for the handle), e.g.: hdl://berkeley.cs/csd-93-712

```
//berkeley.cs/1994.12.05.23.42.12;7
```

The user of this notation is cautioned to avoid confusion with URLs, which name services, while handles name digital objects, not network services.

Various services might exploit semantic conventions to locate an object given its handle, without consulting a handle server. For example, a naming authority may have its own repository and reference server associated with it; the latter might be looked up (perhaps via an additional service), and queried for the location(s) of this particular report.

Users may, of course, attempt to incorporate all manners of semantic or system content in handles. Also,

it is plausible that imposing any content in handles per se could be troublesome. Instead, handles per se could be declared to be uninterpreted, and an additional level of indirection be introduced to interpret them. Additional name services could be created to translate user-oriented **nicknames** to system-oriented handles, as are done for file systems today. We stop short of advocating such a system here, however, assuming that a semantically-motivated convention, such as that which has served for URLs, will continue to be useful at some level, and does not require an additional level of mediation.

5. Conclusion and Summary

This paper provides a method for naming, identifying and/or invoking digital objects in a system of distributed repositories that provides great flexibility and is well-suited to a national-level enterprise. It allows the possibility of locating digital objects without making any presumptions about the object or its locations(s). It also admits value-added conventions that various users may use to their own advantage. For example, a reference server might internally refer to an object by its global handle, and, additionally, keep track of repositories in which this object is believed or known to reside. If a user requests this object, the reference server might look up the repository name or address, determine the repository service, and ask that repository to deliver a version of the object to the user. Alternatively, the server might instead use the object's handle at run time to query syntactically a handle server for the name of repositories or services that house the object.

This system also allows for **public** and **private** naming authorities. Many naming authorities will be private, and only assign identifiers to their chosen clientele (e.g., department members eligible to produce technical reports); however, public naming authorities could provide a service whereby they generate an identifier to anyone who requests one. Individual citizens not associated with any official body might use a public naming authority to generate identifiers for objects they wish to store for private purposes or for public dissemination on their own (this is an example of a situation in which the originator does not control the naming authority.)

In the CS-TR project, CNRI is providing the global naming authority plus a handle management service that accepts handles with and without semantics. This service does not make use of handle semantics; however participants are able to take advantage of handle semantics, if any, to access objects directly. Each participating institution would be free to propose or request names of its own choice. Each of these names may also have associated with them a non-semantic identifier (such as a date-time-stamp) which is not otherwise specified in this document.

Acknowledgments

This research was supported by the Advanced Research Projects Agency under Grant No. MDA-972-92-J-1029 with the Office of Naval Research. We would like to thank Jerry Saltzer, Michael Stonebraker, Jim Davis, Carl Lagoze, Bill Arms, Hector Garcia-Molina, Jim Gray, Patrice Lyons, David Ely, Judy Grass, Barry Leiner, John Garrett and all the members of the CS-TR project for their many helpful comments on and insights into this work.

cnri.dlib/tn95-01

wya

5/13/95



Digital Object Architecture Project

CNRI's program of research and development in digital libraries has a number of inter-related activities that overlap and build upon each other. The work includes development of core technology that is used in several testbeds and implementation projects, with funding from a variety of sources.

The Digital Object Architecture Project continues the architectural work of the DARPA-funded [Computer Science Technical Reports Project](#) (CSTR). This was a cooperative project, led by CNRI, with five major universities. The project developed network access to archives of technical information in the domain of computer science, and carried out related research, with the goal of evolving knowledge in the field of information storage, search, and retrieval.

The objectives of the Digital Object Architecture project are to enhance the architecture, to continue development of the core technology, and to demonstrate and evaluate them in a number of large-scale testbeds.

Support for the Digital Object Architecture project is provided by DARPA, the Library of Congress, and the Defense Technical Information Center (DTIC), through DARPA grant MDA972-92-J-1029.

Technology

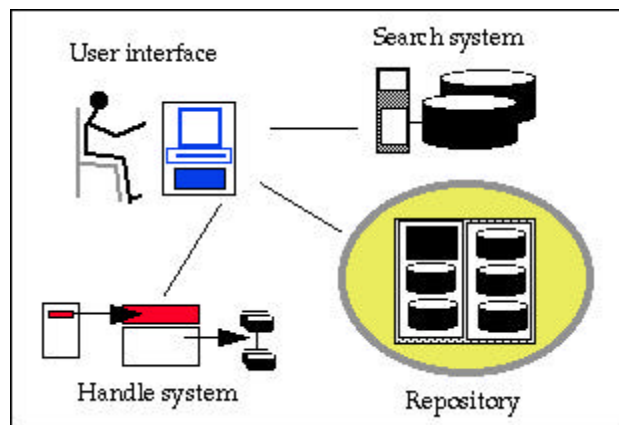


Figure 1

Figure 1 shows the principal system components. CNRI's research concentrates on the concept of digital objects, the Handle System for identifying digital objects, and the Repository for storing them and making them available over the Internet. The Registry is a specialized repository that is used to authenticate digital objects.

The Handle System is a system for providing persistent names for Internet resources. It is a highly reliable, high performance, distributed system.

The Repository Provides network based storage and access to digital objects. All access to digital objects passes uses a simple repository access protocol and is subject to access controls established by the manager of the repository.

The Registry is a specialized repository that provides secure registration and authentication of digital objects.

Applications, Testbeds, and Partners

U. S. Copyright Office (CORDS). This system provides copyright registration and deposit of digital materials over the Internet. When completed, it will integrate the Registry, Handle System, and Repository with the production systems at the Library of Congress.

Library of Congress (NDLP). The National Digital Library Program is a large-scale program to digitize and make available over the Internet materials from the historic collections at the Library of Congress. CNRI is providing the Repository to manage the collections and the Handle System to identify the digital objects.

Defense Virtual Library. CNRI is working in partnership with the Defense Technical Information Center (DTIC) to design and development a digital library for DTIC's extensive collection of report literature.

Papers

A Framework for Distributed Digital Object Services by Robert Kahn and Robert Wilensky, May 1995

Key Concepts in the Architecture of the Digital Library by William Y. Arms, D-Lib Magazine, July 1995

"Implementation Issues in an Open Architecture Framework for Digital Object Services" by Carl Lagoze and David Ely. Cornell Computer Science Technical Report TR95-1540

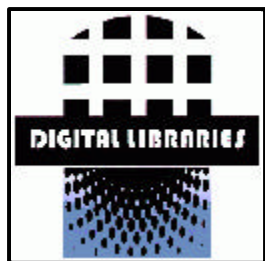
"A Design for Inter-Operable Secure Object Stores (ISOS)" by Carl

Lagoze, Robert McGrath, Ed Overly, Nancy Yeager. Cornell Computer Science Technical Report TR95-1558

[Uniform Resource Names: A Progress Report](#) The URN Implementors, D-Lib Magazine, February 1996

[An Architecture for Information in Digital Libraries](#) William Y. Arms, Christophe Blanchi, Edward A. Overly, D-Lib Magazine, February 1997

[[home](#) | [about CNRI](#) | [programs & activities](#) | [publications](#)]



[THE CENTER](#)

[FACILITIES](#)

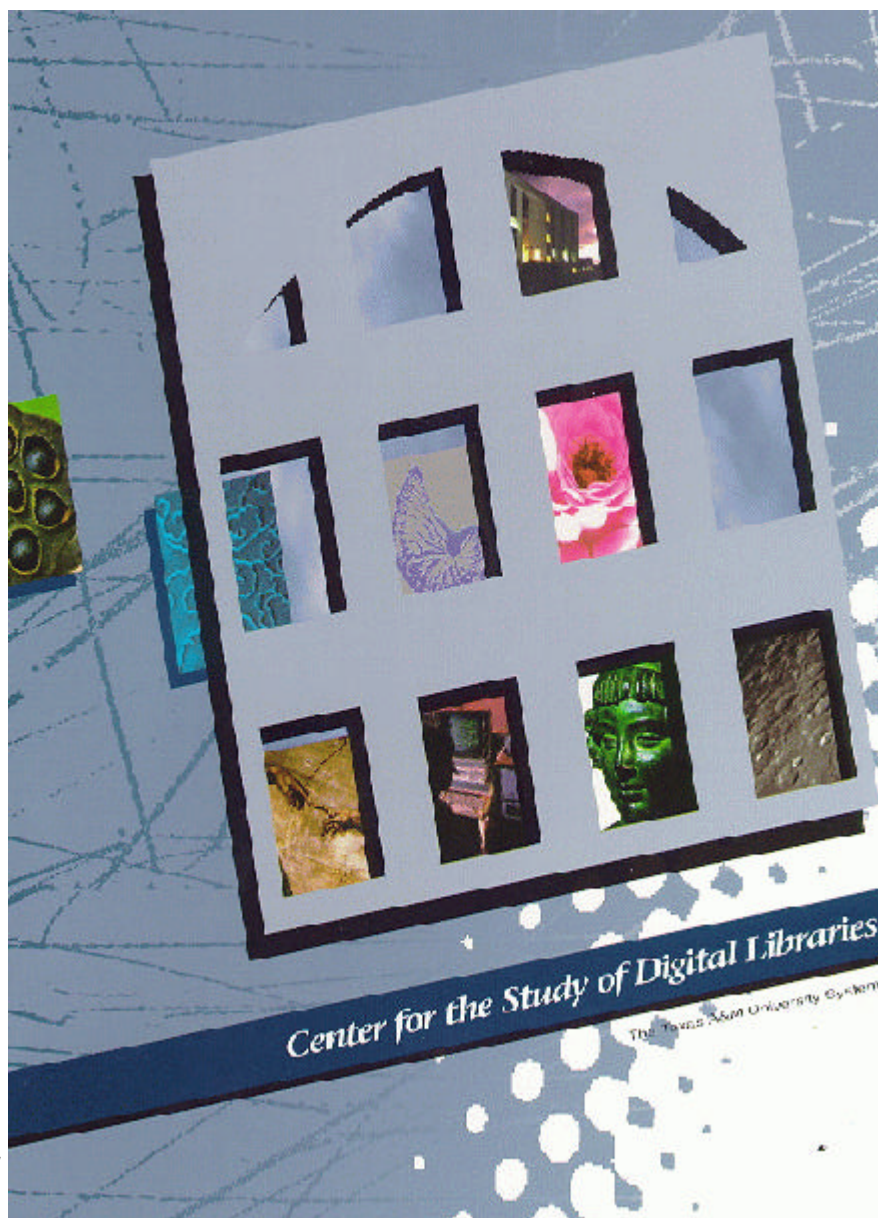
[RESEARCH](#)

[FACULTY/STAFF](#)

[COURSES](#)

[PUBLICATIONS](#)

[CONFERENCES](#)



Center for the Study
of Digital Libraries
Texas A&M
University
College Station,
Texas, USA
77843-3112
Telephone:
01-409-862-3217
Fax:
01-409-847-8578
csdl@csdl.tamu.edu



HEWLETT
PACKARD



The Center for the Study of Digital Libraries gratefully acknowledges the corporate support of the
Hewlett-Packard Company; Informix Software, Inc.; and Knowledge Systems, Inc.

[home](#)[feedback](#)[join/renew](#)[go shopping](#)[search](#)

- [Tables of Contents](#)
- [Search the Digital Library](#)
- [Content Updates](#)
- [Digital Library FAQ](#)
- [Organization of the Digital Library](#)
- [Usage of the Digital Library](#)
- [ACM Members: Create an Account](#)
- [Visitors: Register](#)
- [Download Acrobat® Reader 3.0 to read full-text articles](#)

Tap into the ACM Digital Library, a vast resource of bibliographic information, citations, and full-text articles. Access to full-text is by subscription only: ACM members who are Digital Library subscribers have access to all full-text articles. Members and nonmembers who subscribe to electronic publications (but not to the entire Library) have full-text access to their subscriptions only.

If you are not yet a subscriber, you can still use the Digital Library: As a service to the computing community, the Digital Library will continue to offer its search and bibliographic database resources to all visitors, for free. All you need to do is register with us.

Highlights:

[How to subscribe to the Digital Library](#)

Rate information on joining ACM and subscribing to the Digital Library

[Journal & Magazine Issues](#)

[Not Yet in Full Text](#)

Listing of [Conference Proceedings by Sponsoring SIG](#)

[ACM Document Delivery Service](#)

To purchase articles published by ACM prior to 1991

What's New:

- [1997 Proceedings in the Digital Library](#)
- Please Read: [Subscription-based Access](#)

[SEARCH THE DIGITAL LIBRARY](#) || [TABLES OF CONTENTS](#) || [REGISTER WITH US](#)
[ABOUT THE DIGITAL LIBRARY](#) || [FEEDBACK](#) || [DIGITAL LIBRARY HOMEPAGE](#)

Welcome to the IEEE Computer Society Digital Library News (DLN) Archive

This web site will provide access to back issues of the IEEE Computer Society Digital Library Newsletter. Details on subscribing to the newsletter are given below.

Presentation

[Presentation of IEEE Computer Society DLN](#)

Archives

- [Initial Announcement of IEEE Computer Society DLN](#)
 - [June/July 1997 Volume 1 Number 1](#)
 - [January 1998 Volume 1 Number 2](#)
-

Points of Contact

Articles of Interest

and general comments may be directed to:

Sue Feldman
Datasearch
170 Lexington Dr.
Ithaca, NY 14850
607-257-0937 Phone/Fax
sef2@cornell.edu

Subscription Information

To subscribe to DLN, send an e-mail message to: ieeedln@cimic.rutgers.edu with the contents:

subscribe ieedln Your Real Name

Technical Questions

about this web server or the mailing list can be directed to: holowcza@cimic.rutgers.edu

Page: <http://cimic.rutgers.edu/~ieeedln/index.html> Updated: Thu Jun 26 11:58:21 EDT 1997

Technical Activities Forum

Technical Activities Board (TAB) coordinator: Deborah Scherrer, 3061 Palomares Rd., Castro Valley, CA 94552; voice (510) 881-4489; dscherrer@computer.org

Digital Library Task Force

Nabil R. Adam and Richard Holowczak
Rutgers University, CIMIC

Milton Halem and Nand Lal
NASA Goddard Space Flight Center

Yelena Yesha
UMBC/Center for Excellence in Space Data and Information Sciences

In the past, global networks have usually transported textual information, but there is a growing need for these networks to transport other forms of information such as images, video, and audio. Until recently, electronic information sources served mainly specialized clients, but now these sources will be accessed by a wide range of users, ranging from computer specialists, discipline experts, engineers, and the general public, including novice computer users and students at all levels.

These trends have created an emerging, important discipline: digital libraries. Several US agencies, including NASA, ARPA, and NSF, have made available over the past few years a considerable amount of money to support research in this field. Other countries, including Canada, the UK, France, Italy, and the Netherlands have also invested in digital library development:

National Library of Canada Electronic Collection: <http://www.nlc-bnc.ca/eppp/e-coll-e.h>
Initiative for Access—British Library Board: <http://portico.bl.uk/access/overview.html>
International Institute for Electronic Library Research (involved in several projects):
<http://ford.mk.dmu.ac.uk>.
Elite Project (Italy): <http://cosimo.ing.unifi.it/research/elite/elitinfo.html>

As a result of these activities, a number of recent symposia, workshops, and conferences have been recently devoted to digital library issues, and several journals have published editorial about digital libraries, including Computer and IEEE Transactions on Knowledge and Data Engineering.

Technical challenges

Digital library development faces challenges in several areas, including the subdisciplines we summarize here.

- **Storage.** A digital library's storage system must be capable of storing a large amount of data in a variety of formats and accessing this data as quickly as possible. Text-only documents—stored in formats such as ASCII, LaTeX, HTML, SGML, and PostScript—are by far the easiest to store. Digital audio and video are more difficult to store because they require significantly more storage

space and their delivery is time-dependent.

A typical digital library uses a variety of database-management systems. Current DBMSs range from relational and extended relational systems to object-oriented database systems. Relational DBMSs are most often used for the storage of metadata and indexes with attributes that contain pointers to files in a file system. Most of the commercial RDBMSs also support the storage of Binary Large Objects (BLOBs); in an Oracle RDBMS, BLOBs can be as large as 2 Gbytes. Object-oriented database systems are slowly gaining acceptance and overcoming earlier performance and implementation problems. An OODBS can make it easier to model, store, and work with real-world objects such as images or maps.

Compression techniques save storage. For text-only documents, the Unix compress or freeware gzip utilities provide anywhere from 10- to 60-percent compression. Several compression standards exist for digital images (JPEG), audio (uLaw), and video (MPEG).

Digital library collections that are too large to store entirely on a disk use hierarchical storage mechanisms. In an HSM, the most frequently used data is kept on fast disks while less frequently used data is kept in near-line such as an automated (for example, robotic) tape library. Using data-usage statistics, the HSM can automatically migrate data from tape (near-line) storage to disk (on-line) and back, as required.

- **User interface.** The user interface, perhaps the most important digital library component, must incorporate a wide variety of techniques to afford rich interaction between users and the information they seek. For computer workstations, graphical user interfaces such as X-Windows, Microsoft Windows, and Macintosh System 7 interfaces are the status quo.

A user interface for digital libraries must display large volumes of data effectively. Typically the user is presented with one or more overlapping windows that can be resized and rearranged. In digital libraries, a large amount of data spread through a number of resources necessitates intuitive interfaces for users to query and retrieve information. The ability to smoothly change the user's perspective from high-level summarization information down to a specific paragraph of a document or scene from a film remains a challenge to user interface researchers.

- **Classification and indexing.** Classification and indexing schemes are used to collect related content into groups that are intuitive to a user. Classifying and indexing objects is filled with pitfalls, however, because individual perceptions vary. Another complicating factor in indexing and classifying is the tremendous amount of potential content that remains to be indexed. It is clear that manual methods for classification are insufficient for all but the most trivial digital library.

Automated classification systems differ significantly in their approaches, depending on the type of content under consideration. Classifying short stories is quite different from classifying maps, both in terms of the mechanics involved and the appropriate classes. These distinctions make current automated classification efforts highly domain-specific.

Automated document classification methods can be grouped into two general approaches, but neither can yet capture the meaning of words in the documents. Image classification approaches are conceptually different from those used for text classification. Although many domain-specific systems allow "content-based" querying, most are relegated to a very narrow range of images and

may require the services of human classifiers. Video classification and indexing requires systems that can parse video into manageable portions, typically called camera shots. As with image classification, the type of classification and indexing performed on video is driven by the types of queries posed by users. The classification of audio, musical notation, and maps presents additional research challenges.

- **Information retrieval.** The concepts underlying information retrieval were conceived long before computers and information systems were employed to store library materials. In the digital library domain, there are a variety of information-retrieval techniques, including metadata searching, full-text document searching, and content searching for other data types.

The success of information retrieval can be measured in terms of the percentage of relevant and extraneous information retrieved. It is difficult to pinpoint quantitatively the effectiveness of information retrieval; only an individual user can determine what is truly useful. Techniques to improve retrieval effectiveness include preprocessing documents to extract additional metadata before storing them in a document database.

Several researchers are focusing on automating the creation and maintenance of user profiles and applying these profiles to information retrieval. Software agents are an extension of filtering techniques, although filtering tends to imply passive mechanisms whereas the use of agents implies a more proactive approach. Many people have put forth definitions of software agents, ranging from an adaptable information filter to an autonomous program that works in conjunction with or on behalf of a human user. Software agents also embody the notion of improving over time as they record additional user actions and reactions.

- **Content delivery.** Once an item of interest has been located in a digital library, it may be delivered in several ways. If the content is small, such as 100 pages of text or a 50-Kbyte image file, it may be delivered through the same channel used for information retrieval and querying. Content such as movies and software, however, demands much higher bandwidth. In these cases, delivery is over dedicated leased lines (for example, cable TV or videoconferencing systems) or satellite-based systems such as the Hughes Network Systems project (<http://www.hns.com>)

Increased demands for networking bandwidth come from two main fronts. First, the number of digital library users will undoubtedly increase. If the Internet is any indication, exponential growth in the number of users will be the rule. Second, as the delivery of multimedia data becomes the norm, the demands for high bandwidth increase. However, high bandwidth, in and of itself, is not enough to support digital libraries. The intelligent use of bandwidth and the ability to guarantee bandwidth for a given time period are also required.

Today's open networking standards such as TCP/IP and the ever-growing Internet make it clear that successful digital libraries must be built on an open, interoperable networking infrastructure. Current digital libraries may be run exclusively on a single computer, on several computers connected on a LAN, or on a large number of computers spread out over a wide area network. Delivery systems that require high bandwidth such as video and image libraries are predominantly installed using LANs that run at 10 to 100 Mbits per second. In contrast, the Internet's major backbones run at 1.5 Mbps to 150 Mbps, while links to individual organizations fall in the 56 Kbps to 1.5 Mbps range. Individual users typically connect to the Internet through service providers, local universities, or other organizations at 2.4 Kbps to 28.8 Kbps.

- **Presentation.** Users of a traditional library usually want to read a book or watch a videotape; other uses are rare. With digital technology, it now becomes possible to listen to a book being read, watch a video of a musical performance alongside the original score, or hold a mechanical hand as it forms American Sign Language. Other possible uses are highly personal—individuals may dream up many distinct variations. A digital library's presentation systems must be flexible and highly customizable. They must also be aware of the output hardware's capabilities and limitations, automatically adjusting to deliver the best possible presentation quality at all times.
- **Administrative.** Traditional libraries store a final copy of a book or other documents. Digital libraries store several versions of a document in a way that makes multiple revisions by multiple authors possible. In addition, the content for a digital library may have multiple owners in terms of the sources of the content and annotations made to the content of the library. An administrative system ensures that materials intended for public viewing can indeed be viewed by anyone while private collections and personal annotations may only be viewed by a select group or single individual. And data-versioning techniques track the history of such revisions.

There may also be times when a small group of individuals want access to a portion of digital library content such as when authors are preparing initial drafts of a document. In these cases, security mechanisms must be put into place to ensure that only authorized users gain access. Current digital libraries employ the basic security measures offered by the supporting operating systems. For example, any digital library running on Unix can restrict access using username and password authentication and protect files using group membership and file-access rights. This basic security will not meet the demands of large-scale digital libraries.

Finally, digital libraries must protect the identity of their users, who may wish to browse content that may be embarrassing.

Task Force on Digital Libraries

In 1995, the IEEE Computer Society established the Task Force on Digital Libraries as a first step leading to a full-fledged Technical Committee. The task force is to promote research in the theory and practice of all aspects of digital libraries.

The task force sponsors activities that benefit its members and profession. Such activities include sponsoring and cosponsoring symposia, sessions in large conferences, tutorials, and a newsletter, edited by Prof. Erich Neuhold, GMD-IPSI. Send newsletter contributions (news, brief articles, conference announcements) to neuhold@ darmstadt.gmd.de. The task force cosponsored the Forum on Research and Technology Advances in Digital Libraries, held last May at the Library of Congress and is cosponsoring the [International Journal of Digital Libraries](#), which Springer Verlag will begin publishing this year.

The executive committee of the task force includes Nabil R. Adam (chair), Rutgers University; David Choy, IBM Almaden Research Center; Milton Halem, NASA Goddard Space Flight Center; Nahum Gershon, Mitre; Erich Neuhold, GMD-IPSI; and Yelena Yesha, UMBC/CESDIS.

Membership in the Task Force on Digital Libraries is free. We invite you to join and contribute ideas, suggestions, comments, and time. For more information, see our home page at http://cimic3.rutgers.edu/ieee_dltf.html or through the IEEE Computer Society's home page at <http://www.computer.org>, or send e-mail to adam@adam.rutgers.edu.



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

[News](#)[About OCLC](#)[OCLC Services](#)[Support & User Doc.](#)[Contacts & Addresses](#)

OCLC Research

OCLC identifies research and technical advances that are of value to the organization in meeting its corporate purposes. Some research projects may lead directly to the development of services for OCLC members, and others may be of general use to the library and information science community. Much of this research and technology assessment is conducted internally, but OCLC also funds or otherwise supports research conducted at universities or other research centers that furthers the corporate purposes of OCLC. OCLC is engaged in a number of research projects in the following areas:

1. to make more effective use of WorldCat (the OCLC Online Union Catalog);
2. to explore the concept of the electronic library;
3. to determine the feasibility of image processing applications; and
4. to improve interface design and human/computer interaction.

► [Mission Statement](#)

► [Projects and Publications](#)

► [Related Conferences](#)

► [Staff](#)

► [Links to Related Web Sites](#)

► [Search OCLC Research](#)



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

News

About OCLC

OCLC Services

Support & User Doc.

Contacts & Addresses

OCLC Reference Services

Considering SiteSearch?

- [Overview](#)
- [Guided Tour](#)
- [Demonstrations](#)
- [Components](#)
- [Case Studies & Current Users List](#)
- [How to Order](#)

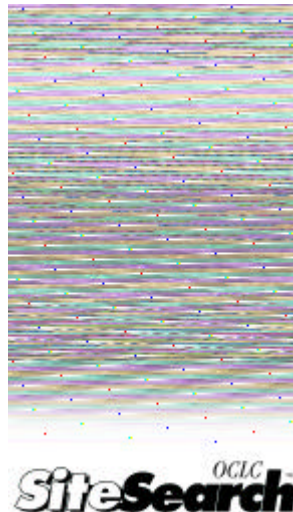
OCLC SiteSearch

The OCLC SiteSearch suite provides a comprehensive solution for managing distributed library information resources in a World Wide Web environment. It offers tools that **integrate** electronic resources under one Web interface, control **access** to resources, and **build** text and image databases locally.

New OCLC SiteSearch was [featured](#) in the January/February 1998 issue of the [OCLC Newsletter](#).

Using SiteSearch?

- [Support](#)
- [News](#)
- [Product Requirements](#)
- [Documentation & Technical Updates](#)
- [Training](#)



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

Interfaces for Information Access

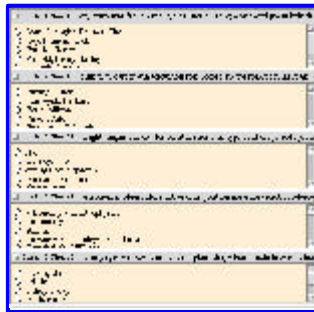
Companion Pages for *Scientific American* Article [*Interfaces for Searching the Web*](#)

The field of Information Access concerns helping people find, use, understand, and create the information they need, often using computer systems as tools. Information can be found in many forms and media, although much of our research has been concerned with text in general, not focusing exclusively on the Web.

Text analysis and user interface technology must be combined with an understanding of how users work with information and computer tools when building systems to support information access.

Currently, these pages provide additional information about some of the ideas discussed in the *Scientific American* article *Interfaces for Searching the Web* by [Marti Hearst](#). There is a great deal of research in Information Access at [Xerox PARC](#), of which this pages show only a small sample.

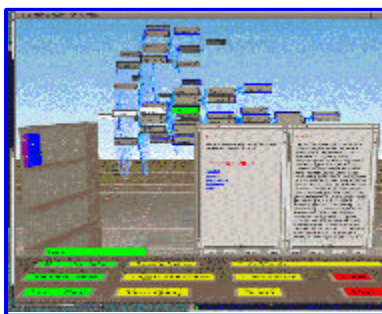
[About Scatter/Gather](#)



[About Tilebars](#)



[About the Cat-a-Cone](#)



Xerox PARC
hearst@parc.xerox.com
6/9/97



SPECIAL REPORT

Interfaces for Searching the Web

The rapid growth of the World Wide Web is outpacing current attempts to search and organize it.

New user interfaces may offer a better approach

by [Marti A. Hearst](#)

SUBTOPICS:

- [The \(Slow\) Speed of Thought](#)
- [Organizing Search Results](#)

[FURTHER READING](#)**[BACK TO THE INTRODUCTION](#)**

How does anyone find anything among the millions of pages linked together in unpredictable tangles on the World Wide Web? Retrieving certain kinds of popular and crisply defined information, such as telephone numbers and stock prices, is not hard; many Web sites offer these services. What makes the Internet so exciting is its potential to transcend geography to bring information on myriad topics directly to the desktop. Yet without any consistent organization, cyberspace is growing increasingly muddled. Using the tools now available for searching the Web to locate the document in Oregon, the catalogue in Britain or the image in Japan that is most relevant for your purposes can be slow and frustrating.

More sophisticated algorithms for ranking the relevance of search results may help, but the answer is more likely to arrive in the form of new user interfaces. Today software designed to analyze text and to manipulate large hierarchies of data can provide better ways to look at the contents of the Internet or other large text collections. True, the page metaphor used by most Web sites is familiar and simple. From the perspective of user interface design, however, the page is unnecessarily restrictive. In the future, it will be superseded by

more powerful alternatives that allow users to see information on the Web from several perspectives simultaneously.

Consider Aunt Alice in Arizona, who connects to the Net to find out what kind of edible bulbs, such as garlic or onions, she can plant in her garden this autumn. Somewhere in the vast panorama of the Web lie answers to her question. But how to find them?

Alice currently has several options, none of them particularly helpful. She can ask friends for recommended Web sites. Or she can turn to Web indexes, of which there are at present two kinds: manually constructed tables of contents that list Web sites by category and search engines that can rapidly scan an index of Web pages for certain key words.

Using dozens of employees who assign category labels to hundreds of Web sites a day, Yahoo compiles the best-known table of contents. To use Yahoo, one chooses from a menu [see illustration at far left] the category that seems most promising, then views either a more specialized submenu or a list of sites that Yahoo technicians thought belonged in that section. The interface can be awkward, however. The categories are not always mutually exclusive: Should Alice choose "Recreation," "Regional" or "Environment"? Whatever she selects, the previous menu will vanish from view, forcing her either to make a mental note of all the alternative paths she could have taken or to retrace her steps methodically and reread each menu. If Alice guesses wrong about which subcategory is most relevant (it is not "Environment"), she has to back up and try again. If the desired information is deep in the hierarchy, or is not available at all, this process can be time-consuming and aggravating.

The (Slow) Speed of Thought

Research in the field of information visualization during the past decade has produced several useful techniques for transforming abstract data sets, such as Yahoo's categorized list, into displays that can be explored more intuitively. One strategy is to shift the user's mental load from slower, thought-intensive processes such as reading to

faster, perceptual processes such as pattern recognition. It is easier, for example, to compare bars in a graph than numbers in a list. Color is very useful for helping people quickly select one particular word or object from a sea of others.

Another strategy is to exploit the illusion of depth that is possible on a computer screen if one departs from the page model. When three-dimensional displays are animated, the perceptual clues offered by perspective, occlusion and shadows can help clarify relations among large groups of objects that would simply clutter a flat page. Items of greater interest can be moved to the foreground, pushing less interesting objects toward the rear or the periphery. In this way, the display can help the user preserve a sense of context.

Such awareness of one's virtual surroundings can make information access a more exploratory process. Users may find partial results that they would like to reuse later, hit on better ways to express their queries, go down paths they didn't think relevant at first--perhaps even think about their topic from a whole new perspective. Aunt Alice could accomplish a lot of this by jotting down notes as she pokes around Yahoo, but a prototype interface developed by my colleagues at the Xerox Palo Alto Research Center aims to make such sense-making activities more efficient.

Called the [Information Visualizer](#), the software draws an animated 3-D tree that links each category with all its subcategories. If Alice searches the Yahoo tree for "garden," all six areas of Yahoo in which "garden" or "gardening" is a subcategory will light up. She can then "spin" each of these categories to the front to explore where it leads. When one path hits a dead end, the roads not taken are just a click away.

When Alice finds useful documents, this interface allows her to store them, along with the search terms that took her to them, in a virtual book. She can place the book on a virtual bookshelf where it is readily visible and clearly labeled. Next weekend, Alice can pick up where she left off by reopening her book, tearing out a page and using it to resubmit her query.

Our interface does not offer much help to the Sisyphean attempt to organize the contents of the entire Web. Because new sites appear on the Web far faster than they can be indexed by hand, the fraction listed by Yahoo (or any other service) is shrinking rapidly. And sites, such as Time magazine's, that contain articles on many topics often appear under only a few of the many relevant categories.

Search engines such as Excite and [AltaVista](#) are considerably more comprehensive--but this is their downfall. Poor Aunt Alice, entering the string of key words "garlic onion autumn fall garden grow" into Excite will, as of this writing, retrieve 583,430 Web pages, which (at two minutes per page) would take more than two years to browse through nonstop. Long lists littered with unwanted, irrelevant material are an unavoidable result of any search that strives to retrieve all relevant documents; conversely, a more discriminating search will almost certainly exclude many useful pages.

The short, necessarily vague queries that most Internet search services encourage with their cramped entry forms exacerbate this problem. One way to help users describe what they want more precisely is to let them use logical operators such as AND, OR and NOT to specify which words must (or must not) be present in retrieved pages. But many users find such Boolean notation intimidating, confusing or simply unhelpful. And even experts' queries are only as good as the terms they choose.

When thousands of documents match a query, giving more weight to those containing more search terms or uncommon key words (which tend to be more important) still does not guarantee that the most relevant pages will appear near the top of the list. Consequently, the user of a search engine often has no choice but to sift through the retrieved entries one by one.

Organizing Search Results

A better solution is to design user interfaces that impose some order on the vast pools of information

generated by Web searches. Algorithms exist that can automatically group pages into certain categories, as Yahoo technicians do. But that approach does not address the fact that most texts cannot be shoehorned into just one category. Real objects can often be assigned a single place in a taxonomy (an onion is a kind of vegetable), but it is a rare Web page indeed that is only about onions. Instead a typical text might discuss produce distributors, or soup recipes, or a debate over planting imported versus indigenous vegetables. The tendency in building hierarchies is to create ever more specific categories to handle such cases ("onion distributors," for example, or "soup recipes with onion," or "agricultural debates about onions," and so on). A more manageable solution is to describe documents by whole sets of categories that apply to them, along with another set of attributes (such as source, date, genre and author). Researchers in Stanford University's digital library project are developing an interface called [SenseMaker](#) along these lines.

At [Xerox PARC](#), we have developed an alternative scheme for grouping the list of pages retrieved by a search engine. Called [Scatter/Gather](#), the technique creates a table of contents that changes along with a user's growing understanding of what kind of documents are available and which are most relevant.

Imagine that Aunt Alice runs her search using Excite and retrieves the first 500 Web pages it suggests. The Scatter/Gather system can then analyze those pages and divide them into groups based on their similarity to one another [see upper illustration on next page]. Alice can rapidly scan each cluster and select those groups that appear interesting.

Although evaluation of user behavior is an inexact process that is difficult to evaluate, preliminary experiments suggest that clustering often helps users zero in on documents of interest. Once Alice has decided, for example, that she is particularly keen on the cluster of 293 texts summarized by "bulb," "soil" and "gardener," she can run them through Scatter/Gather once again, rescattering them into a new set of more specific clusters.

Within several iterations, she can whittle 500 mostly irrelevant pages down to a few dozen useful ones.

By itself, document grouping does not solve another common problem with Web-based search engines such as Excite: the mystery of why they list the documents they do. But if the entry form encourages users to break up their query into several groups of related key words, then a graphical interface can indicate which search topics occurred where in the retrieved documents. If hits on all topics overlap within a single passage, the document is more likely to be relevant, so the program ranks it higher. Alice might have a hard time spelling out in advance which topics must occur in the document or how close together they must lie. But she is likely to recognize what she wants when she sees it and to be able to fine-tune her query in response. More important, the technique, which I call [TileBars](#), can help users decide which documents to view and can speed them directly to the most relevant passages.

The potential for innovative user interfaces and text analysis techniques has only begun to be tapped. Other techniques that combine statistical methods with rules of thumb can automatically summarize documents and place them within an existing category system. They can suggest synonyms for query words and answer simple questions. None of these advanced capabilities has yet been integrated into Web search engines, but they will be. In the future, user interfaces may well evolve even beyond two- and three-dimensional displays, drawing on such other senses as hearing to help Aunt Alices everywhere find their bearings and explore new vistas on the information frontier.

Further Reading

Rich Interaction in the Digital Library. Ramana Rao, Jan O. Pedersen, Marti A. Hearst and Jock D. Mackinlay *et al.* in Communications of the ACM, Vol. 38, No. 4, pages 29-39; April 1995.

The WebBook and the Web Forager: An Information Workspace for the World-Wide Web. Stuart K. Card, George G. Robertson and William

York in Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems, Vancouver, April 1996. Available on the [World Wide Web](#)

[Selected publications by Marti Hearst](#)

["The WebBook and the Web Forager: An Information Workspace for the World-Wide Web."](#)
Stuart K. Card, George G. Robertson and William York in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, April 1996.

["Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results."](#) Marti A. Hearst and Jan O. Pedersen in *Proceedings of the 19th Annual International ACM/SIGIR Conference*, Zurich, August 1996.

["SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests."](#) Michelle Q. Wang Baldonado and Terry Winograd in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Atlanta, 1997 (in press).

[Research in Support of Digital Libraries at Xerox PARC](#)

The Author

[MARTI A. HEARST](#) has been a member of the research staff at the Xerox Palo Alto Research Center since 1994. She received her B.A., M.S. and Ph.D. degrees in computer science from the University of California, Berkeley. Hearst's Ph.D. dissertation, which she completed in 1994, examined context and structure in text documents and graphical interfaces for information access.

>

Transfer interrupted!

A Scatter/Gather Example

Here we demonstrate the use of Scatter/Gather on a collection of encyclopedia articles. Our query is very simple:

Retrieve the top 250 documents that contain the word *star* .

Here we show that Scatter/Gather text clustering does a reasonably good job at organizing the documents into meaningful themes or topics.

We ask Scatter/Gather to place the 250 documents into 5 groups. Here is what results. (Bear in mind that encyclopedia articles are well-written and uniform format. The [next example](#) shows the results of a more complicated query on a more unruly text collection.)

<input type="checkbox"/> Cluster 1 Size: 8	key army war francis spangle banner air song scott word poem british
<input type="radio"/> Star-Spangled Banner, The <input type="radio"/> Key, Francis Scott <input type="radio"/> Fort McHenry <input type="radio"/> Arnold, Henry Harley <input type="radio"/> ...	
<input type="checkbox"/> Cluster 2 Size: 68	film play career win television role record award york popular stage p
<input type="radio"/> Burstyn, Ellen <input type="radio"/> Stanwyck, Barbara <input type="radio"/> Berle, Milton <input type="radio"/> Zukor, Adolph <input type="radio"/> ...	
<input type="checkbox"/> Cluster 3 Size: 97	bright magnitude cluster constellation line type contain period spectr
<input type="radio"/> star <input type="radio"/> Galaxy, The <input type="radio"/> extragalactic systems <input type="radio"/> interstellar matter <input type="radio"/> ...	
<input type="checkbox"/> Cluster 4 Size: 67	astronomer observatory astronomy position measure celestial telescop
<input type="radio"/> astronomy and astrophysics <input type="radio"/> astrometry <input type="radio"/> Agena <input type="radio"/> astronomical catalogs and atlases <input type="radio"/> ...	
<input type="checkbox"/> Cluster 5 Size: 10	family specie flower animal arm plant shape leaf brittle tube foot hor
<input type="radio"/> blazing star <input type="radio"/> brittle star <input type="radio"/> bishop's-cap <input type="radio"/> feather star	

Shown here are the clusters' sizes (how many documents they contain), a list of topical terms, and a list of document titles. One can see from the topical terms of Cluster 1 that this cluster contains documents that involve stars as symbols, as in military rank and patriotic songs.

Cluster 2 has 68 documents that appear mainly to be about movie and tv stars.

Cluster 3 contains 97 documents that having to do with aspects of astrophysics.

Cluster 4 contains 67 documents also about astronomy and astrophysics. This cluster contains many articles about people who are astronomers (this is apparent when the list is scrolled down).

Cluster 5 contains all the articles that discuss animals or plants, and that happen to contain the word star, for example, star fish.

If we ask Scatter/Gather to re-cluster the 68 documents that appear in Cluster 2, the one that discusses movie and tv stars, and place the results into three clusters, we see the following clusters:

☐ **Cluster 1 Size: 14** player league hit game national set bat average season history basebal

- ☐ Musial, Stan
- ☐ Bench, Johnny
- ☐ Carew, Rod
- ☐ Robertson, Oscar
- ☐ Beliveau, Jean
- ☐ Casper, Billy
- ☐ Chinese checkers
- ☐ Best, George
- ☐ Beamon, Bob

☐ **Cluster 2 Size: 47** role stage broadway comedy performance actress production musical

- ☐ Burstyn, Ellen
- ☐ Stanwyck, Barbara
- ☐ Berle, Milton
- ☐ Bankhead, Tallulah
- ☐ Murphy, Eddie
- ☐ Walsh, Raoul
- ☐ Martin, Mary
- ☐ Zukor, Adolph
- ☐ Cosby, Bill

☐ **Cluster 3 Size: 7** music country jazz folk pop paul cowboy leader williams hampton boy

- ☐ Williams, Hank
- ☐ Crosby, Bing
- ☐ Campbell, Glen
- ☐ Belafonte, Harry

This re-clustering reveals that in actuality this cluster had more kinds of documents than we originally thought, based on the topical terms. These three clusters can be rather neatly summarized as containing articles about (Cluster 1) people who are sports stars, (Cluster 2) stars of film, tv, and theatre, and (Cluster 3) musicians.

Now if we back up a step and re-cluster Cluster 3 from the original set, placing the results into four clusters, we see the following:

<input type="checkbox"/> Cluster 1 Size: 12	black white nuclear hole reaction helium neutron gravitational collap
<input type="radio"/> stellar evolution <input type="radio"/> gravitational collapse <input type="radio"/> black hole <input type="radio"/> main sequence <input type="radio"/> carbon cycle <input type="radio"/> mass–luminosity relation	
<input type="checkbox"/> Cluster 2 Size: 49	galaxy type distance stellar variable spectral interstellar brightness ga
<input type="radio"/> star <input type="radio"/> extragalactic systems <input type="radio"/> Galaxy, The <input type="radio"/> interstellar matter <input type="radio"/> cluster, star <input type="radio"/> population, stellar	
<input type="checkbox"/> Cluster 3 Size: 29	constellation northern hemisphere sky locate dipper celestial double r
<input type="radio"/> constellation (astronomy) <input type="radio"/> Auriga <input type="radio"/> Big Dipper <input type="radio"/> Cassiopeia <input type="radio"/> Cygnus <input type="radio"/> Taurus	
<input type="checkbox"/> Cluster 4 Size: 7	fraunhofer designate map joseph frown fur wollaston english von davi
<input type="radio"/> Fraunhofer lines <input type="radio"/> Fraunhofer, Joseph von <input type="radio"/> Star Carr <input type="radio"/> Star of David <input type="radio"/> Star Chamber <input type="radio"/> Hubble's evolutionary model	

The contents of these four clusters can be glossed as general astrophysics, galaxies and stars, constellations, and a cluster of leftover, or outlying documents.

This example suggests the potential power of the system for automatically grouping documents according to themes. It also shows some issues that remain to be addressed. First, we need to determine automatically what the best number of clusters is at each phase. Currently we have the user make the decision of how many clusters to show for each document subcollection. We are working on how to make this choice automatically, based on the characteristics of the subcollection. Second, sometimes the summary is misleading or incomplete in terms of what documents are to be found in the cluster. We saw this with the cluster about film and tv stars -- it also contained documents about sports and music stars, although these were in the minority. We are working on determining how to indicate to the user when there are hidden topic areas in the cluster.

Click [here](#) for another example on a more complex query.

[Back to Scatter/Gather Overview](#)

Xerox PARC

2/13/97

References:

- [Courses](#): Digital Library and related courses being offered at various Universities.
- [Conferences/Workshops](#): Links to various conferences/workshops that have been held in the recent past or will be held in the near future.
- [Journals](#): Digital Library related journal information with links.
- [Repositories & Bibliographies](#): contains information and links to some of the repositories maintained by various organizations such as the [D-Lib Magazine](#).
- [Books](#): Some books that contain valuable information on Digital Libraries (along with links to some publishers)

[\[Main\]](#) [\[Contents\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Digital Library and related courses:

- [Digital Library course offered at Pittsburgh](#)
- [Michael Lesk's Digital Library course at Columbia University](#)
- Virginia Tech
 - [CS6604 \(1997\) Digital Libraries](#)
 - [UH3004 Fall 1997 Honors 3004 - Digital Libraries](#)
 - [CS5604 Information Storage and Retrieval](#)
 - [CS4624 Multimedia, Hypertext and Information Access](#)
 - [CS6604 \(1995\) Interactive Accessibility](#)
- CSEI: [NSF CS Education Innovation](#) - projects around the nation
- Furman University:
 - [Exploring the Digital Domain](#)
 - [Web site on creating WWW pages](#)
- [Cyberspace Law for Non-Lawyers](#): This is an electronic course : a "real" course in the "real world" This site includes a discussion function which will allow you, if you are so inclined, to post your own comments and reactions to the individual messages that the instructors have mailed out.
- [Digital Library \(Alexandria\) Online Tutorial at UCSB](#)

[\[Main\]](#) [\[Contents\]](#) [\[References\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



LIS 2970: Digital Libraries

[School of Information Sciences](#)
[University of Pittsburgh](#)

Fall Term, August-December 1997

[Christinger Tomer](#)
Associate Professor

Economics of Digital Libraries: E6998-043

Columbia University
Mudd Rm 327
Tuesdays, 6:10-8pm, 16 January 1996 to 23 April 1996

[Course Outline & References](#)

Assignments

[First assignment: Write publishers](#)

[Second assignment: Estimate course needs](#)

[Third assignment: Prepare a pre-proposal](#)

Lecture viewgraphs

[Course Summary: 16 January viewgraphs](#)

[Library & IR history: 23 January viewgraphs](#)

[Value of information: 30 January viewgraphs](#)

[Copyrights & patents: 6 February viewgraphs](#)

[The course is now running 1/2 session late]

[Buying clubs; sales on Internet: 13 February viewgraphs](#)

[Text and image storage & retrieval: 20 February viewgraphs](#)

[Formats & Image analysis: 27 February viewgraphs](#)

[Scanning resolution: 5 March viewgraphs](#)

[Well, we caught up but 12 March was spring break.]

[Interface evaluation; US Digital library projects: 19 March viewgraphs](#)

[World DL efforts: 26 March viewgraphs](#)

[Archiving & preservation: 2 April viewgraphs](#)

[Cryptography: 2 April viewgraphs](#)

[Social & research issues: 9 April viewgraphs](#)



CS6604 - Digital Libraries - Fall 1997

Table of Contents

- **Contacts:** [Instructor](#), [GTA](#)
- [Department and Class Policies](#)
- [DLI Articles in IEEE Computer](#)
- [Explore on WWW](#)
- [HyperNews](#)
- [Labs](#)
- **Lectures:** [970911](#), [Multimedia in Digital Libraries](#)
- [Listserv](#)
- [News / Announcements](#) (updated 971203@2am)
- **Overview:** [WWW](#), [PDF - Part 1](#), [PDF - Part 2](#)
- **Projects:** [Grading](#); [Guidelines](#), [old examples](#); [List of project ideas](#); [Projects and People](#)
- **Quizzes:** [Password request](#), [Take a quiz](#) -- Let [Patrick](#) know if you have any problems.
- [Syllabus](#)
- [Topics](#)
- [UNIX Use Hints \(for video accounts\)](#)
- [Materials on reserve in Newman Library](#)

Related Courses

- UH3004 Fall 1997 [Honors 3004 - Digital Libraries](#)
 - CS5604 [Information Storage and Retrieval](#)
 - CS4624 [Multimedia, Hypertext and Information Access](#)
 - CS6604 (1995) [Interactive Accessibility](#)
-

Please send comments and suggestions to: fox@fox.cs.vt.edu

Education Innovation

NSF CISE funds Education Innovation projects, like our own local [EI project](#). Since some of these relate to digital libraries, and all to CS education, it is worthwhile examining key aspects.

- [Home Page for CSEI](#)
- Using modules: [Brooklyn College](#) - Distributed Processing
- Integrating research: [Evergreen State College, OGI](#) - software engineering of scientific systems; formal methods and higher order logics; and neural networks applied to speech recognition
- Multimedia Support: [Georgia Tech](#) - including STABLE, a case library of projects in Smalltalk
- Scheme Programming: [Indiana U.](#) - including a [Scheme Repository](#)
- Parallel and Distributed Computing: [Louisiana Tech University](#) - Java Concurrency Simulator
- Intro Labs: [Oberlin College](#) - [HtX](#) tool to generate WWW pages, [labs, etc.](#)
- Networking Labs: [Ohio State U.](#) - [7 labs](#), [software](#)
- Multi-semester projects: [ODU](#) - [Computer Productivity Initiative \(CPI\)](#)
- Software Design & Development: [RPI](#) - [Design Conference Room](#), Standard Template Library (STL) since incorporated into standard C++ library
- Programming: [Rice](#) - [14 labs with Scheme and Java](#)
- HCI: [San Jose State U.](#) - [devices and distance education](#)
- High-Performance Scientific Computing: [U. Colorado Boulder](#) - [tutorials, etc.](#)
- Wireless Comm.: [VT/UMR](#) - [3 courses](#)
- TeleMentoring: [U. Penn.](#) - ATM distance education, seminars
- Literacy: [Utah State U.](#) - multimedia modules

Conferences/Workshops:

- ACM DL'98: Pittsburgh, June 23-26 <http://fox.cs.vt.edu/confs/DL98.txt>
- ACM DL'97: Philadelphia, July 23-26 <http://www.lis.pitt.edu/~diglib97/>
- DL'96: Bethesda, March (1st ACM ...) <http://fox.cs.vt.edu/DL96/>
- DL'95: Austin, June <http://csdl.tamu.edu/DL95/>
- DL'94: Texas A&M University
- Santa Fe Workshop, Digital Knowledge Work Environments, March 9-11, 1997
<http://www.si.umich.edu/SantaFe/>
- UCLA Workshop, Social Aspects of Digital Libraries, Feb. 16-17, 1996
<http://www-lis.gseis.ucla.edu/DL/>
 - [life cycle](#)
- [IITA Digital Libraries Workshop, 1995](#)
- Allerton, 1996 <http://edfu.lis.uiuc.edu/allerton/96/> and [map](#)
- Allerton, 1995 <http://edfu.lis.uiuc.edu/allerton/95/>
- ADL 96, Forum on Research and Technology Advances in Digital Libraries May 13-15, 1996, Washington, D.C.
- KOLISS DL 96, Proc. Int'l Conf. on Digital Libraries and Information Services for the 21st Century, Sept. 10-13, 1996, Seoul, Korea
- D-Lib supported meetings, conferences and workshops <http://www.dlib.org/groups.html>

[\[Main\]](#) [\[Contents\]](#) [\[References\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Journals:

Selected special issues include:

- Commun. ACM
 - [April 1995](#): 38(4)
 - [April 1998](#): 41(4)
- [IEEE Computer, May 1996](#)
- J. American Society for Information Science, Sept. 1993: 44(8)
- J. of Visual Communication and Image Representation, 7(1), March 1996

There also are closely related journals like:

- J. of Digital Information (British Computer Society)

[\[Main\]](#) [\[Contents\]](#) [\[References\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Repositories & Bibliographies:

- **D-Lib** <http://www.dlib.org/>
 - Articles (by author) <http://www.dlib.org/author-index.html>
 - Articles (by title) <http://www.dlib.org/title-index.html>
 - Research Projects (incl. DLI) <http://www.dlib.org/projects.html>
 - D-Lib Working Groups <http://www.dlib.org/groups.html>
 - Metadata <http://www.dlib.org/metadata/overview.html>
 - Naming <http://www.dlib.org/naming/overview.html>
 - Repository Interfaces <http://www.dlib.org/repository/overview.html>
 - Social Aspects <http://www.dlib.org/social/overview.html>
 - D-Lib Magazine Articles on Key Topics
 - Agents <http://www.dlib.org/dlib/July95/07birmingham.html>
 - Architecture (incl. handles) <http://www.cnri.reston.va.us/home/dlib/July95/07arms.html>
 - Metadata <http://www.dlib.org/dlib/July95/07weibel.html>
 - Uniform Resource Names (URNs) <http://www.dlib.org/dlib/february96/02arms.html>
 - Use <http://www.dlib.org/dlib/october95/10bishop.html>
 - Informedia <http://www.dlib.org/dlib/july96/07wactlar.html>
 - Variations <http://www.dlib.org/dlib/june96/06fenske.html>
 - Access Control: [Articles by Gladney et al.](#)
- UIUC Pointers to Publications <http://dli.grainger.uiuc.edu/pubs/natsynch.htm>
- **Annotated Bibliography for Digital Libraries** - Christine Woerner, 1996 (Case Studies, DL as Place, Archive/Organization/Preservation, Librarianship, Mediation/Interaction, Authoring/Authenticity/Originality)
- Virginia Tech
 - [Digital Library Research Laboratory Publications](#)
 - [BibTeX file](#) for article: E. Fox and O. Sornil. Digital Libraries. Chapter 11 in Modern Information Retrieval, AWL England, 1999: Ricardo Baeza-Yates and Berthier Ribeiro-Neto, eds., to appear.
 - misc ptrs <http://scholar.lib.vt.edu/digilib/>

[\[Main\]](#) [\[Contents\]](#) [\[References\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



► **D-Lib Magazine**

► **Current Issue**

► **Back Issues**

► **Ready Reference**



► **D-Lib Metrics**

► **Activity Hosted by D-Lib**

► **NSF SMETE-Lib Study**

Welcome to D-Lib, an evolving community of research interests in digital libraries.

From here, you may visit **D-Lib Magazine**, a monthly compilation of contributed stories, commentary, and briefings. Or you may find Ready Reference a convenient clearinghouse of pointers to other sites on the web of interest to researchers and users of digital libraries. Finally, the D-Lib Working Group on Digital Library Metrics addresses the thorny problem of developing appropriate metrics.

Research in digital libraries is an expanding horizon with applications for today and implications for tomorrow. We invite you to read, observe, and participate.



Last Updated 6/15/98 af:cr



Site for publications from all six Digital Libraries Initiative projects.

- ⊙ **UNIVERSITY OF CALIFORNIA
at Berkeley**
- ⊙ **UNIVERSITY OF CALIFORNIA
at Santa Barbara**
- ⊙ **CARNEGIE MELLON UNIVERSITY**
- ⊙ **UNIVERSITY OF ILLINOIS
at Urbana-Champaign**
- ⊙ **UNIVERSITY OF MICHIGAN**
- ⊙ **STANFORD UNIVERSITY**

Comments to [Susan Harum](#)

Last updated 5/26/98

An Agent-Based Architecture for Digital Libraries

William P. Birmingham
The University of Michigan
Electrical Engineering and Computer Science Department
School of Information Science and Library Studies
Ann Arbor, MI 48109
wpb@eecs.umich.edu

D-Lib Magazine, July 1995

-
- [Introduction](#)
 - [Agents](#)
 - [What the architecture provides](#)
 - [The Conspectus and the conspectus language](#)
 - [Status and summary](#)
 - [Acknowledgements](#)
 - [References](#)
-

A button with the text "d-Lib forum" in a blue, stylized font, enclosed in a rounded rectangular border.

A button with the text "d-Lib magazine" in a blue, stylized font, enclosed in a rounded rectangular border.

Introduction

One of the most exciting promises of digital libraries is access to a great variety of information and *services* that transcend what is available today through on-line services, such as the World-Wide Web (WWW). A library is more than just stacks of materials on shelves; it is also highly trained people that provide valuable services. These services include such things as *organization and cataloging*, research, notification of new publications, and so forth. Indeed, one of the greatest assets of libraries are these high-valued services. The WWW, while it probably contains more information than any single traditional library, is arguably not as useful as a traditional library because it lacks these services (particularly organization and sophisticated search support). No one is dismantling their libraries because of the WWW yet. The University of Michigan Digital Library Project (UMDL) [1,2] believes that a successful digital library needs to provide both access to a wide variety of valuable content and services.

Because the range of both content and services that are possible for a digital library are potentially large (we cannot even imagine what will be available or needed in the future), there will be no single, complete digital-library solution. Rather, we expect that as editing tools become better and access to networks becomes easier and cheaper, there will be millions of content suppliers; "everyman" can

become a vanity press on the information superhighway. We believe that the days of centralized suppliers of information (e.g., large publishing houses and traditional libraries) are numbered, and that the traditional notion of a "collection" will span multiple databases, each residing in a different place in cyberspace.

Furthermore, the creativity of users of digital libraries will spawn thousands of different, specialized services (e.g., notification and translation, even special collections of information). Perhaps most importantly, methods of organizing information will transcend a single "digital library," in that it is unlikely that a single indexing or naming scheme (e.g., the Dewey Decimal System) will be used across the multiplicity of digital libraries that are sure to emerge. Thus, we must create flexible software architectures that can federate as many content suppliers, information-organizational schemes, and service providers as possible, and yet scale to the extremely large size needed to support the digital libraries of the future.

Considering this view of digital libraries, we have developed some guidelines and objectives for our system. First, the guidelines:

- Given that many digital libraries will emerge, we want to make ours as attractive to users and content and service providers as possible. Thus, we intend to make the fewest and least-restrictive standards.
- The only way to ensure intellectual property in the future is to provide incentives for its creation. Thus, we intend to make support for economic incentives an integral part of the UMDL architecture. This covers a wide range of issues, from definition and protection of intellectual property rights through payment for the use of intellectual property. Please note: *we are not establishing policies related to rates or payment, rights of users to access the contents of the library, or other related issues such as "fair use"*. We simply plan to have the machinery in place to support whatever policies may arise in the future.
- The elements of the library (services and collections of information) are autonomous in that these elements will make decisions based on their own perspective. In other words, there is no central authority that can press an element into service. All elements are considered peers, and thus interaction is achieved entirely through negotiation processes.

Broadly speaking, the objectives of the architecture are to provide services that fall under the following categories:

- Registration: maintaining a comprehensive list of all the agents (collections, user interface, and others) in the UMDL.
- Brokering and teams of agents formation: finding potential information sources and support services (e.g., translation of query languages) to fulfill a user's information needs.
- Commerce support: providing mechanisms to support commerce for information goods, and protecting intellectual property and privacy.

Furthermore, we require that the architecture have the following properties:

- Modular, in that new elements can be added or removed without effecting the operation of other elements;
- Scaleable, to allow for the potential of millions of constituent elements;
- Extensible, to allow new elements (collections, data types, services, etc.) to be easily added to the digital library.

- In the remainder of this paper, an overview of the UMDL architecture is given. We describe the notion of software agents and types of agents in UMDL, and then describe how agents interact to provide service.

Agents

The architecture is based on the notion of a software agent. An agent represents an element of the digital library (collection or service), and is a highly encapsulated piece of software that has the following special properties:

- **Autonomy:** the agent represents both the capabilities (ability to compute something) and the preferences over how that capability is used. Thus, agents have the ability to reason about how they use their resources. In other words, an agent not have to fulfill every request for service, only those consistent with its preferences. A traditional computer program does not have this reasoning ability.
- **Negotiation:** since the agents are autonomous, they must negotiate with other agents to gain access to other resources or capabilities. The process of negotiation can be, but is not required to be, stateful and will often consist of a "conversation sequence", where multiple messages are exchanged according to some prescribed protocol, which itself can be negotiated.

Autonomy is critical to scaling UMDL to a large size because autonomy implies local or decentralized control. As a result, we do not have to update some "master" program everytime a new agent is added to UMDL. The effects of adding or removing an agent are propagated locally using a set of protocols. Thus, there is no need for global coordination among all agents [4]. The notion of decentralized control of autonomous agents is similar to the way our economy works. Each of us is similar to an agent in that the decision about how money is spent is done individually. These spending decisions do not require communication across the entire economy (e.g., when ones buy a car, she do not need to tell the whole country or even the car manufacturer, just the car dealer), nor does one need to get permission from a central authority. Similarly, UMDL agents can make decisions and form teams at a local level, without requiring interaction with all agents in the system or with a central authority.

Negotiation is complementary to autonomy, in that autonomous agents must be capable of making binding commitments for the system to work. Thus, when agents negotiate and strike a deal (i.e., something of value is exchanged for something else of value), the agents are bound to fulfill that deal. It is possible, and even likely, that some deals will allow agents to back out. This "feature", however, must be explicitly negotiated in our system.

The UMDL is populated by three classes of agents:

- **UIAs (User Interface Agents)** provide a communication wrapper around a user interface. This wrapper performs two functions. First, it encapsulates user queries in the proper form for the UMDL protocols. Second, it *publishes* a profile of the user to appropriate agents, which is used by mediator agents to guide the search process.
- **Mediator agents** [8], of which there are many types, perform a variety of functions: essentially, all tasks that are required to refer a query from a UIA to a collection, monitor the progress of the query, transmit the results of a query, and perform all manner of translation and bookkeeping. Presently, two types of mediators populate the UMDL. Registry agents capture the address and contents of each collection. Query-planning agents [5] receive queries and route them to collections, possibly consulting other sources of information to establish the route. Another special

class of mediators currently being developed, called facilitators [7], mediate negotiation among agents [3].

- CIAs (Collection Interface Agents) provide a communication wrapper for a collection of information. While performing translation tasks similar to those performed by the UIA for a user interface, the CIA also publishes the contents and capabilities of a collection in the *conspectus* language (described in the next section, "What the architecture provides").

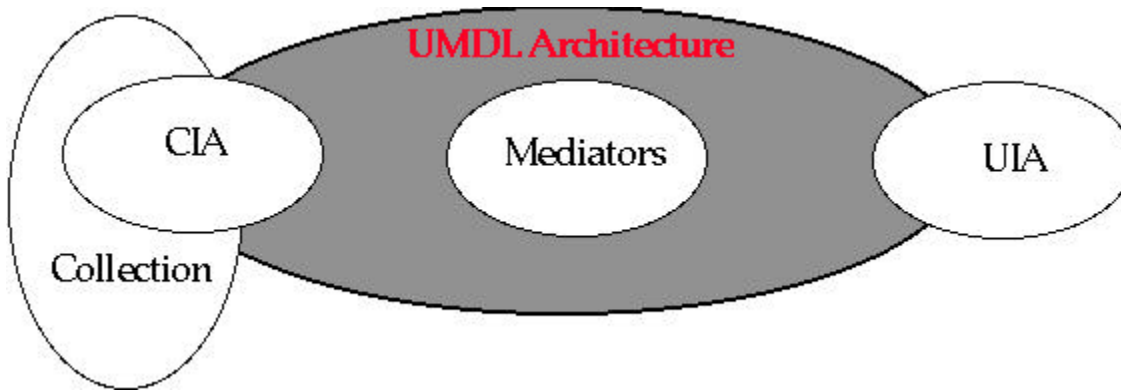


Figure 1: UMDL agent types

As the architecture is developed, the broad classes of agents depicted in Figure 1 will be continually refined; specialized agents will be added to the system as needed (the modularity property). For example, we can create user interfaces that are customized to a particular class of users, rather than to a particular collection or access mechanism (e.g., Boolean search over controlled vocabulary). In addition, the ability to *team* agents (as described in the next section, "What the architecture provides") dynamically creates new services with new agents, which is especially important since we anticipate the agent population will be constantly changing.

What the architecture provides

From a user's perspective, the types of high-level support that make a digital library worth using, such as searching, will be performed by a team of agents. For example, consider Figure 2, where a user (through the UIA) is searching for all articles by "Joan Q. Publique". Assuming that all agents have registered with the registry agent, the UIA contacts a query planner by first requesting the registry for a query planner that knows about author searching. The query planner then goes to the registry to get the addresses of a name authority (meta data that gives variations of Joan Q. Publique) and a name index (a partial listing of collections that contain works sorted by author). The planner then interrogates the authority, and then the index, finally determining the address of a particular collection. The collection is then accessed by the UIA using a protocol specific to the CIA.

It is easy to image how this process can be extended for different types of search by adding new types of agents (e.g., subject indexes and new kinds of query planners). The teaming methods gives the architecture a dynamic planning ability[5] that is critical for finding the best way to perform some service, as well as easily incorporate new types of search methods. There is, however, a cost.

This cost is coordinating the agents, which includes communication and negotiation of which negotiation may prove the more expensive. Communicating among agents certainly takes more time than would be required by a monolithic system. We believe, however, that improved network technology will ameliorate these costs, making them insignificant. The major overhead may come from negotiation.

Recall that agents are autonomous and cannot be coerced into responding to a request for service. What is not shown in Figure 2 is interaction with facilitators. It is possible that the UIA-query planner, query planner-name authority, query planner-name index, and UIA-CIA interactions will all require some type of negotiation (we assume that registration is "free"). Striking deals among all these agents could require significant time and computation resource. This overhead, however, can be minimized by prearranged deals among agents. For example, a UIA could buy a *token* that specifies certain access privileges for a cartel of CIAs; similar arrangements can be made among other agents.

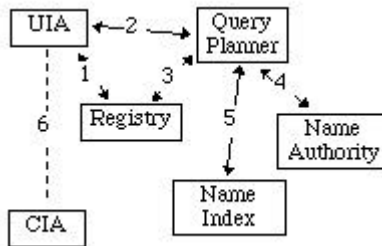


Figure 2: Example search by author

We separate the activities of agents UMDL into two types: that used to organize agents to perform the team building (called *architectural*), and that used to perform the actual task (called *task*), such as actually querying a database. Strictly separating these activities allows us to reduce the commitments that an agent must make to operate in our system (i.e., a CIA is not required to support all query languages used in UMDL, only those it chooses to support.). Thus, we require only that agents use a language, called the *conspectus language*, designed to support architectural activities (see the next section, "The Conspectus and the conspectus language"); the decision to support any particular task language is left up to individual agents.

The distinction between architecture and task has advantages and disadvantages. The advantages include minimal standards, and therefore increased flexibility in creating agents. Furthermore, the agents themselves are smaller, and therefore easier to build and maintain. A disadvantage is that not all agents will have access to all other agents. For example, if a CIA supports only Z39.50, but a UIA uses some other language X (and no mediator exists that can translate X to Z39.50), then that UIA cannot access the CIA. We see, however, no practical solution to this problem at this time.

Since it is impossible to create an architecture that has everything, we prefer flexibility over guaranteed interoperability among all agents. Task languages that will undoubtedly evolve over time, as we learn more about digital libraries. By being neutral on which languages are supported, we avoid having to rewrite significant portions of our software as the languages change.

The Conspectus and the conspectus language

The space of information in UMDL is potentially enormous, as is the possibility of bringing the system to its knees with rogue query processes. To limit queries to potentially applicable CIAs, we reason about the contents of each collection to derive an estimate of their likely usefulness. This leads us to a two-level partition of the information space:

- Conspectus: includes, among other things, the content of the collection, the search capabilities of

the search engine(s) associated with the collection, and the structure of the material (documents) in the collection.

- Collection: the set of actual documents in a collection. These documents are in native formats, and the search engines are engaged through native query languages.

The conspectus is an abstracted description of the aggregate of collections populating the UMDL. Additionally, the conspectus is a *normalized* description of content. This is important, as various collections will have different methods for describing the same thing (e.g., title as TI or TL). To help normalize terms, we are using a variety of thesauri developed by various researchers around the world.

The conspectus is written in a language that we have defined (the UMDL conspectus language, UCL). Although we retain complete control over the UCL, the actual conspectus expressed in UCL will be specified by the separate collections. Our aim is that UCL (and its associated resources, such as various thesauri and cataloging systems) provide sufficient structure for developing compatible representations of collections. Thus, the conspectus provides interoperability for various search and retrieval methods through a common representation over collections.

Since the conspectus will be large both in scope and in size, it will be distributed and hierarchically organized. We expect to create special mediator agents whose sole responsibility is to maintain the integrity of the conspectus.

Agents communicate using patterns of messages, where the content of the message is specified by UCL and sets of *performatives* describing the purpose of the communication (e.g., to ASK or TELL something) [6]. The messages transmitted between the agents describe capabilities, services, and other primitives. For example, all agents use the ASK performative to make requests to the registry for notification about classes of agents with certain capabilities. The registry agent continues sending information about these agents, as they come on-line, until the UNASK performative is received.

Another example performative set is TELL, which is typically used in response to an ASK. The registry agent uses TELL to send the names of agents that correspond to some capability specification. The registry agent uses the UNTELL performative to express that an agent is no longer available, or that its capabilities have changed.

Protocols specify communication patterns among agents. In order to participate in UMDL, an agent must use our protocols. Since these protocols are minimally restrictive in how a task is accomplished, we believe they are not a significant impediment to the development of agents by third parties. Standardizing the protocols, but not the task languages, strikes a balance between flexibility and ease of integration into the UMDL environment.

The agent-identification protocol (used by both the CIA and query planner in the example depicted in Figure 2) provides a way for agents to locate other agents with specific capabilities (Figure 3). The requesting agent (*R*) uses the ASK performative to describe the specific capabilities to the registry agent. The registry agent executes a lookup operation to match the specifications to the agents it knows about. Any matches are sent to the requesting agent via the TELL performative. The ASK performative implies a standing request for information about agents, so that the registry agent continues to send *R* information about other agents as they advertise their capabilities. When *R* receives information about an agent (*A*) from the registry agent, it has the option of storing that information in its local knowledge base for future use.

If R no longer wants to receive information about an agent, then it uses the UNASK performative to communicate this desire to the registry agent. Upon receipt of the UNASK performative, the registry agent stops sending information to R . If A is no longer available, or has a change in capabilities, then the registry agent sends the UNTELL performative to all agents who received a TELL performative about A . Thus, the registry agent must keep track of the agents to which it sent the TELL performative.

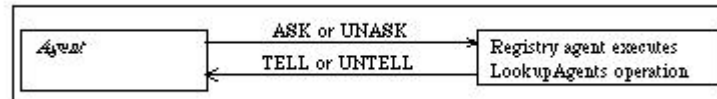


Figure 3: Agent-identification protocol.

The performative and protocol features of the UMDL architecture are general enough to accommodate a variety of actions within the library. As illustrated here, the same protocol can be used by several different agents to achieve their objectives. We expect that once we have established a basic set of protocols, including those for negotiations about intellectual property, they will become relative stable even though the variety of information and services in the library will grow enormously. In fact, the stability of these protocols is the foundation for growth of the system.

Status and summary

The UMDL is operational, and can be accessed through <http://www.sils.umich.edu/Catalog/UMDL.html>. The current system has about 50 CIAs and basic search support. We expect to have subscription, notification, and known-item search running by the end of the calendar year. Two task languages are supported: Z39.50 and FTL (a locally created query language).

The current system demonstrates that the agent architecture approach outlined in this article is viable, and paves the way for more interesting experiments with scaling both the total number of agents as well as the types of services and collections available. It is interesting to note that the architecture was able to handle the addition of new services (new collections and a notification service) without modifications to existing agents and protocols, thus demonstrating properties of scalability, extensibility, and modularity.

Acknowledgments

The members of the UMDL architecture group contributed many of ideas presented here. In particular, Fritz Freiheit provided helpful suggestions to drafts of this paper.

The UMDL project is funded under a joint initiative of NSF/ARPA/NASA; we are grateful for their financial support, and their enthusiasm for the initiative. The views expressed in this paper are those of the author only, and do not necessarily represent the views of the funding agencies (nor of the UMDL project).

References

1. Birmingham, W. P., K. M. Drabenstott, C. O. Frost, et al. (1994). The University of Michigan Digital Library: This is not your father's library. *Digital Libraries '94*, College Station, TX.
2. Birmingham, W. P., E. H. Durfee, T. Mullen, et al. (1995). The distributed agent architecture of

- the University of Michigan Digital Library. *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, Stanford, CA, AAAI Press.
3. D'Ambrosio, J. and W. P. Birmingham (1995). Preference-directed design. *AI in Engineering, Design, Analysis, and Manufacture*. To appear.
 4. Darr, T. P., and W. P. Birmingham (1994). Automated design for concurrent engineering. *IEEE Expert* **9**(5): 35-42.
 5. Durfee, E. H. and T. A. Montgomery (1991). Coordination as distributed search in a hierarchical behavior space. *IEEE Transactions on Systems, Man, and Cybernetics*, Special Issue on Distributed Artificial Intelligence, **21**(6):1363-1378.
 6. Finin, T., R. Fritzson, D. McKay, et al. (1994). KQML as an agent communication language. *Third International Conference on Information and Knowledge Management*, ACM Press.
 7. Mullen, T., and M. P. Wellman (1995). A simple computational market for network information services. *First International Conference on Multi-agent Systems*, San Francisco, CA.
 8. Wiederhold, G. (1992). Mediators in the architecture of future information systems. *Computer* **26**(3): 38-49.

Copyright © 1995 William P. Birmingham

A button with a blue border and rounded corners containing the text "d-Lib forum" in blue.A button with a blue border and rounded corners containing the text "d-Lib magazine" in blue.

hdl:cnri.dlib/july95-birmingham

Key Concepts in the Architecture of the Digital Library

William Y. Arms
Corporation for National Research Initiatives
Reston, Virginia
warms@cnri.reston.va.us

D-Lib Magazine, July 1995

Introduction

For the past two years, the Computer Science Technical Reports project (CS-TR) has been developing an architecture for a digital library with funding from the Department of Defense's Advanced Research Projects Agency (ARPA). This is a general purpose framework for a digital library in which very large numbers of objects, comprising all types of material, are accessible over national computer networks. It is described in a paper by Robert Kahn and Robert Wilensky (cnri.dlib/tn95-01).

This introduction describes the author's view of eight general concepts that emerged from the discussions. These concepts are key issues in the transition to a true digital library from the network services that we have today. The Kahn/Wilensky paper contains a comprehensive framework for resolving the issues.

General Principles

- [1. The technical framework exists within a legal and social framework](#)
- [2. Understanding of digital library concepts is hampered by terminology](#)
- [3. The underlying architecture should be separate from the content stored in the library](#)
- [4. Names and identifiers are the basic building block for the digital library](#)
- [5. Digital library objects are more than collections of bits](#)
- [6. The digital library object that is used is different from the stored object](#)
- [7. Repositories must look after the information they hold](#)
- [8. Users want intellectual works, not digital objects](#)
- [Reference](#)

General Principles

1. The technical framework exists within a legal and social framework

Early networked information systems were developed by technical and professional communities, concentrating on their own needs. The emphasis was on making information available to colleagues and the public, without charge. The digital library of the future will exist within a much larger economic, social and legal framework.

For example, musical works and their performance represent the livelihood of composers and musicians. Their artistic reputations often depend on their work not being changed in storage or transmission. They require payment, as do recording studios and concert halls. Such work will only be part of the digital library, if the library supports their interests.

The legal system's task is to codify this rapidly changing economic and social framework. The relevant areas of law include copyright, performance, and other intellectual property, libel and obscenity, communications law, privacy, and international law.

The Kahn/Wilensky architecture can not write the law, but it provides a technical design that matches the legal structure that is expected to emerge. The architecture respects the creators and owners of intellectual property. It allows the preservation of rights that can last for more than one hundred years, and recognizes that digital works may include material from many sources, with separate property rights.

Society expects the creators of works to be responsible for their content, and for those who make decisions about content to behave responsibly. However, the digital library will not thrive if legal liability for content is placed upon parties whose only function is storage and transmission. Therefore, the architecture establishes clear boundaries between the areas of responsibility of the various parties.

2. Understanding of digital library concepts is hampered by terminology

Terminology proves to be a barrier in describing a digital library. Some words have such strong social, professional, legal, or technical connotations that they obstruct discussion between people of varying backgrounds. Simple words mean different things to different people. For example, the words "copy" and "publish" have different meanings to computing professionals, publishers, and lawyers. Common English usage is not the same as professional usage, and the versions of English around the world have subtle variations of meaning.

Certain words cause such misunderstandings that they are best expunged from any precise discussion of the on-line digital library. The list includes "copy", "publish", and "document". Other words have to be used very carefully and their exact meaning made clear whenever they are used. An example is "content".

In the Kahn/Wilensky architecture, items in the digital library are called "digital objects". They are stored in "repositories" and identified by "handles". Information about the digital object is known as "properties" or "metadata".

3. The underlying architecture should be separate from the content stored in the library.

A conventional research library stores more than books, and the digital library stores more than digitized text. Almost every type of information can be represented in digital form, including text, pictures, musical works, computer programs, databases, models and designs, video programs, and compound works combining many types of information.

The underlying architecture of the digital library, as described by Kahn and Wilensky, specifies those characteristics that apply to all types of material. For example, every object needs to have a name or identifier; the actions of adding objects to the library or deleting them apply to all material; general purpose methods of security can be provided.

This underlying architecture is a base for extensions that can be tailored for various types of information. The extensions typically include specific formats, protocols, and rights management that are appropriate for the type of material. For example, the extensions for digitized movies will be very different from those for video games; texts are usually described by bibliographic terms, such as author and title, which are of little relevance to a computer program; a protocol designed for interaction with a database is unlikely to be useful in manipulating graphic designs.

Separating general functions from those specific to the type of content has other benefits. It encourages different markets to emerge, and allows a legal framework in which storage, transmission and delivery of digital objects is separate from activities to create and manage the intellectual content.

4. Names and identifiers are the basic building block for the digital library

Names are a vital building block for the digital library. Names are needed to identify digital objects, to register intellectual property in digital objects, and to record changes of ownership. They are required for citations, for information retrieval, and are used for links between objects.

These names must be unique. This requires an administrative system to decide who can assign them and change the objects that they identify. They must last for very long time periods, which excludes the use of an identifier tied to a specific location, such as the name of a computer. Names must persist even if the organization that named an object no longer exists when the object is used. There need to be computer systems to resolve the name rapidly, by providing the location where an object with a given name is stored.

The Corporation for National Research Initiatives has implemented a handle system which satisfies these requirements. A "handle" is a unique string used to identify digital objects. The handle is independent of the location where the digital object is stored and can remain valid over very long periods of time. A global handle server provides a definitive resource for legal and archival purposes, with a caching server for fast resolution. The computer system checks that new names are indeed unique, and supports standard user interfaces, such as Mosaic. A local handle server is being added for increased local control.

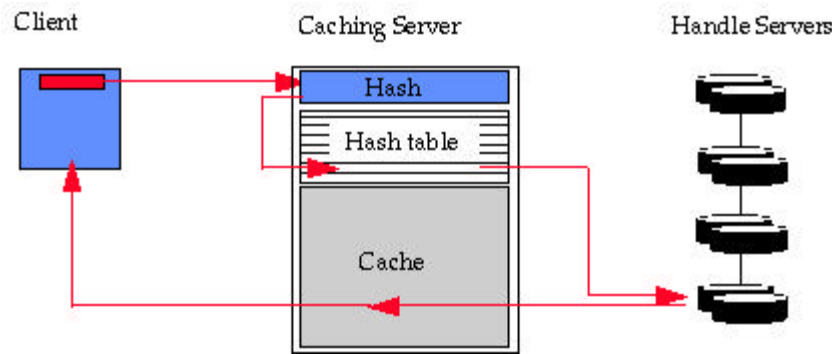


Figure 1. The CNRI handle system

5. Digital library objects are more than collections of bits

In the digital library, information is stored as "digital objects". A primitive idea of a digital object is that it is just a set of bits, but this idea is too simple. The content of even the most basic digital object has some structure, and information, such as intellectual property rights, must be associated with the digital object. Figure 2 shows that a digital object in a repository has two parts, content and associated data, sometimes called "metadata".

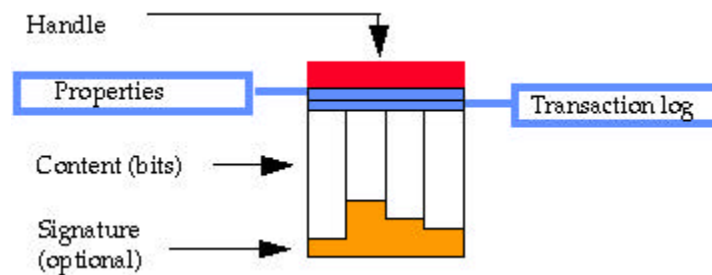


Figure 2. Parts of a digital object

To enable the content to represent useful information, its type must be known. Thus part of the content may be of type text (perhaps encoded in a mark-up language), while another part may be of type audio. A single digital object may contain many types of content. It turns out that arbitrarily complex data types can be constructed from a few basic types, notably bit-sequences, handles and other digital objects. By combining these in various combinations, any digital content can be represented.

To manage valuable intellectual property, certain metadata is required. This is shown in the figure. It always includes a unique identifier (the handle). It may also include properties such as rights and access methods. For example, one property states whether a digital object is mutable, in that it may be altered after being placed in a repository. Another is a digital signature or other method of validating that an object has not been changed. Frequently, it is useful to keep a log of all transactions associated with each digital object.

6. The digital library object that is used is different from the stored object

In the digital library, what you store is not what you get. The architecture must distinguish carefully between digital objects as they are created by an originator, digital objects stored in a repository, and

digital objects as disseminated to a user.

The user receives the result of executing a program on the stored object. This may be a simple program, such as a file transfer program, or something very complex. For example, an image is stored in a library as a set of wavelets. To use it, the stored wavelets are used to generate an image with the characteristics requested. This is transmitted over the network to a user's computer, where it can be further processed or displayed.

Some classes of digital objects can be provided it to a user in more than one way. For example, the score of a musical work is held in the library. One form of use is to transmit a representation of the score to the user's computer. Alternatively, the user could request the repository to execute a synthesizer program, which would perform the score, and transmit the digitally encoded audio over the network. For some types of object, such as a data base or a video game, the use consists of an interaction between the user and the execution of the program.

Legal scholars see an interesting parallel between the computer viewpoint of executing a program to supply a digital object to a user and the legal concept of performance. This may prove to be the correct framework for managing rights in a digital library.

7. Repositories must look after the information they hold

A repository stores digital objects, both the content and the metadata.

A digital object as stored in a repository may be very different from the digital object that is made available to users' computers. Different repositories will have very different internal organizations, but for each digital object every repository will have a properties record, which holds attributes of the object, and a transaction log.

Since digital objects may contain valuable intellectual property, the stored form of a digital object within the repository includes information that allows for it to be managed within economic and social frameworks. The repository maintains this information, provides basic reference information, and provides security to ensure that only valid operations are carried out on the digital objects.

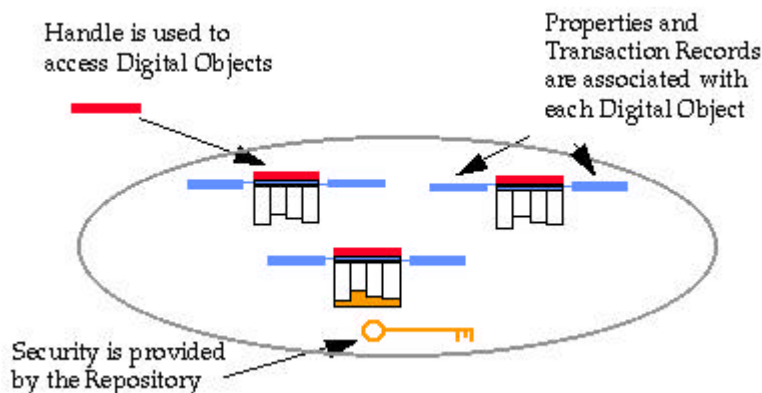


Figure 3. A repository

The internal organization of a repository and the way that digital objects are stored are hidden from the user. A simple protocol is provided for interactions with the repository. This protocol is called the

"repository access protocol." The basic commands in this protocol are those to access a digital object and its metadata, and the service request to disseminate a digital object. In addition there are commands to add and delete digital objects.

8. Users want intellectual works, not digital objects

Digital objects are the basic building blocks of the digital library, but users of the library usually want to refer to items at a higher level of abstraction. Common English terms, such as "technical report", "computer program", or "musical work", often refer to many digital objects that can be grouped together. The individual objects may have different formats, minor differences of content, different usage restrictions, and so on, but certain users are willing to consider them as equivalent.

Which digital objects should be grouped together can not be specified in a few dogmatic rules. The decision depends upon the context, the specific objects, their type of content and sometimes the actual content. The underlying architecture has to support two main needs. It must provide methods for grouping digital library objects and must provide means for retrieval.

The Kahn/Wilensky architecture supports these higher level ideas in several ways. One is to have a digital object containing several digital objects. Thus several formats of a text might be assemble into a single digital object. Another approach is to have these variants stored as separate digital objects, each with its own handle. These handles are contained in a digital object, known as a "meta-object", which acts like a catalog record. It contains a list of the variants with their handles and information about the differences amongst them.

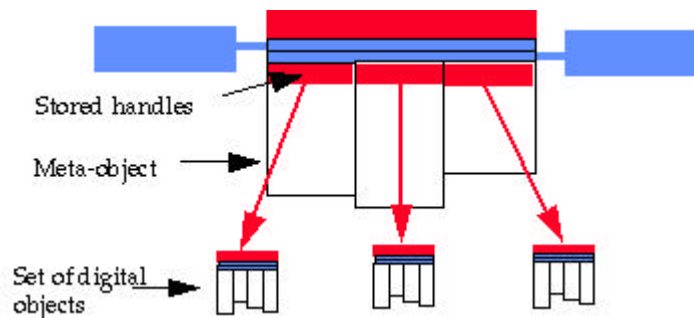


Figure 4. A digital object used as a catalog record

Reference

[hdl:cnri.dlib/tn95-01](http://hdl.cnri.dlib/tn95-01) Kahn, Robert and Wilensky, Robert. "A framework for distributed digital object services". May, 1995. (<http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html>)

Copyright © 1995 Corporation for National Research Initiatives

d-Lib forum

d-Lib magazine

Metadata: The Foundations of Resource Description

Stuart Weibel

Office of Research, OCLC Online Computer Library Center, Inc.

weibel@oclc.org

D-Lib Magazine, July 1995

This paper is an abbreviated version of the [Summary Report of the OCLC/NCSA Metadata Workshop](#). It sets forth a proposal for the content of a simple resource description record (the Dublin Core Metadata Element Set) and outlines a series of further steps to advance the standards for the description of networked information resources.

- [Introduction](#)
 - [Underlying Assumptions](#)
 - [Implementations](#)
 - [Next Steps](#)
 - [References](#)
-

d-lib forum

d-lib magazine

Introduction

The explosive growth of interest in the Internet in recent years has created a digital extension of the academic research library for certain kinds of materials. Valuable collections of texts, images and sounds from many scholarly communities -- collections that may even be the subject of state-of-the-art discussions in these communities--now exist only in electronic form and may be accessible from the Internet. Knowledge regarding the whereabouts and status of this material is often passed on by word of mouth among members of a given community. For outsiders, however, much of this material is so difficult to locate that it is effectively unavailable.

Why is it so difficult to find items of interest on the Internet or the World Wide Web? A number of well-designed locator services, such as Lycos [\[http://lycos.cs.cmu.edu/\]](http://lycos.cs.cmu.edu/), are now available that automatically index many of the resources available on the Web and maintain up-to-date databases of locations. But indexes are most useful in small collections within a given domain. As the scope of their coverage expands, indexes succumb to problems of large retrieval sets and problems of cross disciplinary semantic drift. Richer records, created by content experts, are necessary to improve search

and retrieval. Formal standards such as the [TEI Header](#) and [MARC](#) cataloging) will provide the necessary richness, but such records are time consuming to create and maintain, and hence may be created for only the most important resources.

An alternative solution that promises to mediate these extremes involves the creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create such records, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover these descriptions and collect them.

Can a simple metadata record be defined that sufficiently describes a wide range of electronic objects? The Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) convened the invitational Metadata Workshop on March 1-3, 1995, in Dublin, Ohio to address this issue. Fifty-two librarians, archivists, humanities scholars and geographers, as well as standards makers in the Internet, Z39.50 and Standard Generalized Markup Language (SGML) communities, met to identify the scope of the problem, to achieve consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas, and to lay the groundwork for achieving further progress in the definition of metadata elements that describe electronic information.

Goals

Goals of the workshop included fostering a common understanding of the problems and potential solutions among the stakeholders and promoting a consensus on a core set of metadata elements to describe networked resources.

Scope

Since the Internet contains more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves. The major task of the Metadata Workshop was to identify and define a simple set of elements for describing networked electronic resources. To make this task manageable, it was limited in two ways. First, only those elements necessary for the discovery of the resource were considered. It was believed that resource discovery is the most pressing need that metadata can satisfy, and one that would have to be satisfied regardless of the subject matter or internal complexity of the object.

Secondly, the discussion was further restricted to the metadata elements required for the discovery of what were called **document-like objects**, or **DLOs** by the workshop participants. It was believed that DLOs are still the most common type of resource sought in the Internet and that whatever solution could be proposed for DLOs could be extended to other kinds of resources. More importantly, the likelihood of making progress on this challenging problem would be increased if attention could initially be restricted to something familiar.

DLOs were not rigorously defined, but were understood by example. For example, an electronic version of a newspaper article or a dictionary is a DLO, while an unannotated collection of slides is not. Of course, the crux of the problem is that in a networked environment, DLOs can be arbitrarily complex because they can consist of text with callouts to images, audio or video clips, or to other hypertext

documents. The Metadata Workshop participants made no attempt to limit the complexity of DLOs, except to say that the intellectual content of a DLO is primarily text, and that the metadata required for describing DLOs will bear a strong resemblance to the metadata that describes traditional printed texts.

As a result of the restricted focus of the workshop, certain issues required for a complete description of DLOs, such as cost, archival status and copyright information, were eliminated from the scope of the discussion. Elements required for the description of objects other than DLOs, such as the elements required for the description of complex geological strata in a geospatial resource, were also beyond the scope of the discussion. The goal was to define a core set of metadata elements that would allow authors and information providers to describe their work and to facilitate interoperability among resource discovery tools. But because the core elements do not yield a complete description of objects in a networked environment, careful consideration was also given to mechanisms for extending the element set.

The primary deliverable from the workshop was a set of thirteen metadata elements, named the **Dublin Core Metadata Element Set** (or Dublin Core, for short). The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.

Below is a brief description of the elements in the Dublin Core **Dublin Core Element Description**

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The physical manifestation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object

To make this discussion concrete, consider an electronic a record created with the relevant portions of the Dublin Core, and a sample syntax, that describes an electronic version of Maya Angelou's poem "On the Pulse of Morning". This description is based on a record created by the University of Virginia Library's Electronic Text Center. (For a description of that project, see Gaynor [\[Gaynor\]](#).)

- **Subject:** Poetry
- **Title:** On the Pulse of Morning
- **Author:** Maya Angelou
- **Publisher:** University of Virginia Library Electronic Text Center
- **OtherAgent:** Transcribed by the University of Virginia Electronic Text Center
- **Date:** 1993

- **Object:** Poem
- **Form:** 1 ASCII file
- **Identifier:** AngPuls1
- **Source:** Newspaper stories and oral performance of text at the presidential inauguration of Bill Clinton
- **Language:** English

Underlying Assumptions

The discussions at the Metadata Workshop revealed several principles that should guide the further development of the element set. Adherence to these principles increases the likelihood that the core element set will be kept as small as possible, that the meanings of the elements will be understood by most users, and that the element set will be flexible enough for the description of resources in a wide range of subject areas. These principles are intrinsicality, extensibility, syntax independence, optionality, repeatability, and modifiability.

Intrinsicality

The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. The focus on intrinsic data in no way demeans the importance of other varieties of data, but simply reflects the need to keep the scope of deliberations narrowly focussed.

Extensibility

In addition to its use in dealing with extrinsic data, extension mechanisms will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements.

Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields. In addition, the specification of the Dublin Core itself will change over time, and the extension mechanism will allow revisions while maintaining some backward compatibility with the originally defined element set.

Syntax Independence

Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs.

Optionality

All the elements are optional. The Dublin Core may eventually be applied to objects for which some elements have no meaning (who is the author of a satellite image?). It also seems counterproductive to mandate complex descriptions if the creators of the content are expected to provide the descriptive material. A simple description is better than no description at all.

Repeatability

All elements in the Dublin Core are repeatable. For example, multiple author elements would be used when a resource has multiple authors.

Modifiability

Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier. If no qualifier is present, the element has its common-sense meaning; otherwise, the definition of the element is modified by the value of the qualifier.

Qualifiers will be typically derived from well-known conventions in the library community or from the field of knowledge appropriate to the resource. Qualifiers are important because they give the Dublin Core a mechanism for bridging the gap between casual and sophisticated users. For example, the data in the **Subject** element consists of any word or phrase that describes the object's content. However, a professional cataloger may wish to supply the name of the authoritative source from which the subject terms are taken. In such a case, the element may be written as **Subject (scheme=LCSH)**, indicating that the subject terms are taken from the Library of Congress Subject Headings.

Implementations

One of the goals of the OCLC/NCSA Metadata Workshop was to promote prototype resource description projects based on a common model of resource description. A number of Metadata Workshop conferees represent organizations that have ongoing activities or are starting activities that will be influenced by the results of the workshop. These include:

- The OCLC Spectrum Project
Contact:Diane Vizine-Goetz, vizine@oclc.org
- [The OCLC Internet Resources Cataloging Project](#)
Contact:Erik Jul, jul@oclc.org
- Library of Congress
Contact:Rebecca Guenther, rgue@loc.gov
- O'Reilly Associates
Contact:Terry Allen, terry@ora.com
- Los Alamos National Laboratory and Indiana University
Contact:Ron Daniel Jr., rdaniel@acl.lanl.gov
Contact:Pete Percival, percival@bronze.ucs.indiana.edu
- Bunyip Systems
Contact:Chris Weider, clw@bunyip.com
- Georgia Institute of Technology
Contact:Michael Mealling, michael.mealling@oit.gatech.edu, <http://www.gatech.edu/iir>
- SoftQuad
Contact: Yuri Rubinsky, yuri@sq.com
- Concordia University
Contact:Bipin Desai, bcdesai@cs.concordia.ca,
<http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html>

Next Steps

Refinement and standardization of the metadata element set defined in this document will be an ongoing, dynamic process involving many stakeholder communities. No single forum will suffice to air all concerns and no single standard can be expected to accommodate the needs of all communities. The problem must be divided into manageable chunks and the process must engage the relevant stakeholder communities. Implicit in the present activity is the proposition that there are core elements common to many object types, and that a simple, extensible framework of such elements can be defined to support more complete resource descriptions.

The initial objective--the specification of elements for the discovery of document-like objects--can be extended in a variety of directions:

- Expansion of the Dublin Core to include other object types, such as services or collections.
- Expansion of the Dublin Core to embrace functionality other than resource discovery, such as archival control and the authentication of users and charging mechanisms.
- Establishing standardized methods for extensibility.
- Refinement of existing work. The Dublin Core is an untested approach to the description of resources that will need to be modified with experience.

OCLC and NCSA will establish a workshop series to address aspects of this agenda. A Metadata Workshop Steering Committee will be established to define topics and assure appropriate representation of stakeholders. Design groups of perhaps a dozen or fewer individuals will be solicited to prepare discussion papers to focus workshop activities. Participants will be invited based on their publicly evident accomplishments in relevant areas or by reviewed application. Workshops will be limited to 50 or fewer participants and conducted in roughly the style of the March 1995 Workshop.

Other work will be done in coordination with IETF working group on Uniform Resource Identifiers (URIs) to assure that the results can be integrated into the emerging protocols for resource location and persistent naming.

Finally, active promotion of results will be carried out by establishing liaison with formal associations of stakeholders. In the library community, MARC standards evolve under the guidance of the Machine-Readable Bibliographic Information Committee (MARBI), composed of representatives of the Library of Congress and other stakeholders in the library community. A close relationship should be sustained between this committee and the Metadata Work Group. Relationships should also be established with publishers, document vendors, SGML vendors and theoreticians working on the problem of text encoding. Other communities also have requirements that must be accommodated in any framework for resource description. These communities include the GIS community, government information providers and business communication groups.

References

[MARC]

Network Development and MARC Standards, Office, ed. 1994. USMARC Format for Bibliographic data. 1994. Washington, DC: Cataloging Distribution Service, Library of Congress.

MAGAZINE

Informedia Digital Video Library

Technology Outreach

Howard D. Wactlar
Carnegie Mellon University
Howard.Wactlar@cs.cmu.edu

D-Lib Magazine, July/August 1996

ISSN 1082-9873

Background

The [Informedia Digital Video Library at Carnegie Mellon University](http://www.dlib.org/dlib/july96/07wactlar.html) is one of the NSF/DARPA/NASA jointly funded Digital Library Initiative projects, established in 1995. This particular effort focuses on search and discovery in the video medium. The Informedia project will establish a large, on-line digital video library by developing intelligent, automatic mechanisms to populate the library and allow for full-content and knowledge-based search and retrieval via desktop computer and metropolitan area networks. Initially, the library will be populated with several thousand hours of raw and edited video drawn from licensed public television documentaries and broadcast news and special events. The library is being deployed in testbeds at local area K-12 schools, at Carnegie Mellon University, and as demonstration systems at government sponsors.

The distinguishing feature of our technical approach is the integrated application of speech, language and image understanding technologies for efficient creation and exploration of the library. Using a high-quality speech recognizer, the sound track of each videotape or broadcast, combined and aligned with closed-captioning information when available, is converted to a textual transcript. A language understanding system then analyzes and organizes the transcript and stores it in a full-text information retrieval system. Likewise, image understanding techniques are used for segmenting video sequences by automatically locating boundaries of shots, scenes, and conversations. The system thus partitions video into small-sized segments and provides alternate representations and abstractions of video content to better support information retrieval and manipulation. Exploration of the library is based on these same techniques.

Component and Content Availability

Present

The highly modular system structure and implementation of the Informedia Digital Video Library system is itself a fertile testbed for researchers in many disciplines. Any of the component systems (e.g., speech recognition, image sequence segmentation; user interface display and control tools; text indexing, search and retrieval; video servers; network streaming protocols; dynamic pricing algorithms) can be exported for use in other research projects elsewhere. It is our intent to encourage investigation by DLI researchers

who have interests in any of the components as well as the overall system use and application. We can also import components from DLI members to incorporate into the Informedia system (such as natural language processing, speech recognition, or image segmentation systems, etc.), if built to our interfaces and data types. One application, News on Demand, has already been described in this magazine ([September 1995](#)) and a discussion of some of the education-related applications will be forthcoming in the fall.

Future

External research groups will have much the same set of opportunities, with restricted licensing and a different cost structure. Requests for involvement by external researchers will be evaluated by the project's principal investigators. Criteria include anticipated impact on the performance or function of the overall system and costs to integrate and verify their contributions if implementation is involved.

Maturing Informedia into a universally-usable system will enable easier access to researchers. We are currently moving towards an HTML Informedia client interface, utilizing commonly available technology to allow access over the Internet. To date, the interface has been a customized, proprietary, Windows 95 application. Research into Informedia's data and networking architecture will lead ultimately to using emerging commercial servers for data distribution, and satisfying their standards and protocols. Data and derived metadata in the Informedia library are collected under license, and can be licensed by others. We are now pursuing public domain data as well. [NetBill](#), our network billing component, is a separable body of code (both in client and server) that is being made available to other DLI sites for use as desired.

The Informedia library will continue to exist beyond the end of the current project; we expect that user support and services will be provided by third parties. We anticipate future applications of the technology in the health field, education and training, etc. Work on the various components of the Informedia Digital Video Library system (such as speech, language processing, and image understanding) will continue at Carnegie Mellon for related research efforts. We will maintain the infrastructure for creation and dissemination of digital video content, with network access as appropriate.

An important and explicit goal of this project is to accelerate acceptance of Informedia Library technologies by seeding the network community and priming the providers, both non-profit and commercial. We have assembled the project partners and organized the project structure with this goal in mind. The [partnerships](#) we have established for resources, field testing, and productization will enable us to achieve a more pervasive impact and potential commercial realization, and ultimately allow the Informedia Digital Video Library system to survive beyond its research infancy.

Copyright © 1996 Howard D. Wactlar



<hdl://cnri.dlib/july96-wactlar>

MAGAZINE

The VARIATIONS Project at Indiana University's Music Library

David E. Fenske
Head, Music Library
VARIATIONS Project Director
Indiana University
fenske@indiana.edu

Jon W. Dunn
VARIATIONS Project Technical Director
Indiana University
jwd@indiana.edu

D-Lib Magazine, June 1996

ISSN 1082-9873

History, Context and Background

The VARIATIONS Project is best known for the distribution of high-quality digital audio via an ATM network from servers and storage systems having some special characteristics to Intel-based and Macintosh clients. The evolution of this project from its beginnings in the late 1980's to its initial operational state today is inextricably connected with the design and construction of a new [School of Music Library](#) at Indiana University, and with the opportunities presented by a new design. It also addresses some pedagogical and library preservation problems. This article describes the motivation for the project and its history, its operation and experiences to date, and its future goals. Although the project is now operational, this report should not be viewed as a final one. VARIATIONS is a work in progress and represents several partnerships within Indiana University and our partnership with IBM. Information can be obtained about our internal partnerships by following links to the Indiana University [School of Music](#), the [Indiana University Libraries](#), Indiana University's [University Computing Services](#). The VARIATIONS Project, as a result of its partnership with IBM, uses many of the [IBM Digital Library](#) technologies. Information about the Indiana University School of Music Library's relationship to [IBM](#)'s plans is publicly available at the IBM Digital Library site.

Common knowledge has it that university buildings take a long time to accomplish. We can validate this observation. The first internal documents for a new music library were written in 1977. The officially endorsed proposal was first produced in 1983 with subsequent revisions in 1986 and 1989. It was with the 1989 version that the new Music Library was built.

In the earlier versions, a traditional library of the time was envisioned. The principal issue was providing twenty years of collection growth without compromising the available number of readers and listeners. The debate in 1983 was over allocating space to the listeners versus the readers.

The 1989 plan addressed the same issues for collection growth: twenty years of growth and the need to unify collections, particularly score and recorded sound collections. However, the 1989 plan also completely reexamined the issue of patron spaces. The new patron spaces envisioned a unification of listening and computing spaces (not even mentioned in the 1983 plan) and the ability to reallocate reader spaces to digital library spaces as the need arose.

Why the change? Starting in the mid-1980's, the Music Library had asked the question: if information was going to become increasingly digital in the future, what would be required to continue the place of the Indiana University Music Library at the center of an information hub in the School of Music? Starting in about 1987, the Music Library installed its first Novell server. Distributing information over a network (as opposed to standalone workstations) seemed to us the only appropriate choice for a library. Initially, this network served only a few public workstations and Music Library faculty and staff computing.

The Novell-based server combined public and staff computing and continued to evolve over the years, gradually extending to all six buildings of the School of Music complex. During these years we found new ways to distribute text-based sources, computer-assisted programs and music notation sources. Since about 1990, CD-ROM products have also become an important part of this program.

Supported by a computing vision inherent in the 1989 version of the building program, we realized that we had not yet succeeded in distributing sound nor video sources. Recorded sound had accounted for more than 50% of the items used in the Music Library for the previous 20 years. We realized that we could not move into a digital environment until we addressed the central issue of the network distribution of time-dependent data (e.g. sound).

The term VARIATIONS was first used in a joint paper (David Fenske and Michael Burroughs) presented to the International Computer Music Association at its meeting in Glasgow, Scotland, in 1990. The term has a clear musical allusion to the form, theme and variations. The term was also meant to imply the musician's need for various data formats--text, sound, video, music notation and images--in an integrated setting. Instruction and research are dependent on the aural analysis of music while simultaneously reading a score. This analytical act is supported by text-based and music notation-based research.

The technical challenges in distributing sound over a network became the focus of the VARIATIONS Project from 1990 until its successful operational deployment on April 1, 1996. In several respects, the new Music Library building and the VARIATIONS Project are both focused on the same issue differing only in the environments: unifying and integrating collections of information principally in text, score and recorded sound formats.

From 1992, the VARIATIONS Project examined server, network and client technologies from all of the principal computing companies. We found many worthy products addressing one or more of our requirements. It became clear to us, however, that our concept was, in 1992, beyond the capabilities of technology. We were encouraged that parts of our vision had, then recently, come into existence and that all of these companies were emphasizing at least some of the concepts that then came to be known as the digital library.

One of operating principles in examining technology from many companies was that it had to be shown to work at Indiana University including servers, networks and clients. The computing environment on the Bloomington campus and in the School of Music is heterogeneous. Intel-based and Macintosh-based

machines abound in about equal number. The campus network supports a variety of network protocols, IP, IPX Appletalk and others. UNIX and Novell servers are common throughout the campus and in the Music Library. UNIX workstations from a variety of vendors are more common elsewhere on the campus than they are in the School of Music.

Many products were examined that were by themselves exciting but failed our needs for a networked-based distribution of information in a time-dependent form (i.e., sound). We examined a UNIX-based workstation that had better sound support than any other platform, but added nothing to the network distribution solution. Regrettably, this company no longer exists. We examined UNIX-based products from other vendors some of which did address our need to distribute information over a network. Most of these products failed to integrate well into our campus network or they failed to scale to a level meeting our needs.

For a couple of years, the solutions seemed to lie with UNIX-based clients and servers and it looked as though our problem would be to entice our users away from their Intel and Macintosh-based computers. The reasons were the networking tools native to the UNIX environment and the early deployment of high-level sound manipulation tools combined with high-quality sound. However, this proved to be impossible, as many of the applications needed by our users, especially music-related applications, were only available for Intel and Macintosh platforms. Windows and Macintosh emulators for UNIX workstations could not deal well with sound or MIDI (Musical Instrument Digital Interface) connections to synthesizers. Our examination gradually shifted from one focused primarily on clients to one focused on the network and the servers, with Macs and PC's as the clients. In the process, the contending technology companies were quickly narrowed down to two and then one.

The existing Ethernet-based campus networking solutions present on the campus did not deal well with real-time audio or video streams, due to the fact that their bandwidth is shared in a building by potentially hundreds of stations. For our new building, we had to look to switched networking technologies to accommodate our needs. We considered a number of networking schemes but only two were serious contenders: ATM and switched Ethernet. Switched Ethernet had several initial advantages. It substantially increased the bandwidth dedicated to each workstation. It could be combined with yet-higher bandwidth building backbones even involving the promise of ATM's eventual quality of service and resource reservation from the server to the switch. Switched Ethernet was at the time a more established technology and would have been the more conservative choice. There were some who also argued that it was cheaper than ATM to deploy.

We chose ATM over switched Ethernet for several reasons. While switched Ethernet does provide sufficient bandwidth for some of our immediate needs, there were questions about how long switched Ethernet would serve our purposes. Because of the scale of the VARIATIONS Project one of our choices was use of ATM in the building backbone. As 25 Megabit/second (Mbps) ATM adapters were released and dropped in price, the issue became one of ATM to the desktop. The ability of ATM to reserve bandwidth via quality of service guarantees also formed part of this argument. Sound alone is a more critical network problem than video despite today's video-driven development of networking technologies. Video over a network degrades for a while before stopping altogether. During this degradation, annoying as it may be, information context is not lost. High quality audio only does not degrade gradually, it simply stops. In less than a second, information context is lost when audio over a network breaks. In view of this critical observation, ATM was the only networking technology that promises guaranteed service through resource reservation. Based on the functional requirements addressed in this paragraph, ATM came out somewhat ahead of switched Ethernet, but there were other issues as well.

Even given the declining costs of 25 Mbps adapters, an ATM adapter is more expensive than Ethernet adapter for switched Ethernet. The same comparison holds true for the rest of the networking environment. The question for us became: Was switched Ethernet really the most economical choice? As a wiring plant, we chose category 5 unshielded twisted pair to the desktop, already a more economic choice than many new buildings built only a few years earlier. (They often chose fiber optic to the desktop, which costs more as a wiring plant and as a desktop device, but delivers high bandwidth.) Although 25 Mbps will work over the category 3 twisted pair common in most buildings on the Bloomington campus, category 5 meant that we could deploy higher-speed ATM in the future without rewiring and that we would not have to install fiber to the desktop. (Fiber does connect the ATM switches and the VARIATIONS Project's servers.) Still, switched Ethernet would have worked over category 5 as well and even over category 3 wiring.

The critical components of the economic question became long-term bandwidth needs indicating category 5 and ATM and what might be called the replacement factor. Switched Ethernet might have won the economic argument if we were retrofitting an existing building and needed to make the minimum amount of physical alteration in order to increase bandwidth to the desktop. Installing switched Ethernet in a new facility combined with the functional arguments articulated previously suggested that we would want to replace switched Ethernet within a couple of years for functional reasons. The combined costs of installing switched Ethernet and then replacing it within a short period of time was judged much more expensive as well as unlikely to succeed in a university context. In short, ATM, although initially more expensive, provides a much longer service life than switched Ethernet and was, therefore, for us the economical choice.

There was also another argument in favor of ATM: technology development. While we were in the process of the preceding network examination, we were also carrying out an examination of servers (and to a lesser extent clients). IBM could offer the greatest number of components in essentially an end-to-end installation. IBM stayed with us through years of examination and allowed us to influence their choices in the digital library environment. So many of the decisions we were making generally were high risk ones. The ability to influence technology development and to reduce the vendor -to-vendor finger pointing typical of mixed vendor deployments made IBM the logical choice. Although IBM could have provided either a switched Ethernet solution or an ATM one, it was clear the future belonged with ATM.

The end-to-end solution became additionally important when one also considers the technological challenges involved with serving audio and video data and with storing this data. The VARIATIONS Project, as a result of its partnership with IBM, uses many of the [IBM Digital Library](#) technologies. Information about the Indiana University School of Music Library's relationship to [IBM](#)'s plans is publicly available at the IBM Digital Library site.

Technical Overview

In describing how VARIATIONS works, we can break the system into three primary parts: content creation, content storage, and content distribution.

Content creation

Student workers (under the direction of Constance Mayer, Head of Circulation Services) use specially-equipped personal computers to create CD-quality sound files in Microsoft's .WAV format

from original analog or digital media. We are using a 16-bit sample size at a sampling rate of 44.1 KHz, the same quality used by audio compact discs and typical commodity sound cards for personal computers. In the case of CD's, the sound is already in digital form and can be transferred directly from CD to hard disk without any loss of quality using Microtest's [Disc-to-Disk](#) software on Macintosh and Intel-based workstations equipped with CD-ROM drives. Records, cassette tapes, open-reel tapes, and other analog media must be converted to digital form using Intel-based workstations equipped with Turtle Beach and Roland sound cards. Analog recordings require more attention in order to get good quality results, as one must carefully set recording levels and monitor recording progress.

In addition to simply creating a sound file copy of the original recording, the students also enter the index or band information from the original recording. This is information which is not available in the existing online library catalog record for the item, but is necessary to provide a level of access for the patron which approaches that of having the actual item with CD booklet or record jacket in hand. For this task as well, CD's are easier to work with; a locally-written Macintosh program can extract precise index timing information from the CD itself, requiring the worker to only input the description of each track from the CD booklet. Analog recordings require that the worker carefully identify the exact locations of track breaks and enter this timing information for each track in addition to the descriptions.

After creating a sound file and a track description file on one of the digitizing workstations, these files are transferred via FTP to a central IBM RS/6000 archive server (discussed further below). At night, a batch job runs which compresses these files into [MPEG](#) format, using a 3.6:1 compression ratio. MPEG audio compression works by eliminating frequencies in the sound which cannot be perceived by the human ear and mind. Most listeners have found the MPEG-compressed audio to be of more than acceptable quality for day-to-day use, and the original full-quality uncompressed files are always kept for preservation purposes. Another advantage of MPEG beyond the decreased file size is that it provides a common file format for Intel, Macintosh, and UNIX workstations.

Content Storage

There are two primary servers in the VARIATIONS system for storage of digital audio: a playback server and an archive server. The playback server consists of an IBM RS/6000 Model 59H with 120GB of hard disk storage. This server can store over 600 hours of MPEG-compressed CD-quality audio on file systems managed by an IBM software product known as Multimedia Server for AIX. Via a filesystem technology known as [Tiger Shark](#), Multimedia Server provides for striping of audio and video files across multiple disks, which provides load balancing and guaranteed real-time delivery of these files.

The archive server is an IBM RS/6000 Model J30 with an attached IBM 3494 Optical Tape Library Dataserver containing two IBM 3590 tape drives, which is managed by IBM's [ADSTAR Distributed Storage Manager](#) software. This library can hold up to two terabytes of content, or over 9000 hours of compressed audio. The 3590 drives, with a nine megabyte/second transfer rate, allow for fast access and retrieval of large multimedia files. Currently, this server is being used to archive the uncompressed sound files and store backups of the compressed files residing on the playback server. Later this year, software will be added so that the archive server will be able to transfer MPEG-compressed audio files to the playback server on demand to provide a larger amount of online storage for audio files being accessed by patrons. At that point, the playback server will essentially be acting as a most recently used cache for the sound files residing in the archive server.

Tape was chosen over optical technology for this application because of its higher transfer rates and

better cost/megabyte ratio. While optical storage media offer the advantage of faster seek times than tape, their data transfer rates and seek times are so much slower than disk that sound files would still have to be copied to disk in order to provide multiple simultaneous access to the same file.

Content Distribution

Library patrons access sound recordings in the system from 45 IBM Pentium computers located throughout the library and in a teaching classroom/cluster on the third floor of the library. These stations all currently run Microsoft Windows 3.1, but an upgrade to Windows NT is anticipated in the near future. Each of these stations is equipped with a sound card (IBM Mwave), MPEG audio decoder software from [Xing Technology](#), CD-ROM drive, Kurzweil K2000 synthesizer/keyboard, and headphones. Beyond the access to digital audio, these stations also deliver general computing functions (word processing, e-mail, spreadsheets, etc.), library computing functions (access to CD-ROM databases and the library catalog), and music computing functions (ear training, music notation, composition).

Two scenarios exist for locating and playing recordings in VARIATIONS. The first case is that of a student who needs to listen, for a class assignment, to a particular recording which has been placed on reserve by the instructor. The student sits down at a workstation and launches Netscape, which is set to use the Music Library home page as its starting page. From this page, the student selects "Course reserves," which takes the student to a list of courses being offered in the current semester. The student selects the proper course to obtain a list of recordings on reserve for that course, and then selects the recording to which he or she wishes to listen. This launches a locally-written VARIATIONS Player application which begins playing the sound file from the playback server across the building ATM network. The student has full control over playback of the recording, with the ability to stop, start, rewind, and fast forward. He or she can easily move through the tracks of the recording to get to the particular work or section desired.

The second case is that of a patron who wishes to listen to a particular recording independent of any course assignment. In this case, the user would select IUCAT, Indiana University's NOTIS-based online catalog system, from the Music Library home page, and perform a search of the catalog to find the item desired via the standard NOTIS terminal-based online public catalog interface. If the item is available online, a URL pointing to the online copy of that item will be displayed along with the rest of the catalog record. The user can then cut and paste that URL from the terminal window into Netscape to access the item. Indiana University is planning to implement Ameritech Library Services' [WebPac](#) World Wide Web to [Z39.50](#) gateway software to provide a true web interface to the catalog later this year. With WebPac, the user will simply be able to click on the URL when viewing the catalog record in order to access the online copy of the item.



A screen shot of the VARIATIONS Player application

A Word about Networks

One of the reasons, along with copyright, that VARIATIONS is only accessible within the new Music Library building is that of networking. The existing Ethernet-based campus and building networks at Indiana University are not capable of dealing with large numbers of real-time multimedia sessions.

In the [building](#), we are using an IBM ATM network with a combination of 100 and 155 megabit/second links over fiber-optic cabling to servers, and 25 megabit/second links over copper unshielded twister pair wiring to client PC's. ATM was chosen as the network technology for the new building due to its long-term advantages for real-time multimedia traffic. Currently, audio data is delivered from server to client via the NFS (Network File System) protocol running over ATM via Ethernet LAN Emulation. We hope to be able to transition to using native ATM services with the ability to reserve bandwidth via quality-of-service guarantees. This will require, however, that Multimedia Server product be adapted to support this and that API's and drivers which support quality-of-service become available for Windows and UNIX operating systems.

Our experience in running a production ATM network has been, for the most part, positive. We have not run into any significant management or stability problems, although it is admittedly more difficult to troubleshoot problems when they do occur due to lack of diagnostic tools and the extensive pool of knowledge which has been built up for older technologies such as Ethernet.

Experiences

VARIATIONS was up and running for public access for the first time on April 1, 1996, delivering [course reserves](#) for two undergraduate Music Theory classes, one containing about 20 students and the other containing about 150 students. Training sessions were conducted (by Jon Dunn and Constance Mayer) for both classes; a hands-on session was used for the smaller class while a demonstration/lecture was used for the larger one. In both cases, [step-by-step instructional handouts](#) were provided. Students seemed to be able to pick up quickly on how to use the system, most having had some computing experience (word processing, e-mail, web browsing) previously. By the end of final exams in early May, sound files were being launched over 1000 times per day. For the summer session beginning in June, we plan to provide at least fifty percent of reserve listening materials via VARIATIONS.

A feedback form is provided on the web for students to submit questions, comments, or problem reports regarding the system. Most questions have been of the form, "Why can't I access the recordings from my home/dorm room/favorite campus computer lab?" Now that students have had a taste of what electronic access to sound recordings can provide, their desires for more capability have increased faster than technology can respond. Many faculty members have also been intrigued by the possibilities of VARIATIONS. A number of faculty members, with varying degrees of computer background, are very interested in using VARIATIONS in their instruction.

Future goals

There are a number of library-related aspects not necessarily apparent in a discussion driven by technology: access and preservation. For the first time, digital preservation practices mean greatly

improved analysis and restoration capabilities and increased access to information in all formats.

Digital preservation standards are still evolving. For music, as for all areas in the humanities, preserving information is a crucial component. Readers whose disciplines lie outside of the humanities may not always appreciate the fact that information, for the humanist, retains its research value for extremely long periods of time. It is axiomatic that as publication activities passed through the Industrial Revolution in the early 19th century, the longevity of publications actually decreased due to changes in paper manufacturing processes.

Society at large may generally regard recorded sound as largely entertainment. For the musician, it represents nearly 100 years of changing performance practice now available for research. With the exception of the compact disc, all recorded sound media are regarded as fragile since they deteriorate with each use even under the best of conditions. Even compact discs are not indestructible. Digital preservation captures manuscript, print and recorded sounds in their current state. In all of these cases, we see now the development of digital tools to restore the image or the sound so that nuances crucial to the scholar can again be observed.

The VARIATIONS Project as a digital library project not only means better instructional and research tools, it also means improved access to information: 1) retrieval of the full information object is linked to its corresponding bibliographic record in the online catalog; and 2) in most many cases particularly with graphic images and textural data, the information can be distributed to users elsewhere on the campus, on other campuses and potentially the world.

There are a number of immediate term goals we are pursuing in the library information delivery phase. The solutions to these goals are not yet known:

1. On-demand digitization. We plan soon to digitize all recordings in the order in which they are requested and the to link the digitized file to the online catalog. While there is no problem involved with this process if the patron requests the material in advance, such planning on the part of patrons is the exception. On-demand digitization means that we allow the patron to passively listen to the material as it is being digitized with another stream going to the server for storage and for linking to the online catalog. The technology to split a real-time stream, directing it to two places, is not yet in place.
2. Video. While we do not anticipate any further problems with the serving nor network distribution of video data, we have only recently begun acquiring equipment supporting video digitization. Video is a format of secondary importance for us and one which still requires unusual amounts of computing power for compression and standards-compliant, high-quality client software. From these perspectives, it has been less affordable and has more fluid standards than those for audio.

Since we are still early in the operational deployment of the VARIATIONS Project, the reader may have the impression that this is a project driven largely by pedagogical and library information delivery goals in a single building. This impression would be largely incorrect. It is merely the point where we needed to start.

Having accomplished the network distribution of high-quality sound data within a single-building ATM network, we will investigate wide area distribution. This distribution ranges from campus academic buildings (and some dormitories) attached to the network usually via Ethernet and FDDI to services delivered to other campuses of the university via an ATM wide-area network, to services delivered via modem connections. Eventually, the VARIATIONS Project's services will range from guaranteed under

ATM or other network technologies providing quality of service guarantees, to best effort for high quality information over non-switched Ethernet and to degraded quality over modems. In order for the project to deliver these services, we will need to develop more intelligent software determining the network capabilities of the requesting user as well as tools to gracefully degrade the quality of the sound to match the quality of the network connection. We have actively discussed these plans with our technology partners and will be pursuing solutions in the coming months and years.

One aspect of the above wide area distribution problem is technical (as described in the preceding paragraph) and another is mission-oriented. Our purpose in distributing information is to support the educational mission of Indiana University in the School of Music, on the Bloomington campus, and on other campuses of Indiana University as well as distance-based education. We do not provide free copies of content to users even in this environment. Only a couple of minutes of sound are in memory at any one time. The VARIATIONS Project never distributes a full copy of a work for use by the end user.

For end-user desktops on the campus network, we check the location of the user by the Internet Protocol address before allowing use of commercially-produced sound. Degraded quality content distributed via modems will not be any more attractive to users than are low quality images from art museums. The problem will come with inter-library sound requests now in the digital environment. (It should be noted that we presently honor most inter-library loan requests for out-of-print recordings by sending a cassette copy: a common practice.) The inter-library loan request for a digital copy (such requests have already been received) of an out-of-print recording are inevitable and would require the transmission of a full copy. Notice that we have restricted this discussion to out of print material since in print material should always, in our view, be purchased by the requesting library. In order for out-of-print digital material to be distributed with the minimum of difficulties between copyright owners and libraries, we must have this data encrypted with a time limited (such as 30 days) key. After the key has expired, the data is useless. The borrowing library will presumably erase the file since it is useless. Note that this arrangement provides for tighter controls in the digital world than those of the analog world.

The VARIATIONS Project will create databases of score notation. The primary difficulty is scanning musical scores is the size of the publication, which often exceeds the standard sized 8.5" X 11" scanner. Scanned images of musical scores are useful for reserves, incorporation into HTML files, etc. While these are useful activities, they do not themselves particularly advance research. Within the last two years, a few music character recognition programs have come into existence. One of these, from AR Editions of Madison, WI, does a particularly good job of converting images of printed music into notational files. Unfortunately, the resulting notational file is thus far readable only by AR Edition's music editor. The goal of most music character recognition programs is to reproduce as completely as possible the printed page including not only pitch, meter and rhythm, but also many other facets of music publication such as slurs, accents, ornamentation, etc. The objective is to facilitate further editorial work or publication-related activities.

Also in the past two years, Prof. David Huron (from the University of Waterloo) has released his [Humdrum Toolkit](#) permitting queries of notational files in a variety of notational file formats and in a variety of operating environments.

All of the above are promising signs. One of the VARIATIONS Project's goals is to be able to convert large amounts of printed works into a database. Queries could be formed requesting stylistic information from a large data-set on a scale not previously attempted. The requirements for these activities are different from the kind of ongoing activities listed above:

1. Much of the development effort in present music character recognition programs is directed towards converting all of the information including a great deal of editorial nuance such as slurs, accents, ornamentation, etc. While these are of interest to a publisher, they are not always of interest to the scholar who is trying to form queries based largely on pitch, meter and rhythm. (Most accents, slurs and even ornamentation represent the editorial interpretation incorporated in the modern publication. They are much less frequently the work of the composer.)
2. The history of database software shows that eventually data gets to a size and a complexity requiring database software to manage the data, reports and queries. It is likely that music notational information also shares this characteristic as well.

The VARIATIONS Project will create or assemble tools to directly analyze, and query sound databases. Musical scholars at least since the initiation of musicology in the latter part of the nineteenth century have focused on the score when studying or analyzing music. When members of the general public study a piece, the reference is usually to the score. In both of the preceding instances, the musical sound, live or recorded, is often used to reinforce or confirm score-based observations.

Music score notation, despite the best efforts of composers and other musicians is, at best, an approximation of the composer's intention. Musical conventions (performance practices) have an immediate impact on the process, then and now. The focus of much of musical scholarship for the last 150 years has been to recreate a critical text that will interpret the composer's text and the then contemporary performance practices for modern musicians, who are themselves the product of a differing set of musical conventions.

We now have nearly 100 years of recorded music. Each fixed performance is itself an interpretation of a musical work which differs by necessity and by intent from all other recorded (and live) performances of that work. This observation is also true of the composer performing his/her own works. We are no more likely to identify the "perfect" performance, than the perfect and final critical text. By being able to more directly query recorded sound without any reference to the score, we will gain a view of performance practice and of interpretation that is event driven and takes into account the defining characteristic of music, sound.

While we know how to represent sound, we have not had subtle enough tools to query representations for the small nuances in frequency, duration and amplitude that allows us to study one event as different from another of the same musical work in ways that are insightful. In short, we do not have the tools that allow us to study with any level of discrimination approaching the level of the human ear and intellect. While not trying to displace the role of the human mind, we are limited in our abilities to recall accurately large amounts of sound. Database software that allows us to manage, organize, compare, query and report sound directly promises the potential for new perspectives and new research.

Conclusion

The VARIATIONS Project at Indiana University's School of Music Library has recently achieved initial operational success. The Project will be addressing issues of wider distribution of sound and data in other formats and of score and sound database creation.

Related links

- [Indiana University Music Library home page](#)
- [Indiana University VARIATIONS Project home page](#)

- [IBM Digital Library home page](#)
- [IBM Networking home page](#)
- [MPEG frequently asked questions](#)

June 1996

Copyright © 1996 David E. Fenske and Jon W. Dunn



hdl://cnri.dlib/june96-variations

Annotated Bibliography - Digital Libraries

Case Studies:

- UC Berkeley, Digital Library Project
Van House, N.A., Butler, M.H., Ogle, V., & Schiff, L. (1996, February). User-Centered iterative design for digital libraries: the Cypress experience. D-Lib Magazine 2 (2),(On-line). Available: <http://www.dlib.org/dlib/february96/02vanhouse.html>
This paper reports on the experiences and results of the project's design and evaluation process. After the image-retrieval system Cypress is introduced, the method of assessment and evaluation is described. Important findings for interface design are reported, such as reduction of cognitive load and error prevention, and a table is provided to demonstrate the changes made to the interface.
- Van House, N., Butler, M., & Schiff, L. (1996). Needs assessment and evaluation of a digital library environment: the Berkeley experience, (On-line). Available: <http://info.sims.berkeley.edu/~vanhouse/dl96.html>
This report describes the design approach and context and preliminary findings of the assessment and evaluation study. A number of desired uses were identified, such as locating information, analyzing data, dissemination, publishing and re-use of information.
- University of California, Santa Barbara (Alexandria Project)
Alexandria Digital Library Project Team (1996). 1996 annual report, (On-line). Available: <http://alexandria.sdc.ucsb.edu/public-documents/annual-report/>
Detailed report about the progress of the ADL until February 1996 and plans for the future. Includes sections on the WWW prototype development and the testbed components, management issues (organizational structure, staff), various research efforts (such as metadata, interface design, evaluation, image processing), collaborations with outside institutions and research projects, and finally, educational activities.
- Frew, J., Carver, L., Fischer, C., Goodchild, M., Larsgaard, M., & Smith, T. (1995). The Alexandria rapid prototype: building a digital library for spatial information. 1995 ESRI User Conference Proceedings, (On-line). Available: <http://www.esri.com/resources/userconf/proc95/to300/p255.html>
Describes the first testbed developed at ADL, the Alexandria Rapid Prototype. In particular, the user interface, the prototype architecture (catalog, collection, software, hardware) and future plans are introduced.
- Smith, T.R. (1996). A digital library for geographically referenced materials. Proceedings of Untangling the Web 1996, Santa Barbara, California, (On-line). Available: <http://www.library.ucsb.edu/untangle/frew.html>
This paper describes the different elements of the ADL, such as its general architecture, the catalog component, the user interface, and image processing. For each component technical and methodological issues are discussed.
- University of Illinois at Urbana-Champaign
Bishop, A.P. (1995). Working towards an understanding of digital library Use. D-Lib Magazine 1 (3),(On-line). Available: <http://www.dlib.org/dlib/october95/10bishop.html>
Describes the joint efforts of the user-research working groups of six DLI projects.

Synchronization meetings and creation of a mailing list are means to achieve a joint research program. The aspects of evaluation include adequacy, search and retrieval performance and behavior, effect on work of users and implications on public policy.

Bishop, A.P., Star, S.L., Neumann, L., Ignacio, E., Sandusky, R.J., & Schatz, B. (1995). Building a university digital library: Understanding implications for academic institutions and their constituencies. In Higher Education and the NII: From vision to reality. Proceedings of the Monterey Conference 1995, Washington, DC, (On-line). Available:

http://anshar.grainger.uiuc.edu/dlisoc/monterey.final_copy.html

Report of the user-evaluation team of the DLI project. This study examined general information seeking behavior, computer use, journal use and library use, of undergraduate and graduate students, faculty and middle and high school students. The team focused on the impacts on individuals, institutions, the virtual community and on workflow. Therefore, interviews and observations were conducted.

Schatz, B., Mischo, W., Cole, T., Hardin, J., Bishop, A., & Chen, H. (1996, May). Federating repositories of scientific literature: The Illinois Digital Library Project. IEEE Computer, May 1996, (On-line). Available: <http://www.grainger.uiuc.edu/dli/ieeecom.htm>

Reports on the research efforts of this project, which aims at providing multiple views for distributed repositories. It includes a description of the testbed architecture, methods and techniques for indexing of materials, usage evaluation, interface and server architecture via multiple protocols and scalable semantic retrieval.

- Carnegie Mellon University (Informedia Project)

Christel, M. (1995). Addressing the contents of video in a digital library. ACM Workshop on Effective Abstractions in Multimedia, 1995, (On-line). Available:

<http://www.cs.tufts.edu/~isabel/christel/christel.html>

This paper describes the challenges the media video and audio raise for DLs. The authors explain the methods and techniques used to provide searching and browsing of segmented video and audio. Video paragraphing, information visualization and precision and recall in the DL environment are discussed.

Wactlar, H., Kanade, T., Smith, M., & Stevens, S. (1996). Intelligent access to digital video: The Informedia Project. IEEE Computer, 29 (5), (Online). Available:

<http://www.computer.org/pubs/computer/dli/r50046/r50046.htm>

Latest report on the progress of the Informedia project. Explains the major obstacles in speech recognition for audio material, video content representation for retrieval and the user interface architecture. Nicely enhanced by graphics.

- University of Michigan

Birmingham, W.P. (1995). An agent-based architecture for digital libraries. D-LIB Magazine, 1 (7), (On-line). Available: <http://www.cnri.reston.va.us/home/dlib/July95/07birmingham.html>

First part describes the proposed services of the UMDL and the system specifications. In the second part of the paper the agents of the UMDL are described in more detail, but focus is on their way of interacting to provide service.

Birmingham, W.P., Drabenstott, K.M., Frost, C.O., Warner, A.J., & Willis, K. (1994). The University of Michigan Digital Library: This is not your father's library. Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries 1994, College Station, Texas,

(On-line). Available: <http://www.cSDL.tamu.edu/DL94/paper/umdl.html>

Provides an outline of the UMDL, an earth and space science DL model. Includes the design principles of the system architecture, description of agents used, protocols, testbed design and user evaluation. Only reports on plans, not findings.

Wellman, M., Birmingham, W.P., Durfee, E.H., Rundensteiner, E.A., Glover, E., & Mullen, T. (1996). Building the University of Michigan Digital Library. IEEE Computer, special issue on building large-scale Digital Libraries, May 1996, (On-line). Available:

<http://ai.eecs.umich.edu/people/wellman/pubs/Building-UMDL.html>

Detailed description of the projects distributed agent architecture with regard to its application in education and research. Discusses, in particular, various agent types, their tasks and collaboration, possible search forms, and the challenges for teachers and students to use the system.

- Stanford University

Note: Papers in this series are in development and are not in a final form for publication or general dissemination. They are subject to change. Please do not quote or further distribute them without explicit permission from the authors.

Cousins, S.B. (1995). A task-oriented interface to a digital library. Working Paper, (On-line).

Available: <http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC44.html>

In the first part of the paper the goals of the interface design are described (support of user tasks, time feedback, extensibility). The second part provides a brief description of the InfoBus prototype.

Reich, V., & Winograd, T. (1995). Working assumptions about the digital library. Working Paper, (On-line). Available: <http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC10.html#what>

As a first step in the design process of the Stanford digital library an outline of its objectives were identified. These include a general definition of the DL, target user group, its content and platforms.

Winograd, T. (1995). Conceptual models for comparison of digital library systems and approaches. Working Paper, (On-line). Available:

<http://www-diglib.stanford.edu/diglib/WP/PUBLIC/DOC13.html>

The author discusses different conceptual models of digital library systems in order to provide a tentative conceptual framework for DL research. He primarily distinguishes between 'data models' and 'activity models' and discusses applications and operational consequences for each model.

- Columbia University

Cartolano, R., Gertz, J., & Klimley, S. (n.d.) Oversized color images: addressing issues of preservation and access, (On-line). Available: <http://www.columbia.edu/~klimley/oversized1.html>

Final report of the Oversized Color Image Project Phase 1. Issues of digitization, access, evaluation and bibliographic control are discussed. It is concluded that digitized images from the original and the micromedia version have equal quality.

- Library of Congress, National Digital Library Project

American Memory User Evaluation Team (1993). Final report of the American Memory user evaluation 1991-1993, (On-line). Available: <http://lcweb2.loc.gov/ammem/usereval.html>

Very detailed report on the 2 year field study conducted to determine the project's most appropriate primary audience and to measure user reactions, needs and expectations. It was found

that the project had its best acceptance in secondary schools and was mainly used to teach research and critical thinking skills.

Arms, C.R. (1996, April). Historical collections for the National Digital Library: lessons and challenges at the Library of Congress, part 1. D-Lib Magazine, 4, (On-line).

Available: <http://www.dlib.org/dlib/april96/loc/04c-arms.html>

Part one of a report on the experiences and progress of the project. Gives the background and history of the project, but focuses on the digitization process and its problems, as well as on project management issues, workflow and file description issues.

Arms, C.R. (1996, May). Historical collections for the National Digital Library: lessons and challenges at the Library of Congress, part 2. D-Lib Magazine, 5, (On-line).

Available: <http://www.dlib.org/dlib/may96/loc/05c-arms.html>

Part two of the paper focuses on the problem of facilitating navigation and interfaces to a diverse user community and the organization of the collection. Bibliographic records, support of browsing of subject terms, full-text search, finding aids and access via the library's OPAC are provided as navigation aids. The problem of vocabulary aids and the limitations of computer screens are addressed. A description of the storage system and process and the naming schemes for collection organization follows.

Fleischhauer, C. (1994). Organizing digital archival collections: American Memory's experiences with bibliographic records and other finding aids. Proceedings of the Seminar on Cataloging Digital Documents 1994, Charlottesville, Virginia, (On-line). Available:

<http://lcweb.loc.gov/catdir/semdigdocs/carl.html>

Describes the experiences of the American Memory Project with the organization of a digital collection. Focuses on the electronic records for large multipart and multifformat collections, copyright information in the electronic record, the digitization process and the management of processing activities.

Fleischhauer, C., & Erway, R.L. (1992). Observations on the reproduction of various library and archival material formats for access and preservation, (On-line). Available:

<http://lcweb.loc.gov/pub/american.memory/white.papers/reprod.txt>

Describes the digitization policies and procedures of the American Memory project. Facsimile vs. searchable text as reproductions, as well as factors of reproduction quality for different media types and efficiency of reproduction are discussed.

- Intercat

Caplan, P. (1993). Cataloging Internet resources. The Public-Access Computer Systems Review, 4 (2), 61-66, (On-line). Available: <http://www.nlc-bnc.ca/documents/libraries/cataloging/caplan.txt>

Describes developments that led to the creation of the 856 field of the Marc record. The author also addresses the problem of including the location of the electronic document in the Marc record. Finally, the problem of distinction between electronic documents and online services is addressed marginally. Paper precedes the OCLC project, but contributed to its creation.

Olson, N.B. (1995). Cataloging Internet Resources: A manual and practical guide, (On-line).

Available: <http://www.oclc.org.oclc/man/9256cat/toc.htm>

Guidelines for participants of the OCLC project. The manual is to be used with the OCLC document 'Bibliographic Formats and Standards'. Covers bibliographic description and bibliographic access and gives examples.

Shieh, J. (n.d.). Does it really matter? The cataloging format, the sequential order of note fields, and the specifics of field 856. OCLC Internet Cataloging Project Colloquium, (On-line).

Available: <http://www.oclc.org/oclc/man/coloq/shieh.htm>

Introduces the University of Virginia Library's method of Internet cataloging. In particular, selection of the serials format, order of 5xx fields in different catalog records (OCLC and VIRGO), and the handling of the 856 field are discussed.

- IDEAL by Academics Press

[About IDEAL](#)

Short, one-page introduction to the International Digital Electronic Access Library (IDEAL) by Academics Press. Describes what IDEAL can do for you as a customer, but does not give any technical or organizational specifications.

- IBM Digital Library

Anderson, L.C., & Lotspiech, J.B. (1995). Rights management and security in the electronic library. *Bulletin of the American Society for Information Science*, 22 (1), 21-23.

Description of rights management and security in the ISI Electronic Library Project. Gives details about the underlying system architecture, rights management by user authentication and session encryption, watermarking, digital fingerprints to secure authenticity of documents and addresses user privacy issues.

Lunin, L.F. (1995). IBM announces electronic copyright solutions. *Information Today*, 12 (5), 1,3,5.

Brief report on recent developments around the IBM DL and related research projects. Introduced are IBM's latest techniques for marking, encryption, metering and billing. Finally, the architecture and functions of the IBM DL are described.

Mintzer, F.C., Boyle, L.E., Cazes, A.N., Christian, B.S., Cox, S.C., & Giordano, F.P. (1996). Towards online, world wide access to Vatican Library Materials. *IBM Journal of Research and Development*, 40, 2, 139-162.

Very comprehensive description of the Vatican DL project. Provides information on their requirements, workflow of cataloging of scanning, system specifications (server and interface), the digital watermark technique, image display and conclusions of the project.

- TULIP by Elsevier Science

Tulip final report, (On-line). Available: <http://www.elsevier.nl/info/projects/trmenu.htm>

The final report of the TULIP project first gives a general project description, addresses the experiences with technical issues at Elsevier Science and the participating universities. Results of the user evaluation, organizational and economical issues and consequences for future digital library research are also discussed. The report concludes that economical and technical obstacles as well as user needs are central and must not be underestimated.

Lynch, C.A. (1995). The TULIP project: context, history, and perspective. *Library Hi Tech*, 13 (4), 25-30.

Summary of the TULIP project, describing the technological and historical context, by one of the project implementors at UC. He addresses selection criteria for content, obstacles in format and resolution, organizational issues and a number of challenges encountered during the implementation at the participating institutions.

Mostert, P. (1995). TULIP at Elsevier Science. *Library Hi Tech*, 13 (4), 25-30.

Summary of the issues involved with the large-scale production of journal material in electronic form at Elsevier Science. Describes in particular, format issues, production and customization methods and problems of file delivery via ftp. Concludes that storage, bandwidth and Internet transfer were the major challenges.

Worona, S., Saylor, J. (1995). TULIP at Cornell University. *Library Hi Tech*, 13 (4), 61-4.

Brief description of the TULIP project from the viewpoint of one of the participating universities. Describes problems of incorporating the TULIP material into the pre-existing library environment and the file transfer method, as well as the move to a Web based environment.

Imagining the Digital Library:

Library as Place:

Arnold, K. (1995). Virtual transformations, the evolution of publication media. *Library Trends*, 43 (4), 609-626.

The author discusses the changing state of publication media in its latest development such as the WWW and digital libraries. The end of the print medium has been announced frequently throughout the last decades and men like Bush and Nelson have developed scenarios, that we now come close to. This development shows its first effects on the academic world with the use of e-mail, the WWW and finally, digital libraries.

Billington, J.H. (1996). A technological flood requires human navigators. *American Libraries*, 27 (6), 39-40.

After examining the public libraries' contribution to the democratization of society, the author argues that information society will evolve backwards in the evolutionary process unless libraries provide for the transition of raw-data to knowledge and librarians will be knowledge navigators. In addition, he stresses the importance of the library as a vital place for community interaction.

LaRue, J. (1993). The library tomorrow: a virtual certainty. *Computers in Libraries*, 13 (2), 14-17.

In this essay the author lets the reader reconsider the term 'virtual library' with regard to the public library environment in the digital age. He concludes that the physical public library provides many essential services that will and cannot be replaced by the virtual or virtual reality library.

Levy, D.M., Marshall, C.C. (1995). Going digital: A look at assumptions underlying digital libraries. *Communications of the ACM* (38) 4, 77-84.

It is argued that the current definitions of the DL are too narrow and consequently will not satisfy the needs of future users. Therefore the authors demand integration of media (hybrid documents), version management tools, collaboration tools, development of new cataloging and collections maintenance methods.

Miksa, F., & Doty, P. (1994). Intellectual realities and the digital library. In *Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries 1994*, College Station, Texas (On-line). Available: <http://www.csd.tamu.edu/DL94/paper/miksa.html>

The authors examine the definition of the library and its validity for the concept of the digital library. The idea of the collection, of information sources and of place are discussed in order to question the concept of the DL as a library.

Wiederhold, G. (1995). Digital libraries, value, and productivity. *Communications of the ACM*, (38) 4, 85-96.

The author discusses new ways of publishing, traditional library services that will disappear and new ones that will be created. He identifies electronic copyright management, image search, dynamic books and ensuring of access as new services provided by future libraries.

Archival, Organization and Preservation:

Ackerman, M.S., & Fielding, R.T. (1995). Collection development in the digital library. *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries 1995*, Austin, Texas (On-line). Available: <http://csdl.tamu.edu/DL95/papers/ackerman/ackerman.html>

The authors distinguish between a narrowly-constructed library with a single collection and the broadly-constructed library, with distributed and multiple collections. They argue that in the first the traditional collection management methods can be used. In the latter, however, automated collection maintenance mechanisms are necessary, such as agents. Therefore, the authors describe some of those agents.

Anderson, G., Lasher, R., Reich, V. (1996). The Computer Science Technical Report (CS-TR) Project: A pioneering digital library project viewed from a library perspective. *The Public-Access Computer Systems Review*, 7 (2), (On-line). Available: <http://info.lib.uh.edu/pr/v7/n2/ande7n2.html>

This article addresses the challenges encountered in one of the first DL projects. A bibliographic record format for exchange was developed, in addition, a distributed delivery protocol, scanning procedures, copyright issues, collaboration efforts are introduced.

Cole, T. and Kazmer, M. (1995). SGML as a Component of the digital library. *Library High Tech*, 13(4), 17-90.

Description of SGML characteristics and its significance in the networked environment. Also examines issues that librarians need to consider when planning to use SGML. A second part discusses search, retrieval and display issues.

Corrado, E. (1996). *Annotated Bibliography on Cataloging Internet Resources*, (On-line). Available: <http://www.scils.rutgers.edu/~ecorrado/cataloging/index2.html>

Fleischhauer, Carl (1994). Organizing digital archival collections: American Memory's experiences with bibliographic records and other finding aids. *Proceedings of the Seminar on Cataloging Digital Documents 1994*, Charlottesville, Virginia, (On-line). Available:

<http://lcweb.loc.gov/catdir/semdigdocs/carl.html>

For annotation see National Digital Library Project

Mitchell, S. (1996). Library of Congress Subject Headings as subject terminology in a virtual library: The INFOMINE example. *Proceedings of Untangling the Web 1996*, Santa Barbara, California, (On-line). Available: <http://www.library.ucsb.edu/untangle/smitch.html>

This paper describes the Infomine Project's reasons for using LCSHs, its techniques, software and hardware environment. Infomine uses a copy-and-paste technique to transfer relevant subject headings into the Infomine database. Therefore the Windows environment proved to be especially suitable for this project.

Levy, D.M. (1995). Cataloging in the digital order. *Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries 1995*, Austin, Texas, (On-line). Available:

<http://csdl.tamu.edu/DL95/papers/levy/levy.html>

Computer scientist explains and summarizes cataloging, introduces current challenges and comes to the conclusion that 'cataloging as order-making' will still be necessary in the digital environment.

Olson, N.B. (1995). Cataloging Internet resources: A manual and practical guide, (On-line). Available:

<http://www.oclc.org/oclc/man/9256cat/toc.htm>

For annotation see the InterCat project.

Saffady, W. (1995). Digital library concepts and technologies for the management of library collections: an analysis of methods and costs. *Library Technology Reports*, 31 (3), 223-383.

The article begins with an examination of DL definitions and gives a very good and comprehensive discussion of its history. A history of digital library projects is also provided. The main part of the report is concerned with text-based and image-based implementations of DLs (hardware, software, document conversion and cost calculation).

The Digital Library in the Information Society; Politics of Librarianship and the Digital Library:

Braman, S. (1994). The autopoietic state: communication and democratic potential in the Net. *Journal of the American Society for Information Science*, 45 (6), 358-368.

Discusses chaos theory, second-order cybernetics, organizational sociology and theories of state to demonstrate the relationship between information and power. Develops theory of the autopoietic (self-organizing) state in which the citizen is facilitated with an increased democratic potential and how this potential could be increased with the correct policy of information.

Broering, N.C. (1995). Changing Focus: Tomorrow's virtual library. *Serials Librarian*, 25 (3/4), 73-94.

A large part of the article is about the IAIMS project of the NLM, but the author discusses in addition to it strategic planning and the role of librarians in the academic virtual library from a idealistic point of view.

Dervin, B. (1994). Information <-> democracy: an examination of underlying assumptions. *Journal of the American Society for Information Science*, 45 (6), 369-385.

The information <-> democracy narrative is examined within six 'stereotypes' (authority, naturalism, cultural, relativity, constructivism, postmodernism, communitarianism) views from both an ontological and an epistemological point of view.

England, M., & Shaffer, M. (1994). Librarians in the digital library. *Proceedings of First Annual Conference on the Theory and Practice of Digital Libraries 1994*, College Station, Texas, (On-line).

Available: <http://www.csdl.tamu.edu/DL94/position/england.html>

Brief position paper holding the viewpoint that the roles of librarians will shift from acquisition, preservation and storage towards teaching, consulting, researching and ensuring access. Design and maintenance are also suggested in cooperation with computer scientists.

Furuta, K. (1994). Librarianship in the digital library. *Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries 1994*, College Station, Texas, (On-line). Available:

<http://www.csdl.tamu.edu/DL94/position/kfuruta.html>

Short position paper, in which the author argues that the traditional roles of librarians, such as reference service and collection development, will continue in the digital library environment, but does not give any reasons why he thinks so.

Kahin, B. (1994). Institutional and policy issues in the development of the digital library, (On-line).

Available: <http://www.press.umich.edu/jep/works/kahin.dl.html>

The article focuses on such issues as resource sharing, cost licensing, control of information via copyrights and patent laws. Digital libraries are placed in the context of the research and higher education community.

Kochtanek, T.R. (1995). On the role of libraries in a virtual landscape. In Proceedings of the 16th National Online Meeting 1995, New York, New York, 223-231.

The first part of the article examines to what extent the terms 'virtual' and 'digital libraries' are reflected in the literature and discusses shortly its definition in comparison to the traditional library. In the second part the 'evolutionary steps' of libraries towards digital libraries are described. Finally, the author discusses fields of professional involvement for librarians in DLs.

Lievrouw, L.A. (1994). Information resources and democracy: Understanding the paradox. Journal of the American Society for Information Science, 45 (6), 350-357.

Discusses types of democracies, and argues that information environment fails to increase democracy, because it is not an 'involving' information environment. Potential of this 'involving' information environment is given by technology, but the current tendency is towards an 'informing' information environment.

Mullin, D.I. (1996). The First Amendment and the Web: The Internet porn panic and restricting indecency in cyberspace. Proceedings of Untangling the Web 1996, Santa Barbara, California, (On-line). Available: <http://www.library.ucsb.edu/untangle/mullin.html>

For annotation see copyright section.

Spink, A. (1995). Digital libraries and sustainable development? Proceedings of the Second Annual Conference on the Theory and Practice of Digital Libraries 1995, Austin, Texas, (On-line). Available: <http://csdl.tamu.edu/DL95/papers/spink/spink.html>

This paper addresses digital library research within the context of social change. It raises questions of how DL research is or should in future be related to the sustainable development debate.

Issues:

Meditation and Interaction in the Digital Library:

Ackerman, M.S. (1994). Providing social interaction in the digital library. Proceedings of the First Annual Conference on the Theory and Practice of Digital Libraries 1994, College Station, Texas, (On-line). Available: <http://www.csdl.tamu.edu/DL94/position/ackerman.html>

The author argues that social interaction is a valuable and important information source in the research environment. Therefore, methods need to be found to provide means of social interaction within digital libraries. The Cafe ConstructionKit, that provides such interaction, is introduced.

Brewer, A., Ding, W., Hahn, K., & Komlodi, A. (1996). The role of intermediary services in emerging digital libraries. Proceedings of the First International Conference on Digital Libraries 1996, Bethesda, Maryland, pp.29-35.

Discussion of intermediation services in the DL environment. The view is expressed that services such as search tools, classification, filtering, translation and publishing are essential and value-added.

Dillon, A. (1995). What is the shape of information? Human factors in the development and use of

digital libraries. Proceedings of the 37th Allerton Institute 1995, Monticello, Illinois, (On-line). Available: <http://edfu.lis.uiuc.edu/allerton/95/s4/dillon.html>

Interim report on the study of HCI and information seeking behavior in the WWW environment. The study focuses on users' conceptualizations of navigation in information spaces. Findings suggest that experience contributes to ease of navigation, and determines accuracy, but not speed.

Ehrlich, K., & Cash, D. (1994). Turning information into knowledge: Information finding as a collaborative activity. Proceedings of the First Annual Conference on the Theory and Practise of Digital Libraries 1994, College Station, Texas, (On-line). Available:

<http://www.csd1.tamu.edu/DL94/paper/lotus.html>

Article reports on a study conducted of a workgroup of customer supporters using Lotus Notes, a commercial 'digital library model'. The authors conclude that face-to-face interaction and gatekeepers will continue to play an important role in the digital library environment and that technology will be used to support rather than supplant human work.

Flynn, K.M. (1995). The knowledge manager as a digital librarian: An overview of the knowledge management pilot program at the MITRE Corporation. Proceedings of the Second Annual Conference on the Theory and Practise of Digital Libraries 1995, Austin, Texas, (On-line). Available:

<http://csdl.tamu.edu/DL95/flynn/flynn.html>

This paper reports on the experiences of the Mitre Corporation regarding new roles of librarians in the networked environment. These roles include collection development by selection of reliable WWW sites, resource organization, bibliographic instruction and electronic reference service

Kling, R., & Elliott, M. (1994). Digital library design for usability. Proceedings of the First Annual Conference on the Theory and Practise of Digital Libraries 1994, College Station, Texas, (On-line).

Available: <http://www.csd1.tamu.edu/DL94/paper/kling.html>

The authors discuss interface and organizational usability with the examples of Mosaic and Gopher. The second part of the article discusses five models of computer systems design as cultural models, including the 'organizationally sensitive model' promoted by the authors

Marchionini, G. (1995). User-centered methods for library interface design. Proceedings of the 37 Allerton Institute 1995, Monticello, Illinois, (On-line). Available:

<http://edfu.lis.uiuc.edu/allerton/95/s4/marchio.html>

Outline of the design process of a digital library interface using a workgroup approach (participatory design). The author stresses the importance of a design driven by user needs, not system constraints. In conclusion, a number of characteristics important for interface design are provided.

Marchionini, G., & Maurer, H. (1995). The roles of digital libraries in teaching and learning. Communications of the ACM, 38(4), 67-76.

Examines how digital libraries might change teaching and learning. It is argued that DLs are useful for informal, formal and professional learning, will blur the boundaries of learning resources, types of learning and distinctions between teaching and learning. In addition, new ways of teaching and learning suitable for the DL environment need to be developed, because this environment allows for new enhanced interactions.

Nilan, M. (1995). Ease of user navigation through digital information spaces. Proceedings of the 37 Allerton Institute 1995, Monticello, Illinois, (On-line). Available:

<http://edfu.lis.uiuc.edu/allerton/95/s4/nilan.html>

Short discussion of research issues on design and organization and classification of content in DLs to

facilitate easy navigation. The author proposes a high-order concept. He suggests among other things, that the representation of content for the user should ideally 'match' the users' concepts in order to avoid learning arbitrary and complicated mechanisms.

Rao, R., Pedersen, J.O., Hearst, M.A., Mackinlay, J.D., Card, S.K., & Masinter, L. (1995). Rich interaction in the digital library. *Communications of the ACM*, 38 (4), 29-40.

Represents the computer scientists' view of rich interaction in the DL: The introduction of interfaces will facilitate efficient and effective use of the DL by multiple search methods, multiple viewing methods and other mechanisms.

Authority, Authenticity, Originality and Intellectual Freedom in the Digital Library:

Copyright and Intellectual Property Resources of the IFLA

<http://www.nlc-bnc.ca/ifla/II/copyright.htm>

Arnold, K. (1995). The body in the virtual library: Rethinking scholarly communication. *Journal of Electronic Publishing*, 1 (1), (On-line). Available:

<http://www.press.umich.edu/jep/works/arnold.body.html>

The author discusses future organizational models of scholarly communication and intellectual property protection. As he assumes the breakdown of scholarly communication as we know it, the author discusses possible roles of university presses and libraries in an electronic scholarly community.

Barlow, J.P. (1994, March). The economy of ideas: A framework for rethinking patents and copyrights in the digital age. *Wired*, Issue 2 (3), 1994, (On-line). Available:

<http://www.nlc-bnc.ca/documents/infopol/copyright/jpbarlow.htm>

Wired article full of speculations about the future of intellectual property. The author discusses in metaphorical language his scenario of intellectual property in the digital age. (Copyright protects bottle not wine, so what to do in a bottleless age?)

The Commission on Preservation and Access and the Research Libraries Group. (1995). Task force on archiving of digital information proposed charge, (On-line). Available:

<http://www.oclc.org:5046/~weibel/archtf.html>

Graham, P.S. (1994). Intellectual preservation: electronic preservation of the third kind, (On-line).

Available: <http://aultnis.rutgers.edu/texts/epaintpres.html>

The author is concerned with intellectual preservation regarding preservation of digital documents and correspondence and possible (illegal) alterations. He introduces the method of digital time-stamping and discusses implementation models and future questions to be solved.

Graham, P.S. (1994). Intellectual preservation and electronic intellectual property. IP Workshop Proceedings 1994, (On-line). Available: <http://www.cni.org/docs/ima.ip-workshop/www/Graham.html>

In this paper preservation issues in the electronic environment are discussed. First, medium preservation, in particular, the storage obsolescence problem, is addressed. Preservation of storage technology, migration to the next generation of storage media or to hard copy are suggested solutions. Second, problems of intellectual preservation, accidental or intended change of contents, are addressed. Here encryption, hashing and digital time-stamping are suggested.

Mullin, D.I. (1996). The First Amendment and the Web: The Internet porn panic and restricting indecency in cyberspace. *Proceedings of Untangling the Web 1996*, Santa Barbara, California,

(On-line). Available: <http://www.library.ucsb.edu/untangle/mullin.html>

This paper examines the contribution of various studies to the perception of the Internet as a pool of indecent material, and how these developments led to the Communications Decency Act.

National Research Council. (1994). Rights and responsibilities of participants in networked communities, (On-line). Available: <http://www.nap.edu/nap/online/rights/>
Comprehensive discussion of ethical and legal questions and concerns raised by the electronic communication. The report is based on a workshop and public forum including participants from the government, academia, industry, technology and law. Discussed are free speech, electronic vandalism, intellectual property and privacy. Concludes that future social norms in the environment will be merged from existing laws and new social rules, such as the 'netiquette'.

Lyons, P.A., & Garrett, J.R. (1993). Toward an electronic copyright management information system. *Journal of the American Society for Information Science*, 44 (8), 468-73.
The authors concentrate on copyright management of literary works in the electronic environment and how this process could be automated. First, they introduce the readers to the context of digital libraries and copyright law. Then they discuss features and requirements of a successful ECMS and the rights that can be licensed through such a system.

Perritt, H. (1993). Knowbots, permission headers and contract law. Conference on Technological Strategies for Protecting Intellectual Properties in the Networked Multimedia Environment 1993, (On-line). Available: <http://www.nlc-bnc.ca/documents/infopol/copyright/perh2.txt>
Comprehensive paper addressing intellectual property protection in the digital library, here defined as a 'multiplicity of hosts'. Suggests ensuring intellectual property protection by implementation of 'permission headers' that indicate potential users (includes electronic signatures and electronic contracting).

Rosenbaum, H. (1996). In the trenches of the digital revolution: Intellectual freedom and the 'public' digital library. ASIS 1996 MidYear Conference, (On-line). Available: <http://silver.ucs.indiana.edu/~hrosenba/Papers/ASIS963.html>
General discussion of all policy issues for libraries regarding Internet access. In particular examines intellectual freedom, free speech, privacy, access, intellectual property. Also raises questions on how libraries should handle current and upcoming issues.

Samuelson, P. (1995). Copyright and digital libraries. *Communications of the ACM*, 38 (4), 15-21.
Begins with an introduction to the history of copyright, arguing that the purpose of copyright has 'historically been to promote public access to learning'. Argues that copyright might become unnecessary if other cost-effective means of assuring the availability of adequate supplies of information to the public can be found. Introduces a pay-per-use scenario of copyrighted material in comparison to related services and developments.

Any suggestions and comments welcome.

Last Modification July 11, 1996

Christine Woerner

DLRL Related Publications

Table of Contents

- [1](#) BOOKS
 - [2](#) JOURNAL ARTICLES
 - [3](#) BOOK CHAPTERS
 - [4](#) REFEREED CONFERENCE/WORKSHOP PAPERS
 - [5](#) REFEREED POSTERS
 - [6](#) UNREFEREED INVITED CONFERENCE/WORKSHOP PAPERS
 - [7](#) KEYNOTE / BANQUET / DISTINGUISHED SPEAKER PRESENTATIONS
 - [8](#) ORAL PRESENTATIONS
 - [9](#) TECHNICAL REPORTS, OTHER PUBLICATIONS
-

1. BOOKS

- E. Fox and G. Marchionini, eds., Proceedings of the First ACM International Conference on Digital Libraries, DL'96, Bethesda, MD, March 20-23, 1996.
 - E. Fox, ed. Sourcebook on Digital Libraries: Report for the National Science Foundation, TR-93-35, VPI&SU Computer Science Dept., Dec. 1993, Blacksburg, VA. Available by anonymous FTP from directory pub/DigitalLibrary on fox.cs.vt.edu, over 400 pages.
-

2. JOURNAL ARTICLES

- E. Fox and G. Marchionini. Toward a Worldwide Digital Library. Guest Editors' Introduction to special section (pp. 28-98) on Digital Libraries: Global Scope, Unlimited Access. Commun. of the ACM, Apr. 1998, 41(4): 28-32. <http://purl.lib.vt.edu/dlib/pubs/CACM199804>
- E. Fox. Networked Digital Library of Theses and Dissertations: An International Collaboration Promoting Scholarship. ICSTI Forum, Quarterly Newsletter of the International Council for Scientific and Technical Information, No. 26: 8-9, Nov. 1997. <http://www.icsti.nrc.ca/icsti/>
- E. Fox, R. Hall, and N. Kipp. NDLTD: Preparing the Next Generation of Scholars for the Information Age. The New Review of Information Networking (NRIN), 3: 59-76, 1997.
- Edward A. Fox, Robert Hall, Neill A. Kipp, John L. Eaton, Gail McMillan, and Paul Mather. NDLTD: Encouraging International Collaboration in the Academy. In Special Issue on Digital Libraries, DESIDOC Bulletin of Information Technology, 17(6): 45-56, Nov. 1997.
- E. Fox, J. Eaton, G. McMillan, N. Kipp, P. Mather, T. McGonigle, W. Schweiker, and B. DeVane. Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources. D-Lib Magazine, The Magazine of Digital Library Research, ISSN 1082-9873, Sep. 1997.
- E. Fox, J. Eaton, G. McMillan, N. Kipp, L. Weiss, E. Arce, S. Guyer. National Digital Library of

- Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources. D-Lib Magazine, The Magazine of Digital Library Research, ISSN 1082-9873, Sep. 1996.
- S. Chen and E. Fox. Guest Editors' Introduction to Special Issue on Digital Libraries, Journal of Visual Communication and Image Representation, 7(1), March 1996, Academic Press.
 - E. Fox and L. Kieffer. Multimedia Curricula, Courses and Knowledge Modules, ACM Computing Surveys, Dec. 1995, 27(4): 549-551.
 - F. Can, E. Fox, C. Snively, and R. France. Incremental Clustering for Very Large Document Databases: Initial MARIAN Experience. Information Systems, 84:101-114, 1995.
 - W. C. Dougherty and E. A. Fox. TULIP at Virginia Tech. Library Hi Tech, 13(4): 54-60, 1995.
 - E. Fox. Hypermedia Support for a Digital Library in CS. SIGLINK Newsletter, Sept. 1995, 4(2), Special Issue on Digital Libraries.
 - N. J. Belkin, P. Kantor, E. A. Fox and J. A. Shaw. Combining the Evidence of Multiple Query Representations for Information Retrieval. Information Processing & Management, 31(3), 431-448, May-June 1995.
 - E. Fox, R. Akscyn, R. Furuta, and J. Leggett. Guest Editors' Introduction to Digital Libraries. Commun. of the ACM, Apr. 1995, 38(4):22-28.
 - E. Fox. World-Wide Web and Computer Science Reports. Commun. of the ACM, Apr. 1995, 38(4):43-44.
 - L. Heath, D. Hix, L. Nowell, W. Wake, G. Averboch, and E. Fox. Envision: A User-Centered Database from the Computer Science Literature. Commun. of the ACM, Apr. 1995, 38(4):52-53.
 - J. French, E. Fox, K. Maly, and A. Selman. Wide Area Technical Report Service --- technical reports online. Commun. of the ACM, Apr. 1995, 38(4):45.
 - E. Fox. Digital Libraries ("hot topics" section), IEEE Computer, Nov. 1993, 26(11): 79-81.
 - E. Fox and L. Lunin. Introduction and Overview to Perspectives on Digital Libraries. Journal of the American Society for Information Science (JASIS), Sept. 1993, 44(8): 441-443. (Guest editor's introduction to special issue)
 - E. Fox, D. Hix, L. Nowell, D. Brueni, W. Wake, L. Heath, and D. Rao. Users, User Interfaces, and Objects: Envision, a Digital Library. Journal of the American Society for Information Science (JASIS), Sept. 1993, 44(8): 480-491.
 - E. Fox. How to Proceed Toward Electronic Archives and Publishing. Psychological Science, Nov. 1990, 1(6): 355-8.
 - E. Fox. ACM Press Database and Electronic Products -- New Services for the Information Age. Commun. of the ACM, Aug. 1988, 31(8): 948-951.
-

3. BOOK CHAPTERS

- E. Fox. How to make intelligent digital libraries. In Methodologies for Intelligent Systems, Proceedings of the 8th International Symposium, ISMIS'94, Charlotte, NC, Oct. 1994. Lecture Notes in Artificial Intelligence 869, Springer-Verlag, Berlin, 27-38.
-

4. REFEREED CONFERENCE/WORKSHOP PAPERS

- G. Abdulla, E. A. Fox, M. Abrams, "Shared User Behavior on the World Wide Web", in Proc. WebNet97 Conference, Toronto, Canada November 1997, <http://www.AACE.org/conf/webnet>
- M. Kirschenbaum, E. Fox. 1997. Electronic Theses and Dissertations in the Humanities. In Proc.

- Joint Annual Conf. of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, ACH-ALLC'97, June 3-7, 1997, Queen's Univ., Kingston, Ontario.
- L. Tinoco, E. Fox, N. D. Barnette. Online Evaluation in WWW-based Courseware. In SIGCSE '97, Proceedings of the 28th SIGCSE Technical Symp. on Computer Science Education, San Jose, 2/27 - 3/1/97, 194-198. <http://ei.cs.vt.edu/papers/sigcse97.pdf>
 - L. Tinoco, E. Fox, R. Ehrich, H. Fuks. 1996. QUIZIT: An interactive online quiz system for WWW-based instruction. In VII Proceedings of the Symposium of Educational Technology: Belo Horizonte, MG, Brazil, Nov 1996. <http://ei.cs.vt.edu/papers/QUIZIT9611.pdf>
 - S. Williams, M. Abrams, C. Standridge, G. Abdulla, E. Fox. Removal Policies in Network Caches for World-Wide Web Documents, in Proc. ACM SIGCOMM '96, Stanford U., Palo Alto, CA, Aug. 28-30, 1996, 293-305. See full paper and errata page at, respectively:
<http://ei.cs.vt.edu/~succeed/96sigcomm/96sigcomm.html>
<http://www.cs.vt.edu/~chitra/docs/96sigcomm/Errata.ps>
 - L. Nowell, R. France, D. Hix, L. Heath, and E. Fox. Visualizing Search Results: Some Alternatives to Query-document Similarity, in Proc. SIGIR'96, Zurich, Switzerland, Aug. 18-22, 1996, 67-75.
 - M. Abrams, C. R. Standridge, G. Abdulla, S. Williams and E. A. Fox. Caching Proxies: Limitations and Potentials, in Proc. 4th International World-Wide Web Conference, Boston, Dec. 1995. URL: <http://ei.cs.vt.edu/~succeed/WWW4/WWW4.html>
 - W. Wake and E. Fox. SortTables: A Browser for a Digital Library. In Proc. 4th Int. Conf. on Information and Knowledge Management, CIKM '95, Baltimore, MD, Nov. 28 - Dec. 2, 1995, 175-181.
 - M. Abrams, S. Williams, G. Abdulla, S. Patel, R. Ribler and E. Fox. Multimedia Traffic Analysis Using Chitra95. In Proc. 3rd Int. Multimedia Conf. and Exhibition, Multimedia '95, San Francisco, Nov. 5-9, 1995, 267-276.
 - E. Fox and D. Barnette. Improving Education through a Computer Science Digital Library with Three Types of WWW Servers. In Proc. Second International WWW '94: Mosaic and the Web, WWW'94, Chicago, IL, Oct. 17-20, 1994.
 - K. Maly, J. French, A. Selman and E. Fox. The Wide Area Technical Report Service. In Proc. Second International WWW '94: Mosaic and the Web, WWW'94, Chicago, IL, Oct. 17-20, 1994, 523-533.
 - H. Gladney, E. Fox, Z. Ahmed, R. Ashany, N. Belkin, and M. Zemankova. Digital Library: Gross Structure and Requirements: Report from a March 1994 Workshop. Digital Libraries '94, June 19-21, 1994, College Station, TX, ed. J. Schnase, J. Leggett, R. Furuta, T. Metcalfe, 101-107.
 - E. Fox and G. Abdulla. Digital Video Delivery for a Digital Library in Computer Science. High-Speed Networking and Multimedia Computing Workshop, IS&T/SPIE Symposium on Electronic Imaging Science and Technology, Feb. 6-10, 1994, San Jose, CA, 7 pages.
 - E. Fox, R. France, E. Sahle, A. Daoud, and B. Cline. Development of a Modern OPAC: From REVTOLC to MARIAN. Proc. 16th Annual Intern'l ACM SIGIR Conf. on R & D in Information Retrieval, SIGIR '93, Pittsburgh, PA, June 27 - July 1, 1993, 248-259.
 - D. Brueni, B. Cross, E. Fox, L. Heath, D. Hix, L. Nowell, and W. Wake. What if there were desktop access to the Computer Science Literature?, Proc. 21st Annual Computer Science Conference, ACM CSC '93, Feb. 16-18, 1993, Indianapolis, IN, 15-22.

5. REFEREED POSTERS

- E. Fox, G. Abdulla, W. Heagy. Quantitative Analysis and Visualization Regarding Interactive Learning with a Digital Library in Computer Science. ACM Digital Libraries'97, Philadelphia, PA, July 23-26, 1997
-

6. UNREFEREED INVITED CONFERENCE/WORKSHOP PAPERS

- E. Fox. "Future of Electronic Publishing." Invited seminar in Virginia Tech Symposium on Scholarship in the Electronic World, text and audio online, paper to appear in published proceedings, Research and Graduate Studies, Blacksburg, VA, April 13, 1998
- E. Fox. "Update on the Networked Digital Library of Theses and Dissertations (NDLTD)." In Proc. 35th Annual Clinic on Library Applications of Data Processing, GSLIS 98, March 22-24, 1998, ISSN 0069-4789, <http://edfu.lis.uiuc.edu/dpc98/>, to appear
- E. Fox. "Effects on Education, and a Proposal for Collection of Tools" IR Tools Workshop, <http://www.sis.pitt.edu/~erasmus/workshop.html>, U. Pittsburgh, March 19-21, 1998, <http://fox.cs.vt.edu/talks/IRtools.htm>
- E. Fox. "A Scalable Digital Ecology for a Networked Digital Library of Theses and Dissertations (NDLTD)." Winter Workshop of the Human-Computer Interaction Consortium (HCIC), Fraser, Colorado, March 4-8, 1998
- E. Fox. Networked Digital Library of Theses and Dissertations (NDLTD): A Worldwide Initiative to Advance Scholarship, invited presentation for session on Progress on Digital Dissertation Initiatives, Chair: Joan K. Lippincott, for Coalition for Networked Information Fall Meeting, Minneapolis, MN, Oct. 25-27, 1997 <http://www.ndltd.org/talks/CNIF97.pdf>
- E. Fox. Digital Libraries and Virtual Universities, invited presentation for "Information research for designing and planning virtual universities" Seminar at Centro Universitario de Investigaciones Bibliotecolgicas Universidad Nacional Autnoma de Mexico Cd. Universitaria, Mexico, D.F. (Library and Information Research Center, National University of Mexico), Aug. 11-15, 1997, <http://fox.cs.vt.edu/talks/Mexico97.html>
- J. Eaton, E. Fox, G. McMillan. Electronic Theses and Dissertations (ETDs) and Their Contribution to Graduate Education. Proc. 53rd Annual Meeting Midwestern Association of Graduate Schools, MAGS, 1997, 73-78
- E. Fox. Interactive Learning with a Digital Library in Computer Science. Invited paper for Proc. Frontiers in Education - FIE'96, Salt Lake City, Utah, Nov. 6-9, 1996.
- E. Fox. Digital Libraries, WWW, and Educational Technology: Lessons Learned. Invited paper for Proc. ED-MEDIA 96, World Conference on Educational Multimedia and Hypermedia, Boston, MA, June 17-22, 1996, 246-251. See also <http://ei.cs.vt.edu/~fox/EDMEDIA96/>
- E. Fox. Digital Library Support for Education - A Case Study of Advanced Networked Information Systems, Invited presentation for Dagstuhl Workshop on Networked Information Systems - Discovery, Retrieval, Dissemination, Feb. 26 - March 1, 1996, Dagstuhl, Germany.
- J.M. Carroll, M.B. Rosson, E.A. Fox, A.M. Cohill, K.W. Schmidt & N.A. Kipp. Community Networks as Working Memory, Invited presentation for HCIC (HCIC Consortium), Feb. 1996.
- E. Fox, N. Barnette, C. Shaffer, L. Heath, W. Wake, L. Nowell, J. Lee, D. Hix, and H. Hartson. Progress in Interactive Learning with a Digital Library in Computer Science. Invited paper for Proc. ED-MEDIA 95, World Conference on Educational Multimedia and Hypermedia, Graz, Austria, June 17-21, 1995, pp. 7-12.
- E. Fox. Seamless Multimedia Integration for Digital Libraries. Invited position paper for Dagstuhl Seminar on Fundamentals and Perspectives of Multimedia Systems, International Conf. and Research Center for Computer Science, Dagstuhl Castle, Germany, July 4-8, 1994, 118-123.

- E. Fox. A digital library connecting Envision, KMS, and Mosaic with interfaces, communications, and data interchange. Invited presentation for 1994 Workshop on Digital Libraries: Current Issues, sponsored by: Rutgers and Purdue Universities, AT&T and Bellcore, at Rutgers Univ., Newark, NJ, May 19-20, 1994. Abstract in SGOIS Bulletin, Aug. 1994, 15(1):6.
 - E. Fox. A User-Centered Hypermedia Database from the Computer Science Literature. Invited presentation for Advances in Data Management for the Scientist and Engineer Session 2, AAAS '93, Feb. 11-16, 1993, Boston. Abstract in Proc. 159th National Meeting of the AAAS, AAAS'93: Science and Engineering for the Future, p. 145. Longer paper: E. Fox, L. Heath, and D. Hix, A User-Centered Database from the Computer Science Literature, in Proceedings of the NSF Scientific Database Projects, 1991-1993, eds. W. Chu, A. Cardenas, and R. Taira, Feb. 14-16, 1993, Boston, 70-75.
 - E. Fox. An Electronic Publishing / Information Storage and Retrieval Perspective on the Management of Scientific Databases. Invited presentation for National Science Foundation sponsored Workshop on Scientific Databases, March 12-13, 1990, Charlottesville, VA, 2 pages. MD, 10 pages.
-

7. KEYNOTE / BANQUET / DISTINGUISHED SPEAKER PRESENTATIONS

- Digital Libraries: Their Educational Applications and Uses. Opening presentation, keynote session for Web Week and first talk in the Rice University Fondren Library and Information Technology 1997 Lecture Series "Rethinking Information Access in the Digital Age", March 17, 1997.
 - Digital Libraries through Information Retrieval for Education. Presentation for Distinguished Lecture Series, Univ. Utah Dept. of Computer Science, March 7, 1996.
 - Rethinking libraries in the information age: lessons learned with 5 digital library projects, Henderson Lecture, UNC Chapel Hill, Feb. 1, 1996.
 - Images of Digital Libraries. Invited opening keynote address. INFO Conference: Digital transfer of images, Helsinki, Finland, Nov. 10-11, 1994, 2 pg. extended abstract for conference plus longer paper for proceedings.
 - Multimedia in Education and Digital Libraries. Sole keynote speaker for Multimedia Systems: Technology and Applications, Oct. 12-13, 1994, Ottawa.
 - How to make intelligent digital libraries. Invited plenary presentation for 8th Int'l Symp. on Methodologies for Intelligent Systems (ISMIS'94), Charlotte, NC, Oct. 16-19, 1994.
 - Toward a Widely Used Hypermedia Digital Library in Computer Science. Invited plenary presentation for EG-MM '94, First Eurographics Symposium and Workshop on Multimedia: Multimedia/Hypermedia in Open Distributed Environments, June 6-9, 1994, Graz, Austria.
 - Digital Libraries: Why People Use Tools, Not AI. Invited plenary presentation for Tenth IEEE Conf. on Artificial Intelligence for Applications (CAIA), San Antonio, March 1-4, 1994.
 - From Information Retrieval to Networked Multimedia Information Access, keynote address. In G. Knorz, J. Krause, C. Womser-Hacker, eds., Information Retrieval '93, Proc. der 1. Tagung Information Retrieval '93, 13-15 September, 1993, University of Regensburg, Germany, Univ. of Konstanz Press, 116-124.
 - Multimedia Systems and Electronic Libraries, IBM/FAU Distinguished Lecture Series, April 8, 1993, Florida Atlantic Univ., Boca Raton, FL.
 - Building a User-Centered Database from the ACM Literature. Invited plenary presentation for Symposium on Document Analysis and Information Retrieval, March 16-18, 1992, Tropicana Hotel, Las Vegas, Nevada, 235-246.
-

8. ORAL PRESENTATIONS

- Improving Education through the Networked Digital Library of Theses and Dissertations (NDLTD) and the Computer Science Teaching Center (CSTC), invited presentation for seminar the Gore-Chernomyrdin Telecommunications Working Group process, Moscow, April 16-17, 1998, to be given
- Digital Libraries to Enhance Learning: Case Studies in Computer Science and Graduate Education. Invited presentation at Pacific Northwest National Laboratory, Pasco, WA, April 3, 1998, to be given.
- Helping Learners through Digital Libraries: The Networked Digital Library of Theses and Dissertations (NDLTD) and the Computer Science Teaching Center (CSTC), invited seminar for Dept. of Computer Science, Univ. of Mass., Amherst, March 12, 1998. Seminar series page is <http://www.cs.umass.edu/csinfo/colloquia/DEPT/Seminars1997-98.html>
- The Worldwide Electronic Thesis and Dissertation Initiative: Joining the Networked Digital Library of Theses and Dissertations seminar hosted by University Denver for all neighboring universities, Denver, CO, March 9, 1998
- Progress on the National Digital Library of Theses and Dissertations (NDLTD), invited presentation for Computers in Libraries '98, Hyatt Regency, Crystal City, Arlington, VA, March 2, 1998
- Digital Libraries: Preparing the Next Generation of Scholars in session "Driving Influences: An Update on Issues, Trends and Current Developments that will Affect Secondary Publishers", NFAIS'98, Four Seasons, Philadelphia, PA, Feb. 24, 1998 <http://www.pa.utulsa.edu/nfais.html>
- The Networked Digital Library of Theses and Dissertations: Improving Graduate Education, invited seminar for Drexel U. College of Information Science and Technology, Feb. 23, 1998
- The Worldwide Electronic Thesis and Dissertation Initiative: Building the Networked Digital Library of Theses and Dissertations, presentation with John Eaton for National Agricultural Library, MD, Feb. 20, 1998
- The Worldwide Electronic Thesis and Dissertation Initiative: Joining the Networked Digital Library of Theses and Dissertations presentation with John Eaton, James Madison U., Harrisonburg, VA, Feb. 19, 1998
- Digital Libraries, presentation arranged by Assistant Professor of IS, Youngjin Yoo, Weatherhead School of Management, Case Western Reserve U., through videoconference for MIDS411: Advances in Information Technology, Feb. 5, 1998, 6-8pm
- The Worldwide Electronic Thesis and Dissertation Initiative: Joining the Networked Digital Library of Theses and Dissertations, invited videoconference presentation, Florida International U., 1pm Jan. 27, 1998, hosted by Jackie Zelman, Director, UCS
- The Worldwide Electronic Thesis and Dissertation Initiative: Joining the Networked Digital Library of Theses and Dissertations, invited presentation, VCU, Richmond, VA, Jan. 12, 1998
- The Networked Digital Library of Theses and Dissertations, invited presentation for Seminar on Information Access, School of Information Management & Systems, University of California, Berkeley, CA, Jan. 6, 1998
- Distributed Learner Spaces with Digital Libraries: Future digital library technologies across high-speed distributed systems. Presentation for signing of Memorandum of Understanding between VPI&SU and the Institute of Systems Science of Singapore, as part of the ceremony launching the SINGAREN-vBNS connection, Washington, D.C., Nov. 7, 1997. <http://www.ndltd.org/talks/singapore>
- Implications of the Electronic Thesis and Dissertation (ETD) Initiative, invited presentation for Electronic Publishing session of DTIC's Annual Users Conference, November 5, 1997, Arlington,

VA

- Improved Education with a Digital Library: the NDLTD Case Study, invited session and presentation (with Gail McMillan) for FIPSE PI meeting, Nov. 2, 1997, Arlington, VA
- The Worldwide Electronic Thesis and Dissertation Initiative: Joining the Networked Digital Library of Theses and Dissertations, invited presentation at Fall CNI Meeting, Minneapolis, MN 10/26-27/97 <http://www.ndltd.org/talks/CNIF97.pdf>
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at U. Michigan, Oct. 8, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at Tech. Univ. of Lisbon, Oct. 2, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at U. Illinois Urbana-Champaign, Sept. 26, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at U. Virginia, Aug. 29, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at Florida Inst. of Tech, Aug. 21, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at University N. Florida, Aug. 18, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at U. Pennsylvania, July 29, 1997
- Panel discussion "NCSTRL: Experience with a Global Digital Library" on Networked CS Technical Report Library, D-Lib Panel on Interoperability I, ACM DL'97, July 24, 1997, Philadelphia, PA
- Renaissance Consortium Report on the Worldwide ETD Initiative / Networked Digital Library of Theses and Dissertations (NDLTD), IBM Almaden Lab, San Jose, CA, July 11, 1997.
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), informal presentation at Stanford U., July 11, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at Naval Postgraduate School, July 9, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at U. Ca. Berkeley, July 8, 1997
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), informal presentation at U. Ca. Santa Barbara, July 7, 1997
- Information Retrieval, Digital Libraries, Education Innovation, Theses and Dissertations, and WWW Traffic Analysis/Modeling: Related Work at Virginia Tech. Presentation July 3, 1997, ISS, Singapore.
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at National U. Singapore, July 1-3, 1997
- Powerful Interactivity in a Networked World. Presentation to the Computer Society of Singapore, July 1, 1997, Singapore.
- The Digital Era: Implications for Librarians, Information Providers and Users. Presentation for National Computer Board Digital Libraries Cluster. June 30, 1997, ISS Auditorium, Singapore.
- The Worldwide Electronic Thesis and Dissertation Initiative: Joining the Networked Digital Library of Theses and Dissertations seminar in conjunction with NSF DLI briefing, Carnegie-Mellon U., June 3, 1997
- Networked Digital Library of Theses and Dissertations, or, the ETD Initiative, presentation at Univ. Waterloo, June 2, 1997, <http://www.lib.uwaterloo.ca/~uw-etpt/Flyer.html>
- E. Fox, J. Eaton, G. McMillan. National Digital Library of Theses and Dissertations, invited session for CAUSE/CNI regional conference, Univ. of Del., May 22, 1997.

- Networked Digital Library of Theses and Dissertations, presentation at NYU, NYC, May 20, 1997.
- Networked Digital Library of Theses and Dissertations, presentation for the AAP PSP Executive Council, AAP, NY, May 20, 1997.
- Networked Digital Library of Theses and Dissertations. Invited videoconference presentation at Univ. S. Florida, April 11, 1997 (w. J. Eaton, G. McMillan, N. Kipp)
- Publishers and Electronic Theses, project briefing at Coalition for Networked Information Spring Meeting, Crystal City, VA, April 1-2, 1997.
- The Worldwide ETD Initiative: Joining the Networked Digital Library of Theses and Dissertations (NDLTD), invited presentation at Rutgers, The State University, New Brunswick, NJ, March 6, 1997
- Networked Digital Library of Theses and Dissertations. Invited presentation at Clemson U., March 5, 1997 (w. J. Eaton)
- Networked Digital Library of Theses and Dissertations. Invited presentation at U. Georgia, Athens, March 5, 1997 (w. J. Eaton)
- Networked Digital Library of Theses and Dissertations. Invited presentation at U. Alabama, Tuscaloosa, March 4, 1997 (w. J. Eaton)
- Networked Digital Library of Theses and Dissertations. Invited presentation at U. Alabama, Birmingham, March 4, 1997 (w. J. Eaton)
- Networked Digital Library of Theses and Dissertations. Invited presentation at U. Tennessee, Knoxville, March 3, 1997 (w. J. Eaton)
- Networked Digital Library of Theses and Dissertations. Informal presentation at San Jose State U., Feb. 28, 1997
- Multimedia, Hypertext and Information Access. Panel presentation in session on "Defining Multimedia Courses within a Computer Science Education", ACM SIGCSE'97, San Jose, 2/27 - 3/1/97.
- Digital Libraries: Theory, Educational Applications, and Use. Invited colloquium series talk for Dept. of Computer & Information Science, Ohio State Univ., Feb. 11, 1997.
- Networked Digital Library of Theses and Dissertations. Invited presentation at Ohio State U., Feb. 10, 1997
- Networked Digital Library of Theses and Dissertations. Invited presentation at Vanderbilt U., Jan. 17, 1997 (w. J. Eaton, G. McMillan)
- National Digital Library of Theses and Dissertations. Invited presentation, AT&T Laboratories, Murray Hill, NJ, Dec. 27, 1996.
- National Digital Library of Theses and Dissertations. Invited presentation and chairing of working group session on this topic, NSF/DARPA/NASA DLI (Digital Library Initiative) meeting, Dec. 16-17, 1996
- Electronic Theses and Dissertations, project briefing at Coalition for Networked Information Fall Meeting, San Francisco, Dec. 6-7, 1996
- The Future Role of Publishers in the New Educational Dynamic. Chaired plenary luncheon for Proc. Frontiers in Education - FIE'96, Salt Lake City, Utah, Nov. 7, 1996.
- Interactive Learning with a Digital Library in Computer Science. Invited long presentation for NSF CISE EI PIs workshop at FIE'96, Salt Lake City, Utah, Nov. 5-8, 1996.
- IR Curriculum: Information Engineering to Digital Libraries. Invited presentation for Drexel University hosted Workshop/Symposium sponsored by the W.K. Kellogg Foundation "Information Retrieval 2000 --- Workplace Needs and Curricular Implications", Marriott Hotel, Philadelphia PA, May 24, 1996.
- Virginia Tech's Digital Library Project. Invited short presentation for IBM Digital Library Consortium Meeting, Case Western Reserve University, Cleveland, May 2, 1996.

- Digital Libraries and CS Education. Invited Colloquium, George Mason U., Nov. 13, 1995.
- Education and Research with a Digital Library in Computer Science. Invited presentation for IBM Almaden Research Center, Nov. 3, 1995.
- Education and Research with a Digital Library in Computer Science. Invited presentation for IBM Watson Research Lab, Oct. 27, 1995.
- Electronic librarians, intelligent network agents, and information catalogues. Invited presentation for Reconnecting Science and Humanities in Digital Libraries, sponsored by the Univ. of Kentucky and the British Library, 19-21 October 1995, Lexington, KY.
- Computer Science Technical Reports Library. Invited presentation for Unlocking University Information, sponsored by SURA/SOLINET, Atlanta, Sept. 6-7, 1995.
- IR is at the Heart of Digital Libraries and the Global Information Infrastructure. Invited presentation for A SMART Celebration, Cornell Univ., Ithaca, NY, April 22, 1995.
- Improving CS Education with Digital Libraries. Invited Colloquium for Dept. of Computer Science, SUNY Buffalo, Dec. 2, 1994.
- Requirements for Knowledge Workers and Education. Invited presentation for Digital Library Academy Workshop, sponsored by IBM Academy of Technology, Edith Macy Conference Center, Briarcliff Manor, NY, Sept. 12-13, 1994.
- Digital Libraries. Norfolk State University Seminar, Norfolk, VA, March 28, 1994.
- Electronic Publishing Experiments: Report on Digital Libraries and their Relationship to SIG Activities. Presentation for ACM SIG Chairs Meeting, Phoenix, AZ, March 6, 1994.
- Background to Current Work on Digital Libraries. Invited presentation for Workshop on Intelligent Access to On-line Digital Libraries, in connection with IEEE CAIA '94, San Antonio, TX, March 1, 1994
- Electronic Theses and Dissertations. Invited presentation for Workshop: Innovative Uses of High-Tech in Graduate School Operations, CSGS'94, Conf. of Southern Graduate Schools, 23rd Annual Meeting, Clearwater Beach, FL, Feb. 18-21, 1994.
- Digital Library Related Research. Seminar for Case Western Reserve University, Cleveland, OH, Dec. 21, 1993.
- Interactive Learning with a Digital Library in Computer Science. Invited presentation, as part of this year's program on Applications of Computers in Instruction, for Virginia Tech Chapter of Sigma Xi, Nov. 30, 1993, Blacksburg, VA.
- Worldwide Digital Libraries. Invited presentation for Virginia Tech Center for the Study of Science and Society, Nov. 18, 1993, Blacksburg, VA.
- Research In Digital Libraries and Networked Multimedia Information Access. Invited presentation for panel on Academic Research Trends in Information Storage and Retrieval, sponsored by SIG/SRT and SIG/MGT, ASIS'93, October 25-29, 1993, Columbus, OH.
- Improving the Undergraduate Experience through Research in Interactive Accessibility and Digital Libraries. Invited presentation for ACM Student Chapter, Virginia Tech, Blacksburg, VA, Sept. 29, 1993.
- Envision-ing a Computer Science Digital Library. Invited presentation and chair for panel on Digital Libraries of the Future, ACM Multimedia 93, Aug. 4-6, 1993, Anaheim, CA.
- Technical Reports / Dissertations. Invited project presentation for Monticello Electronic Library meeting, sponsored by SURA, SURAnet, SOLINET, and NSF, July 29-30, 1993, Atlanta.
- Electronic Dissertation Project. Presentation for Invited SIGIR Panel on Information Retrieval, and panel chair, ACH-ALLC93, Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing, June 16-19, 1993, Georgetown Univ., Washington, D.C..
- An Information Retrieval and Digital Library Perspective. Invited Presentation for Panel on Multimedia Databases, Joint Conf.: 1993 ACM SIGMOD International Conference on

Management of Data and Twelfth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS), Washington, D.C., May 25-28, 1993.

- Envision, the Future. Invited presentation for Digital Libraries of the Future session, Online Publishing '93, Pittsburgh, PA, March 22-24, 1993.
 - Electronic Libraries. Invited presentation for session on Impacts of Application Environments on Hyperbase Systems, NSF Hyperbases Workshop, Oct. 14-16, 1992, Washington.
 - Project Envision & Information Retrieval Invited presentation for NSF Workshop on Electronic Libraries, July 20-21, 1992, Washington, D.C.
 - Constructing and Distributing Information Retrieval Databases. Invited presentation for "Databases for Research and Testing in Document Analysis and Information Retrieval" panel at Symposium on Document Analysis and Information Retrieval, March 16-18, 1992, Tropicana Hotel, Las Vegas, Nevada.
 - Developing Interactive Digital Multimedia Applications and Archives. Invited seminar for Graduate Center, CUNY, NY, Feb. 6, 1992. Approaches to Improving Information Access. Invited seminar for AT&T Bell Laboratories Murray Hill, NJ, Feb. 6, 1992.
 - Developing Interactive Digital Multimedia Applications and Archives. Invited presentation for Bell Communications Research, Morristown, NJ, Dec. 23, 1991.
 - Building an Archive of Computer Science Literature. Invited seminar for Department of Computer Science, The Ohio State University, Nov. 5, 1991.
-

9. TECHNICAL REPORTS, OTHER PUBLICATIONS

- O. Balci, C. Ulusarac, P. Shah and E. Fox. A Library of Reusable Model Components for Visual Simulation of the NCSTRL System. TR-98-02, Virginia Tech Dept. of Computer Science Technical Report, Jan. 1998. URN: [ncstrl.vatech_cs/TR-98-02](http://ncstrl.vatech.cs/TR-98-02)
- G. Abdulla, A.H. Nayfeh and E. Fox. Modeling Correlated Proxy Web Traffic Using Fourier Analysis. TR-97-19, Virginia Tech Dept. of Computer Science Technical Report, Nov. 1997. URN: ncstrl.vatech_cs/TR-97-19
- G. Abdulla, B. Liu, R. Saad and E. Fox. Characterizing World Wide Web Queries. TR-97-04, Virginia Tech Dept. of Computer Science Technical Report, March 1997. URN: ncstrl.vatech_cs/TR-97-04
- G. Abdulla, E. Fox, M. Abrams and S. Williams. WWW Proxy Traffic Characterization with Application to Caching. TR-97-03, Virginia Tech Dept. of Computer Science Technical Report, March 1997. URN: ncstrl.vatech_cs/TR-97-03
- G. Abdulla, M. Abrams and E. Fox. Scaling the World-Wide Web. TR-96-06, Virginia Tech Dept. of Computer Science Technical Report, March, 1996.
- E. Fox. Virginia Tech Department of Computer Science Information Access Laboratory. ACM SIGIR Forum, Lab Report Special Section, 30(1), Spring 1996.
- H. Gladney, Z. Ahmed, R. Ashany, N. Belkin, E. Fox and M. Zemankova. Digital Library: Gross Structure and Requirements (Report from a Workshop). IBM Research Report RJ9840, IBM Almaden Research Center, May, 1994. Virginia Tech Dept. of Computer Science Technical Report 94-25, June, 1994.
- K. Maly, J. French, A. Selman and E. Fox. Wide Area Technical Report Service, TR_94_13, Old Dominion Univ. Dept. of Computer Science, June 1994.
- L. Nowell, E. Fox, L. Heath, D. Hix, W. Wake and E. Labow. Seeing Things Your Way: Information Visualization for a User-Centered Database of Computer Science Literature, TR-94-06, VPI&SU Computer Science Dept., Jan. 1994, Blacksburg, VA.

- K. Dalal and E. Fox. Document Translation: Dissertations and Technical Reports, TR-93-31, VPI&SU Computer Science Dept., Sep. 21, 1993, Blacksburg, VA.
- E. Fox, R. France, E. Sahle, A. Daoud, and B. Cline. Development of a Modern OPAC: From REVTOC to MARIAN, TR-93-06, VPI&SU Computer Science Dept., Feb. 17, 1993, Blacksburg, VA.
- S. Betrabet, E. Fox, and Q. Chen. A Query Language for Information Graphs, TR-93-03, VPI&SU Computer Science Dept., Jan. 1993, Blacksburg, VA.
- K. Maly, E. Fox, J. French, and A. Selman. Wide Area Technical Report Service, TR_92_44, Old Dominion Univ. Dept. of Computer Science, Dec. 1992
- D. Brueni, E. Fox, L. Heath, D. Hix, L. Nowell, and W. Wake. What if there were Desktop Access to the Computer Science Literature?, TR-92-42, VPI&SU Computer Science Dept., Aug. 12, 1992, Blacksburg, VA.

*nkipp**Revised: Wed May 20 17:39:23 EDT 1998*[*foxpubs.sl*](#)

Books:

There is only one really good book on digital libraries:

- Michael Lesk, [Practical Digital Libraries](#), Morgan Kaufmann, 1997, San Francisco

For a history of many digital library activities through Fall 1993, including reports on key workshops, see:

- Digital Library Source Book, Edward Fox, ed., 1993 <http://fox.cs.vt.edu/DLSB.html>

In the related field of Information Retrieval the best set of readings is:

- Karen Sparck Jones and Peter Willett, [Readings in Information Retrieval](#), Morgan Kaufmann, 1997, San Francisco

Some miscellaneous related works include:

- Elsevier, [TULIP Final Report](#), 1996, New York. This booklet was distributed after completion of the TULIP digital library prototype [project](#) by [Elsevier](#), and led to their current digital library effort, [EES](#).
- Hermann Maurer, ed., Hyper-G/Hyperwave: The Next Generation Web Solution, Addison Wesley Longman, 1996, Harlow, England
- Setrag Khoshafian, A. Brad Baker, MultiMedia and Imaging Databases, Morgan Kaufmann, 1996, San Francisco
- V.S. Subrahmanian, Sushil Jajodia, eds., Multimedia Database Systems: Issues and Directions, Springer, 1996, Berlin

[\[Main\]](#) [\[Contents\]](#) [\[References\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Topics:

- [Search, retrieval, resource discovery](#) (See Chapter 2 in Dr. Lesk's book.)
 - [Multimedia, representations](#) (See Chapter 4 in Dr. Lesk's book.)
 - [Architectures](#) (See Chapter 6 in Dr. Lesk's book.)
 - [Interfaces](#) (See Chapter 7 in Dr. Lesk's book.)
 - [Metadata](#)
 - [Electronic publishing, SGML](#)
 - [Database issues](#)
 - [Agents](#)
 - [Commerce, economics, publishers](#) (See Chapter 9 in Dr. Lesk's book.)
 - [Intellectual property rights, copyright laws & security](#) (See Chapter 10 in Dr. Lesk's book.)
 - [Social issues](#) (See Chapters 11, 12 in Dr. Lesk's book.)
-

Pedagogy:

We recommend that the topics be covered in the order given above, with the reader examining the material in the book by Dr. Lesk before visiting the online information. Topics that do not correspond to chapters in the book have been included as supplementary material that seemed to be of special interest to students at Virginia Tech, and/or where there is keen interest and progress by the digital library community. However, these can be skipped by novices interested in a general overview.

[\[Main\]](#) [\[Contents\]](#)

Search, retrieval, resource discovery:

Searching - LoC

- [LoC Home Page](#)
- [The WWW Virtual Library arranged by LoC standards](#)
- [UNDERSTANDING AND COMPARING WEB SEARCH TOOLS](#)
- [Matrix of WWW Indices: A comparison of Internet indexing tools](#)

Federated search

- [UIUC Federation Across Heter. DBs](#)
- [STARTS](#)
- [INFOSEEK patent](#)
- [TSIMMIS](#)
- [Virginia Tech Federated Search Demonstration for NDLTD \(theses, dissertations\)](#)

CyberStacks (WWW, Classification, Catalogs, Reviews/Clearinghouses)

- [Home Page](#)
- [Net Projects](#)
- [Alphabetical topics vs. LC ranges](#)
- [Call for contributions](#)
- Question: Which efforts are far along? What demonstrations can you find that are the most informative / explanatory? How well does the Library of Congress classification system fit for WWW resources?
- Related work: [OCLC's Scorpion Project](#); [DDC](#)

Columbia

- [D-Lib Article on Images/Video](#)
- [WebSeek Home Page](#)

BioKleisli

- [project](#)
- [demos](#)

[Filtering](#)

[Cross-Language Information Retrieval Resources](#)

- [Eurospider Demos](#)
- [Analogical Language Processor Demo](#)
- [Mundial](#) - English and Spanish Demo
- Questions:
 - What languages are covered?
 - How well are phrases handled?

[Stanford DL info finding projects](#)

[Berkeley documents and queries](#) (please study carefully, answering questions)

[UCSB spatial indexing and retrieval](#)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

UNDERSTANDING AND COMPARING WEB SEARCH TOOLS

[Beyond Surfing: Tools and Techniques for Searching the Web](#)

by Kathleen Webster & Kathryn Paul, January 1996

[General Internet Resource Finding Tools:](#)

A Review and List of Those Used to Build INFOMINE, March 1996

[How to Search the Web - A Guide to Search Tools](#)

by Terry A. Gray

[Jacob Hausauer's Page for Search Engines](#)

March, 1996

[Just the Answers, Please](#)

©, Susan Feldman

Searcher Magazine, 1997

note: this link will expire in May, 1998

[Librarians' Index to the Internet](#)

Lots of useful links about searching. Be sure to check "about"

[Literature about search services](#)

by Traugott Koch

January, 1996; updated Nov., 1996

[Precision among World Wide Web Search Services \(Search Engines\): Alta Vista, Excite, Hotbot, Infoseek, Lycos](#)

By H. Vernon Leighton and Dr. Jaideep Srivastava,
June 1997

[Reviews of Search Engines](#) from the Search Page.

June 1996; updated, November 1996

[Search the Net: Top Internet Searching Resources Reviewed](#)

by Tracy Marks

February, 1997; updated October, 1997

[Signal Detection Analysis of WWW Search Engines](#)

by Carsten Schlichting & Erik Nilsen, Lewis & Clark College
1996

[Top keyword Resources of the Web](#)

by John December, November, 1996

[Tips on Popular Search Engines](#)

by Karen Campbell, March 1997

[Top keyword Resources of the Web](#)

by John December, November, 1996

[Search Engine Reference List](#)

by Rowan Brownlee, April 1996
from Web4Lib

[Understanding WWW Search Tools](#)

Jian Liu, September 1995, February 1996
This page describes some of features and drawbacks of various search tools

[UCB Library Internet Search Tool Details](#)

Library, University of California, Berkeley, November 1995; updated September 1996
This list includes information on the size of each search engine's database.

[World Wide Web Indexes: a study](#)

H. Vernon Leighton, June 1996
This paper compares the performance of four major search engines: Lycos, Infoseek, WebCrawler, and WWWorm.

[World Wide Web Searching Tools, An Evaluation](#)

Ian Winship, 1995
This evaluation compares four search engines, Lycos, WebCrawler, WWWorm, and Harvest, and two Subject Trees with search engines, Yahoo and EiNet Galaxy



[Return to Library](#)



[Return to Searching the Internet](#)

© Bush Library, Hamline University, 1995

This document may be freely distributed in its entirety for educational purposes only

Karen Campbell, April 1996

Updated: January, 1997



CONTENTS



[Minutes](#)



[Presentations](#)



[AGENDA](#)



[ATTENDEES](#)

FEDERATION ACROSS HETEROGENEOUS DATABASES

April 3-4, 1997

Grainger Engineering Library Information Center

University of Illinois at Urbana-Champaign
1301 W. Springfield Ave., Urbana, IL

Welcome to the official site for the UIUC Digital Library
Initiative Spring '97 Partners Workshop.

Please contact Susan Harum dli@uiuc.edu for any questions
or comments about the workshop.

[Go back to the DLI workshop page](#)

STARTS

Stanford Protocol Proposal for Internet Search and Retrieval

STARTS is the result of an informal "standards" effort that we ([Luis Gravano](#), [Kevin Chang](#), [Hector Garcia-Molina](#), [Carl Lagoze](#), and [Andreas Paepcke](#)) are coordinating at Stanford. This project developed a simple protocol that text search engines should follow to facilitate searching and indexing multiple collections of text documents.

[Final writeup](#) of the *STARTS* protocol ([PostScript version](#))

[A reference-implementation](#) of *STARTS* by Carl Lagoze

[A more readable description](#) of the *STARTS* protocol that appeared in Sigmod'97

[List of participants](#) of the *STARTS* Workshop, Stanford, August 1st, 1996

Slides of the talk that Prof. Hector Garcia-Molina gave at the *STARTS* workshop ([Powerpoint Version](#))

Slides of the talk that Luis Gravano gave at the *STARTS* workshop ([Powerpoint Version](#))

[Luis Gravano](#)
gravano@cs.stanford.edu



Distributed Search Patent

[Patents](#)

The Infoseek Distributed Search patent is a novel technique for performing full-text searches over distributed databases. The technique is directly applicable to searching web sites on the Internet, as well as geographically distributed databases within corporate Intranets. The patent, US Patent Number 5,659,732, entitled "Document Retrieval over networks wherein ranking and relevance scores are computed at the client for multiple database documents," was issued on August 19, 1997.

[Press release](#)

The official corporate press release announcing the patent

[Background information](#)

An expanded press release containing more technical information and additional background information

[News Articles](#)

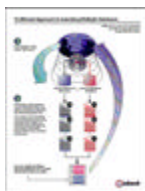
News articles appeared in [The New York Times](#), [Inter@ctive Week](#), and [CNET](#).

[Patent text](#)

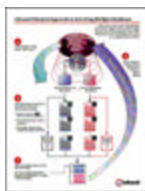
Text of the patent

[Graphic](#)

Illustration of traditional vs. Infoseek patent approach. This file contains two pages: the first page depicts the traditional approach, while the second page portrays the Infoseek patented approach. Available in GIF (left) or PDF (below) format.



[Traditional Approach](#)
(GIF)



[Infoseek Approach](#)
(GIF)



[PDF format](#)



Why the name?

As an acronym, TSIMMIS stands for "*The Stanford-[IBM](#) Manager of Multiple Information Sources.*" In addition, TSIMMIS is a Yiddish word for a stew with "heterogeneous" fruits and vegetables integrated into a surprisingly tasty whole.

Short Project Description

The goal of the TSIMMIS Project is to develop tools that facilitate the rapid integration of heterogeneous information sources that may include both structured and semistructured data. TSIMMIS has components that:

- translate queries and information (source wrappers);
- extract data from World Wide Web sites;
- combine information from several sources (mediator);
- allow browsing of data sources over the Web.

The TSIMMIS project is funded by [DARPA](#).

TSIMMIS Links

- TSIMMIS [publications](#)
- [People](#) in the TSIMMIS project
- [Developer's page](#) (restricted access)

TSIMMIS Related Links

- [LORE](#), an OEM repository
- [I3 Initiative Projects Home Page](#)
- [DARPA Progress Reports](#)
- [Garlic](#), our sister project at IBM

Demo And Source Code

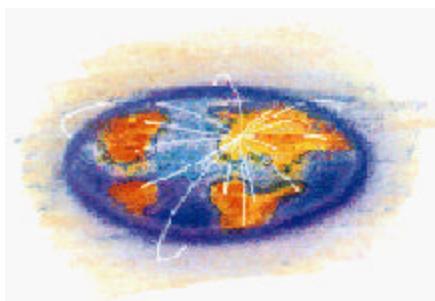
An overview of [MOBIE](#) used for the demo.

- Run a [Stock mediator](#) demo
- Run a [Other sources\(weather source, bibliographic sources\)](#) demo
- [Download source code](#)

ETD Digital Library

[About ETD Federated Search](#)

[Search or Browse the Catalog](#)



CyberStacks(sm)

Welcome To CyberStacks(sm)!

CyberStacks(sm) is a *centralized, integrated, and unified* collection of significant World Wide Web (WWW) and other Internet resources categorized using the Library of Congress classification scheme. Resources are organized under one or more relevant Library of Congress class numbers and an associated publication format and subject description. The majority of resources incorporated within its collection are monographic or serial works, files, databases or search services. All of the selected resources in *CyberStacks(sm)* are *full-text, hypertext, or, hypermedia*, and of a research or scholarly nature.

Using an abridged Library of Congress call number, *Cyberstacks(sm)* allows users to browse through a virtual library stacks to identify potentially relevant information resources. Resources are categorized:

- * first within a broad classification,
- * then within narrower subclasses,
- * and then finally listed under a specific classification range and associated subject description that best characterize the content and coverage of the resource.

For each resource, a brief summary is provided, and when necessary, specific instructions on using the resource are also included. Where appropriate, the mode of access to the resource is noted, as is the subject coverage and scope; notable features, where applicable, are also included. At present, *CyberStacks(sm)* is a prototype demonstration service and is limited to significant WWW and other Internet resources in selected fields of Science and Technology.

A systematic effort is now underway to also identify and review resources that relate to the missions of the Center for Indigenous Knowledge for Agriculture and Rural Development (CIKARD) and the International Institute of Theoretical and Applied Physics (IITAP), two international research centers based at Iowa State University. While most of the current collection consists of Reference works, a number of scholarly journal Tables of Contents were recently added to its Title Index. Selected full-text serial titles and non-Reference monographic works, with subject coverage relevant to the interests of IITAP and CIKARD, have also been included.

Significant studies, essays, reports, proceedings, or other unique information sources of potential value to the efforts of CIKARD and IITAP, are also listed. Selected table of contents for non-Reference monographic works have also been included if such works are particularly relevant to the interests of these research centers and their associates.

Thank You For Visiting CyberStacks(sm)!

BROWSE and SEARCH		
Main Menu	Cross-Classification Index	Title Index

UNDER CONSTRUCTION				
Nominations	Participation	Virtual Advisory Boards	Web Publication Suggestion	Planned Enhancements

This site is NetScape 3.0 enhanced using HTML 3.0.

Acknowledgements and Disclaimers	Support	Warranties and Liabilities

[Special Thanks](#)

News and Net Publications							
D-Lib Magazine	IATUL Proceedings (New Series)	Intelligence	Internet Trend Watch for Libraries	InterNIC News	Issues in Science and Technology Librarianship	OCLC Internet Cataloging Project Colloquium	Untangling the Web

RECOGNITION
Map to Navigating the Web: Web Indexes
PC Computing
MDLink Approved
MDLink: Maclean Hunter Medical Publishing & Communications Group

[NET PROJECTS](#)

gerrymck@iastate.edu

April 20, 1998



The Scorpion Project



[Scorpion](#) is a project of the [OCLC Office of Research](#) exploring the indexing and cataloging of electronic resources. Since subject information is key to advanced retrieval, browsing, and clustering, the primary focus of Scorpion is the building of tools for automatic subject recognition based on well known schemes like the [Dewey Decimal System](#).

Scorpion Documentation

- [A brief introduction to Scorpion](#)
- [Evaluating Dewey Concepts](#)
- [Evaluating Scorpion Results](#)
- [Measures for Evaluating ...](#) **NEW**
- [Clustering](#) **NEW**
- [AMIGOS 97](#) (full image [version](#))
- [Scorpion helps catalog the Web](#)
- [Dewey Database Design](#)
- [ESS Field Label Descriptions](#)
- [Example ESS Record](#)
- [SMART Weighting Schemes](#)

Automatic Subject Assignment

- [Change User Password](#)
- [Simple URL Input Form](#) **NEW**
- [Simple Text Input Form](#) **NEW**
- [Advanced Input Form](#) **NEW**

Related Work

- [Online Classification: Implications for Classifying and Document\[-like Object\] Retrieval](#)
 - [Using Library Classification Schemes for Internet Resources](#)
 - [Dewey 2000](#)
 - [Cataloguing Rules and Conceptual Models](#)
 - [The Dublin Core](#)
 - [Prototype Dublin Core Metadata System](#)
 - [Electronic classification schemes](#)
 - Pharos ([demo](#), [publications](#))
-



Comments/suggestions to shafer@oclc.org
Scorpion [contributors](#)

 Dewey Home Search Site Map What's New Feedback**Content Areas**[About the DDC](#)[News](#)[DDC Updates](#)[DDC Users](#)[Products](#)[Research](#)[Dewey for Windows](#)**Popular Pages**

- [Dewey Screen Saver](#)
- [DDC Summaries](#)
- [LCSH/DDC](#)
- [New & Changed Entries](#)
- [DFW Guide](#)



OCLC Forest Press
Dewey
Decimal Classification

 Search Dewey Web Site

For more advanced searching, select the Search icon in the toolbar above

Last updated [June 15, 1998](#). ©1997, 1998 OCLC
[View Text Only](#)

OCLC Forest Press
6565 Frantz Road, Dublin, OH
43017-3395, USA
+1-614-764-6000
oclc@oclc.org

**[OCLC Online Computer
Library Center, Inc.](#)**
6565 Frantz Road, Dublin, OH
43017-3395, USA
+1-614-764-6000
oclc@oclc.org

For more contact information, see [Contacts & Addresses](#).

MAGAZINE

Finding Images/Video in Large Archives

Columbia's Content-Based Visual Query Project

Shih-Fu Chang, John R. Smith
Horace J. Meng, Hualu Wang, and Di Zhong
Department of Electrical Engineering and
Center for Telecommunications Research
Columbia University

{sfchang,jrsmith,jmeng,hwang,dzhong}@ctr.columbia.edu

D-Lib Magazine, February 1997

ISSN 1082-9873

Table of Contents

- [An Application Driven Problem](#)
 - [State of the Art](#)
 - [Research Strategies](#)
 - [Prototype Systems](#)
 - [Testbed Support and User Evaluation](#)
 - [Open Issues](#)
 - [References](#)
-

An Application Driven Problem

How do we find a photograph from a large archive which contains thousands or millions of pictures? How does a CNN video journalist find a specific clip from the myriad of video tapes, ranging from historical to contemporary, from sports to humanities? How do people organize and search the content of personal video tapes of family events, travel scenes, or social gatherings?

The era of "the information explosion" has brought about the wide dissemination and use of visual information, particularly, digital images and video, which we are also seeing in combination with text, audio, and graphics. The development of tools and systems that enhance image functionalities, such as searching and authoring, is critical to the effective use of visual information in the new media applications.

The current research and development of images and video search tools is driven by practical applications. We are seeing the establishment of large digital image and video archives, such as the Corbis catalog, which includes the Bettman Archive; the Picture Exchange, which is a joint venture between Kodak and Sprint; and many digital video libraries in various domains (e.g., environment,

politics, arts), such as the on-line CNN news archives.

The systems for the search and retrieval of images and video from these archives require the development of efficient and effective image query tools.

State of the Art

The use of comprehensive textual annotations provide one method for image and video search and retrieval. Today, text-based search techniques are the most direct, accurate, and efficient methods for finding "unconstrained" images and video. Text annotation is obtained by manual effort, transcripts, captions, embedded text, or hyperlinked documents. In these systems, keyword and full text searching may be enhanced by natural language processing techniques to provide great potential for categorizing and matching images.

The searching of images by their visual content complements the text-based approaches. Very often, textual information is not sufficient. Visual features of the images and video also provide a description of their content. By exploring the synergy between textual and visual features, these image search systems may be further improved. However, it is a significant challenge to automatically reconcile inconsistency between input from these features.

Many content-based image search systems have been developed for various applications. There has been substantial progress in developing powerful tools which allow users to specify image queries by giving examples, drawing sketches, selecting visual features (e.g., color and texture), and arranging spatial structure of features. Using these approaches, the greatest success is achieved in specific domains, such as remote sensing and medical applications. The reason is that in constrained domains, it is easier to model the users' needs and to restrict the automated analysis of the images, such as to a finite set of objects.

The integration of computer vision and image processing promises a wealth of techniques for solving the image and video search problems. But new challenges remain. In unconstrained images, the set of known object classes is not available. Also, use of the image search systems varies greatly. Users may want to find the most similar images, find an appropriate class of images, browse the image collection quickly, and so on. One unique aspect of image search systems is the active role played by users. By modeling the users and learning from them in the search process, we can better adapt to the users' subjectivity. In this way, we can adjust the search system to the fact that the perception of the image content varies between individuals, or over time.

The general system architecture for a content-based visual query system is included in Figure 1. The analysis of images and feature extraction plays important roles in both off-line and on-line processes. Other important aspects of the system include the closed interaction loop (including users), the supporting database components for retrieval and indexing, the integration with multimedia features, and the efficient user interfaces for query specification and image browsing.

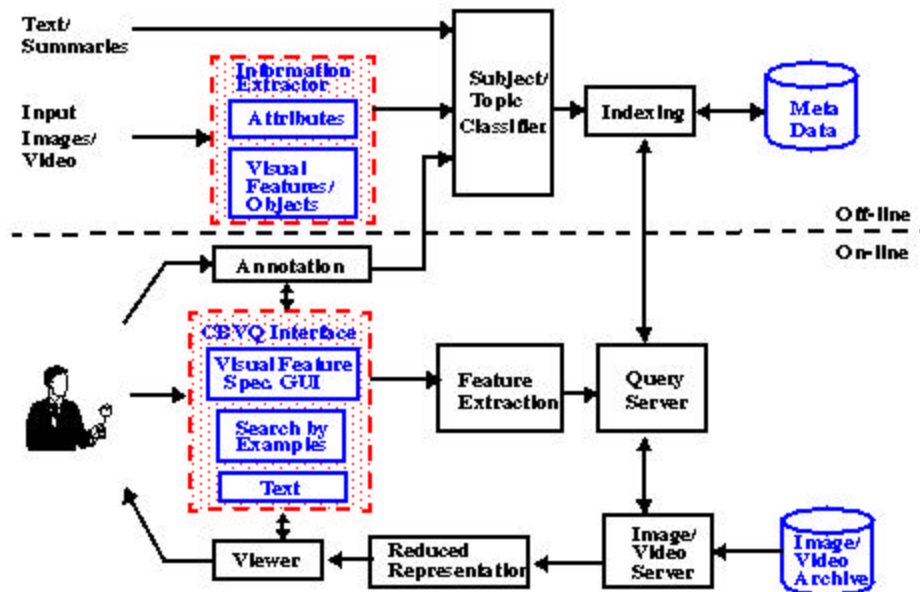


Figure 1. A general CBVQ system architecture.

The search of images is an emerging field with many exciting research challenges. The research tasks are practical, important, but not easy. In the following, we present our research strategies, prototype systems for image/video search, and our views on the important open research issues.

Research Strategies

We present our strategies for tackling the above challenging issues in this section.

Create a visual feature library by automatic image analysis

Although today's computer vision systems cannot recognize high-level objects in unconstrained images, we are finding that low-level visual features can be used to partially characterize image content. These features also provide a potential basis for abstraction of the image semantic content. The extraction of local region features (such as color, texture, face, contour, motion) and their spatial/temporal relationships is being achieved with success. We argue that the automated segmentation of images/video objects does not need to accurately identify real world objects contained in the imagery. Our goal is to extract the "salient" visual features and index them with efficient data structures for fast and powerful querying. Semi-automated region extraction processes and use of domain knowledge may further improve the extraction process.

In the later sections, we discuss the use of automatically extracted spatial/color regions for image search, and the integration of multiple visual features for video object indexing. We use a hierarchical object based schema for feature indexing and high-level object abstraction [4] (see Figure 2). The fusion of multiple visual features improves the region extraction process. We also show that the aggregation of regions into higher level objects is influenced by the spatial/temporal relationships of the regions. For example, Figure 3 shows the results of automatic video object segmentation and tracking. The visual features and spatial/temporal attributes of regions generate an index for searching for the video objects stored in the archive.

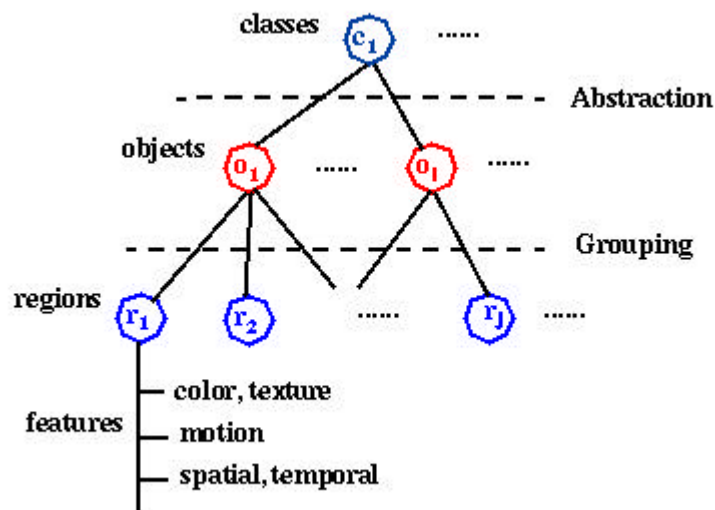


Figure 2. A hierarchical object based schema for images/video.

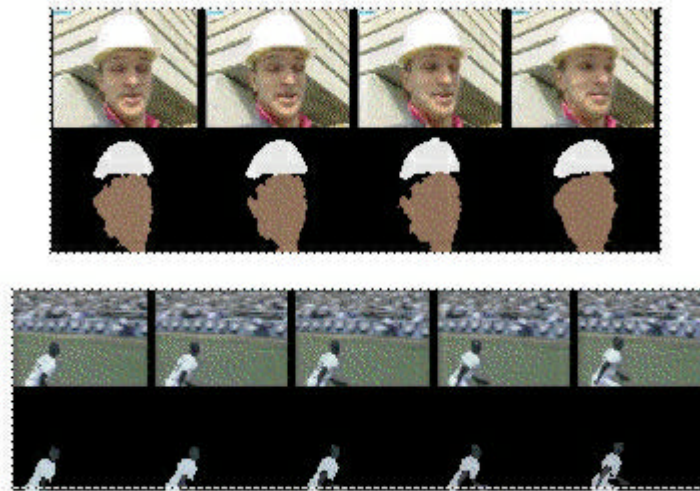


Figure 3. Examples of automatically segmented and tracked video objects.

Explore the synergy between compression and functionalities

It's impossible to anticipate the users' needs completely at the feature extraction and indexing stage. The ideal solution is that images and video are represented (for compression also) in a way that is amenable to dynamic feature extraction. Today's compression standards (such as JPEG, MPEG-1, MPEG-2), are not suited to this need. The objective in the design of these compression standards was to reduce bandwidth and increase subjective quality. Although many interesting analysis and manipulation tasks can still be achieved in today's compression formats (as described later), the potential functionalities of the images were not considered. However, recent trends in compression, such as MPEG-4 and object-based video, have shown interest and promise in this direction. The goal is to develop a system in

which the video objects are extracted, then encoded, transmitted, manipulated, and indexed flexibly with efficient adaptation to users' preference and system conditions.

Learn from users and domain ontologies

To break the barrier of decoding semantic content in images, user-interaction and domain knowledge is needed. These systems learn from the users' input as to how the low-level visual features are to be used in the matching of images at the semantic level. For example, the system may model the cases in which low-level feature search tools are successful in finding the images with the desired semantic content. In this way, the categories can be monitored and better analyzed by the system. Learning and other techniques in artificial intelligence provide great potential for these systems.

If the applications require the definition of specific semantic subjects, the feature models of images in these classes are constructed by hand and then used to match objects in the unknown images/video. This object recognition and subject classification method provides a system for on-line information filtering. We see great potential for improving image search systems to link the low-level visual features with high-level semantics. However, in unconstrained application domains, we expect only moderate success early on.

Integrate visual and other multimedia features

Exploring the association of visual features with other multimedia features, such as text, speech, and audio, provides another potentially fruitful direction. Our experience indicates that it is more difficult to characterize the visual content of still images compared to video. Video often has text transcripts and audio that may also be analyzed, indexed, and searched. Also, images on the World Wide Web typically have text associated with them. In this domain, the use of all potential multimedia features enhances image retrieval performance.

Prototype Systems

We have developed several content-based visual query prototype systems. WebSEEk and VisualSEEk explore the problem of efficiently searching large image archives. WebClip focuses on browsing, search, and content editing of networked video.

In WebSEEk, the images and video are analyzed in two separate automatic processes:

- (1) visual features (such as color histograms and color regions) are extracted and indexed off-line,
- (2) the associated text is parsed, and utilized to classify the images into subject classes in a customized image taxonomy (including more than 2000 classes).

More than 650,000 unconstrained images and video clips from various sources have been indexed in the initial prototype implementation. Users search for images by navigating through subject categories, or by using content-based search tools. The details of the system design and operation are described in [1].

One objective of WebSEEk is to explore the synergy between visual features and text. We also demonstrate the feasibility of image searching in a large scale testbed, the World Wide Web. We are developing more sophisticated content-based image search techniques in the VisualSEEk system [2].

VisualSEEk enhances the search capability by integrating the spatial query (like those used in geographic information systems) and the visual feature query. Users ask the system to find images/video that include regions of matched features and spatial relationships. Figure 4 shows a query example in which two spatially arranged color patches were issued to find images with blue sky and open grass fields.

File Edit View Go Bookmarks Options Directory Window Help

Back Forward Home Edit Reload Images Open Print Find Stop

Location:

SaFe Spatial and Feature query system

VisualSEEk Photographs Database: Spatial and Feature Query

2346 matches for REGION 1
2171 matches for REGION 2

0 [364] (479.16) 1 [2841] (511.52) 2 [2788] (528.12)

3 [99] (541.52) 4 [1606] (558.76) 5 [372] (609.52)

6 [1141] (600.24) 7 [1128] (601.60) 8 [1741] (600.40)

Query

Grid

A B C D E F G
H I J K L M N
O P Q R S T U
V W X Y Z 1 2

Spatial Query:
☒ Absolute ☐ Relative

Query Weights:
Spatial 10
Feature 10
Size 10
Region 10

Figure 4. An example query using VisualSEEk.

For video, we have developed a system called WebClip [3], which allows for efficient browsing and editing of compressed video over the Web. One objective is to demonstrate the benefits of using compressed video without full decoding during the content analysis and manipulation stages. Visual features (like scene changes, foreground motion objects, and icon streams) can be extracted directly from the compressed video. Web users do not need high-end video decoding or processing facilities like those used in professional studios. Another objective of WebClip is to integrate the search and editing functionalities in the same environment. Tools developed in image search systems (like the above mentioned WebSEEk and VisualSEEk systems) are being ported to the video system. We are also adding new tools for searching by motion feature and temporal characteristics. After retrieval of matched video clips, the users use web-based tools to edit the video and compose new presentations with various video special effects.

Figure 5 shows the functionality components of WebClip. The compressed video sequences are parsed to obtain visual features and objects. The browsing and search interface provides a tree-structure hierarchical scene-based interface. This display can be adapted to different browsing modalities:

- (1) the time-based model,
- (2) the story-based model, and
- (3) the feature-based model.

The time-based model hierarchically lays out the icons of key frames from each video scene. This allows for rapid inspection of video content according to a sequential order of time. The story-based model recognizes (automatically or manually) the story structure within the video (e.g., a complete news story) and groups all video scenes belonging to the same story under a single node in the tree. The feature-based model clusters all video scenes to classes within each of which all video scenes have similar visual features. We have also undertaken new efforts to extend the joint spatial/feature query tools of the VisualSEEK system to the video domain. Video is indexed and searched by spatial/temporal relationships and visual features of video objects contained in the video sequence.

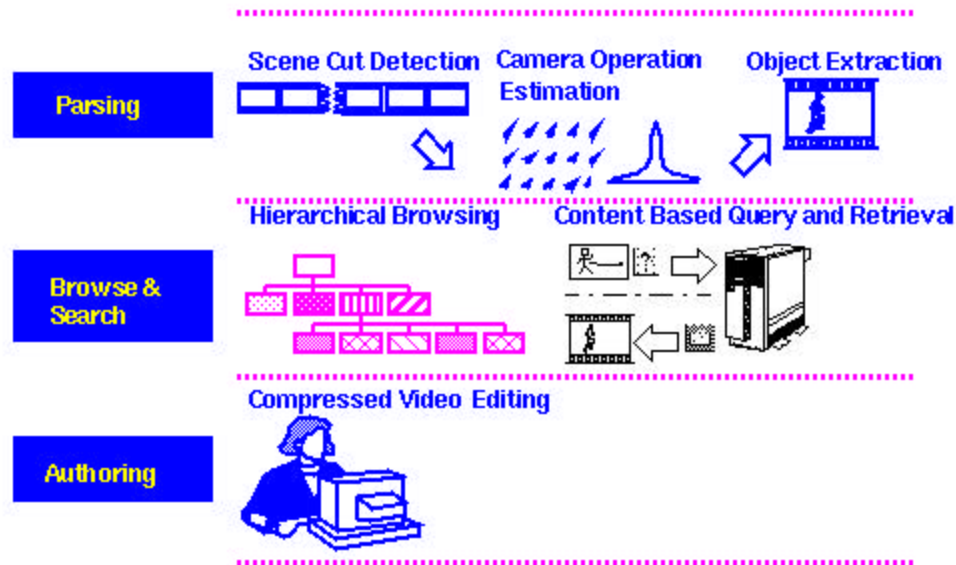


Figure 5. Functionality components of WebClip.

Testbed Support and User Evaluation

Most of the test images and video in our testbed are collected from the public domain, including data on the Web, copyright free photograph stock from commercial CD's, MPEG simulation test video, and proprietary content from local research groups. Features extracted from these images are stored in our SGI ONYX-based server, which has 50GB storage space on disk arrays, and 50GB tertiary space on a tape archive.

Network facilities include standard Internet connections (via a T-3 line to outside), ATM connections within the campus and with external wide area networks (NYNET), and internal wireless networks running mobile IP. A video-on-demand (VoD) system which supports software-based video servers, MPEG-2 transport, and heterogeneous client terminals has been developed in the Image and Advanced TV lab. We envision the integration of our search systems with the VoD system soon to provide integrated image services.

An important work plan for the near future is the collaboration with faculty and students in the School of Journalism and at Teachers College, Columbia University. User studies and performance evaluation are being conducted in the news and education domains. One example is the Columbia Digital News Systems group [5], which integrates our efforts with others on information tracking, natural language processing, and multimedia briefing.

Open Issues

Image/video searching is a relatively new field, but it has many exciting research issues. It requires close interaction between multiple technical disciplines and applications users. Researchers have made great progress in recent years, but a few critical issues have still not been addressed adequately. In particular, we believe that further breakthroughs need to be made in the following areas before image search systems can make significant impacts on real applications.

Effective evaluation metrics and testset

Today, there are no satisfactory methods for measuring the effectiveness of image search techniques. Precision/recall types of metrics have been used in some of the literature but are impractical due to the tedious process of measuring image relevancies. There are no standard image corpus or benchmark procedures. We believe that resolution of this issue is of top priority for researchers and users in this field.

Dynamic extraction and matching of visual features

As mentioned earlier, the image indexing and search schemes must adapt to dynamic user needs, resource conditions, and input data. In particular, the user needs and application requirements vary over time. A static set of features and matching schemes is limited. Efficient, if not real-time, methods should be developed to perform dynamic feature extraction, matching and abstraction. Real-time is defined in three different aspects:

- (1) fast enough to process live information (like live video),
- (2) fast enough to process a large amount of new information on-line (like on-line information filtering), and
- (3) fast enough to re-process existing data in the archive.

The degree of time urgency decreases in the same order. All these aspects demand breakthroughs in image/video representation and dynamic content analysis.

Linking low-level features to high-level semantics

Today's content-based image search systems allow for image queries based on image examples, feature specification, and primitive text-based search. The WebSEEk system uses automatically extracted text in image subject classification. Other researchers have also shown some success in using newspaper photograph captions and video transcripts to assist visual content analysis. Adaptive visual feature organization through user interaction has also been proposed. But the linkage between low-level visual features and high-level semantics is still very weak. Non-technical, general users tend to expect the same level of functionalities as those seen in today's text search systems. We admit that this is a difficult objective. But, as they are driven by critical application needs, image search systems will benefit from any breakthrough made in this direction.

Acknowledgements

This project is supported in part by the ADVENT industry partnership project at the Image and Advanced TV Lab of CTR, Columbia University, Columbia Digital Library project, and National Science Foundation (IRI-9501266). We appreciate the research collaboration in this area with Dr. Chung-Sheng Li of IBM, Dr. Kenrick Mock of Intel, Dr. Harold Stone of NEC, Dr. HongJiang Zhang of Hewlett-Packard, and Mr. Jan Stanger.

References

1. J. R. Smith and S.-F. Chang, "Searching for Images and Videos on the World-Wide Web," to appear in IEEE Multimedia Magazine, Summer, 1997. (also Columbia University CU/CTR Technical Report

#459-96-25). Demo: <http://www.ctr.columbia.edu/webseek>
<ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96e.ps>

2. J. R. Smith and S.-F. Chang, "VisualSEEK: A Fully Automated Content-Based Image Query System," ACM Multimedia Conference, Boston, MA, Nov. 1996. Demo:
<http://www.ctr.columbia.edu/VisualSEEK>
<ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96f.ps>

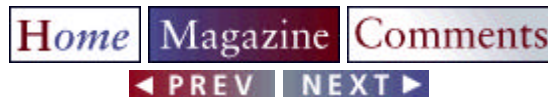
3. J. Meng and S.-F. Chang, "CVEPS: A Compressed Video Editing and Parsing System," ACM Multimedia Conference, Boston, MA, Nov. 1996. Demo: <http://www.ctr.columbia.edu/WebClip>
<ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/meng96c.ps>

4. D. Zhong and S.-F. Chang, "Video Object Model and Segmentation for Content-Based Video Indexing," IEEE Intern. Conf. on Circuits and Systems, June, 1997, Hong Kong. (special session on Networked Multimedia Technology & Application)
<ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/97/zhong97a.ps>

5. A. Aho, S.-F. Chang, K. McKeown, D. Radev, J. Smith, and K. Zaman, "Columbia Digital News Systems," to appear in Workshop on Advances in Digital Libraries, 1997.

Approved for release, February 14, 1997.

Copyright ©1997 Shih-Fu Chang, John R. Smith, Horace J. Meng, Hualu Wang, and Di Zhong



hdl:cnri.dlib/february97-chang



at Columbia University

A Content-Based Image and Video Search and Catalog Tool for the Web

(press here to [Browse](#) all subjects)

Animals

[birds, dinosaurs,](#)
[monkeys, fishes](#)

Cats

[leopards, lions,](#)
[kittens, cheetahs](#)

Horror

[godzilla, aliens,](#)
[skeletons, monsters](#)

Nature

[sunsets, flowers,](#)
[weather, mountains](#)

Architecture

[bridges, lighting, domes](#)
[heating](#)

Celebrities

[bullock, aniston, monroe,](#)
[keanu](#)

Humour

[simpsons, beavis, dilbert,](#)
[ren/stimpy](#)

Sports

[baseball, basketball,](#)
[swimming, hockey,](#)
[olympics, surfing](#)

Art

[painting, illustr, sketching](#)
[cezanne, monet, vangogh](#)

Dogs

[bulldogs, puppies, coyotes,](#)
[wolves](#)

Movies

[batman, starwars, jurassic,](#)
[python, blade runner, actresses](#)

Transportation

[cars, planes, titanic,](#)
[motorcycles, porsches](#)

Astronomy

[nasa, planets, eclipses,](#)
[space](#)

Food

[apples, beer, pizza, cakes,](#)
[fruits, veges](#)

Music

[beatles, metal, rock, cure,](#)
[zeppelin, guitars](#)

Travel

[asia, europe, newyork,](#)
[paris, australia, mexico](#)

Image/Video Topic

(single word)

Search

☒ all ☐ videos ☐ color photos ☐ gray images ☐ graphics

[\[WebSEEk\]](#)

[\[browse\]](#)

[\[add urls\]](#)

[\[postcards\]](#)

[\[info\]](#)

[\[credits\]](#)

WebSEEk has catalogued **665115** images and videos

All licensing inquiries may be directed to [Dr. Joseph R. Flicek](#) of the Columbia Innovative Enterprise.

BioInformatics Centre

Research Unit

Hosted at [Kent Ridge Digital Labs](#)

BioKleisli is a collaboration with Kent Ridge Digital Labs (previously known as the Institute of Systems Science) and the University of Pennsylvania. It is a tool for the broad scale integration of databanks. It offers flexible access to biological information sources that are highly heterogeneous, geographically scattered, highly complex, constantly evolving and high in volume.

- BioKleisli offers high-level flexible access to human genome and other molecular biological sources that are
 - Highly heterogeneous
 - Geographically scattered
 - Highly complex
 - Constantly evolving
 - High in volume
 - Flexible Access = Migrate + Integrate + Restructure
 - Deployed at the HGP Philadelphia Center for Chromosome 22 for
 - Integration of private and public data banks
 - Integration of data banks, analysis software, and visualization tools
- Major benefit: Typical query implementation time is reduced from weeks to days (some times, hours).**
- Why is BioKleisli Special?
 - Self-describing data model for complex structured data. Beyond the reach of relational DBs.
 - High-level query language for data transformation. ``Internet SQL" as opposed to simple navigational IR.
 - Flexible yet precise control in ad-hoc queries.
-

For a recent poster on BioKleisli, click [here](#).

For more information on the underlying technology, click [here](#).

For some examples queries powered by BioKleisli, click [here](#).

For list of some bioinformatics sources BioKleisli talks to, click [here](#).

For the history of the project, click [here](#).

BioInformatics Centre

Research Unit

Hosted at [Kent Ridge Digital Labs](#)

Here some links to see demonstrations of some of our work. (Some machines are only accessible from within KRDL.)

These are demo machines. Not for heavy duty use. If you want to do large-scale queries, please contact us. (We kill external processes left-right-and-center when they interfere with our daily work.)

- **BioKleisli** ([@adenine](#) [@cytosine](#) [@guanine](#) [@thymine](#) [@uracil](#) [@alanine](#) [@arginine](#))
A collaboration with Kent Ridge Digital Labs and the University of Pennsylvania.
- **SeqIndex** ([@adenine](#) [@cytosine](#) [@guanine](#) [@thymine](#) [@uracil](#) [@alanine](#) [@arginine](#))
A collaboration with Kent Ridge Digital Labs.
- **ViewBLAST** ([@adenine](#) [@cytosine](#) [@guanine](#) [@thymine](#) [@uracil](#) [@alanine](#) [@arginine](#))
A collaboration with Kent Ridge Digital Labs and Institute of Molecular and Cell Biology.
- **Web PHYLIB** ([@adenine](#) [@cytosine](#) [@thymine](#) [@uracil](#) [@alanine](#) [@arginine](#))
A collaboration with Kent Ridge Digital Labs.
- **DNA Chip** ([@adenine](#) [@cytosine](#) [@thymine](#) [@uracil](#) [@alanine](#) [@arginine](#))

[Limsoon Wong](#) / BioInformatics Center and [Kent Ridge Digital Labs](#), 21 Heng Mui Keng Terrace, Singapore 119613 / Limsoon@Saul.CIS.UPenn.EDU, Limsoon@KRDL.Org.SG

Information Filtering

U. Md. Information Filtering

- [Defn](#)
- Fast Data Finder: [Genetic sequence analysis](#)

Questions:

- What is *information filtering*? How does it differ from information retrieval?

Information Filtering Resources

This page is designed as a resource for people conducting research in [information filtering](#). It was developed as part of the [Information Filtering Project](#) at the University of Maryland, and is maintained as a part of my ongoing [research program](#). The first section may be of interest to a wider audience, though, since it contains links to working systems for a variety of operating systems which are freely available on the net, and the third section should become more valuable to a wide audience as links to more commercial systems are added.

This page lists all known internet-accessible information filtering resources. If you are aware of resources which do not appear here, please [send mail to Doug Oard](#). Since it's impossible to look at everything here, I'm often asked where to start. I've tried to answer that in my recent [paper](#). Although I know of no comprehensive bibliography of papers, theses and dissertations on information filtering which exist only in printed form, scanning the bibliographies in the first two sections will reveal almost every source that I am aware of. There is also a good start at a comprehensive bibliography at the University of North Carolina in the second section.

[Freely Available Information Filtering Systems](#)

Working information filtering systems which are publicly available. In most cases, papers describing the theory and/or implementation details are also available.

[Commercial Information Filtering Systems](#)

An incomplete sample of commercial systems that might be of interest to researchers in the field.

[Information Filtering Papers and Project Descriptions](#)

Descriptions of experimental or proprietary systems for which the software is not being distributed. If you're looking for papers you should also check the previous section because research results included there are not repeated in this section.

[Other People Interested in Information Filtering](#)

Home pages of people who have either published work on information filtering that is not presently available on the net or who are actively working in the field. This is not a good place to look for papers, but a great way to learn how to contact people who don't appear in either of the previous two sections.

[Related Resource Pages](#)

Web pages which collect links to resources that may be of interest to information filtering researchers.

Last modified: Wed Jan 28 20:10:19 1998

[Doug Oard](#) oard@glue.umd.edu

Information Filtering Defined

A universally accepted definition of information filtering is, unfortunately, still lacking. So here is my personal definition, which I have used to build the Information Filtering Resources [web page](#). Generally, the goal of an information filtering system is to sort through large volumes of dynamically generated information and present to the user those which are likely to satisfy his or her information requirement.

In order to sharpen this definition, a distinction should be drawn between information collection and information filtering. In some domains (e.g. USENET News) the collection effort is minimal because the information comes to you. In other domains (e.g. the World Wide Web) the collection effort can be considerable because no mechanism exists to draw new information to the attention of a filtering system. The point to be made here, though, is that information collection is an interesting area in its own right, but I do not propose to include it in my definition of information filtering. In my view, the information filtering problem begins only after you have gained access to the new information.

Information filtering has been applied to a several domains using a variety of technical approaches. The original methods were manual alerting services that brought new information to the attention of users of research and special libraries. At the time this was referred to as Selective Dissemination of Information (SDI), a name which fell from favor about the time the Strategic Defense Initiative (SDI) was introduced in the United States :-). A few modern systems have adopted this remarkably descriptive name for the filtering process, however, and the interest in information filtering that has resulted from the present research thrusts in digital libraries arises at least in part from this tradition.

With the growth of the internet and other networked information, research in automatic filtering of networked information has exploded in recent years. Because of their low cost, large volume, and ease of recognizing new information, the most popular domains for research systems have been USENET News and electronic mail. The recent explosive growth of the World Wide Web has made this an interesting domain which has attracted some good research, although the information collection problem appears to make this a more difficult domain in which to conduct basic research on information filtering techniques. Another domain which has attracted considerable research interest is the annual Text REtrieval Conference (TREC) in which a standard text collection is used and a carefully controlled evaluation methodology is enforced. In TREC the information filtering task is referred to as "routing," adding somewhat to the confusion of terminology in this field. In fact, TREC recently adopted a special interest "filtering" track which adopts a different evaluation methodology but which conforms to the definition of filtering presented above. Commercial systems which filter newswire articles and other specialized information sources are becoming available as well. Filtering techniques will likely be applied to other domains such as images, sound and video in the future.

The distinction between information filtering and the more established field of information retrieval has proven to be the source of some confusion as well. Information retrieval broadly deals with the selection of information, and many of the features of information retrieval system design (e.g. representation, similarity measures or boolean selection, document space visualization) are present in information filtering systems as well. If one considers information retrieval from a very general "information selection" viewpoint, information filtering is simply a special case in which the information space is very dynamic. If, on the other hand, your personal definition of information retrieval involves selection of relatively static information in response to relatively dynamic queries, then information filtering is best viewed as the dual problem to information retrieval. Regardless of which viewpoint you take, though, it is clear that researchers in information filtering will likely benefit from familiarity with the legacy of

research in various aspects of information retrieval. For practical reasons I have not attempted to compile a comprehensive listing of network-accessible resources on information retrieval, however, so the interested researcher should refer to the Related Web Pages section of the Information Filtering Resources web page for some starting points on information Retrieval.

[*Doug Oard*](#)

Last modified: Tue Dec 12 15:33:26 1995

Cross-Language Information Retrieval Resources

This page is designed as a resource for people conducting research in [cross-language information retrieval](#). It is intended to collect references to all information on information retrieval systems which can accept queries in one language and return documents in another. It is maintained by the [Digital Library Research Group](#) of the [College of Library and Information Services](#) at the University of Maryland. If you are aware of resources that are within the scope of this page but do not appear here, please [send mail to Doug Oard](#).

[December 1997 D-lib Magazine Article](#)

A recently written introduction to cross-language information retrieval.

[Conferences](#)

The best single source for information in the field. This page includes links to the full proceedings of every major cross-language information retrieval workshop as well as to a fairly complete list of upcoming conferences and workshops that include some treatment of cross-language information retrieval.

[Cross-Language Information Retrieval Papers and Project Descriptions](#)

Another excellent place to look for information. Here you will find descriptions of experimental work on cross-language text retrieval that may not have been presented at one of the major workshops

[Working Systems](#)

Here you will find links to experimental and commercial cross-language information retrieval systems that you can either obtain or use over the net. Some carry a fairly hefty price tag, others are free.

[Bibliography](#)

A fairly comprehensive bibliography of published work on cross-language information retrieval in BibTeX form, last updated on July 3, 1997. The bibliography is also available in [postscript](#). Most of the references are described in at least one of my survey [papers](#) on cross-language information retrieval.

[Related Resource Pages](#)

Web pages which collect links to resources that may be of interest to cross-language information retrieval researchers. None of these pages are devoted solely to cross-language information retrieval.

Last modified: Wed Dec 24 00:49:55 1997

[Doug Oard](#) oard@glue.umd.edu



Information Finding Projects in the Stanford Digital Library

One of the major research thrusts of the Stanford Digital Library project is helping users to find information. We have initiated a number of projects in this area, most related to our over-arching theme of interoperability. We have looked at ways that search tools can be used across multiple sources that use different syntaxes or languages. We have also looked at tools to provide statistical or collaborative filtering to locate relevant articles.

[FAB](#)

FAB is an adaptive multi-agent information retrieval system which finds interesting pages on the web.

"[An Adaptive Agent for Automated Web Browsing](#)"

- [Marko Balabanovic](#)

[GLOSS](#)

The Glossary Server of Servers (GLOSS) project is designed to locate relevant information sources for your query.

"[Generalizing GLOSS to Vector-Space Databases and Broker Hierarchies](#)"

- [Luis Gravano](#)

[Query Translator](#)

Databases have different query syntax and different capabilities, even for simple Boolean queries. Translation allows a single query to be mapped into the native format appropriate for each database.

- [Chen-Chuan K. Chang](#)

[SenseMaker](#)

SenseMaker helps users iteratively reformulate their information needs through multi-dimensional organizing and active gathering of search results.

"[SenseMaker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests](#)"

- [Michelle Q Wang Baldonado](#)
-

Grassroots

Groupware for information finding, combines mail, news, and web in a single environment with distribution lists

" [Grassroots: A System Providing a Uniform Framework for Communicating, Structuring, Sharing Information, and Organizing People](#)"

- [Kenichi Kamiya](#)
 - [Martin Röscheisen](#)
-

The Stanford Digital Library Metadata Architecture

Services need to provide

- metadata about their offerings to help users decide when they should be invoked
- protocol metadata to figure out how they should be invoked, and
- collection metadata for what they should be invoked upon.

The metadata architecture provides a system organization to provide these metadata in a uniform, scalable way.

Metadata for Digital Libraries: Architecture and Design Rationale

- [Michelle Q Wang Baldonado](#)
 - [Chen-Chuan K. Chang](#)
 - [Luis Gravano](#)
 - [Andreas Paepcke](#)
-

STARTS: Stanford Protocol Proposal for Internet Retrieval and Search

A set of informal standards negotiated among the major search vendors and users to facilitate interoperation.

- [Chen-Chuan K. Chang](#)
 - [Hector Garcia-Molina](#)
 - [Luis Gravano](#)
 - [Andreas Paepcke](#)
-

Machine Learning for Information Retrieval

Statistical AI techniques allow the extraction of minimal sets of meaningful search terms

"[Toward Optimal Feature Selection](#)"

- [Mehran Sahami](#)
 - [Daphne Koller](#)
-

[BackRub](#)

BackRub is a web crawler which is designed to store the connection graph for the web. In other words BackRub stores which pages every web page links to. Currently we are developing techniques using this link data to improve web search engines as well as understand the structure of the web.

- [Larry Page](#)
-

[ComMentor](#)

Third-Party Annotations on web pages provide for ways to share information, rate content, and keep notes

["A Platform for Third-Party Value-Added Information Providers: Architecture, Protocols, and Usage Examples"](#)

- [Martin Röscheisen](#)
 - [Christian Mogensen](#)
 - [Terry Winograd](#)
-

[InterOp Protocol](#)

The heart of the "InfoBus", this protocol describes access methods to search collections, acquire results, and find out about sources.

- [Steve Cousins](#)
 - [Prof. Hector Garcia-Molina](#)
 - [Scott Hassan](#)
 - [Andreas Paepcke](#)
-

[SCAM: The Stanford Copy Analysis Mechanism](#)

Making a perfect digital copy of a copyrighted work is easy in a networked world. How can the intellectual property rightsholders be protected? By detecting attempted distribution of illegal copies. Duplicate detection has other uses in information finding as well. An earlier, related project was known as COPS: The Copyright Protection Scheme.

["Building a Scalable and Accurate Copy Detection Mechanism"](#)

- [Prof. Hector Garcia-Molina](#)
 - [Narayanan Shivakumar](#)
-

[InterBib](#)

InterBib is a tool for maintaining bibliographic information. Capable of reading from and writing to many different formats, it acts as a unified, searchable repository of bibliographic records.

[Information on InterBib](#)

- [Andreas Paepcke](#)

Multimedia, Representations:

The Basics:

- [text file formats](#)
- [graphic file formats](#)
- [hypermedia & multimedia](#)

ACM DL'97 Tutorial: [Multimedia Information and Systems](#)

Digital Video

- [An iconic visual language for video annotation](#) (Table of Contents. Segmented Annotation: hierarchical annotation structure on top of physical video stream)
- [Jacob: GUI iteratively refine query till satisfying result](#) (Note: Query specification can be direct, by example, or iteratively.)
- [CNN uses Quicktime for WWW daily news clips](#)

MHIA Courseware and Curricula

- [MHIA Home Page](#)
- [SIGIR 96 Workshop](#)
- [Drexel 96 Workshop](#)
- [IR Courses](#)
- [Information Engineering - European Commission: Work Program, Pilots](#)
- [MM 96 Workshop](#)
- [Lisbon Workshop](#)
- Questions:
 - What is the need for education related to information? What jobs?
 - What subjects should be covered in such education programs?
 - How should those subjects be ordered into each specific program?

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

ACM DL'97 Tutorial: Multimedia Information and Systems

Edward A. Fox

Content

- [Outline with References: Image and Video Processing, Retrieval](#)
- [CS4624: Multimedia, Hypertext, Information Access](#)
- **CS4624 highlights**
 - [Outline](#)
 - [Lab Sessions](#)
 - [Lecture Notes](#)
 - [Trips and Special Events](#)
 - [Computers and Tools](#)
 - [Figures](#)
 - [Figures with Captions](#)
 - **Glossary:** [Local](#), [from IMA](#)
 - [Index](#)
 - [Link Sets](#)
 - [Readings and References](#)
 - Syllabus
 - [Calendar](#)
 - [Department and Class Policies](#)
 - [Instructor and GTA](#)
 - [Syllabus Details: Grading, etc.](#)

References

- [Links for ACM DL'97 Tutorial](#)
- [Printed References on Image and Video Processing, Retrieval](#)
- [References from NSF EI Project](#)

[PDFs](#)

Copyright 1997 Edward A. Fox, all rights reserved

MHIA: Multimedia, Hypertext and Information Access Curriculum & Courseware

Welcome to WWW pages for a multi-institution, multi-association effort in the broad area of Multimedia, Hypertext and Information Access to develop curriculum and courseware, and related materials, services, and initiatives to promote education and training in this important field.

For an idea of the potential benefits and required efforts, see our [pre-proposal](#) or the [excerpt](#) of key parts from the 10/8/96 full proposal.

For more information, or if you would like to offer to help, send mail to fox@vt.edu to reach [Ed Fox](#).

Important Events:

- *Courseware, Education and Curriculum in Information Retrieval*
Aug. 22, 1996 - [workshop](#), part of the [workshop program](#) at [ACM SIGIR'96](#).
- *Courseware, Education and Curriculum in Multimedia*
Nov. 19, 1996, Room 301 - [ACM MM'96 Workshop](#) and [Conf. Info.](#)
For discussion, send email to mm96wk@fox.cs.vt.edu
- Workshop for ACM DL'97 and SIGIR'97
- Proposals to NSF:
 1. Multimedia Curricula and Repositories
 2. Computer Science Lab Repository and Registry
- National Academy of Science / NSF Workshop Aug. 1997 on National Digital Library for SME&T
- [Lisbon Workshop](#)

References:

- E. Fox and L. Kieffer. *Multimedia Curricula, Courses and Knowledge Modules*, *ACM Computing Surveys*, Dec. 1995, 27(4): 549-551.
- [ACM SIGIR WWW pages](#)
- [Multimedia Educational Materials](#) from [ACM SIGMM](#)'s education committee (thanks to Brian Smith at Cornell!)
- E. Fox, [presentation](#) for: *Information Retrieval 2000 --- Workplace Needs and Curricular Implications*
A Workshop/Symposium sponsored by the W.K. Kellogg Foundation
Hosted by Drexel University
May 24, 1996
Marriott Hotel, Philadelphia PA

Architectures:

Core topics include:

- [D-Lib article on architecture](#)
- [Other CNRI activities](#)
- **Naming**
 - [PURL](#)
 - [Handles](#)
- [Networks](#): online notes of Dr. Lesk

Other topics of general interest, that are being studied by the [D-Lib Metrics Group](#) include:

- **Distributed processing (client/server)**
- **Interoperability** (see [IITA workshop on Interoperability ...](#))
- **Performance**

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).


(c) Copyright 1998, Edward A. Fox, Rajat Gupta



Handles and Web Browsers

This note describes how to use a World Wide Web browser to access items identified by handles and other Uniform Resource Names (URNs).

Adding Handle Resolution to a Browser

The  [CNRI Handle System Resolver](#) is a web browser extension, that enables Netscape and Internet Explorer (version 3+) for the Windows and Macintosh platforms, to recognize the handle protocol and communicate directly with the Handle System to resolve handles into their associated URLs.

Resolving Handles Using a Proxy Server

Without the extension, web browsers must be directed to a Proxy Server, which understands the handle protocol, in order to resolve the handle to a URL. CNRI maintains two public proxy servers addressed by "hdl.handle.net" and "dx.doi.org", and one private proxy server, "hdl.loc.gov". Embedding a handle in a URL that uses one of the proxy URLs will permit any browser to resolve that handle.

For example, the February 1998 issue of *D-Lib Magazine* has an article by William Y. Arms with the handle (URN) "cnri.dlib/february98-arms". Currently the article is stored on three web servers, one maintained by CNRI, the others a mirror site at UKOLN in England and a mirror site at the Australian National University Sunsite. The story is formatted in html and accessible using the http protocol. The handle is the permanent name of the article. It will not be changed if *D-Lib Magazine* moves to different computers, the story is moved from a web server to some other type of storage, the magazine changes publisher, or even if CNRI and D-Lib change name or disappear. For these reasons, the best long-term way to cite the article is by the handle.

To read this article from a web browser which does not have the CNRI Handle Resolver extension, direct the browser to open the following URL, which combines a proxy server address with the handle "cnri.dlib/february98-arms", as shown in the example below:

<http://hdl.handle.net/cnri.dlib/february98-arms>


The article will then be displayed by the browser. The browser's "Address" or "Location" panel shows the location from which the article was read.

Handle Resolution Using a Form

You may also use the web form at

 <http://www.handle.net/docs/gethdl.html>

to see what URL a handle resolves to.

The  [Digital Object Identifier System](#) web site uses a proxy server to resolve Digital Object Identifiers (DOIs) into URLs.

NOTE: To demonstrate the Handle System, the hyperlinks on this site use handles. The small red arrows accompanying the text are hyperlinks containing handles; these require the [CNRI Handle System Resolver](#) to be installed on your computer. Underlined text contains handles embedded in URLs that are resolved using a [proxy server](#). These work with any browser. Links to sites not controlled by CNRI are conventional URLs.

[[home](#) | [introduction](#) | [documentation](#) | [download](#) | [administration](#) | [CNRI home page](#)]

Updated: 11 May 98



D-Lib Working Group on Digital Library Metrics

This Working Group is aimed at developing a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment. Initial emphasis will be on (a) information discovery with a human in the loop, and (b) retrieval in a heterogeneous world.

[Working Group Charter](#)

[Other Working Group Documents](#)

[Working Group Private Area](#)

This is an open working group, and anyone interested in the subject and in contributing to the work of the group is encouraged to join. For further information or to join the group, contact Barry Leiner <BLEiner@cnri.reston.va.us>.

The next meeting of the Working Group will be held the morning of 24 June 1998, just prior to the [DL'98 conference](#). The meeting will be open but those planning on attending are asked to contact [Jeanette Bartolomeo](#) to assure there will be sufficient space.

[Meeting Information](#)

The Working Group is also sponsoring a [Workshop on Digital Library Metrics](#), to be held 27 June 1998, just after the [DL'98 conference](#). Interested parties should contact the workshop organizers: [Bill Pottenger](#) and [Bob McGrath](#).

prepared by [Barry Leiner](#)

last modified 5/26/98

Interfaces:

[Stanford DL user interface projects](#)

Xerox Interfaces for Information Access

- [Home Page](#)
- [Scientific American article](#)
- [Cat-a-Cone figures](#)
- [Scatter/Gather examples](#)
- Questions:
 - Compare
 - What are the various interfaces built? How do they compare? What is the best use of each?
 - Scatter/gather
 - Explain clustering, relate it to scatter/gather.
 - What are special problems with large category systems and how can they be solved?

[Envision](#) : ENVISION project at Virginia Tech ...

[Berkeley:](#) TileBars, Multivalent documents

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Cat-a-Cone: An Interactive Interface for Specifying Searches and Viewing Retrieval Results using a Large Category Hierarchy

Marti A. Hearst

Xerox PARC

3333 Coyote Hill Rd

Palo Alto, CA 94304

hears@parc.xerox.com

Chandu Karadi

School of Medicine, M121

Stanford University

Stanford, CA 94305

karadi@leland.stanford.edu

This paper appears in the proceedings of 20th Annual International ACM/SIGIR Conference, Philadelphia, PA, July 1997.

An unpublished [Appendix](#) to this paper contains additional figures which were omitted from the proceedings due to space limitations.

Abstract:

This paper introduces a novel user interface that integrates search and browsing of very large category hierarchies with their associated text collections. A key component is the separate but simultaneous display of the representations of the categories and the retrieved documents. Another key component is the display of *multiple* selected categories simultaneously, complete with their hierarchical context. The prototype implementation uses animation and a three-dimensional graphical workspace to accommodate the category hierarchy and to store intermediate search results. Query specification in this 3D environment is accomplished via a novel method for painting Boolean queries over a combination of category labels and free text. Examples are shown on a collection of medical text.

In the *Proceedings of the Twentieth Annual International ACM SIGIR Conference*, Philadelphia, PA, July 1997, to appear.

© Copyright 1997 by ACM, Inc.

INTRODUCTION

There exist today many large online text collections to which category labels have been assigned. MEDLINE, a huge collection of biomedical articles, has associated with it Medical Subject Headings (MeSH) consisting of approximately 16,000 categories [[Lowe & Barnett1994](#)]. The Association for Computing Machinery (ACM) has developed a hierarchy of approximately 1200 category (keyword) labels (<http://www.acm.org/class/1991/cr91.html>). Yahoo!, one of the most popular search sites on the World Wide Web, organizes web pages into a hierarchy consisting of thousands of category labels (<http://www.yahoo.com>). And traditional online bibliographic systems have for decades assigned subject

headings to books and other documents [[Svenonius1986](#)].

The meanings of the category labels differ somewhat among collections, but usually they are intended to help organize the documents and to aid in query specification. Unfortunately, as reported in a recent paper from the library sciences community, users of online bibliographic catalogs rarely use the available subject headings [[Drabenstott & Weller1996](#)]. These and other authors put much of the blame on poor (command line-based) user interfaces which provide little aid for selecting subject labels and force users to scroll through long alphabetic lists.

Although many researchers have investigated techniques for automatically augmenting word-based queries with category labels, there has been surprising little research on advanced user interfaces for browsing in and selecting from large category hierarchies for the purposes of information access. There has been still less on how to integrate category hierarchies with retrieval results, and work is especially lacking in the support of search when *multiple* categories have been assigned to each document.

This paper describes an interactive user interface that integrates search and browsing of very large category hierarchies with their associated text collections. The prototype system, called the Cat-a-Cone, uses existing 3D+animation interface components, applied in a novel way, to support browsing and search of text collections and their category hierarchies.

(Cat-a-Cone integrates *category* hierarchies into *Cone*Trees. The name is almost a homophone of *catacomb*, a word whose secondary definition is *a complex set of interrelated things*.)

A key component of the interface is the separation of the graphical representation of the category hierarchy from the graphical representation of the documents. This separation allows for a fluid, flexible interaction between browsing and search, and between categories and documents. It also provides a mechanism by which a *set* of categories associated with a document can be viewed along with their hierarchical context.

Another key component of the design is assignment of first-class status to the representation of text content. The retrieved documents are stored in a 3D+animation book representation [[Card et al.1996](#)] that allows for compact display of moderate numbers of documents. Associated with each retrieved document is a page of links to the category hierarchy and a page of text showing the document contents. The user can "ruffle" the pages of the book of retrieval results and see corresponding changes in the category hierarchy, which is also represented in 3D+animation. All and only those parts of the category space that reflect the semantics of the retrieved document are shown with that document.

In summary, this interface is designed to exhibit the following features (see Figure 1):

- Make large hierarchical category sets easier to view and understand:
 - Allow for viewing of entire hierarchy in one window (occluded portions unveiled through animation)
 - Aid in the understanding of the meanings of unfamiliar terms
 - Help disambiguate terms that occur in several places
- Couple viewing of categories with search and display of retrieval results:
 - Separate category labels from documents, but link both together to allow for a flexible two-way interaction
 - Assign first-class status to the display of document content, rather than reducing a document to a small glyph or title alone

- Introduce a new, simple method for specifying complex Boolean queries consisting of free text plus category subtrees
- Be compatible with recent advances in ranking algorithms and with some other GUIs for information access.

Our interaction model is similar to that described in [Agosti *et al.*1992]. These authors define a two-level architecture for linking documents and their "auxiliary data". However, the implementation and that used in a followup study [Belkin *et al.*1993] use a text-based interface which does not provide most of the affordances listed above. [Ingwersen & Wormell1986] also describes a text-based interface that allows the user to alternate between free text and thesaurus terms.

Before describing the interface in more detail, Section 2 discusses our view of the role of categories in information access and related work. Then Section 3 describes the Cat-a-Cone and examples of its use, Section 4 discusses other graphical information access interfaces, and Section 5 summarizes the paper.

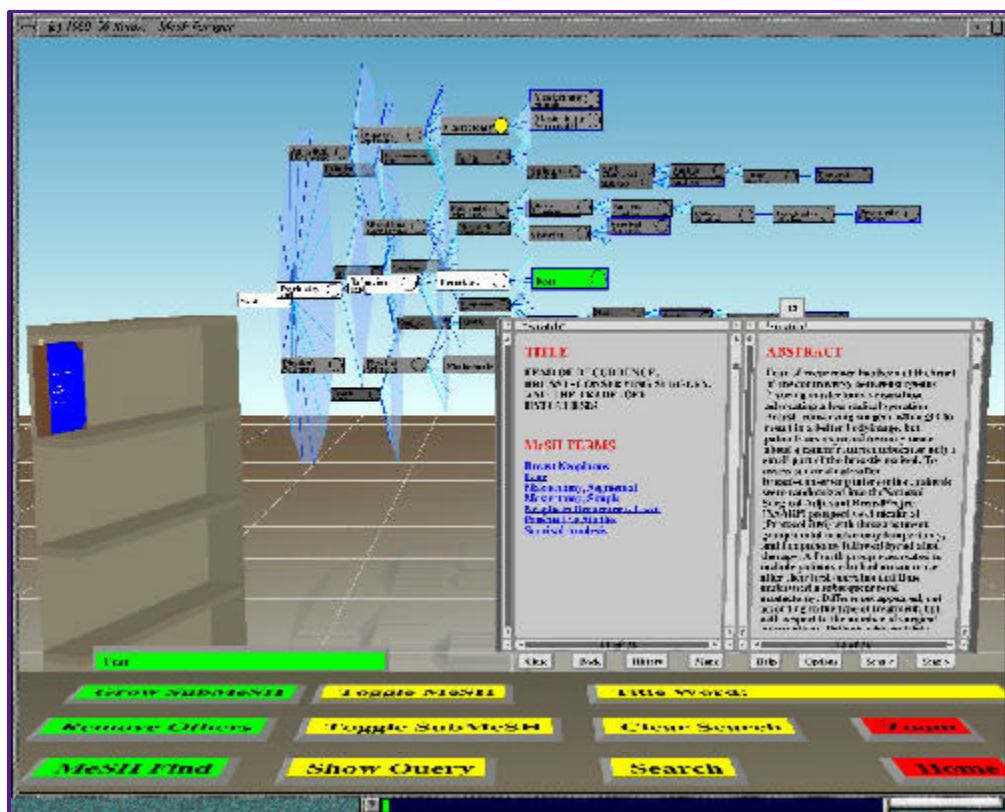


Figure 1: The Cat-a-Cone interface. Shown are the results of a search on category labels *Mastectomy* and *Radiation Therapy* in conjunction with the text word "lumpectomy" on a breast cancer subset of the MEDLINE collection. A ConeTree displays category labels and a WebBook shows retrieval results. The lefthand page shows the title and the category labels associated with the document. The righthand page shows the abstract associated with the document. Books that are the results of previous searches are stored in the workspace on the bookshelf, thus acting as a memory aid.

All and only those categories (and their ancestors) present in the current document are displayed in the category hierarchy. When the user clicks on a category label on the lefthand page, the corresponding category label in the ConeTree is rotated to the foreground and the labels of its ancestors are highlighted. When the user "turns" the pages of the book, the subtrees of the category hierarchy rotate, expand, and contract appropriately (categories that were present on the previous page and are present on the new page remain constant, categories that were on the previous page but are not on the new page are pruned away, and new categories that were not on the previous page but are on the new page are expanded out).

CATEGORIES AND INFORMATION ACCESS

In this section we discuss what is meant by category labels in this work, in particular contrasting categories with other kinds of meta-information and with thesaurus terms. We also motivate our design decisions by first describing results of recent experiments in improving document ranking by automatically combining category labels with free text, and then discussing the special considerations that accompany the search and display of documents to which multiple categories have been assigned.

Categories vs. External Meta-Information

Most documents have some kinds of meta-information associated with them - that is, information that characterizes the external properties of the document, that help identify it and the circumstances surrounding its creation and use. These attributes include author(s), date of publication, length of document, publisher, and document genre. Several research groups are exploring issues associated with user interfaces for exploiting this kind of information, usually employing variations on a table format, e.g., [[Fox et al.1993](#), [Baldonado & Winograd1997](#)].

By contrast, this paper focuses on category sets that have been assigned in an attempt to characterize the content or meanings found within the text of documents. Content-oriented category sets are difficult to depict graphically because they are abstract and there are potentially more meaningful combinations of content categories than external meta-information.

Thesaurus Term Associations

Many researchers have investigated automatic thesaurus creation, most often using word co-occurrences, e.g., [[Crouch1990](#), [Ruge1991](#), [Evans et al.1991](#), [Grefenstette1994](#)]. Thesaurus terms are related to category labels in information access systems, in that both kinds of information are used to improve recall. However, there are several important differences between the two.

Category labels are used to classify documents' contents according to general subject areas and other semantic attributes. Categories are instantiated by the set of documents they are assigned to, and are represented by their labels. These labels are used as a kind of meta-information, and typically category labels are matched against other labels when used in search.

By contrast, thesaurus terms are usually used as alternative ways to express a concept. Thesaurus terms can compensate for the occasions when the actual words used by the user in the query do not match the way the concepts are expressed in the document. When a user adds terms to the query from a thesaurus (or from relevance feedback for that matter), this usually means adding the actual words from the thesaurus into the query in hopes that they will appropriately match words in the documents of interest. Thesaurus terms are typically represented by the words themselves,

This distinction does not hold in all cases. Techniques such as Latent Semantic Indexing [[Deerwester et al.1990](#), [Schütze1993](#)] map the meaning of one set of words to another by means of a vector representation. Thus it attempts to find general themes which could be thought of as categories although the representation is nonstandard. However, thinking of the terms as thesaurus terms rather than category labels tends to influence how they are used in the user interface.

Combining Category Labels and Free Text Search

Library catalog systems have long provided categorization information in the form of subject headings. Researchers have reported that these kinds of headings often mismatch user expectations [[Svenonius1986](#), [Lancaster1986](#)]. However, there is also evidence that when such subject heading information is combined with free text search, results are improved over using either categories or free text alone, although usually these improvements are small [[Markey et al.1982](#), [Henzler1978](#), [Lancaster1986](#)]. Most of this work was done in the bibliographic context and did not employ modern user interface technology.

Studies on the biomedical category set, MeSH, usually do not achieve strong improvements on MEDLINE searches with automatic addition of category labels over non-MeSH searches [[Srinivasan1996a](#)]. Only very small improvements were found in two studies [[Hersh et al.1994](#), [Aronson et al.1994](#)] on a larger collection (using the Metathesaurus, a much larger hierarchy than MeSH [[Schuyler et al.1993](#)]). Larger improvements were found in another study [[Yang & Chute1994](#)], but under the assumption that a new query will match against a query that had already been seen and analyzed extensively with training examples.

However, recent careful studies [[Srinivasan1996b](#), [Srinivasan1996c](#)] have shown that if done appropriately, adding in MeSH terms can lead to significant improvements in precision and recall simultaneously, even over an initial high baseline. Importantly, these improvements were not found by using the standard technique of automatically mapping natural language queries into semantically equivalent MeSH terms, as is done with most attempts to automatically improve ranking with category labels.

Rather, improvements were found by retrieving documents based on the free-text version of the query, taking the top-ranked documents for relevance feedback, and adding to the query the MeSH category labels that appeared in the top-ranked documents. This suggests that first finding some good example documents, and then examining the category labels assigned to them, and using these to revise the query is a more effective way to improve rankings via category labels. This is precisely the kind of interaction that our framework is designed to support.

The Case for Multiple Category Representation

Many graphical user interfaces for text collections place documents within one point in semantic space,

usually based on some measure of inter-document similarity (see Section 4). We assert there is a problem with assuming that documents can be placed into a single category or a single point in semantic space. Although real-life objects can (arguably) be assigned one place in a taxonomy (a truck is a kind of vehicle), the content of text is usually not so simply classified.

Consider for example a biomedical journal article entitled *Fear of Recurrence, Breast-Conserving Surgery and the Trade-Off Hypothesis*. This article has been manually assigned the MeSH category labels: (Items in parentheses represent subheading modifiers for the main category labels. Our subset of MEDLINE documents have on average eight category labels assigned to them.)

Breast Neoplasms
Fear
Mastectomy (Segmental)
Mastectomy (Simple)
Neoplasm Recurrence (Local)
Prospective Studies
Survival Analysis

The article discusses a statistical study of the effects of a patient's fear of recurrence of breast cancer after a partial mastectomy versus the improved self image the patient has from retaining part of the breast. Thus, at a high level the human categorizers have placed the article into the semantic space at the intersection of Surgery, Statistics, and Psychology, since all three areas help characterize the complex subject matter of the document. At a more detailed level, the document content rests along the axes of a contrast between particular surgical procedures, a particular kind of statistical study and analysis, and the measurement of a particular psychological attribute.

When documents are clustered or grouped according to overall similarity, the distinctions about which axes they are similar on are not visible to the user. Placing the example document within a cluster of others might be done because any subset of the topic areas discussed were held in common. A cluster-based representation can be useful for some purposes, such as getting an overview of a collection's contents, but a user interface that shows the inter-relations among documents according to their category labels (or in general, according to orthogonal semantic attributes) allows the user an alternative view of document similarity.

Existing Category-based Interfaces

Most interfaces that depict category hierarchies graphically do so by associating a document directly with the node of the category hierarchy to which it has been assigned. For example, clicking on a category link in Yahoo! brings up a list of documents that have been assigned that category label. Conceptually, the document is stored within the category label. When navigating the results of a search in Yahoo!, the user must look through a list of category labels and guess which one is most likely to contain references to the topic of interest. A wrong path requires backing up and trying again, and remembering which pages contain which information. If the desired information is deep in the hierarchy, or not available at all, this can be a time-consuming and frustrating process.

The MeSHBROWSE system [Korn & Shneiderman1995] allows users to interactively browse a subset of semantically associated links in the MeSH hierarchy. From a given starting point, clicking on a category causes the associated categories to be displayed in a 2D tree representation. The interface has the space limitations inherent in a 2D hierarchy display and does not provide mechanisms for search

over an underlying document collection.

Internet Grateful Med (<http://igm.nlm.nih.gov:80/>) is a World Wide Web-based service that allows an integration of search with display and selection of MeSH category labels. The site is designed to support many simultaneous users and is constrained by the limitations of HTML. After the user types in the name of a potential category label, a long list of choices is shown in a page. To see more information about a given label, the user selects a link (e.g., Radiation Injuries). The current context is lost, and a new web page appears showing the ancestors of the term and its immediate descendants. If the user attempts to see the siblings of the parent term (Wounds and Injuries) then the context is changed again. Radiation Injuries appears as one of many siblings and the context of its children is lost. To go back to the previous list of choices, this page is obliterated.

One recent interface focuses on displaying faceted category sets [Allen1995]. In this work, hierarchical category labels corresponding to the Dewey Decimal system are shown indented in a scrollable window, in a focus-plus-context manner similar to that used in the Superbook table of contents [Egan *et al.*1989]. All documents that have been assigned a selected category are listed in another scrollable window. When the user issues a search, all categories that have a title with a hit in them are displayed in a scrollable window, with a number showing how many hits fall into each category. Thus this interface does not support display of multiple labels per document.

Our earlier work [Hearst1994] emphasized many of the same points as in the present work. However, the interface for choosing and displaying categories could not support hierarchies of categories and did not provide a good mechanism for viewing the retrieved documents. The Cat-a-Cone interface represents a dramatic improvement.

THE CAT-A-CONE

We assert that separating the documents from the category hierarchy can open the door to more powerful search and display strategies, especially for collections in which documents have multiple category assignments.

Our approach consists of three main components. The first is a better representation of the category space, the second is a separate, compact representation of retrieved documents, and the third is a model of interaction that makes use of these visualizations in novel ways. The instantiation of these ideas is in a prototype interface that makes use of particular visualization technologies, but does not exclude using alternative technologies that support the same or similar functionalities.

Our prototype implementation (written in Common LISP, running on Silicon Graphics IRIX machines), makes use of the ConeTree 3D+animation visualization from the Xerox PARC Information Visualizer (IV) [Robertson *et al.*1993]. (Processing speeds and monitor quality are increasing rapidly enough to make general availability realistic within the next few years.) The ConeTree allows for the display of a very large category hierarchy all in one window. Those categories that are farther away and less legible can be rotated to the foreground with a simple click on the leftmost (highest ancestor) category label. The interface allows users to choose from menus, or gesture-based "gardening" commands for growing and pruning subportions of the hierarchy, as well as keyboard accelerators and traditional buttons on the "desktop" portion of the workspace for specifying searches.

If a label is terminated with a triangle, this signals the existence of a subtree. The user can expand the category one level down by a right-drag gesture, or can select the node with the left mouse button and then click on the "Grow SubMeSH" button on the workspace to expand the entire subtree (this button's label and functionality toggles depending on the state of the selected label).

The examples in the following subsections are drawn from a subcollection of MEDLINE abstracts on breast cancer. This subcollection was further narrowed to include only those 403 documents that contained the word *lumpectomy* in their title or abstract (the equivalent MeSH term is *Mastectomy, Segmental*).

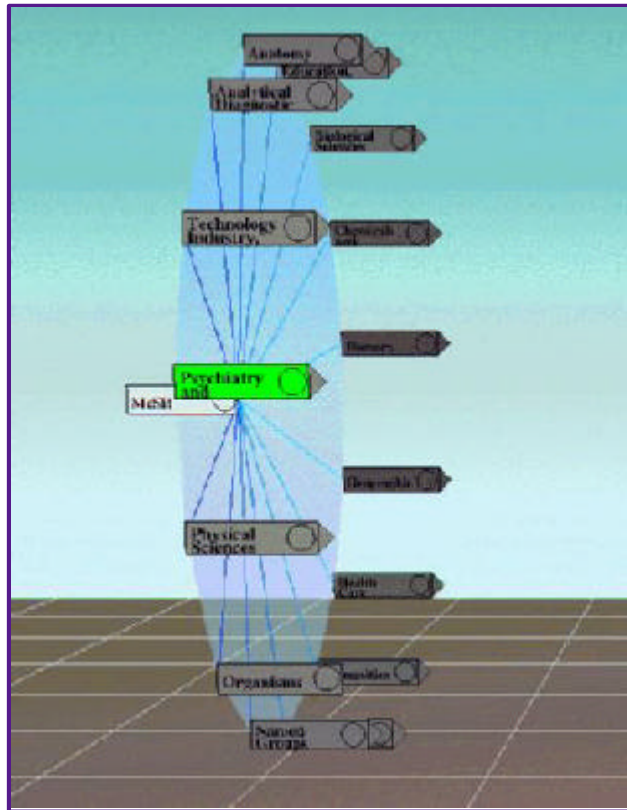


Figure 2: The top-level representation of the MeSH hierarchy as a search starting point.

Starting a Search and Discovering Categories

Search interfaces must provide users with good ways to get started. [Shneiderman1996] advocates a model of: first overview, then zoom and filter, then details, and repeat. In terms of the overview, an empty screen does not provide a good starting point. On the other hand, the contents of the entire hierarchy can be overwhelming, even if it can be made to fit in one window. The model of interaction used here allows for several different starting points and interaction sequences.

One starting point shows all of the top level categories initially, and allows the user to control the subsequent expansions (see Figure 2). The user can select one label to expand in detail, revealing all of its descendants rather than having to navigate laboriously through many pages to see all of the subtrees. If this is still too overwhelming, the user can select a few nodes and issue the command "prune others"

to view a smaller space. Or the system can be programmed to automatically expand subtrees to a depth that has been determined by user studies to be most comprehensible.

An alternative starting point is to have the user type in a category label and see which parts of the hierarchy match or partially match that label. The user can then expand and explore other nearby regions.

As a final Cat-a-Cone starting point type, the user can type in free text, causing the system to retrieve documents containing those words in their titles or abstracts. The user can then view the category labels associated with the retrieved documents. This kind of interaction should be useful in analogy with the findings of the experiments of [Srinivasan1996b, Srinivasan1996c] discussed above, in which good MeSH categories were associated with documents highly ranked against the free-text query.

For example, after a search on the category *Mastectomy* the system retrieves an article that has a link to *Survival Analysis*, a category which the user might not have known in advance. The user can then decide to delve more deeply into this topic by issuing a search on this category label, another category label, and one free-text word, yielding the book of Figure 1. This book in turn can aid the user in discovering that the MeSH hierarchy has a set of category labels pertaining to psychological issues, including *Emotions*.

When indexing new MEDLINE citations, indexers are instructed to use the most specific MeSH term available [Lowe & Barnett1994]. Often these low-level categories are meaningless to a patient or non-specialist. The context-preserving display of ancestor and sibling information provided by this representation can help the user see the general meaning of a term. For example, the category label Doxorubicin is the name of a chemical, and this is indicated by its ancestor labels.

Viewing Retrieved Documents

The Cat-a-Cone also makes use of a modification of WebBooks from the Web Forager project [Card *et al.*1996], an extension of the Information Visualizer project. After a set of documents has been retrieved (in response to a query on free text and/or category labels), the documents are organized into a "book" of pages (see Figure 1). The lefthand page contains the document's title and the list of category labels associated with that document. The righthand page shows the abstract and/or content of the documents (the pages are scrollable).

The cover of the book shows the query responsible for producing the retrieval results. When the book is closed and is the selected focus, the ConeTree can show a representation of all of the categories that have occurrences within the pages of the book. Note that since multiple categories have been assigned to each document, the book will in general contain many more category labels than were present in the original query. Only those parts of the hierarchy that contain categories that are in the book (and their ancestors) are shown automatically; the rest of the hierarchy is pruned away. The user is able to expand or contract any part of the hierarchy at any time.

When the user opens the book, the ConeTree is automatically modified to show only those parts of the hierarchy whose categories appear in the document on the current page. The labels themselves are outlined in blue and the ancestors are shown without outline. This representation shows the space of concepts in which the document resides. When the user flips through the pages of the book, the representation of the tree adjusts accordingly. The use of animation helps retain the context of the category set. Often many category labels are shared among the book's retrieval results, so only a few offshoots of the hierarchy grow or are pruned as the user flips through the retrieved pages. The animation helps the user retain context, showing which parts of the category space differ from document

to document.

Since the user can store the book away and reopen it at any time (making use of the workspace capabilities of IV) there is less reason to worry about getting "lost" or forgetting what happened in a previous session, because the retrieval results can be stored away and reused.

As currently described, the interaction between the category hierarchy and the retrieved documents is not entirely two-way: clicking on category labels does not influence the behavior of the retrieval results. This choice was made in part because there is a many-to-one mapping of category labels to book pages. However, the model could be altered as follows: clicking on a label causes the page of the book to turn to the next article containing that category label, if such an article exists.

In order to help further explain the retrieval results, other information access interface ideas can be incorporated into this representation. For example, we plan to place a TileBars [Hearst1995] representation of the retrieval results into the first page of the book, which can be pulled out and viewed alongside the rest.

Query Specification

Research suggests that users often want to search on an "exploded" part of the MeSH hierarchy to improve recall [Lowe & Barnett1994]. In effect, they wish to specify a conjunction of disjunctions over category hierarchies, that is, require that at least one representative from each of a set of concepts be present in the retrieved documents. Our initial simple approach to query specification is to have the selection of the yellow circle on a category label indicate a disjunction of all of that category's descendants, inclusive. A conjunction is imposed over all selected subtrees. This approach is simple, but for queries such as (*hand OR foot*) AND *arthritis* it requires the user to select higher up in the anatomy subtree than desired.

For this reason, we have devised a novel scheme for specifying Boolean queries using the 3D environment. The user selects colors from a "palette" and "paints" subtrees with these colors, where each color representing members of a disjunction, and different colors indicate different components of a conjunction. Additionally, free text search terms are typed into entry lines of corresponding colors. The users are instructed to think of the query in terms of colors rather than as a Boolean expression. For example, a user can specify a conjunction of disjunctions while thinking something like "I want documents that contain at least one green, at least one yellow, and at least one blue category or word." The retrieval results do not need to employ a strict Boolean filter; a quorum ranking strategy [Salton1989] can be used instead. In this scheme, different colors receive different weight, and those documents represented by hits on more colors are ranked higher than those with fewer colors.

This query specification scheme thus stays within the object-centric paradigm and provides a simple way for users to link subsets of category labels and free text in complex Boolean expressions (with NOT represented by red).

OTHER GRAPHICAL APPROACHES

Graphical Concept Spaces

Several approaches map documents from their high dimensional representation in document space into a 2D representation in which each document is represented as a dot or other small glyph. The functions for transforming the data into the lower dimensional space differ, but the net effect is that documents are placed at one point in a scatter-plot-like representation of the space, and users are meant to detect themes or clusters in the arrangement of the glyphs. These systems include BEAD [Chalmers & Chitson1992], ThemeScapes [Wise *et al.*1995], and the Galaxy of News [Rennison1994]. Other systems, e.g. [Cutting *et al.*1992, Maarek & Wecker1994, Allen *et al.*1993] display inter-document similarity hierarchically. The systems of [Fowler *et al.*1991] and IR [Thompson & Croft1989] display retrieved documents in networks based on interdocument similarity.

Systems such as VIBE [Korfhage1991] and the InfoCrystal [Spoerri1993] ask the user specify the query in terms of k words (although category labels could be used instead) where k is a small number. They then display, for each subset of the k categories, the number of documents that contain that subset of words. These systems show the categories in a graphical concept space, and do not provide a mechanism for choosing which of a large number of words or categories to choose from, nor do they suggest new words or categories, nor associations among category labels, and they do not introduce methods for associating the text of the documents with new words or categories.

The Lyberworld system [Hemmje *et al.*1994] makes use of a ConeTree but uses it quite differently than the Cat-a-Cone. Lyberworld uses the ConeTree to display the navigation path generated by a sequence of search steps, placing the documents associated with the results of the search at the first level of the hierarchy, expanding on the words of a given document's node at the next level, and repeating.

Some researchers, e.g., [Pedersen1993, Carpineto & Romano1996], have employed a graphical depiction of a lattice for query formulation, where the query consists of a set of constraints on a hierarchy of categories (actually, semantic attributes in these systems). This is one solution to the problem of displaying documents in terms of multiple attributes; a document containing terms A, B, C, and D could be placed at a point in the lattice with these four categories as parents. However, if such a representation were to be applied to retrieval results instead of query formulation, the lattice layout would in most cases be too complex to allow for readability.

In the AIR/SCALIR interface [Rose & Belew1991] a connectionist network determines in advance a set of terms that characterize documents from a collection of bibliographic records. The term nodes are connected to the document nodes via edge links, so the user can see which documents are associated with each important term. If there are a large number of links between associated terms and documents, or if the links are not neatly organized, the relationships will be difficult to discern.

Finally, Kohonen's feature map algorithm has been used to create maps that graphically characterize the overall content of a document collection or subcollection [Lin *et al.*1991, Chen *et al.*1997]. The regions of the 2D map vary in size and shape corresponding to how frequently their corresponding themes occur in the collection. Regions are characterized by single words or phrases, and adjacency of regions is meant to reflect semantic relatedness of the themes within the collection. If a document is strongly associated with the region according to the training of the feature map, its title can be viewed via a pop-up window over that region; documents can be associated with more than one region.

A Comparison

A recent evaluation [[Chen et al.1997](#)] compared the Kohonen feature map representation on a browsing task to that of Yahoo!. The results found that some users expressed a desire for a visible hierarchical organization, others wanted an ability to zoom in on a subarea to get more detail, and some users disliked having to look through the entire map to find a theme, desiring an alphabetical ordering instead. Many found the single-term labels to be misleading, in part because they were ambiguous (one region called "BILL" was thought to correspond to a person's name rather than counting money). The subjects like the ease of being able to jump from one area to another without having to back up as is required in Yahoo! and liked the fact that the maps have varying levels of granularity. (The authors concluded that this interface is more appropriate for casual browsing than for search.)

These results all support the design decisions made in the Cat-a-Cone. Hierarchical representation of term meanings is supported, so users can choose which level of description is meaningful to them. Furthermore, they can view different levels of description simultaneously, so more familiar concepts can be viewed in more detail, and less familiar at a more general level. An alphabetical ordering of the categories coupled with a regular expression search mechanism allows for straightforward location of category labels. Retrieved documents are represented as first-class objects, so full text is visible, but in a compact form. Category labels are disambiguated by their ancestor/descendant/sibling representation. Users can jump easily from one category to another and can in addition query on multiple categories simultaneously (something that is not a natural feature of the maps). The Cat-a-Cone has several additional advantages as well, such as allowing a document to be placed at the intersection of several categories, and explicitly linking document contents with the category representation.

CONCLUSIONS

The Cat-a-Cone interface allows for fluid two-way interaction between selection of category labels for search and display of multiple category labels within retrieval results. The ConeTree provides easy selection of subparts of the category hierarchy, to help users understand unfamiliar terms and usages of ambiguous terms by seeing their contexts. The book-based representation of the retrieval results with the corresponding display of multiple subparts of the category hierarchy helps show which categories are a part of the results, what these categories mean, and which new categories might be useful to search on.

Although this is a simple combination of simple ideas, it seems to produce a powerful, intuitive and original way for users to use large category hierarchies to aid them in query specification and understanding of retrieval results. In future work we plan to evaluate the interface using as subjects breast cancer patients and clinicians.

This kind of interaction should also be useful for other tasks, for example, helping authors of articles for ACM publications *choose* which category labels to assign to their documents. The user can tell the system to find documents whose textual contents are similar to their new document's, and then examine the resulting category hierarchy.

The ConeTree as currently implemented does not satisfy fully the needs of the design. For example, the system should reformat subtrees when space becomes available as a consequence of pruning away other subtrees. Additionally, a recent study shows that fisheye, variable zoom algorithms work better than full zoom algorithms for navigating networks [[Schaffer et al.1996](#)], and an improved facility for

focus-plus-context of this kind could be used to help direct user attention. Subtrees should be easily "pulled off" and stored for later use. Finally, the display should reflect the frequency of the categories for a given retrieval result with a visual analogue such as "Read Wear" [Hill *et al.* 1992] or using a Kohonen feature map [Lin *et al.* 1991] as an overview of the highest-level categories.

We view information retrieval as a complex task that requires many different tools. Support for category browsing and search is one important capability, but by no means solves the entire problem. This interface must be combined with others to create an effective information seeking system.

Acknowledgements

Karadi was supported in part by a Medical Scholars Grant from Stanford Medical School. We would like to thank Larry Fagan, Bill York, and Stu Card for their help with and encouragement of this work. Additional figures can be found at <http://www.parc.xerox.com/ia>.

References

Agosti *et al.* 1992

AGOSTI, M., G. GRADENIGO, & P.G. MARCHETTI. 1992. A hypertext environment for interacting with large textual databases. *Information Processing & Management* 28.371-387.

Allen 1995

ALLEN, ROBERT B. 1995. Two digital library interfaces that exploit hierarchical structure. In *Proceedings of DAGS95: Electronic Publishing and the Information Superhighway*, Boston, MA.

Allen *et al.* 1993

--1993. An interface for navigating clustered document sets returned by queries. In *Proceedings of ACM COOCS: Conference on Organizational Computing Systems*, Milpitas, CA.

Aronson *et al.* 1994

ARONSON, A., T. RINDGLESCH, & A. BROWNE. 1994. Exploiting a large thesaurus for information retrieval. In *Proceedings of RIAO '94: Intelligent Multimedia Information Retrieval Systems and Management*, 197-216.

Baldonado & Winograd 1997

BALDONADO, MICHELLE Q. WANG, & TERRY WINOGRAD. 1997. Sensemaker: An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. To appear.

Belkin *et al.* 1993

BELKIN, N., P. G. MARCHETTI, & C. COOL. 1993. Braque - design of an interface to support user interaction in information retrieval. *Information Processing and Management* 29.325-344.

Card *et al.* 1996

CARD, STUART K., GEORGE G. ROBERTSON, & WILLIAM YORK. 1996. The webbook and the web forager: An information workspace for the world-wide web. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, Canada.

Carpineto & Romano1996

CARPINETO, CLAUDIO, & GIOVANNI ROMANO. 1996. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies* 45.553-578.

Chalmers & Chitson1992

CHALMERS, MATTHEW, & PAUL CHITSON. 1992. Bead: Exploration in information visualization. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 330-337, Copenhagen, Denmark.

Chen et al.1997

CHEN, HSINCHEN, ANDREA L. HOUSTON, ROBIN R. SEWELL, & BRUCE R. SCHATZ. 1997. Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Sciences (JASIS)* . To appear.

Crouch1990

CROUCH, C. J. 1990. An approach to the automatic construction of global thesauri. *Information Processing and Management* 26.629-640.

Cutting et al.1992

CUTTING, DOUGLASS R., JAN O. PEDERSEN, DAVID KARGER, & JOHN W. TUKEY. 1992. Scatter/Gather: A cluster-based approach to browsing large document collections. In *Proceedings of the 15th Annual International ACM/SIGIR Conference*, 318-329, Copenhagen, Denmark.

Deerwester et al.1990

DEERWESTER, SCOTT, SUSAN T. DUMAIS, GEORGE W. FURNAS, THOMAS K. LANDAUER, & RICHARD HARSHMAN. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41.391-407.

Drabenstott & Weller1996

DRABENSTOTT, KAREN M., & MARJORIE S. WELLER. 1996. The exact-display approach for online catalog subject searching. *Information Processing and Management* 32.719-745.

Egan et al.1989

EGAN, DENNIS E., JOEL R. REMDE, LOUIS M. GOMEZ, THOMAS K. LANDAUER, JENNIFER EBERHARDT, & CAROL C. LOCHBAUM. 1989. Formative design evaluation of SuperBook. *Transaction on Information Systems* 7.

Evans et al.1991

EVANS, DAVID A., KIMBERLY GINTHER-WEBSTER, MARY HART, ROBERT G. LEFFERTS, & IRA A. MONARCH. 1991. Automatic indexing using selective NLP and first-order thesauri. In *Proceedings of the RIAO*, volume 2, 624-643.

Fowler et al.1991

FOWLER, RICHARD H., WENDY A. L. FOWLER, & BRADLEY A. WILSON. 1991. Integrating query, thesaurus, and documents through a common visual representation. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 142-151, Chicago.

Fox et al.1993

FOX, EDWARD A., DEBORAH HIX, LUCY T. NOWELL, DENNIS J. BRUENI, WILLIAM C. WAKE, LENWOD S.

HEATH, & DURGESH RAO. 1993. Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science* 44.480-491.

Grefenstette1994

GREFENSTETTE, GREGORY. 1994. *Explorations in automatic thesaurus discovery*. Kluwer international series in engineering and computer science. Kluwer Academic Publishers.

Hearst1994

HEARST, MARTI A. 1994. Using categories to provide context for full-text retrieval results. In *Proceedings of RIAO '94; Intelligent Multimedia Information Retrieval Systems and Management*, 115-130.

Hearst1995

--1995. Tilebars: Visualization of term distribution information in full text information access. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, Denver, CO.

Hemmje et al.1994

HEMMJE, MATTHIAS, CLEMENS KUNKEL, & ALEXANDER WILLETT. 1994. LyberWorld - a visualization user interface supporting fulltext retrieval. In *Proceedings of the 17th Annual International ACM/SIGIR Conference*, 249-259, Dublin, Ireland.

Henzler1978

HENZLER, ROLF G. 1978. Free or controlled vocabularies: Some statistical user-oriented evaluations of biomedical information systems. *International Classification* 5.21-26.

Hersh et al.1994

HERSH, WILLIAM R., DAVID H. HICKMAN, BRIAN HAYNES, & K. ANN MCKIBBON. 1994. A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association* 1.51-60.

Hill et al.1992

HILL, WILLIAM C., JAMES D. HOLLAN, DAVE WROBLEWSKI, & TIM MCCANDLESS. 1992. Edit wear and read wear. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 3-9.

Ingwersen & Wormell1986

INGWERSEN, PETER, & IRENE WORMELL. 1986. Improved subject access, browsing, and scanning mechanisms in modern online IR. In *Proceedings of the 9th Annual International ACM/SIGIR Conference*, 68-76, Pisa, Italy.

Korfhage1991

KORFHAGE, ROBERT R. 1991. To see or not to see - is that the query? In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 134-141, Chicago.

Korn & Shneiderman1995

KORN, FLIP, & BEN SHNEIDERMAN. 1995. Navigating terminology hierarchies to access a digital library of medical images. Technical Report HCIL-TR-94-03, University of Maryland.

Lancaster1986

LANCASTER, F. 1986. *Vocabulary control for information retrieval, second edition*. Arlington, VA: Information Resources.

Lin et al.1991

LIN, XIA, DAGOBERT SOERGEL, & GARY MARCHIONINI. 1991. A self-organizing semantic map for information retrieval. In *Proceedings of the 14th Annual International ACM/SIGIR Conference*, 262-269, Chicago.

Lowe & Barnett1994

LOWE, HENRY J., & G. OCTO BARNETT. 1994. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. *Journal of the American Medical Association (JAMA)* 271.1103-1108.

Maarek & Wecker1994

MAAREK, Y. S., & A.J. WECKER. 1994. The librarian's assistant: Automatically assembling books into dynamic bookshelves. In *Proceedings of RIAO '94; Intelligent Multimedia Information Retrieval Systems and Management*.

Markey et al.1982

MARKEY, KAREN, PAULINE ATHERTON, & CLAUDIA NEWTON. 1982. An analysis of controlled vocabulary and free text search statements in online searches. *Online Review* 4.225-236.

Pedersen1993

PEDERSEN, GERT SCHMELTZ. 1993. A browser for bibliographic information retrieval, based on an application of lattice theory. In *Proceedings of the 16th Annual International ACM/SIGIR Conference*, 270-279, Pittsburgh, PA.

Rennison1994

RENNISON, EARL. 1994. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *Proceedings of UIST 94, ACM Symposium on User Interface Software and Technology*, 3-12, New York.

Robertson et al.1993

ROBERTSON, GEORGE C., STUART K. CARD, & JOCK D. MACKINLAY. 1993. Information visualization using 3D interactive animation. *Communications of the ACM* 36.56-71.

Rose & Belew1991

ROSE, DANIEL E., & RICHARD K. BELEW. 1991. Toward a direct-manipulation interface for conceptual information retrieval systems. In *Interfaces for information retrieval and online systems*, ed. by Martin Dillon, 39-54. New York, NY: Greenwood Press.

Ruge1991

RUGE, GERDA. 1991. Experiments on linguistically based term associations. In *Proceedings of the RIAO*, 528-545.

Salton1989

SALTON, GERARD. 1989. *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Reading, MA: Addison-Wesley.

Schaffer et al.1996

SCHAFER, DOUG, ZHENGPING ZUO, SAUL GREENBERG, LYN BARTRAM, JOHN DILL, SHELLI DUBS, & MARK ROSEMAN. 1996. Navigating hierarchically clustered networks through fisheye and full-zoom methods. *ACM Transactions on Computer-Human Interaction* 3.162-188.

Schütze1993

SCHUTZE, HINRICH. 1993. Word space. In *Advances in neural information processing systems* 5, ed. by Stephen J. Hanson, Jack D. Cowan, & C. Lee Giles. San Mateo CA: Morgan Kaufmann.

Schuyler et al.1993

SCHUYLER, P. Y., W. T. HOLE, M. S. TUTTLE, & D. D. SHERERTZ. 1993. The UMLS metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association* 81.217-222.

Shneiderman1996

SHNEIDERMAN, BEN. 1996. The eyes have it: A task by data type taxonomy. In *Proceedings of Visual Languages 96*, Boulder, CO.

Spoerri1993

SPOERRI, ANSELM. 1993. InfoCrystal: A visual tool for information retrieval & management. In *Proceedings of Information Knowledge and Management '93*, Washington, D.C.

Srinivasan1996a

SRINIVASAN, PADMINI. 1996a. Optimal document-indexing vocabulary for medline. *Information Processing and Management* 32.503-514.

Srinivasan1996b

-- 1996b. Query expansion and medline. *Information Processing and Management* 32.431-443.

Srinivasan1996c

--1996c. Retrieval feedback in medline. *Journal of the American Medical Informatics Association (JAMA)* 3.157-167.

Svenonius1986

SVENONIUS, ELAINE. 1986. Unanswered questions in the design of controlled vocabularies. *Journal of the American Society for Information Science* 37.331-340.

Thompson & Croft1989

THOMPSON, R. H., & B. W. CROFT. 1989. Support for browsing in an intelligent text retrieval system. *International Journal of Man [sic] -Machine Studies* 30.639-668.

Wise et al.1995

WISE, JAMES A., JAMES J. THOMAS, KELLY PENNOCK, DAVID LANTRIP, MARC POTTIER, & ANNE SCHUR. 1995. Visualizing the non-visual: Spatial analysis and interaction with information from text documents. In *Proceedings of the Information Visualization Symposium 95*, 51-58. IEEE Computer Society Press.

Yang & Chute1994

YANG, YIMING, & CHRISTOPHER G. CHUTE. 1994. Expert network: Effective and efficient learning from human decisions in text categorization and retrieval. In *Proceedings of the 17th Annual International ACM/SIGIR Conference*, 13-22, Dublin, Ireland.

Metadata:

- [IMS Metadata spec.](#) and [tool](#)
- [Metadata: the Foundations of Resource Description](#)
- [OCLC/NCSA Metadata Workshop Report](#)
- [RFC-1807](#)
- [TEI](#)
- [BASIS article](#)
- [D-Lib Working Group on Metadata](#)
- [STARTS](#)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998. Edward A. Fox, Rajat Gupta




The Instructional Management Systems Project, an Educom NLII initiative, is developing a specification and software for managing online learning resources. Learning resources can include people, educational service companies, content, tools, and activities.

The pages listed below will introduce the concept of meta-data and the specification for its use. Visitors to this site can explore the potential of meta-data by experimenting with the [Meta-data Tool](#). To join the dialogue regarding the development of the IMS Meta-data Specification, please subscribe to the [IMS Meta-data Listserv](#).



 **EXECUTIVE SUMMARY** *A brief description of the information and the organization of this site*

 **INTRODUCTION** *An overview of meta-data: its purpose and development*

- What is Meta-data?
- How is Meta-data Organized?
- How is Meta-data Used?
- Evolution of Meta-data

 **USING META-DATA** *A description of using IMS meta-data for learning modules*

- Searching
- Creating

 **MANAGING IMS META-DATA** *A description of how to manage and enhance meta-data*


 **IMS META-DATA SPECIFICATION** *Technical documents defining meta-data fields and values and their application to different learning resources*

- IMS Meta-data Dictionary
- IMS Meta-data Sets

 **CREDITS AND REFERENCES**

 **META-DATA PRESS RELEASE AND QUOTES**

- Sept. 8, 1997
- March 23, 1998

 **FORUM** *A space for dialogue about IMS meta-data fields and values. This is a section of the larger IMS Public Forum.*

 **IMS Meta-data Tool** *Access the IMS Content Server and Meta-data Tool for experimenting with IMS meta-data. This is a new*

Metadata: The Foundations of Resource Description

Stuart Weibel

Office of Research, OCLC Online Computer Library Center, Inc.

weibel@oclc.org

D-Lib Magazine, July 1995

This paper is an abbreviated version of the [Summary Report of the OCLC/NCSA Metadata Workshop](#). It sets forth a proposal for the content of a simple resource description record (the Dublin Core Metadata Element Set) and outlines a series of further steps to advance the standards for the description of networked information resources.

- [Introduction](#)
 - [Underlying Assumptions](#)
 - [Implementations](#)
 - [Next Steps](#)
 - [References](#)
-

d-lib forum

d-lib magazine

Introduction

The explosive growth of interest in the Internet in recent years has created a digital extension of the academic research library for certain kinds of materials. Valuable collections of texts, images and sounds from many scholarly communities -- collections that may even be the subject of state-of-the-art discussions in these communities--now exist only in electronic form and may be accessible from the Internet. Knowledge regarding the whereabouts and status of this material is often passed on by word of mouth among members of a given community. For outsiders, however, much of this material is so difficult to locate that it is effectively unavailable.

Why is it so difficult to find items of interest on the Internet or the World Wide Web? A number of well-designed locator services, such as Lycos [\[http://lycos.cs.cmu.edu/\]](http://lycos.cs.cmu.edu/), are now available that automatically index many of the resources available on the Web and maintain up-to-date databases of locations. But indexes are most useful in small collections within a given domain. As the scope of their coverage expands, indexes succumb to problems of large retrieval sets and problems of cross disciplinary semantic drift. Richer records, created by content experts, are necessary to improve search

and retrieval. Formal standards such as the [TEI Header](#) and [MARC](#) cataloging) will provide the necessary richness, but such records are time consuming to create and maintain, and hence may be created for only the most important resources.

An alternative solution that promises to mediate these extremes involves the creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create such records, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover these descriptions and collect them.

Can a simple metadata record be defined that sufficiently describes a wide range of electronic objects? The Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) convened the invitational Metadata Workshop on March 1-3, 1995, in Dublin, Ohio to address this issue. Fifty-two librarians, archivists, humanities scholars and geographers, as well as standards makers in the Internet, Z39.50 and Standard Generalized Markup Language (SGML) communities, met to identify the scope of the problem, to achieve consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas, and to lay the groundwork for achieving further progress in the definition of metadata elements that describe electronic information.

Goals

Goals of the workshop included fostering a common understanding of the problems and potential solutions among the stakeholders and promoting a consensus on a core set of metadata elements to describe networked resources.

Scope

Since the Internet contains more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves. The major task of the Metadata Workshop was to identify and define a simple set of elements for describing networked electronic resources. To make this task manageable, it was limited in two ways. First, only those elements necessary for the discovery of the resource were considered. It was believed that resource discovery is the most pressing need that metadata can satisfy, and one that would have to be satisfied regardless of the subject matter or internal complexity of the object.

Secondly, the discussion was further restricted to the metadata elements required for the discovery of what were called **document-like objects**, or **DLOs** by the workshop participants. It was believed that DLOs are still the most common type of resource sought in the Internet and that whatever solution could be proposed for DLOs could be extended to other kinds of resources. More importantly, the likelihood of making progress on this challenging problem would be increased if attention could initially be restricted to something familiar.

DLOs were not rigorously defined, but were understood by example. For example, an electronic version of a newspaper article or a dictionary is a DLO, while an unannotated collection of slides is not. Of course, the crux of the problem is that in a networked environment, DLOs can be arbitrarily complex because they can consist of text with callouts to images, audio or video clips, or to other hypertext

documents. The Metadata Workshop participants made no attempt to limit the complexity of DLOs, except to say that the intellectual content of a DLO is primarily text, and that the metadata required for describing DLOs will bear a strong resemblance to the metadata that describes traditional printed texts.

As a result of the restricted focus of the workshop, certain issues required for a complete description of DLOs, such as cost, archival status and copyright information, were eliminated from the scope of the discussion. Elements required for the description of objects other than DLOs, such as the elements required for the description of complex geological strata in a geospatial resource, were also beyond the scope of the discussion. The goal was to define a core set of metadata elements that would allow authors and information providers to describe their work and to facilitate interoperability among resource discovery tools. But because the core elements do not yield a complete description of objects in a networked environment, careful consideration was also given to mechanisms for extending the element set.

The primary deliverable from the workshop was a set of thirteen metadata elements, named the **Dublin Core Metadata Element Set** (or Dublin Core, for short). The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.

Below is a brief description of the elements in the Dublin Core **Dublin Core Element Description**

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The physical manifestation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object

To make this discussion concrete, consider an electronic a record created with the relevant portions of the Dublin Core, and a sample syntax, that describes an electronic version of Maya Angelou's poem "On the Pulse of Morning". This description is based on a record created by the University of Virginia Library's Electronic Text Center. (For a description of that project, see Gaynor [\[Gaynor\]](#).)

- **Subject:** Poetry
- **Title:** On the Pulse of Morning
- **Author:** Maya Angelou
- **Publisher:** University of Virginia Library Electronic Text Center
- **OtherAgent:** Transcribed by the University of Virginia Electronic Text Center
- **Date:** 1993

- **Object:** Poem
- **Form:** 1 ASCII file
- **Identifier:** AngPuls1
- **Source:** Newspaper stories and oral performance of text at the presidential inauguration of Bill Clinton
- **Language:** English

Underlying Assumptions

The discussions at the Metadata Workshop revealed several principles that should guide the further development of the element set. Adherence to these principles increases the likelihood that the core element set will be kept as small as possible, that the meanings of the elements will be understood by most users, and that the element set will be flexible enough for the description of resources in a wide range of subject areas. These principles are intrinsicality, extensibility, syntax independence, optionality, repeatability, and modifiability.

Intrinsicality

The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. The focus on intrinsic data in no way demeans the importance of other varieties of data, but simply reflects the need to keep the scope of deliberations narrowly focussed.

Extensibility

In addition to its use in dealing with extrinsic data, extension mechanisms will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements.

Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields. In addition, the specification of the Dublin Core itself will change over time, and the extension mechanism will allow revisions while maintaining some backward compatibility with the originally defined element set.

Syntax Independence

Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs.

Optionality

All the elements are optional. The Dublin Core may eventually be applied to objects for which some elements have no meaning (who is the author of a satellite image?). It also seems counterproductive to mandate complex descriptions if the creators of the content are expected to provide the descriptive material. A simple description is better than no description at all.

Repeatability

All elements in the Dublin Core are repeatable. For example, multiple author elements would be used when a resource has multiple authors.

Modifiability

Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier. If no qualifier is present, the element has its common-sense meaning; otherwise, the definition of the element is modified by the value of the qualifier.

Qualifiers will be typically derived from well-known conventions in the library community or from the field of knowledge appropriate to the resource. Qualifiers are important because they give the Dublin Core a mechanism for bridging the gap between casual and sophisticated users. For example, the data in the **Subject** element consists of any word or phrase that describes the object's content. However, a professional cataloger may wish to supply the name of the authoritative source from which the subject terms are taken. In such a case, the element may be written as **Subject (scheme=LCSH)**, indicating that the subject terms are taken from the Library of Congress Subject Headings.

Implementations

One of the goals of the OCLC/NCSA Metadata Workshop was to promote prototype resource description projects based on a common model of resource description. A number of Metadata Workshop conferees represent organizations that have ongoing activities or are starting activities that will be influenced by the results of the workshop. These include:

- The OCLC Spectrum Project
Contact:Diane Vizine-Goetz, vizine@oclc.org
- [The OCLC Internet Resources Cataloging Project](#)
Contact:Erik Jul, jul@oclc.org
- Library of Congress
Contact:Rebecca Guenther, rgue@loc.gov
- O'Reilly Associates
Contact:Terry Allen, terry@ora.com
- Los Alamos National Laboratory and Indiana University
Contact:Ron Daniel Jr., rdaniel@acl.lanl.gov
Contact:Pete Percival, percival@bronze.ucs.indiana.edu
- Bunyip Systems
Contact:Chris Weider, clw@bunyip.com
- Georgia Institute of Technology
Contact:Michael Mealling, michael.mealling@oit.gatech.edu, <http://www.gatech.edu/iir>
- SoftQuad
Contact: Yuri Rubinsky, yuri@sq.com
- Concordia University
Contact:Bipin Desai, bcdesai@cs.concordia.ca,
<http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html>

Next Steps

Refinement and standardization of the metadata element set defined in this document will be an ongoing, dynamic process involving many stakeholder communities. No single forum will suffice to air all concerns and no single standard can be expected to accommodate the needs of all communities. The problem must be divided into manageable chunks and the process must engage the relevant stakeholder communities. Implicit in the present activity is the proposition that there are core elements common to many object types, and that a simple, extensible framework of such elements can be defined to support more complete resource descriptions.

The initial objective--the specification of elements for the discovery of document-like objects--can be extended in a variety of directions:

- Expansion of the Dublin Core to include other object types, such as services or collections.
- Expansion of the Dublin Core to embrace functionality other than resource discovery, such as archival control and the authentication of users and charging mechanisms.
- Establishing standardized methods for extensibility.
- Refinement of existing work. The Dublin Core is an untested approach to the description of resources that will need to be modified with experience.

OCLC and NCSA will establish a workshop series to address aspects of this agenda. A Metadata Workshop Steering Committee will be established to define topics and assure appropriate representation of stakeholders. Design groups of perhaps a dozen or fewer individuals will be solicited to prepare discussion papers to focus workshop activities. Participants will be invited based on their publicly evident accomplishments in relevant areas or by reviewed application. Workshops will be limited to 50 or fewer participants and conducted in roughly the style of the March 1995 Workshop.

Other work will be done in coordination with IETF working group on Uniform Resource Identifiers (URIs) to assure that the results can be integrated into the emerging protocols for resource location and persistent naming.

Finally, active promotion of results will be carried out by establishing liaison with formal associations of stakeholders. In the library community, MARC standards evolve under the guidance of the Machine-Readable Bibliographic Information Committee (MARBI), composed of representatives of the Library of Congress and other stakeholders in the library community. A close relationship should be sustained between this committee and the Metadata Work Group. Relationships should also be established with publishers, document vendors, SGML vendors and theoreticians working on the problem of text encoding. Other communities also have requirements that must be accommodated in any framework for resource description. These communities include the GIS community, government information providers and business communication groups.

References

[MARC]

Network Development and MARC Standards, Office, ed. 1994. USMARC Format for Bibliographic data. 1994. Washington, DC: Cataloging Distribution Service, Library of Congress.

[TEI]

Sperberg-McQueen, C. M., and Leu Burnard, ed. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

[Gaynor]

Gaynor, Edward. 1994. "Cataloging Electronic Texts: The University of Virginia Library Experience." Library Resources and Technical Services 38(4): 403-413 (October 1994).

Copyright © 1995 OCLC

The logo for d-Lib forum, featuring the text "d-Lib forum" in a stylized font inside a rounded rectangular border.The logo for d-Lib magazine, featuring the text "d-Lib magazine" in a stylized font inside a rounded rectangular border.

hdl:cnri.dlib/july95-weibel

OCLC/NCSA Metadata Workshop Report

Stuart Weibel, Jean Godby, Eric Miller

Office of Research, OCLC Online Computer Library Center, Inc.

Ron Daniel

Advanced Computing Lab, Los Alamos National Laboratory

1.0 Executive Summary

The March 1995 Metadata Workshop, sponsored by the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA), convened 52 selected researchers and professionals from librarianship, computer science, text encoding, and related areas, to advance the state of the art in the development of resource description (or metadata) records for networked electronic information objects.

1.1 Goals

Goals of the workshop included (1) fostering a common understanding of the needs, strengths, shortcomings, and solutions of the stakeholders; and (2) reaching consensus on a core set of metadata elements to describe networked resources.

1.2 Scope

The size and complexity of the resource description problem required limiting the scope of deliberations. Given that the majority of current networked information objects are recognizably "documents", and that the metadata records are immediately needed to facilitate resource discovery on the Internet, the proposed set of metadata elements (The Dublin Core) is intended to describe the essential features of electronic documents that support resource discovery. Other important metadata elements, such as those describing cost accounting or archiving information, were excluded from consideration. It was recognized that these elements might be included in a more complete record that could be derived from the Dublin Core by a well-defined extension.

1.3 The Intended Niche

The Dublin Core is not intended to supplant other resource descriptions, but rather to complement them. There are currently two types of resource descriptions for networked electronic documents: automatically generated indexes used by locator services such as Lycos and WebCrawler; and cataloging records, such as MARC, created by professional information providers. Automatically generated records often contain too little information to be useful, while manually generated records are too costly to create and maintain for the large number of electronic documents currently available on the Internet. Records created from the Dublin Core are intended to mediate these extremes, affording a simple structured record that may be enhanced or mapped to more complex records as called for, either by direct extension or by a link to a more elaborate record.

1.4 Next Steps

The work of the 1995 workshop is one of a series of steps being taken to improve the description of

networked information objects. A Metadata Workshop Steering Committee is being formed to extend the work of the March 1995 Workshop through a series of similar activities that will bring together stakeholder communities to focus on discrete parts of the larger problem. As with the initial workshop, promoting effective communication among the communities will be a primary goal. The diversity of implementation efforts influenced by the Dublin Core testifies to the benefits of such communication, and promises to help integrate related activities among librarians, the Internet Engineering Task Force (IETF), text encoding researchers, and other researchers who have substantial investments in resource description.

2.0 Introduction

The explosive growth of interest in the Internet and the World Wide Web in the past five years has created a digital extension of the academic research library for certain kinds of materials. Valuable collections of texts, images and sounds from many scholarly communities--collections that may even be the subject of state-of-the-art discussions in these communities--now exist only in electronic form and may be accessible from the Internet. Knowledge regarding the whereabouts and status of this material is often passed on by word of mouth among members of a given community. For outsiders, however, much of this material is so difficult to locate that it is effectively unavailable.

Why is it so difficult to find items of interest on the Internet or the World Wide Web? A number of well-designed locator services, such as Lycos [<http://lycos.cs.cmu.edu/>] are now available that automatically index every resource available on the Web and maintain up-to-date databases of locations. But it has not yet been demonstrated that indexes contain sufficiently rich resource descriptions, especially if the location databases are large and span many fields of study. Moreover, a huge number of resources on the Internet have no description at all beyond a filename which may or may not carry semantic content. If these resources are to be discovered through a systematic search, they must be described by someone familiar with their intellectual content, preferably in a form appropriate for inclusion in a database of pointers to resources. But current attempts to describe electronic resources according to formal standards (e.g. the TEI header [[TEI](#)] or MARC [[MARC](#)] cataloging) can accommodate only a small subset of the most important resources.

Another solution, not yet implemented, that promises to mediate these extremes involves the manual creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create the record, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover the descriptions and collect them into searchable databases.

What should this hypothetical description contain? Put in a convenient jargon, the question is about **metadata**--literally, data about data--or the contents of a surrogate record that characterize an object. Thus the question can be recast more precisely: how can a simple metadata record be defined that sufficiently describes a wide range of electronic objects? Recognizing the need to answer this and a multitude of associated questions, the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) sponsored the invitational Metadata Workshop on March 1-3, 1995, in Dublin, Ohio. Fifty-two librarians, archivists, humanities scholars and geographers, as well as standards makers in the Internet, Z39.50 and Standard Generalized Markup Language (SGML) communities, met to identify the scope of the problem, to achieve consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas, and to lay the

groundwork for achieving further progress in the definition of metadata elements that describe electronic information. This paper reports on the progress made at that workshop.

3.0 The Dublin Metadata Workshop

Since the Internet will contain more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves, without having to undergo the extensive training required to create records conforming to established standards. As one step toward realizing this goal, the major task of the Metadata Workshop was to identify and define a simple set of elements for describing networked electronic resources. To make this task manageable, it was limited in two ways. First, only those elements necessary for the discovery of the resource were considered. It was believed that resource discovery is the most pressing need that metadata can satisfy, and one that would have to be satisfied regardless of the subject matter or internal complexity of the object.

The discussion was further restricted to the metadata elements required for the discovery of what were called **document-like objects**, or **DLOs** by the workshop participants. It was believed that DLOs are still the most common type of resource sought in the Internet and that whatever solution could be proposed for DLOs could be extended to other kinds of resources. More importantly, the likelihood of making progress on this challenging problem would be increased if attention could initially be restricted to something familiar.

DLOs were not rigorously defined, but were understood by example. For example, an electronic version of a newspaper article or a dictionary is a DLO, while an unannotated collection of slides is not. Of course, the crux of the problem is that in a networked environment, DLOs can be arbitrarily complex because they can consist of text with callouts to images, audio or video clips, or to other hypertext documents. The participants of the Metadata Workshop made no attempt to limit the complexity of DLOs, except possibly to say that the intellectual content of a DLO is primarily text, and that the metadata required for describing DLOs will bear a strong resemblance to the metadata that describes traditional printed texts.

As a result of the restricted focus of the workshop, certain issues required for a complete description of DLOs, such as cost, archival status and copyright information, were eliminated from the scope of the discussion. Elements required for the description of objects other than DLOs, such as the elements required for the description of complex geological strata in a geospatial resource, were also beyond the scope of the discussion. The goal was to define a small, universally understood set of metadata elements that would allow authors and information providers to describe their work and to facilitate interoperability among resource discovery tools. But because the core elements do not yield a description of objects that meets the needs of specialized user communities, careful consideration was also given to mechanisms for extending the element set.

The primary deliverable from the workshop was a set of thirteen metadata elements, named the **Dublin Metadata Core Element Set** (or Dublin Core, for short) by the workshop participants. The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.

Below is an introductory discussion of the elements in the Dublin Core

Element Description

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The data representation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object

To make this discussion concrete, consider an electronic record created with the relevant portions of the Dublin Core, and a sample syntax, that describes an electronic version of Maya Angelou's poem "On the Pulse of Morning". This description is based on a record created by the University of Virginia Library's Electronic Text Center. (For a description of that project, see Gaynor [\[Gaynor\]](#) .)

- **Title:** On the Pulse of Morning
- **Author:** Maya Angelou
- **Publisher:** University of Virginia Library Electronic Text Center
- **OtherAgent:** Transcribed by the University of Virginia Electronic Text Center
- **Date:** 1993
- **Object:** Poem
- **Form:** 1 ASCII file
- **Source:** Newspaper stories and oral performance of text at the presidential inauguration of Bill Clinton
- **Language:** English

Although the goal of the Dublin Metadata Workshop was to develop a simple set of metadata elements, these elements also had to be defined in such a way that they could be mapped into more complex and highly controlled systems such as USMARC. These conflicting demands have been reconciled in two ways. The first was to create a set of metadata elements with definitions that could be understood easily without the need for user training, as well as an approach to modification that meets the needs of specialized communities for more precise descriptions. This is described in Section 5. The second was to provide mechanisms for extending the core element set to describe items other than document-like objects. These mechanisms are discussed in Section 6.

As Guenther [\[Guenther\]](#) points out, a small set of metadata elements such as the Dublin Core would be valuable for at least four reasons. First, it would encourage authors and publishers to provide metadata, in a form that automated resource discovery tools could collect it, when they make data available. Second, it would encourage the creation of network publishing tools that contain a template for metadata elements, further simplifying the task of creating metadata records. Third, a record created with the

Dublin Core could serve as the basis for a more detailed cataloging record if the need arises. Finally, if something like the Dublin Core became a standard, metadata records could be understood across user communities.

Since the difficult work of identifying a simple but useful specification for the description of networked resources has only begun, the major accomplishment of the Metadata Workshop was to define the problem and sketch out a solution. Many details of that solution need further refinement. Accordingly, one goal of this paper is to focus discussion by identifying the areas that need the most work. Section 7 lists the most pressing unresolved issues. The appendices identify further problems that arise when an attempt is made to formalize what has been proposed so far. Section 8 describes projects already underway that increase our understanding of how resources on the Internet should be described and show what is possible if a metadata element set such as the Dublin Core becomes a standard. Finally, Section 9 outlines the steps that are being taken to ensure that progress on this important problem continues to be made.

4.0 Underlying Assumptions

The discussions at the Metadata Workshop revealed several principles that should guide the further development of the element set. Adherence to these principles increases the likelihood that the core element set will be kept as small as possible, that the meanings of the elements will be understood by most users, and that the element set will be flexible enough for the description of resources in a wide range of subject areas. These principles are intrinsicality, extensibility, syntax-independence, optionality, repeatability and modifiability.

4.1 Intrinsicality

The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. Though extrinsic data may be important for a complete description of an object, it is handled by the extension mechanisms described in Section 6.

4.2 Extensibility

In addition to its use in dealing with extrinsic data, the extension mechanism will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements.

Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields. In addition, the specification of the Dublin Core itself may change over time, and the extension mechanism allows revisions while maintaining some backward compatibility with the originally defined element set.

4.3 Syntax-Independence

Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs. The examples in Appendix I show some sample encodings.

4.4 Optionality

For two reasons, all the elements are optional. First, the Dublin Core may eventually be applied to objects for which some elements have no meaning. Who is the author of a satellite image? Second, it seems futile to mandate complex descriptions if the creators of the content are expected to provide the descriptive material. A simple description is better than no description at all.

4.5 Repeatability

All elements in the Dublin Core are repeatable. For example, multiple author elements would be used when a resource has multiple authors.

4.6 Modifiability

Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier. If no qualifier is present, the element has its common-sense meaning; otherwise, the definition of the element is modified by the value of the qualifier.

Qualifiers will be typically derived from well-known conventions in the library community or from the field of knowledge appropriate to the resource. Qualifiers are important because they give the Dublin Core a mechanism for bridging the gap between casual and sophisticated users. For example, the data in the **Subject** element consists of any word or phrase that describes the object's content. However, a professional cataloger may wish to supply the name of the authoritative source from which the subject terms are taken. In such a case, the element may be written as **Subject (scheme=LCSH)**, indicating that the subject terms are taken from the Library of Congress Subject Headings.

5.0 A Detailed Description of the Elements in the Dublin Core

This section presents detailed definitions of the elements in the Dublin Core. The simple and informal records that would result from the application of this element set will help raise the standards of resource description without imposing the overhead normally associated with more exacting standards. But when necessary, the elements in the Dublin Core should support definitions that are precise enough to enable the mapping of records to widely used standards such as USMARC, TEI or FGDC.

To accomplish this goal, each element in the Dublin Core can be qualified with a scheme. Schemes are used whenever it is necessary to describe the rationale for the encoding of the data associated with the element, such as when reference is made to controlled vocabulary, a well-known notation or a published standard. This section describes the use of schemes, but since qualification is still an active topic of discussion, other qualifiers will undoubtedly be proposed in future versions of the Dublin Core. See Appendix II for a sample SGML Document Type Definition (DTD) [[Herwijnen](#)] that defines additional qualifiers for some of the elements.

A comment about syntax is necessary at this point. As stated in Section 4.3, the elements in the Dublin Core were purposely defined in a syntax-independent manner. Nonetheless, providing examples is often the best way to communicate, so the following descriptions include examples represented in a simple syntax that might be suitable for a HyperText Markup Language (HTML) document. This should not be

considered as the definitive representation, but rather as a possible representation.

5.1 Subject

The **Subject** is the field of knowledge to which the work belongs. This may be a general description of a broad discipline, or a series of descriptors of differing scope.

The **Subject** element can be qualified by a scheme, which specifies adherence to a known classification system such as the Library of Congress Subject Headings, the Dewey Decimal System, or the Art and Architecture Thesaurus, to name a few. For example:

- Subject (scheme=LCSH) = UNIX (Computer system)
- Subject (scheme=Dewey Decimal System)=004.251 Supercomputers--systems design

Without the scheme, the **Subject** element is a keyword and may contain any word or phrase that describes the intellectual content of the object. For example:

- Subject = Network Information Discovery and Retrieval
- Subject = Information Retrieval
- Subject = Cataloging of Internet Resources
- Subject = Metadata
- Subject = Description of electronic resources

5.2 Title

Title is defined as the name of the object. Most document-like objects will have an obvious title, but if the Dublin Core is eventually used to describe resources such as satellite images or objects in a museum, there will be no designated character string that is understood as a title. A descriptive phrase might be appropriate instead. For example:

- Title = Statistical Summary of Web Site Usage for NASA's Shoemaker Levy Home Page
- Title = A Description of Networked Information: The Dublin Core Element Set
- Title = Benjamin Franklin's Spectacles

If greater precision is desired, titles may be qualified by a scheme. For example:

- Title (scheme=AACR2) = The structure of language

5.3 Author

The common understanding of **author** is the person(s) or agent primarily responsible for the intellectual content of the work. For example:

- Author = Smith, Jeffrey K.
- Author = World Book Encyclopedia

The **author** element could also be modified by a scheme such as USMARC. In the example below, the author is a person with an honorific title who lived between 1859 and 1930.

- Author (scheme=USMARC) = 100 1 Doyle, Author Conan \$c Sir, \$d 1859-1930

This example illustrates an important problem. As Guenther [\[Guenther\]](#) argues, the Dublin Core does not achieve an unambiguous mapping to a complex scheme such as USMARC with the above syntax because the **Author** element can map to many fields in the USMARC standard. It is also necessary to specify the portion of the external scheme that is being referenced. This is a general problem with all of the Dublin Core elements, and more discussion is required to solve it in a way that is satisfactory to all affected user communities.

5.4 Publisher

Publisher is defined as the the agent or agency responsible for making the object available. For example:

- Publisher = University of Virginia Center for Electronic Text

5.5 OtherAgent

OtherAgents are the persons or organizations other than authors or publishers who have made significant intellectual contributions to the work. Strictly speaking, authors, publishers, and others responsible for intellectual content could all be described by an appropriately qualified element such as **OtherAgent**. The more verbose description adopted here is intended to make clearer the common roles of Author and Publisher. The **OtherAgent** element is intended to describe less common roles, such as editor, illustrator, compiler, convenor, photographer, or any secondary but significant role of content responsibility.

OtherAgent can be further specified using a scheme. For example, the TEI [\[TEI\]](#) standard contains valuable detail regarding the roles involved in creating document-like objects.

- OtherAgent (scheme=TEI) = <pubStmt><role>Creation of machine-readable version</role>
<name>The Ohio State University</name> </pubStmt>

Some workshop participants suggested that it would be useful to define a set of common roles for the **OtherAgent** element so they could be used without referring to an external scheme. If this list were defined, the OtherAgent element could take the following forms:

- OtherAgent (role=Editor) = Weibel, Stuart L.
- OtherAgent (role=Illustrator) = Dennis, Wesley

5.6 Date

The **Date** of publication is intended to reflect the date at which the object became available in its current form.

- Date = May 6, 1995

The scheme may be used to identify the syntactic form of the date. For example:

- Date (scheme=ANSI X3.30-1985) = 950506

One problem with the **Date** element is that it is potentially misleading. Since electronic objects can be easily copied, the date stamp of a particular object may have no significance. It may be necessary to define a way to refer to more meaningful dates, such as latest date of the current controlled version for executable software, or the date of the original for digitized texts. More discussion is required to resolve this issue.

5.7 ObjectType

ObjectType is defined as the abstract category or genre of the object, such as novel, poem, dictionary, thesaurus, executable software, source code, data file, or any other category judged to be useful for the discovery and retrieval of the resource being described. For example:

- **ObjectType** = Internet Draft Informational RFC

The scheme can be used to indicate that the words used to describe the **ObjectType** are taken from controlled vocabulary. For example:

- **ObjectType** (scheme=AACR2) = computer file

5.8 Form

Form is defined as the data representation of the object and is intended to provide an information seeker with information about the hardware or software resources necessary to display or operate the object. Information included should provide a sufficient description to make possible a judgement of usefulness prior to accessing an object. Examples might include Postscript-II document, Windows 3.1 executable file, HTML file, or WordPerfect 6.1 document. In the example below, the form is an Internet Media Type.

- **Form** (scheme=IMT) = text/html

The use of the **scheme** qualifier for the **Form** element is analogous to that defined for the **ObjectType** element.

5.9 Identifier

The **Identifier** is the string or number used to uniquely identify an object. To enhance the usefulness of this data, a scheme value will specify the authority of the identifier. For example:

- **Identifier** (scheme=ISBN) = 0-8230-2355-9
- **Identifier** (scheme=URL) = <http://www.oclc.org/metadata.html>

Non-public identifiers, such as a university department's technical paper number, could also be used.

5.10 Relation

The **Relation** element identifies the object's relationship to other objects, print or electronic, such as other parts of a document hierarchy, other parts of a collection of documents, or another of a series of documents.

The **Relation** element is designed to give the author flexibility when selecting the scope of the resource being described. In a hypertext environment, the resource might reasonably be a paragraph from a larger text, an entry on a Web page, or a slide in a collection. Of course, the resource may also be isolated or freestanding, in which case the **Relation** element is not relevant to the description. The simple record of the Maya Angelou poem in Section 2 omits the **Relation** element for this reason.

The exact form of the **Relation** element needs to be worked out by further discussion, but it needs to contain at least two sub-elements: a description of the relation and a pointer to the related item. In the example below, it is assumed that the sub-elements **Type** and **Identifier** have been defined. The example below contains a URL, establishing the fact that the object being described is part of the proceedings from the Third International World-Wide Web Conference.

- Relation (type=ContainedIn) (identifier = URL) = <http://www.elsevier.nl/>

5.11 Source

The **Source** element identifies the object from which the object being described is derived. This element is intended to connect an electronic object with a previous version, possibly in another medium, that establishes the history of the object. The **Source** element identifies other objects with the same intellectual content as the resource being described, while the **Relation** element identifies objects of different intellectual content with which the resource is logically connected.

The data in the **Source** element is most easily understood as an identifier, named in the scheme, that uniquely points to the record describing the previous version. For example:

- Source (scheme=ISBN) = 0-201-63337-X

If it is more convenient, the **Source** element may contain an extensive bibliographic citation or even a recursive instance of the Dublin Core, as one of the examples in Appendix I illustrates. In such a case, the scheme is "Dublin Core."

5.12 Language

The **Language** element should specify the language of the intellectual content of the object being described. For example:

- Language = English

If an abbreviation is used, the scheme can be used to identify the source. For example:

- Language (scheme=USMARC) = Eng

5.13 Coverage

The **Coverage** element describes the spatial and temporal characteristics of the object and is the key element for supporting spatial or temporal range searching on document-like objects that describe geospatial data. The first example below is a simple, nontechnical use of the **Coverage** element. The second example illustrates spatial coverage, with the scheme identifying the data as latitude/longitude

coordinates; and the third example shows temporal coverage, with the scheme identifying the syntax of the dates.

- Coverage (type = spatial) = The Atlantic Ocean
- Coverage (type = spatial, scheme = LATLONG) = {West = 180, East = 180, North = 90, South = 90}
- Coverage (type = temporal, scheme = ANSI X3.30-1985) = {Begin = 19910101, End = 19930601}

For greater precision, some workshop participants suggested that **Coverage** should be modified by the qualifiers **spatial** and **temporal**. See the sample SGML Document Type Definition in Appendix II.

6.0 Extensions to the Dublin Core

Extensibility is an essential feature of any metadata system because a single set of metadata elements, no matter how large, cannot possibly accommodate all resource types. But increasing the size of the core element set would complicate rather than simplify the problem because a large element set would be less comprehensible to a diverse user population. To reconcile these conflicting demands, the Dublin Core has been designed from the outset to be small but extensible.

This is manifested in three ways. First, local additions are accommodated by allowing elements to be added to the record that describes a resource. These additional fields are not guaranteed to be understood outside the community that proposed them, but they need not cause errors for systems that understand the core element set. The extensions may be an unstructured string of text, a pointer to another record that conforms to an established standard, or even the record itself.

The second mechanism for implementing extensibility is the "scheme" sub-element, described in the previous section. The set of values for "scheme" is open-ended and unspecified in the Dublin Core because these are expected to be supplied by user communities. For some elements, such as **Subject**, the set of values will be small because only a limited number of classification hierarchies are available from which subject terms can be derived. For other elements, the set of scheme values may be large, reflecting the complexity of the resources being described. For example, the **Identifier** element may have schemes of **FTP**, **URL**, **URN**, **ISBN**, or any number of locally assigned schemes. These scheme values must necessarily reflect our current understanding of the difficult problem of name resolution for electronic objects accessible from the Internet.

The third mechanism for extensibility is the labeling of the Dublin Core itself. This can be understood as a version number and it may be changed if new elements are added to the base set or if the semantics of existing elements changes. The current version of the Dublin Core is 0.1.

This discussion describes the principles by which records created with the Dublin Core can be extended, but a separate paper is required for a more rigorous discussion that proposes implementation solutions.

7.0 Unresolved Issues

This document describes the work of a relatively small number of people who have begun to address the difficult issues arising from the need to provide descriptions of resources in order to promote their discovery in a networked environment such as the Internet. Of course, this is a large problem, and the

work has only begun. One of the goals of this document is to record what has been proposed so far--in the Dublin Metadata Workshop, and in followup discussions with the participants. If this effort has succeeded, others need not spend time covering the same ground, and it will be possible to move quickly to the issues that remain. Of course, more discussion and refinement of the core element set is expected. But there are larger problems to be solved, especially those involving versions, extensibility and character sets.

7.1 Versions

The treatment of different versions of an electronic resource has not been addressed in this document because there is no consensus regarding the definition or even the vocabulary used to describe versions of electronic objects. They seem to be fundamentally different from the versions of printed materials, such as editions and printings, which have been well-described by scholars in the library community. Unlike printed versions, new electronic versions often supplant or obliterate older versions. Electronic versions also proliferate more easily, and as a result, differences between versions may be more slight.

It is therefore nontrivial to ask: when are two different versions of an electronic resource the same work? The easy answer is that they are the same if bitwise comparisons reveal them to be identical. But this criterion is too restrictive because two electronic objects, such as a document that exists in WordPerfect and ASCII formats, could have identical intellectual content but fail the test of bitwise identity.

Without a better understanding of electronic versions, it may not be possible to use the Dublin Core to create unambiguous records that describe different versions of a resource. If the creators of records judged two versions to be the same work, they would use the **Source** element to describe the earlier version; otherwise, they would record this data in the **Relation** field. This shows that the important and reasonably well-understood concept of *Edition* can be captured in either of two elements of the Dublin Core.

7.2 Extensibility

Although the principles for extending the Dublin Core are straightforward, much work remains to be done before an important question is answered: how can records be extended in such a way that meets the needs of different communities while maintaining some level of interoperability? The extended records could conform to the specifications of the Dublin Core, yet still contain an infinite variety of data types and standards. There are several ways to solve this problem but none of them are currently feasible.

First, the problem could be eliminated by requiring a single consistent transfer syntax such as SGML. But this would violate one of the underlying assumptions of the Dublin Core and would be impossible to enforce. Second, a requirement could be added to the Dublin Core stipulating that an extended record must contain a pointer to the software that understands it. This violates the rule that all elements in the Dublin Core are optional and assumes the existence of software that understands the heterogeneous records yet to be created. Third, user communities could devise and maintain their own standards. However, there is no guarantee that the independent evolution of standards for the specification of metadata to promote resource discovery in an electronic environment will produce even minimal interoperability.

7.3 Character Sets

ASCII is the most widely deployed representation of text, and in the interest of interoperability, information exchange on the Internet relies on it almost exclusively. However, the Internet reaches communities all over the world. If it is to become a significant cultural force, the needs of languages using non-ASCII character sets will eventually have to be addressed. These issues have been avoided in this document because the intent is to adopt the solutions put forth by other standards makers.

The Multipurpose Internet Mail Extensions (MIME) introduces Internet Media Types, including text representations other than ASCII. HTML, used by most World-Wide Web browsers, is a proposed Internet Media Type as well as an SGML application. In the MIME and SGML specifications, however, character representation is notoriously complex, and the two specifications are inconsistent and incompatible. The Internet Engineering Task Force (IETF), and the MIME_SGML, HTML and HyperTextTransfer Protocol (HTTP) working groups are attempting to rectify these inconsistencies and are discussing the best ways of incorporating text representations other than ASCII.

7.4 "Third-party" Metadata

In the largest sense, "metadata" includes any information which purports to be about other information. Some of the most useful metadata is produced by persons other than librarians and document owners, and it can be found neither in card catalogs nor in self-descriptions of the documents themselves. Many kinds of such "third-party" metadata (e.g., bibliographies) are indispensable aids to information discovery. It should be possible to allow topic-oriented metadata documents with semantic network functionality to be cooperatively authored, interchanged, and integrated into master documents. Such documents (and amalgamated master documents) might resemble traditional catalogs, indexes, thesauri, encyclopediae, bibliographies, etc., with functional enhancements such as the hiding of references that are outside the scope of the researcher's interest, etc. Early work in this area is being done by the [\[CApH\]](#) group.

8.0 Implementations

The OCLC/NCSA Metadata Workshop is one of a series of developments that will lead to more effective resource discovery systems for the Internet. Other developments include the Denver 1992 data element discussions, the Library of Congress 1994 Workshop on the description of electronic resources, the IETF Uniform Resource Locator working group meetings, the Digital Library projects funded by the National Science Foundation, and many other activities in stakeholder communities, as well as the experimental Web crawler "infobots."

It is no small challenge to integrate these disparate activities, each with vocabularies, agendas, and objectives of their own, into a coherent whole. One of the goals of the OCLC/NCSA Metadata Workshop was to bring many of the relevant stakeholders together for such discussions, and some progress has been achieved toward this end.

All of these activities are but the first hesitant steps toward the goal of rational resource description. Whatever advances in understanding and communication among the participants were achieved at the Dublin Workshop, the most important measures of success will be the dissemination of these ideas in the community. But the ultimate proof is in the implementations of these ideas.

A number of Metadata Workshop conferees represent organizations that have ongoing activities or are starting activities that will be influenced by the results of the workshop. These include:

8.1.1 The OCLC Spectrum Project

The primary goal of the Spectrum project is to develop a tool that enables individuals, with or without specialized knowledge of library cataloging or markup to create records for describing and accessing networked electronic resources of various types.

Contact:"Diane Vizine-Goetz" email = vizine@oclc.org

8.1.2 The OCLC Internet Resources Cataloging Project

A group of volunteer libraries is participating in a U.S. Department of Education-funded project to identify, select, and catalog Internet-accessible electronic resources using standards and conventions widely adopted in the library community.

The overall objectives of this project are to employ, evaluate and extend the library catalog model to embrace Internet resources, and to focus the intellectual resources of scores of professions on the attendant problems and opportunities. Records created in this project will conform to USMARC standards and will be suitable for integration within local and national library catalogs.

Contact:Erik Jul, jul@oclc.org

8.1.3 Library of Congress

The Machine-Readabale Bibliographic Information (MARBI) Committee at LC has drafted a discussion paper (DP86: Mapping the Dublin core metadata elements to USMARC [\[Guenther\]](#)) for review at the summer meeting. MARBI is the committee responsible for overseeing changes to the USMARC format.

Contact:Rebecca Guenther, rgue@loc.gov

8.1.4 O'Reilly Associates

O'Reilly & Associates, a leading publisher of Internet books, is exploring online publishing with HTML and SGML. They are supporting a definition of HTML 2.0 containing the META element to permit the inclusion of metadata records such as the Dublin Core.

Contact:Terry Allen, terry@ora.com

8.1.5 Los Alamos National Laboratory and Indiana University

Researchers at LANL and IU are cooperating in the implementation of a META tag implementation for application of the Dublin Core element set in HTML documents.

Contact:Ron Daniel Jr., rdaniel@acl.lanl.gov

Contact:Pete Percival

8.1.6 Bunyip Systems

Bunyip will be indexing the DublinCore in its deployment of WHOIS++. In addition, Bunyip will be advocating the use of the Internet Anonymous FTP Archive (IAFA) templates (a superset of the Dublin

Core) for the indexing of Anonymous FTP archives through its archie service.

Contact:Chris Weider, clw@bunyip.com

8.1.7 Georgia Institute of Technology

Implementors are using the Dublin Core as the set of metadata elements to include in a resource discovery system based on whois++ and the centroids mesh. It will act primarily as a method for transforming different data formats. In particular it will translate from the more complex TEI header into a simpler attribute/value flat list based on the Dublin Core. This flat list will be the basis for limited information discovery.

Contact:Michael Mealling, michael.mealling@oit.gatech.edu, <http://www.gatech.edu/iir>

8.1.8 SoftQuad

SoftQuad Panorama, the company's SGML viewer for the Web, will become metadata-aware as implementations of Dublin Core records become available. It will support, through the use of SGML Document Type Definitions, both the basic core set as well as site-specific extended versions.

Contact:Yuri Rubinsky, yuri@sq.com

8.1.9 Concordia University

The semantic header is a metadata structure which stores indexing information about Internet resources. This information will be stored in distributed databases and will be accessible for search using a GUI. The metadata elements contained in the semantic header have been influenced by the Metadata Workshop.

Contact:Bipin Desai, bcdesai@cs.concordia.ca,
<http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html>

9.0 Next Steps

Refinement and standardization of the metadata element set defined in this document will be an ongoing, dynamic process involving many stakeholder communities. No single forum will suffice to air all concerns and no single standard can be expected to accommodate the needs of all communities. The problem must be divided into manageable chunks and the process must engage the relevant stakeholder communities. Implicit in the present activity is the proposition that there are core elements common to many object types, and that a simple, extensible framework of such elements can be defined to support more complete resource descriptions.

The initial objective--the specification of elements for the discovery of document-like objects--can be extended in a variety of directions.

- Expansion of the Dublin Core to include other object types, such as services or collections.
- Expansion of the Dublin Core to embrace functionality other than resource discovery, such as archival control and the authentication of users and charging mechanisms.
- Establishing standardized methods for extensibility.

- Refinement of existing work. The Dublin Core is an untested approach to the description of resources that will need to be modified with experience.

OCLC and NCSA will establish a workshop series to address aspects of this agenda. A Metadata Workshop Steering Committee will be established to define topics and assure appropriate representation of stakeholders. Design groups of perhaps a dozen or fewer individuals will be solicited to prepare discussion papers to focus workshop activities. Participants will be invited based on their publicly evident accomplishments in relevant areas or by reviewed application. Workshops will be limited to 50 or fewer participants and conducted in roughly the style of the March 1995 Workshop.

Further work will be done at a Birds of a Feather (BOF) meeting planned for the Stockholm IETF Meeting in July of 1995. A BOF discussion is an initial step in the establishment of an IETF working group. The IETF serves best in the capacity of validating design work that is done by smaller, more focused groups. An IETF working group on Metadata will provide an effective way of involving the technical stakeholders in the computer and networking community that must implement and use the standards.

Finally, active promotion of results will be carried out by establishing liaison with formal associations of stakeholders. In the library community, MARC standards evolve under the guidance of the Machine-Readable Bibliographic Information Committee (MARBI), composed of representatives of the Library of Congress and other stakeholders in the library community. A close relationship should be sustained between this committee and the Metadata Work Group. Relationships should also be established with publishers, document vendors, SGML vendors and theoreticians working on the problem of text encoding. Other communities also have requirements that must be accommodated in any framework for resource description. These communities include the GIS community, Government information providers and business communication groups.

References

[CApH]

Conventions for the Application of HyTime (CApH). "Semantic Assignment Module" and "Topic Relationship Module." 1995. Graphic Communications Association, Alexandria, VA. ([ftp.techno.com, pub/HyTime/CApH](ftp://ftp.techno.com/pub/HyTime/CApH))

[FGDC]

Federal Geographic Data Committee. 1994. Content standards for digital geospatial metadata (June 8). Federal Geographic Data Committee. Washington, D.C.

[Gaynor]

Gaynor, Edward. 1994. "Cataloging Electronic Texts: The University of Virginia Library Experience." Library Resources and Technical Services 38(4): 403-413 (October 1994).

[Guenther]

Guenther, Rebecca. 1995. "Mapping the Dublin Core Metadata Elements to USMARC" MARBI Discussion Paper NO. 86.

[Herwijnen]

Herwijnen, Eric van, Practical SGML , Kluwer Academic Publishers, 1994

[MARC]

Network Development and MARC Standards, Office, ed. 1994. USMARC Format for Bibliographic data. 1994. Washington, DC: Cataloging Distribution Service, Library of Congress.

[TEI]

Sperberg-McQueen, C. M., and Leu Burnard, ed. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

Appendix 1.0: Sample Records Encoded Using the Dublin Metadata Core

This Appendix contains two sample Dublin Core records encoded in a simple syntax that might be translated easily to SGML. As in all other examples in this document, the syntax of these records is presented only for clarity of exposition.

The first example was created by a subject-matter specialist who has no library cataloging expertise. It describes an Internet Request for Comment (RFC) found on a Web page containing similar RFCs. The Relation element points to the Web page and to a similar record on the same page.

```
Subject:      IETF, URI, Uniform Resource Identifiers
Title:        A Unifying Syntax for the Expression of Names and Addresses of Objects
              on the Network as used in the World-Wide Web.
Title:        (Subtitle) Universal Resource Identifiers in WWW
Author:       Berners-Lee, T.
Publisher:    CERN
Date:         1994
Object-Type:  Internet RFC
Form (scheme=IMT): text/plain
Identifier(scheme=URL): gopher://gopher.es.net:70/0R0-57601-/pub/rfcs/rfc1630.txt
Relation (type=child)(identifier=URL): http://ds.internic.net/ds/dspglinndoc.html
Relation (type=sibling)(identifier=URL): http://ds.internic.net/rfc/rfc1738.txt">
```

The second Dublin Core record was created by a librarian for a Postscript version of a monograph that is derived from a previous printed edition. It contains references to schemes commonly used in the library community. The link to the printed version is described in the **Source** element, here realized as a recursive instance of the Dublin Core record.

Example 1: Dublin Core record for an electronic version of an OCLC Research Report

Element Name	Content
Subject:	
scheme=LCSH:	Internet (Computer network) Cataloging of computer files Information networks Computer networks Libraries--Communication systems Information storage and retrieval

systems

Title: Assessing Information on the Internet: Toward Providing Library Services for Computer Mediated Communication

Author: Martin Dillon
Author: Erik Jul
Author: Mark Burge
Author: Carol Hickey

Publisher: OCLC

Date: 1994

Identifier:
Scheme=OCLC: 155653163X

Object type:
Scheme=AACR2: monograph

Form:
7 postscript files
1 Unix tar file

Relation: For a Web page listing Internet accessible OCLC research publications go to:
<http://www.oclc.org/oclc/menu/reschdoc.htm>

Language: English

Source(scheme=DublinCore): Subject:
scheme=LCSH:
Internet (Computer network)
Cataloging of computer files
Information networks
Computer networks
Libraries--Communication systems
Information storage and retrieval systems

Title: Assessing Information on the Internet: Toward Providing Library Services for Computer Mediated Communication

Author: Martin Dillon
Author: Erik Jul
Author: Mark Burge
Author: Carol Hickey

Identifier:
scheme=OCLC Technical Report Number:
1234567

Date: 1993
Object type
Scheme=AACR2: monograph

Form:
Scheme=AACR2: 1 v. (various pagings) : ill. ; 29 cm.
Publisher: OCLC

Appendix 2.0: A Proposed Document Type Definition for the Dublin Metadata Core

This sample DTD is included here to make the proposals in this paper more precise and to promote discussion among those wishing to make improvements. The attribute **Scheme** is used when reference is made to an external authority for notation or vocabulary.

As in all examples in this paper, the DTD in this appendix specifies one possible syntax for an SGML version of the Dublin core. It is for illustrative purposes only.

```
<!-- This is the ISO8879:1986 document type definition for the DublinCore URC. -->
<!-- Note: This DTD is subject to discussion and/or modification by the
      participants of the OCLC/NCSA Metadata Workshop.
      95/20/06, eric j. miller, emiller@oclc.org -->

<!-- ===== Parameterizable Lists ===== -->

<!-- MeSH (Medical Subject Heading) Publication Types can be found
      <URL:http://www.sils.umich.edu/~nscherer/Medline/Table1.html> -->

<!ENTITY      % Subject.Scheme
      "LCSH | MeSH | Sears | AAT | INSPEC | ERIC | DDC | Other" >

<!-- TEI Information can be found <URL:http://etext.virginia.edu/TEI.html> -->

<!ENTITY      % Title.Scheme
      "AACR2 | TEI | Other" >

<!ENTITY      % Author.Scheme
      "AACR2 | TEI | Other" >

<!ENTITY      % OtherAgent.Scheme
      "AACR2 | TEI | MARC | Other" >

<!ENTITY      % Publisher.Scheme
      "AACR2 | TEI | Other" >

<!-- ANSI3.30 ::= yyyymmdd (4 for the year, 2 for the month, 2 for the day) -->
<!-- ANSI3.43 ::= hhmmss.f (2 for the hour, 2 for the minute, 2 for the sec
      and 2 for the fraction of the second including the decimal point -->
<!-- ANSI3.51 ::= -->

<!ENTITY      % Date.Scheme
      "ANSI3.30 | ANSI3.43 | ANSI3.51 | Other" >

<!ENTITY      % ObjectType.Scheme
      "NLM | Other">

<!ENTITY      % Form.Scheme
      "IMT | X.400 | Other">

<!ENTITY      % Identifier.Scheme
      "URN | URL | LCCN | ISBN | ISSN | SICI | MessageID | FPI | Other" >

<!ENTITY      % Source.Scheme
      "TEI | Other" >
```

```

<!ENTITY      % Language.Scheme
               "MARC | Other" >

<!ENTITY      % Relationship.Scheme
               "URN | URL | LCCN | ISBN | ISSN | SICI | MessageID | FPI | Other" >

<!ENTITY      % Hierarchy.Link
               "Top | Parent | Child | Sibling | Other" >

<!ENTITY      % Relationship.Type
               "Supersedes | Continues | Continued.From |
               Contained.In | Superseded.By | Cites | Extracted.From |
               Is.Part.Of | Contains | IsIndexOf | IsIndexedBy | GlossaryOf |
               Predecessor | Successor | IsDerivativeOf | Child | Parent |
               Sibling" >

<!ENTITY      % n.spacewindow
               "WestBounding, EastBounding, NorthBounding, SouthBounding" >

<!ENTITY      % n.timewindow
               "Begin | End" >

<!-- ===== Body of the DublinCore Metadata DTD ===== -->

<!-- Element list: Subject to change -->

<!ELEMENT     DublinCore      - -
               (BaseDesc?, Extension*) >
<!ATTLIST     DublinCore      Version          CDATA              #IMPLIED >

<!ELEMENT     BaseDesc        - -
               (Subject | Title | Author | OtherAgent | Publisher |
               Date | ObjectType | Form | Identifier | Relation |
               Source | Language | Coverage)* >

<!ELEMENT     Subject          - -      ANY >
<!ATTLIST     Subject          Scheme      (%Subject.Scheme;)      #IMPLIED >

<!ELEMENT     Title            - -      ANY >
<!ATTLIST     Title            Scheme      (%Title.Scheme;)        #IMPLIED >

<!ELEMENT     Author           - -      ANY >
<!ATTLIST     Author           Scheme      (%Author.Scheme;)       #IMPLIED >

<!ELEMENT     OtherAgent       - -      ANY >
<!ATTLIST     OtherAgent       Scheme      (%OtherAgent.Scheme;)   #IMPLIED >

<!ELEMENT     Publisher        - -      ANY >
<!ATTLIST     Publisher        Scheme      (%Publisher.Scheme;)    #IMPLIED >

<!ELEMENT     Date             - -      ANY >
<!ATTLIST     Date             Scheme      (%Date.Scheme;)         #IMPLIED >

<!ELEMENT     ObjectType       - -      ANY >
<!ATTLIST     ObjectType       Scheme      (%ObjectType.Scheme;)   #IMPLIED >

<!ELEMENT     Form             - -      ANY >
<!ATTLIST     Form             Scheme      (%Form.Scheme;)         #IMPLIED >

<!ELEMENT     Identifier       - -      ANY >

```

<!ATTLIST	Identifier	Scheme		(%Identifier.Scheme;) #IMPLIED >
<!ELEMENT	Relation		- -	ANY >
<!ATTLIST	Relation	Scheme		(%Relationship.Scheme;) #IMPLIED
		Type		(%Relationship.Type;) #IMPLIED >
<!ELEMENT	Source		- -	ANY >
<!ATTLIST	Source	Scheme		(%Source.Scheme;) #IMPLIED >
<!ELEMENT	Language		- -	ANY >
<!ATTLIST	Language	Scheme		(%Language.Scheme;) #IMPLIED >
<!ELEMENT	Coverage		- -	((Spatial Temporal)+) >
<!ELEMENT	Spatial		- -	((WestBounding, EastBounding, SouthBounding, NorthBounding)? Place*) >
<!ELEMENT	Place		- -	ANY >
<!ELEMENT	WestBounding		- -	ANY >
<!ELEMENT	EastBounding		- -	ANY >
<!ELEMENT	SouthBounding		- -	ANY >
<!ELEMENT	NorthBounding		- -	ANY >
<!ELEMENT	Temporal		- -	((Begin, End)? Time*) >
<!ELEMENT	Time		- -	ANY >
<!ELEMENT	Begin		- -	ANY >
<!ATTLIST	Begin	Scheme		(%Date.Scheme;) #IMPLIED >
<!ELEMENT	End		- -	ANY >
<!ATTLIST	End	Scheme		(%Date.Scheme;) #IMPLIED >
<!ELEMENT	Extension		- -	CDATA >

Notes on Metadata and the Web

For an overview paper on related areas, read about the [Warwick Framework](#), a container architecture for aggregating metadata.

These notes are based on the articles that appear in the Oct./Nov. 1997 issue (v. 24 no. 1) of the *Bulletin of the American Society for Information Science* (ASIS). The issue title is *Organizing Internet Resources: Metadata and the Web*.

Some of the key topics considered are:

- Dublin Core, its evolution, its adaptations
- Cataloging, MARC, and their extension to Internet
- Automatic classification: Scorpion
- Naming: URL, URN, URI, URC, DOI

Useful Links by Topic - Alphabetical

The following links are either taken from the articles in the *Bulletin* issue or relate closely and fill in helpful information.

- [InterCat Catalog](#)- proof-of-concept database, made of records extracted from OCLC's WorldCat, demonstrating catalog services plus Web access to resources of the Internet
- [International Conf. on Principles and Future Development of AACR](#)- related papers, on Anglo-American Cataloging Rules, and their revision
- [Persistent URLs](#)- PURLs
- [Dublin Core Home Page](#)
- [Dublin Core Elements](#)
- [Dublin Core element Coverage](#) - proposed standard
- [Center for Electronic Text in the Humanities](#)
- [EAD \(Encoded Archival Description\): SGML for Archival Finding Aids](#)
- [UC Berkeley Finding Aids](#)
- [Cataloging Internet Resources: Manual and Practical Guide, by Nancy B. Olson](#)
- OCLC and its [Research Department](#)
- [Stuart Weibel](#)- senior research scientist at OCLC, leader of Dublin Core efforts
- [Workshop on Metadata for Networked Images](#)
- [RDF Home Page](#)- Resource Description Framework, on metadata architecture on the Web
- [UKOLN Metadata Home Page](#)- summary of pubs, projects, metadata resources from UK and beyond, definitions
- [metadata element sets crosswalks](#)- mappings and relationships between various metadata sets, including Dublin Core
- [Resource Discovery project in Australia](#)
- [Dublin Core Workshop, 4th, official report](#) - held at National Library of Australia - and a [light-hearted account](#)
- [National Library of Australia PANDORA Project](#) (Preserving and Accessing Networked Documentary Resources of Australia)
- [In the Company of Strangers: Challenges and Opportunities in Metadata Implementation](#) paper by Maxine Brodie, policy level issues which impact on metadata implementation at the State Library of New South Wales, Sydney, Australia; also [Implications for Metadata Implementation](#)

- [Architecture for Access to Government Information](#) : report, Australia, 1996
- [ERIN - Environmental Resources Information Network](#), Australia - also runs a metadata listserv
- [Core Data Elements for Land and Geographic Directories in Australia and New Zealand](#)
- [Dataset Publishing - A Means to Motivate Metadata Entry, by S.D. Callahan, B.D. Johnson, and E.P. Shelley](#) - Australian Resources, NPI Theory (choice behavior)
- [meta-searcher called HotOIL that accesses both HTTP and Z39.50 servers - demo](#) - translates user requests, merges results, displays summary
- [MetaWeb project](#) - develop and disseminate metadata tools
- [GEM](#) - educational resources - which calls for adding elements like Resource Needed, Standard, Audience, Pedagogy, Quality - see [elements](#)
- [NetFirst](#) - database/directory, cataloging of Internet (uses Dewey)
- [Canadian Information by Subject](#) - info on Canada in Internet (uses Dewey)
- [BUBL Information Service, Scotland, higher education, with subject tree](#) (uses Dewey)
- [Internet Public Library Youth Division](#) (uses Dewey)
- [Blue Web'n, by Pacific Bell, to organize Web sites for students, educators, ...](#) (uses Dewey)
- [Enhancing the indexing vocabulary of DDC by C.J. Godby](#)
- [Scorpion project at OCLC](#)

Acknowledgements

Thanks are given to the authors of the respective articles, from whose contributions the notes above are derived. All distortions of their content and intention are the fault of E. Fox, who apologizes for any misrepresentation inadvertently resulting from this attempt to summarize a valuable set of interesting articles.

- Guest editors' intro. to Special Section, by Efthimis N. Efthimiadis and Allyson Carlyle
- Cataloging Internet Resources: Survey and Prospectus, by Erik Jul
- The Dublin Core: A Simple Content Description Model for Electronic Resources, by Stuart Weibel
- Uniform Resource Identifiers and the Effort to Bring "Bibliographic" Control" to the Web: An Overview of Current Progress, by Ray Schwartz
- Options for Organizing Electronic Resources: The Coexistence of Metadata, by Sherry L. Vellucci
- Metadata in Australia, by Carmel Maguire
- GEM: Using Metadata to Enhance Internet Retrieval by K-12 Teachers, by Stuart Sutton and Sam G. Oh
- From Book Classification to Knowledge Organization: Improving Internet Resource Description and Discovery, by Diane Vizine-Goetz
- Scorpion Helps Catalog the Web, by Keith Shafer

Please follow the above mentioned links to find answers to the following questions:

- What is metadata?
- How many elements are in the Dublin Core?
- What are some new elements added for educators in GEM?
- Describe TEI briefly and explain how it relates to Dublin Core work.
- Explain *finding aid*.
- Describe EAD briefly and explain how it relates to cataloging archival collections.
- Where are their detailed instructions on how to catalog the internet?

- What is RDF?
- What is happening in UK re metadata?
- What mappings are there between metadata representations?
- What is the Resource Discovery project in Australia?
- What happened at the Australian metadata meeting?
- What is covered by the Dublin Core *coverage* element?
- What metadata is needed for geographic information?
- When you search on "digital library" with HotOIL, what refinements are suggested? What are the results of the default processing of your query and what sources were used? Can you find the abstract of a talk on archiving the Internet?
- What WWW search/browse services use Dewey?
- What systems are available to automatically catalog WWW pages?

Electronic Publishing:

- [The SGML/XML Web Page](#)
- [CS5604 unit on SGML](#): check out the related course notes offered at Virginia Tech.
- [Elsevier](#)
[TULIP](#)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Database Issues:

- [UCB database management](#)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

[\[Berkeley\]](#) [\[EECS Dept.\]](#) [\[CS Division\]](#) [\[Database Research\]](#)



Database Research Group papers

Most technical reports for which we have [PostScript](#) and [PDF](#) are also available from the [Berkeley CS-TR server](#) as TIFF images and/or OCR'd text.

Items marked *image PDF* are copyrighted by the Institute of Electrical and Electronic Engineers ([IEEE](#)) or the Association for Computing Machinery ([ACM](#)) and are restricted to use within the University of California. They are here solely as a convenience to local users, for use when [MELVYL](#) is unavailable. If you're not a UC user, don't even try asking for access.

Items marked *Word PS* were originally generated using Microsoft Word. You may have difficulty previewing the resulting PostScript files; if so, try using the regular PS files or the PDF files. (We make the Word PS originals available because they may be better for hardcopy.)

- [UCB CSD Technical Reports](#)
- [UCB ERL Technical Reports](#)
- [LBL Technical Reports](#)
- [Sequoia 2000 Technical Reports](#)
- [UCB Graduate Theses/Reports](#)
- [Wisconsin Technical Reports](#)
- [Other Published Papers](#)
- [Related Papers from Other Groups on Campus](#)

We also have a [Mike Stonebraker bibliography](#). This is a work in progress; additional contributions are [welcome](#), especially for papers and technical reports from the early 1970s.

COPYRIGHT NOTICE:

The documents in these directories have been submitted by their authors to scholarly technical reports series whose purpose is the non-commercial dissemination of scientific work. They are put on-line to facilitate this purpose. These reports are copyrighted by the authors, and their existence in electronic format does not imply that the authors have relinquished any rights. You may copy a report provided that you agree to respect the author's copyright. In particular, a report may be copied for scholarly, non-commercial purposes, such as research or instruction. Reports may not be excerpted unless due acknowledgement is given the author.

Moreover, we do not know what additional arrangements authors may have made concerning these reports. Therefore, in copying a report, you are assuming whatever legal responsibilities copying any document might entail.

Other restrictions to copying individual reports may apply.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

UCB Computer Science Division Technical Reports

More [CS Division](#) technical reports are available from the [Berkeley CS-TR server](#).

CSD-83-144 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Stonebraker, M.; Woodfill, J.; Ranstrom, J.; Kalash, J.; Arnold, K.; Andersen, E.

Performance analysis of distributed data base systems.

Appeared in: Proceedings Third Symposium on Reliability in Distributed Software and Database Systems. (Proceedings Third Symposium on Reliability in Distributed Software and Database Systems, Clearwater Beach, FL, USA, 17-19 Oct. 1983). Silver Spring, MD, USA: IEEE Comput. Soc. Press, Nov. 1983. p. 135-8.

CSD-83-149 [\[CS-TR\]](#)

Stonebraker, M.; Rowe, L.A.

Data base portals: a new application program interface.

Appeared in: Proc. 1984 VLDB Conference.

CSD-83-150 [\[CS-TR\]](#)

Woodfill, J.; Stonebraker, M.

An implementation of hypothetical relations.

Appeared in: Proc. 1983 VLDB Conference.

CSD-83-151 [\[CS-TR\]](#)

Stonebraker, M.; Woodfill, J.; Andersen, E.

Implementation of rules in relational data base systems.

CSD-88-401 [\[CS-TR\]](#)

Butler, Margaret Helen.

Persistent LISP: storing interobject references in a database.

Ph.D. thesis.

CSD-95-876 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Fontaine, A.M.

Sub-element indexing and probabilistic retrieval in the POSTGRES database system.

CSD-96-908 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Hellerstein, J.M.

The case for online aggregation.

Appeared as: Hellerstein, J.M.; Haas, P.J.; Wang, H.J.

Online aggregation.

(SIGMOD 1997. ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13-15 May 1997). SIGMOD Record, June 1997, vol.26, (no.2):171-82. [\[PS\]](#) [\[PDF\]](#)

CSD-97-932 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Woodruff, A.; Stonebraker, M.

Supporting fine-grained data lineage in a database visualization environment.

Appeared in: Proceedings. 13th International Conference on Data Engineering (Cat. No.97CB36038). (Proceedings. 13th International Conference on Data Engineering (Cat. No.97CB36038) Proceedings 13th International Conference on Data Engineering, Birmingham, UK, 7-11 April 1997). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1997. p. 91-102. [\[PS\]](#) [\[PDF\]](#)

CSD-97-950 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Aoki, P.M.

Generalizing ``search'' in generalized search trees.

Appeared in: Proc. 14th Int'l Conf. on Data Engineering, Orlando, FL, Feb. 1998, p. 380-389. [[PS](#)]

[[PDF](#)] **NEW**

CSD-97-968 [[PS](#)] [[PDF](#)] [[CS-TR](#)] **NEW**

Woodruff, A.; Stonebraker, M.

Visual information density adjuster (VIDA).

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

UCB Electronics Research Laboratory Technical Reports

More [Electronics Research Laboratory](#) technical reports are available from the [Berkeley CS-TR server](#).

ERL-M83-74 [[PS](#)] [[PDF](#)]

Stonebraker, M.

Virtual memory transaction management.

Appeared in: Operating Systems Review, April 1984, vol.18, (no.2):8-16.

ERL-M84-58 [[PS](#)] [[PDF](#)]

Kung, R.M.; Hanson, E.; Ioannidis, Y.; Sellis, T.; Shapiro, L.; Stonebraker, M.

Heuristic search in data base systems.

Appeared in: Expert database systems : proceedings from the first international workshop / Larry Kerschberg, editor. Menlo Park, Calif.: Benjamin/Cummings Pub. Co., c1986, p. 537-548.

ERL-M84-87 [[PS](#)] [[PDF](#)]

Stonebraker, M.; DuBourdieu, D.; Edwards, W.

Problems in supporting database transactions in an operating system transaction manager.

Appeared in: Operating Systems Review, Jan. 1985, vol.19, (no.1):6-14.

ERL-M85-59 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Anton, J.; Hanson, E.

Extending a database system with procedures.

Appeared in: ACM Transactions on Database Systems, Sept. 1987, vol.12, (no.3):350-76.

ERL-M85-67 [[PS](#)] [[PDF](#)]

Stonebraker, M.

Inclusion of new types in relational data base systems.

Appeared in: International Conference on Data Engineering (Cat. No.86CH2261-6). (International Conference on Data Engineering (Cat. No.86CH2261-6), Los Angeles, CA, USA, 5-7 Feb. 1986). Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 262-9.

ERL-M85-95 [[PS](#)] [[PDF](#)] (**Warning: missing figures.**)

Stonebraker, M.; Rowe, L.A.

The design of POSTGRES.

Appeared in: (Proceedings of ACM SIGMOD '86. International Conference on Management of Data, Washington, DC, USA, 28-30 May 1986). SIGMOD Record, June 1986, vol.15, (no.2):340-55.

ERL-M86-06 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Sellis, T.; Hanson, E.

An analysis of rule indexing implementations in data base systems.

Appeared in: Proceedings from the First International Conference on Expert Database Systems.

(Proceedings from the First International Conference on Expert Database Systems, Charleston, SC, USA, 1-4 April 1986). Edited by: Kerschberg, L. Menlo Park, CA, USA:

Benjamin/Cummings, 1987. p. 465-76.

ERL-M86-59 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Object management in POSTGRES using procedures.

Appeared in: Proceedings of the 1986 International Workshop on Object-Oriented Database Systems (Cat. No.86TH0161-0). (Proceedings of the 1986 International Workshop on Object-Oriented Database Systems (Cat. No.86TH0161-0), Pacific Grove, CA, USA, 23-26 Sept. 1986). Edited by: Dittrich, K.; Dayal, U. Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 66-72.

ERL-M87-06 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

The design of the POSTGRES storage system.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 289-300.

ERL-M87-13 [\[PS\]](#) [\[PDF\]](#)

Rowe, L.A.; Stonebraker, M.R.

The POSTGRES data model.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 83-96.

ERL-M87-15 [\[PS\]](#) [\[PDF\]](#) (Warning: missing figures.)

Kumar, A.; Stonebraker, M.

Performance evaluation of an operating system transaction manager.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 473-81.

Appeared as: **Performance considerations for an operating system transaction manager.** IEEE Transactions on Software Engineering, June 1989, vol.15, (no.6):705-14. [\[image PDF\]](#)

ERL-M88-19 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Katz, R.; Patterson, D.; Ousterhout, J.

The design of XPRS.

Appeared in: Proceedings of the Fourteenth International Conference on Very Large Databases. (Proceedings of the Fourteenth International Conference on Very Large Databases, Los Angeles, CA, USA, 29 Aug.-1 Sept. 1988). Edited by: Bancilhon, F.; DeWitt, D.J. Palo Alto, CA, USA: Morgan Kaufmann, 1988. p. 318-30.

ERL-M88-07 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Future trends in database systems.

Appeared in: IEEE Transactions on Knowledge and Data Engineering, March 1989, vol.1, (no.1):33-44. [\[image PDF\]](#)

Appeared in: Proceedings Fourth International Conference on Data Engineering (Cat. No.88CH2550-2). (Proceedings Fourth International Conference on Data Engineering (Cat. No.88CH2550-2), Los Angeles, CA, USA, 1-5 Feb. 1988). Washington, DC, USA: IEEE Comput.

Soc. Press, 1988. p. 222-31.

ERL-M89-16 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Aoki, P.; Seltzer, M.

Parallelism in XPRS.

ERL-M89-17 [[PS](#)] [[PDF](#)]

Stonebraker, M.

The case for partial indexes.

Appeared in: SIGMOD Record, Dec. 1989, vol.18, (no.4):4-11.

ERL-M89-56 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Schloss, G.A.

Distributed RAID-a new multiple copy algorithm.

Appeared in: Sixth International Conference on Data Engineering (Cat. No.90CH2840-7). (Sixth International Conference on Data Engineering (Cat. No.90CH2840-7), Los Angeles, CA, USA, 5-9 Feb. 1990). Los Alamitos, CA, USA: IEEE Comput. Soc, 1990. p. 430-7. *Appeared as:* Schloss, G.A.; Stonebraker, M.

Highly redundant management of distributed data.

Proceedings. Workshop on the Management of Replicated Data (Cat. No.90TH0329-3), (Proceedings. Workshop on the Management of Replicated Data (Cat. No.90TH0329-3), Houston, TX, USA, 8-9 Nov. 1990.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1990. p.91-5.

[[image PDF](#)]

ERL-M89-82 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Hearst, M.; Potamianos, S.

A commentary on the POSTGRES rules system.

Appeared in: SIGMOD Record, Sept. 1989, vol.18, (no.3):5-11.

ERL-M90-11 [[CS-TR](#)]

Sullivan, M.; Stonebraker, M.

Improving software fault tolerance in highly available database systems.

ERL-M90-28 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Stonebraker, M.; Rowe, L.A.; Lindsay, B.; Gray, J.; Carey, M.; Brodie, M.; Bernstein, P.; Beech, D. (as ``The Committee for Advanced Database Function")

Third-generation database system manifesto.

Appeared in: SIGMOD Record, Sept. 1990, vol.19, (no.3):31-44.

Appeared in: Object-Oriented Databases: Analysis, Design and Construction (DS-4). Proceedings of the IFIP TC2/WG 2.6 Working Conference. (Object-Oriented Databases: Analysis, Design and Construction (DS-4). Proceedings of the IFIP TC2/WG 2.6 Working Conference, Windermere, UK, 2-6 July 1990). Edited by: Meersman, R.A.; Kent, W.; Khosla, S. Amsterdam, Netherlands: North-Holland, 1991. p. 495-511.

Appeared in: Computer Standards & Interfaces, Oct. 1991, vol.13, (no.1-3):41-54.

ERL-M90-34 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Rowe, L.A.; Hirohama, M.

The implementation of POSTGRES.

Appeared in: IEEE Transactions on Knowledge and Data Engineering, March 1990, vol.2, (no.1):125-42. [[image PDF](#)]

ERL-M90-36 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Stonebraker, M.; Jhingran, A.; Goh, J.; Potamianos, S.

On rules, procedures, caching and views in database systems.

Appeared in: (1990 ACM SIGMOD International Conference on Management of Data, Atlantic City, NJ, USA, 23-25 May 1990). SIGMOD Record, June 1990, vol.19, (no.2):281-90.

ERL-M91-50 [[CS-TR](#)]

Hong, W.; Stonebraker, M.

Optimization of parallel query execution plans in XPRS.

Appeared in: Proceedings of the First International Conference on Parallel and Distributed Information Systems (Cat. No.91TH0393-4), (Proceedings of the First International Conference on Parallel and Distributed Information Systems (Cat. No.91TH0393-4), Miami Beach, FL, USA, 4-6 Dec. 1991.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1991. p.218-25. [[image](#) [PDF](#)]

Appeared in: Distributed and Parallel Databases, Jan. 1993, vol.1, (no.1):9-32.

ERL-M91-51 [[CS-TR](#)]

Ong, Lay-peng.

Version modeling using production rules in the POSTGRES DBMS.

M.S. report.

ERL-M91-52 [[CS-TR](#)]

Goh, Khoon-San Jeffrey.

Rule processing with query rewrite.

M.S. report.

ERL-M91-56 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Sullivan, M.; Stonebraker, M.

Using write protected data structures to improve software fault tolerance in highly available database management systems.

Appeared in: Proceedings of the Seventeenth International Conference on Very Large Data Bases. (Proceedings of the Seventeenth International Conference on Very Large Data Bases, Barcelona, Spain, 3-6 Sept. 1991). Edited by: Lohman, G.M.; Sernadas, A.; Camps, R. San Mateo, CA, USA: Morgan Kaufmann, 1991. p. 171-80.

ERL-M91-62 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Kemnitz, G.

The POSTGRES next-generation database management system.

Appeared in: Communications of the ACM, Oct. 1991, vol.34, (no.10):78-92.

ERL-M92-02 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Seltzer, M.; Olson, M.

LIBTP: portable, modular transactions for Unix.

Appeared in: Proceedings of the Winter 1992 USENIX Conference. (Proceedings of the Winter 1992 USENIX Conference, San Francisco, CA, USA, 20-24 Jan. 1992). Berkeley, CA, USA: USENIX, 1991. p. 9-25.

ERL-M93-01 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Seltzer, Margo Ilene.

File system performance and transaction support.

Ph.D. thesis.

ERL-M93-05 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Sullivan, Mark Paul.

System support for software fault tolerance in highly available database management systems.

Ph.D. thesis.

ERL-M93-22 [[PS](#)] [[PDF](#)] [[CS-TR](#)]

Stonebraker, M.; Olson, M.

Large object support in POSTGRES.

Appeared in: Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1). (Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1) Proceedings of IEEE 9th International Conference on Data Engineering, Vienna, Austria, 19-23 April 1993). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p. 355-62.

[\[image PDF\]](#)

ERL-M93-25 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Stonebraker, M.

The integration of rule systems and database systems.

Appeared in: IEEE Transactions on Knowledge and Data Engineering, Oct. 1992, vol.4, (no.5):415-23. [\[image PDF\]](#)

ERL-M93-28 [\[PS\]](#) [\[PDF\]](#) [\[CS-TR\]](#)

Hong, Wei.

Parallel query processing using shared memory multiprocessors and disk arrays.

Ph.D. thesis.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

Lawrence Berkeley Laboratory Technical Reports

LBL-TR-32883 [\[PS\]](#) [\[PDF\]](#)

Olken, F.

Random sampling from databases.

Ph.D. thesis.

LBL-TR-34229 [\[PS\]](#) [\[PDF\]](#)

Chandra, R.; Segev, A.; Stonebraker, M.

Implementing calendars and temporal rules in next generation databases.

Appeared in: Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7). (Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7)Proceedings of 1994 IEEE 10th International Conference on Data Engineering, Houston, TX, USA, 14-18 Feb. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 264-73. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

Sequoia 2000 Technical Reports

More [Sequoia 2000](#) technical reports (e.g., those on topics other than databases) are available [elsewhere on this server](#) and from the [Berkeley CS-TR server](#).

S2K-91-01 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Dozier, J.

Sequoia 2000: large capacity object servers to support global change research.

S2K-91-04 [\[PS\]](#) [\[PDF\]](#)

Chen, J.; Larson, R.; Stonebraker, M.

The Sequoia 2000 object browser.

Appeared in: Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1). (Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1), San Francisco, CA, USA, 24-28 Feb. 1992). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p. 389-94. [\[image PDF\]](#)

S2K-91-05 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

An overview of the Sequoia 2000 project.

Appeared in: Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1), (Digest of Papers. COMPCON Spring 1992. Thirty-Seventh IEEE Computer Society International Conference (Cat. No.92CH3098-1), San Francisco, CA, USA, 24-28 Feb. 1992.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p.383-8. [[image](#) [PDF](#)]

S2K-92-12 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Frew, J.; Gardels, K.; Meredith, J.

The SEQUOIA 2000 storage benchmark.

Appeared in: (SIGMOD '93. 1993 ACM SIGMOD. International Conference on Management of Data, Washington, DC, USA, 26-28 May 1993). SIGMOD Record, June 1993, vol.22, (no.2):2-11. [[image](#) [PDF](#)]

S2K-92-13 [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Stonebraker, M.

Predicate migration: optimizing queries with expensive predicates.

Appeared in: (SIGMOD '93. 1993 ACM SIGMOD. International Conference on Management of Data, Washington, DC, USA, 26-28 May 1993). SIGMOD Record, June 1993, vol.22, (no.2):267-77. [[PS](#)] [[PDF](#)] [[image](#) [PDF](#)]

S2K-92-16 [[PS](#)] [[PDF](#)]

Kohl, J.T.; Staelin, C.; Stonebraker, M.

HighLight: using a log-structured file system for tertiary storage management.

Appeared in: USENIX Association. Proceedings of the Winter 1993 USENIX Conference. (USENIX Association. Proceedings of the Winter 1993 USENIX Conference, San Diego, CA, USA, 25-29 Jan. 1993). Berkley, CA, USA: USENIX Assoc, 1993. p. 435-47.

Appeared as: Kohl, J.; Stonebraker, M.; Staelin, C.

HighLight: a file system for tertiary storage.

Proceedings Twelfth IEEE Symposium on Mass Storage Systems. Putting all that Data to Work (Cat. No.93CH3246-6), (Proceedings Twelfth IEEE Symposium on Mass Storage Systems. Putting all that Data to Work (Cat. No.93CH3246-6), Proceedings of 12th IEEE Symposium on Mass Storage Systems, Monterey, CA, USA, 26-29 April 1993.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p.157-61. [[image](#) [PDF](#)]

S2K-92-20 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Chen, J.; Nathan, N.; Paxson, C.

Tioga: providing data management support for scientific visualization applications.

S2K-93-23 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Frew, J.; Dozier, J.

The Sequoia 2000 architecture and implementation strategy.

S2K-93-25 [[PS](#)] [[PDF](#)]

Brodie, M.; Stonebraker, M.

DARWIN: on the incremental migration of legacy information systems.

S2K-93-28 [[PS](#)] [[PDF](#)]

Olson, M.A.

The design and implementation of the Inversion file system.

Appeared in: USENIX Association. Proceedings of the Winter 1993 USENIX Conference. (USENIX Association. Proceedings of the Winter 1993 USENIX Conference, San Diego, CA, USA, 25-29 Jan. 1993). Berkley, CA, USA: USENIX Assoc, 1993. p. 205-17.

S2K-93-29 [[PS](#)] [[PDF](#)]

Stonebraker, M.; Chen, J.; Nathan, N.; Paxson, C.; Wu, J.

Tioga: providing data management support for scientific visualization applications.

Appeared in: 19th International Conference on Very Large Data Bases Proceedings. (19th International Conference on Very Large Data Bases Proceedings Proceeding of 19th International Conference on Very Large Data Bases, Dublin, Ireland, 24-27 Aug. 1993). Edited by: Agrawal, R.; Baken, S.; Bell, D. Palo Alto, CA, USA: Morgan Kaufmann Publishers, 1993. p. 25-38. [\[PS\]](#) [\[PDF\]](#)

Appeared as: **Tioga: A database-oriented visualization tool.** Proceedings Visualization '93. (Cat. No.93CH3354-8). (Proceedings Visualization '93. (Cat. No.93CH3354-8) Proceedings Visualization '93, San Jose, CA, USA, 25-29 Oct. 1993). Edited by: Nielson, G.M.; Bergeron, D. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p. 86-93. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

S2K-93-30 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Olson, M.

Large object support in POSTGRES.

Appeared in: Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1). (Proceedings. Ninth International Conference on Data Engineering (Cat. No.92CH3258-1) Proceedings of IEEE 9th International Conference on Data Engineering, Vienna, Austria, 19-23 April 1993). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1993. p. 355-62. [\[image PDF\]](#)

S2K-93-31 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Aoki, P.M.; Devine, R.; Litwin, W.; Olson, M.

Mariposa: a new architecture for distributed data.

Appeared in: Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7). (Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7) Proceedings of 1994 IEEE 10th International Conference on Data Engineering, Houston, TX, USA, 14-18 Feb. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 54-65. [\[PS\]](#) [\[PDF\]](#)

S2K-93-32 [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.; Stonebraker, M.

Efficient organization of large multidimensional arrays.

Appeared in: Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7). (Proceedings. The 10th International Conference Data Engineering (Cat. No.94CH3383-7) Proceedings of 1994 IEEE 10th International Conference on Data Engineering, Houston, TX, USA, 14-18 Feb. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 328-36. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

S2K-93-38 [\[PS\]](#) [\[PDF\]](#)

Chen, J.; Aiken, A.; Nathan, N.; Paxson, C.; Stonebraker, M.; Wu, J.

Extending a graphical query language to support updates, foreign systems, and transactions.

S2K-94-41 [\[PS\]](#) [\[PDF\]](#)

Woodruff, A.G.; Plaunt, C.

GIPSY: georeferenced information processing system.

Appeared as: **GIPSY: automated geographic indexing of text documents.**

Journal of the American Society for Information Science, Oct. 1994, vol.45, (no.9):645-55.

S2K-94-45 [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.; Stonebraker, M.

Single query optimization for tertiary memory.

S2K-94-48 [\[PS\]](#) [\[PDF\]](#)

Woodruff, A.; Wisnovsky, P.; Taylor, C.; Stonebraker, M.; Paxson, C.; Chen, J.; Aiken, A.

Zooming and tunneling in Tioga: supporting navigation in multidimensional space.

Appeared (extended abstract) in: Proceedings. IEEE Symposium on Visual Languages (Cat.

No.94TH8010). (Proceedings. IEEE Symposium on Visual Languages (Cat. No.94TH8010)Proceedings of 1994 IEEE Symposium on Visual Languages, St. Louis, MO, USA, 4-7 Oct. 1994). Edited by: Ambler, A.L.; Kimura, T.D. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 191-3. [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)
Appeared in: Woodruff, A.; Su, A.; Stonebraker, M.; Paxson, C.; Chen, J.; Aiken, A.; Wisnovsky, P.; Taylor, C.

Navigation and coordination primitives for multidimensional visual browsers.

Visual Database Systems 3. Visual Information Management. Proceedings of the Third IFIP 2.6 Working Conference on Visual Database Systems, 1995. (Visual Database Systems 3. Visual Information Management. Proceedings of the Third IFIP 2.6 Working Conference on Visual Database Systems, 1995Proceedings IFIP 2.6 3rd Working Conference on Visual Database Systems (VDB-3), Lausanne, Switzerland, 27-29 March 1995). Edited by: Spaccapietra, S.; Jain, R. London, UK: Chapman and Hall, 1995. p. 360-71. [\[PS\]](#) [\[Word PS\]](#) [\[PDF\]](#)

S2K-94-49 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.; Devine, R.; Kornacker, M.; Litwin, W.; Pfeffer, A.; Sah, A.; Staelin, C.

An economic paradigm for query processing and data migration in Mariposa.

Appeared in: Proceedings of the Third International Conference on Parallel and Distributed Information Systems (Cat. No.94TH0668-4). (Proceedings of the Third International Conference on Parallel and Distributed Information Systems (Cat. No.94TH0668-4)Proceedings of 3rd International Conference on Parallel and Distributed Information Systems, Austin, TX, USA, 28-30 Sept. 1994). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 58-67. [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

S2K-94-50 [\[PS\]](#) [\[PDF\]](#)

Devine, R.

Design and implementation of DDH: a distributed dynamic hashing algorithm.

Appeared in: Foundations of Data Organization and Algorithms. 4th International Conference. FODO '93 Proceedings. (Foundations of Data Organization and Algorithms. 4th International Conference. FODO '93 Proceedings, Chigago, IL, USA, 13-15 Oct. 1993). Edited by: Lomet, D.B. Berlin, Germany: Springer-Verlag, 1993. p. 101-14.

S2K-94-56 [\[PS\]](#) [\[PDF\]](#)

Banks, D.; Kornacker, M.; Stonebraker, M.

High-concurrency locking in R-trees.

Appeared as: Kornacker, M.; Banks, D.

High-concurrency locking in R-trees.

Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95. Proceedings of the 21st International Conference on Very Large Data BasesProceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 134-45. [\[PS\]](#) [\[PDF\]](#)

S2K-94-58 [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Sequoia 2000-a reflection on the first three years.

Appeared in: Proceedings. Seventh International Working Conference on Scientific and Statistical Database Management (Cat. No.94TH0689-0). (Proceedings. Seventh International Working Conference on Scientific and Statistical Database Management (Cat. No.94TH0689-0)Seventh International Working Conference on Scientific and Statistical Database Management, Charlottesville, VA, USA, 28-30 Sept. 1994). Edited by: French, J.C.; Hinterberger, H. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 108-16. [\[image PDF\]](#)

Appeared in: **Sequoia 2000: a reflection on the first three years.**

- IEEE Computational Science and Engineering, Winter 1994, vol.1, (no.4):63-72. [[image](#) [PDF](#)]
 S2K-94-59 [[PS](#)] [[PDF](#)]
 Anderson, J.T.; Stonebraker, M.
Sequoia 2000 metadata schema for satellite images.
Appeared in: SIGMOD Record, Dec. 1994, vol.23, (no.4):42-8.
- S2K-95-60 [[PS](#)] [[Word PS](#)] [[PDF](#)]
 Woodruff, A.; Stonebraker, M.
Buffering of intermediate results in dataflow diagrams.
Appeared in: Proceedings. 11th IEEE International Symposium on Visual Languages (Cat. No.95TB8105). (Proceedings. 11th IEEE International Symposium on Visual Languages (Cat. No.95TB8105) Proceedings of Symposium on Visual Languages, Darmstadt, Germany, 5-9 Sept. 1995). Edited by: Haarslev, V. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1995. p. 187-94. [[PS](#)] [[PDF](#)] [[image PDF](#)]
- S2K-95-61 [[MS Word](#)]
 Davis, F.; Farrell, W.; Gray, J.; Mechoso, R.; Moore, R.; Sides, S.; Stonebraker, M.
EOSDIS alternative architecture.
- S2K-95-62 [[PS](#)] [[PDF](#)]
 Sidell, J.; Aoki, P.M.; Barr, S.; Sah, A.; Staelin, C.; Stonebraker, M.
Data replication in Mariposa.
Appeared in: Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888). (Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888) Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, LA, USA, 26 Feb.-1 March 1996). Edited by: Su, S.Y.W. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1996. p. 485-94. [[PS](#)] [[PDF](#)] [[image PDF](#)]
- S2K-95-63 [[PS](#)] [[PDF](#)]
 Stonebraker, M.; Aoki, P.M. Pfeiffer, A.; Sah, A.; Sidell, J.; Staelin, C.; Yu, A.
Mariposa: a wide-area distributed database system.
Appeared in: VLDB Journal 5(1), Jan. 1996, p. 48-63. [[PS](#)] [[PDF](#)]
- S2K-95-64 [[PS](#)] [[PDF](#)]
 Aiken, A.; Chen, J.; Stonebraker, M.; Woodruff, A.
Tioga-2: a direct manipulation database visualization environment.
Appeared in: Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888). (Proceedings of the Twelfth International Conference on Data Engineering (Cat. No.96CB35888) Proceedings of the Twelfth International Conference on Data Engineering, New Orleans, LA, USA, 26 Feb.-1 March 1996). Edited by: Su, S.Y.W. Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1996. p. 208-17. [[PS](#)] [[PDF](#)] [[image PDF](#)]
- S2K-95-65 [[PS](#)] [[PDF](#)]
 Brown, P.; Stonebraker, M.
BigSur: a system for the management of earth science data.
Appeared in: Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95. Proceedings of the 21st International Conference on Very Large Data Bases Proceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 720-28.
- S2K-95-66 [[PS](#)] [[PDF](#)] (**Revised** [[PS](#)] [[PDF](#)])
 Aoki, P.M.
Recycling secondary index structures.
-

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

UCB Graduate Theses and Reports

UCB-MS-aoki [\[PS\]](#) [\[PDF\]](#)

Aoki, Paul Masami.

Query processing techniques in XPRS.

M.S. thesis.

UCB-MS-devine [\[tar\]](#)

Devine, Robert J.

Design of Eureka, an extensible query optimizer.

M.S. report. *Contact author to obtain a copy.*

UCB-MS-ginger [\[PS\]](#) [\[PDF\]](#)

Ogle, Virginia E.

Chabot: a system for retrieval from a relational database of images.

M.S. report.

UCB-MS-jtkohl [\[PS\]](#) [\[PDF\]](#)

Kohl, John T.

HighLight: using a log-structured file system for tertiary storage management.

M.S. report.

UCB-MS-mao [\[PS\]](#) [\[PDF\]](#)

Olson, Michael Allen.

Extending the POSTGRES database system to manage tertiary storage.

M.S. thesis.

UCB-MS-paxson [\[tar\]](#)

Paxson, Caroline Marie.

Design and implementation of sets in POSTGRES

M.S. report. *Contact author to obtain a copy.*

UCB-MS-plai [\[PS\]](#) [\[PDF\]](#)

Lai, Peter K.

Analyzing and improving the performance of POSTGRES.

M.S. report.

UCB-MS-sunita [\[tar\]](#)

Sarawagi, Sunita.

Efficient organization of large multidimensional arrays.

M.S. report.

UCB-MS-zfong [\[PS\]](#) [\[PDF\]](#)

Fong, Zelaine.

The design and implementation of the POSTGRES query optimizer.

M.S. report.

UCB-PhD-butler [\[CS-TR\]](#)

Butler, Margaret Helen.

Persistent LISP: storing interobject references in a database.

Ph.D. thesis.

UCB-PhD-hong [\[PS\]](#) [\[PDF\]](#)

Hong, Wei.

Parallel query processing using shared memory multiprocessors and disk arrays.

Ph.D. thesis.

UCB-PhD-olken [\[PS\]](#) [\[PDF\]](#)

Olken, F.

Random sampling from databases.

Ph.D. thesis.

UCB-PhD-seltzer [[PS](#)] [[PDF](#)]

Seltzer, Margo Ilene.

File system performance and transaction support.

Ph.D. thesis.

UCB-PhD-sullivan [[PS](#)] [[PDF](#)]

Sullivan, Mark Paul.

System support for software fault tolerance in highly available database management systems.

Ph.D. thesis.

UCB-PhD-sunita [[PS](#)] [[PDF](#)]

Sarawagi, Sunita.

Query processing in tertiary memory databases.

Ph.D. thesis.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

Univ. of Wisconsin, Madison Computer Science Technical Reports

More [Wisconsin CS](#) technical reports are available from the [Wisconsin CS-TR server](#).

UW-CS-TR-1252 [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Pfeffer, A.

The RD-tree: an index structure for sets.

UW-CS-TR-1274 [[PS](#)] [[PDF](#)]

Hellerstein, J.M.; Naughton, J.F.; Pfeffer, A.

Generalized search trees for database systems.

Appeared in: Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95. Proceedings of the 21st International Conference on Very Large Data Bases Proceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 562-73. [[PS](#)] [[PDF](#)]

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

Other Published Papers

avi98-density [[PS](#)] [[PDF](#)] **NEW**

Woodruff, A.; Landay, J.; Stonebraker, M.

Constant information density in zoomable interfaces.

Proc. Int'l Working Conf. on Advanced Visual Interfaces (to appear).

cacm91-opp [[image PDF](#)]

Silberschatz, A.; Stonebraker, M.; Ullman, J.

Database systems: achievements and opportunities.

Communications of the ACM, Oct. 1991, vol.34, (no.10):110-20.

chi98-zoom [\[PS\]](#) [\[PDF\]](#) **NEW**

Woodruff, A.; Landay, J.; Stonebraker, M.

Goal-directed zoom.

Proc. ACM SIGCHI Conf. on Human Factors in Computing (to appear).

compcon86-object [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Object management in a relational data base system.

Digest of Papers. COMPCON Spring 86. Thirty-First IEEE Computer Society International Conference (Cat. No.86CH2285-5). (Digest of Papers. COMPCON Spring 86. Thirty-First IEEE Computer Society International Conference (Cat. No.86CH2285-5), San Francisco, CA, USA, 3-6 March 1986). Edited by: Bell, A.G. Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 336-41.

comp95-chabot [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Ogle, V.E.; Stonebraker, M.

Chabot: retrieval from a relational database of images.

IEEE Computer, Sep. 1995, vol.28, (no.9):40-48.

debull93-s2k [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

The Sequoia 2000 project.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Mar. 1993, vol. 16 (no.1):24-28.

debull96-ordbms [\[PS\]](#) [\[PDF\]](#)

Olson, M.A.; Hong, W.M.; Ubell, M.; Stonebraker, M.

Query processing in a parallel object-relational database.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Dec. 1996, vol. 19 (no.4):3-10.

debull97-online [\[PS\]](#) [\[PDF\]](#) **NEW**

Hellerstein, J.M.

Online processing redux.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Sep. 1997, vol. 20 (no. 3): 20-29.

debull97-reduction [\[PS\]](#) [\[PDF\]](#) **NEW**

Barbara, D.; DuMouchel, W.; Faloutsos, C.; Haas, P.J.; Hellerstein, J.M.; Ioannidis, Y.; Jagadish, H.V.; Johnson, T.; Ng, R.; Poosala, V.; Ross, K.A.; Sevcik, K.C.

The New Jersey data reduction report.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Sep. 1997, vol. 20 (no. 4): 3-45.

dtj95-s2k [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

An overview of the Sequoia 2000 project.

Digital Technical Journal, 1995, vol.7, (no.3):39-49.

dtj95-repo [\[PS\]](#) [\[PDF\]](#)

Larson, R.R.; Plaunt, C.; Woodruff, A.G.; Hearst, M.A.

The Sequoia 2000 electronic repository.

Digital Technical Journal, 1995, vol.7, (no.3):50-65.

ftc92-dbos [\[image PDF\]](#)

Sullivan, M.; Chillarege, R.

A comparison of software defects in database management systems and operating systems.

Digest of Papers. FTCS-22. The Twenty-Second International Symposium on Fault-Tolerant

Computing (Cat. No.92CH3155-9), (Digest of Papers. FTCS-22. The Twenty-Second International Symposium on Fault-Tolerant Computing (Cat. No.92CH3155-9), Boston, MA, USA, 8-10 July 1992.) New York, NY, USA: IEEE, 1992. p.475-84.

ftps92-dbos [[image PDF](#)]

Sullivan, M.; Chillarege, R.

A comparison of software defects in database management systems and operating systems.

Digest of Papers. The 1992 IEEE Workshop on Fault-Tolerant Parallel and Distributed Systems (Cat. No.92TH0449-9), (Digest of Papers. The 1992 IEEE Workshop on Fault-Tolerant Parallel and Distributed Systems (Cat. No.92TH0449-9), Amherst, MA, USA, 6-7 July 1992.) Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p.475-84.

hpca98-io [[PS](#)] [[PDF](#)] NEW

Arpaci-Dusseau, R.H.; Arpaci-Dusseau, A.C.; Culler, D.E.; Hellerstein, J.M.; Patterson, D.A.

The architectural costs of streaming I/O: a comparison of workstations, clusters, and SMPs.

Proc. 1998 Symp. on High Performance Computer Architecture (to appear).

hpts85-nothing [[PS](#)] [[PDF](#)]

Stonebraker, M.

The case for shared nothing.

Proc. 1985 Symp. on High Performance Transaction Systems.

icde92-nobtree [[PS](#)] [[PDF](#)]

Sullivan, M.; Olson, M.

An index implementation supporting fast recovery for the POSTGRES storage system.

Eighth International Conference on Data Engineering (Cat. No.92CH3097-3). (Eighth International Conference on Data Engineering (Cat. No.92CH3097-3), Tempe, AZ, USA, 2-3 Feb. 1992). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p. 293-300. [[image PDF](#)]

info97-greatwave [[HTML](#)]

Stonebraker, M.

Object-relational DBMS: the next great wave.

Informix white paper.

info97-middleware [[HTML](#)]

Stonebraker, M.; Brown, P.

Objects in middleware: how bad can it be?

Informix white paper.

info97-options [[PDF](#)]

Stonebraker, M.

Architectural options for object-relational DBMSs.

Informix document 000-21460-70, Feb. 1997.

info97-simulating [[PDF](#)]

Stonebraker, M.

Performance penalties for simulating object-relational DBMSs.

Informix document 000-21451-70, Feb. 1997.

mss94-s2k [[image PDF](#)]

Dozier, J.; Stonebraker, M.; Frew, J.

Sequoia 2000: a next-generation information system for the study of global change.

Proceedings Thirteenth IEEE Symposium on Mass Storage Systems. Towards Distributed Storage and Data Management Systems. First International Symposium (Cat. No.94CH3457-9).

(Proceedings Thirteenth IEEE Symposium on Mass Storage Systems. Towards Distributed Storage and Data Management Systems. First International Symposium (Cat.

No.94CH3457-9)Proceedings Thirteenth IEEE Symposium on Mass Storage Systems. Toward Distributed Storage and Data Management Systems, Annecy, France, 12-16 June 1994). Los


Alamitos, CA, USA: IEEE Comput. Soc. Press, 1994. p. 47-53.

mss95-tert [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Sarawagi, S.

Database systems for efficient access to tertiary memory.

Proceedings of the Fourteenth IEEE Symposium on Mass Storage Systems. Storage - At the Forefront of Information Infrastructures (Cat. No.95CB35860). (Proceedings of the Fourteenth IEEE Symposium on Mass Storage Systems. Storage - At the Forefront of Information Infrastructures (Cat. No.95CB35860) Proceedings of IEEE 14th Symposium on Mass Storage Systems, Monterey, CA, USA, 11-14 Sept. 1995). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1995. p. 120-6.

ngits97-control [\[PS\]](#) [\[PDF\]](#) 

Hellerstein, J.M.

Towards a crystal ball for data retrieval.

The Third International Workshop on Next Generation Information Technologies and Systems, Neve Ilan, Israel, July 1997.

pod97-index [\[PS\]](#) [\[PDF\]](#)

Hellerstein, J.M.; Koutsoupias, E.; Papadimitriou, C.H.

On the analysis of indexing schemes.

1997 PODS Conference, 249-256.

sigirf91-index [\[PS\]](#) [\[PDF\]](#)

Aoki, P.M.

Implementation of extended indexes in POSTGRES.

SIGIR Forum, Spring 1991, vol.25, (no.1):2-9.

sigmod81-views.ps [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Hypothetical data bases as views.

Proc. 1981 SIGMOD Conf.

sigmod91-multilevel [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Stonebraker, M.

Managing persistent objects in a multi-level store.

(1991 ACM SIGMOD International Conference on Management of Data, Denver, CO, USA, 29-31 May 1991). SIGMOD Record, June 1991, vol.20, (no.2):2-11.

sigmod91-segment [\[image PDF\]](#)

Kolovson, C.P.; Stonebraker, M.

Segment indexes: dynamic indexing techniques for multi-dimensional interval data.

(1991 ACM SIGMOD International Conference on Management of Data, Denver, CO, USA, 29-31 May 1991). SIGMOD Record, June 1991, vol.20, (no.2):138-47.

sigmod93-miro [\[image PDF\]](#)

Stonebraker, M.

The Miro DBMS.

(SIGMOD '93. 1993 ACM SIGMOD. International Conference on Management of Data, Washington, DC, USA, 26-28 May 1993). SIGMOD Record, June 1993, vol.22, (no.2):439.

sigmod96-magic [\[PS\]](#) [\[PDF\]](#)

Seshadri, P.; Hellerstein, J.M.; Leung, T.Y.C.; Pirahesh, H.; Ramakrishnan, R.; Srivastava, D.; Stuckey, P.J.; Sudarshan, S..

Cost-based optimization for magic: algebra and implementation.

(1996 ACM SIGMOD International Conference on Management of Data, Montreal, Que., Canada, 4-6 June 1996.) SIGMOD Record, June 1996, vol.25, (no.2):435-46.

sigmod97-nowsort [\[PS\]](#) [\[PDF\]](#)

Arpaci-Dusseau, A.C.; Arpaci-Dusseau, R.H.; Culler, D.E.; Hellerstein, J.M.; Patterson, D.A.

High-performance sorting on networks of workstations.

(SIGMOD 1997. ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13-15 May 1997). SIGMOD Record, June 1997, vol.26, (no.2):243-54.

sigmod97-gist [\[PS\]](#) [\[PDF\]](#)

Kornacker, M.; Mohan, C.; Hellerstein, J.M.

Concurrency and recovery in generalized search trees.

(SIGMOD 1997. ACM SIGMOD International Conference on Management of Data, Tucson, AZ, USA, 13-15 May 1997). SIGMOD Record, June 1997, vol.26, (no.2):62-72.

sigmodr87-selec [\[PS\]](#) [\[PDF\]](#)

Kumar, A.; Stonebraker, M.

The effect of join selectivities on optimal nesting order.

SIGMOD Record, March 1987, vol.16, (no.1):28-41.

sigmodr90-ucb [\[PS\]](#) [\[PDF\]](#)

Stonebraker, M.

Data base research at Berkeley.

SIGMOD Record, Dec. 1990, vol.19, (no.4):113-18.

sigmodr94-industry [\[PS\]](#) [\[PDF\]](#)

Blakeley, J.A.; Fishman, D.; Lomet, D.; Stonebraker, M.; Barbara, D.

The impact of database research on industrial products. (panel summary)

SIGMOD Record, Sept. 1994, vol.23, (no.3):35-40.

tods98-xfunc [\[PS\]](#) [\[PDF\]](#) NEW

Hellerstein, J.M.

Optimization techniques for queries with expensive methods.

ACM Transactions on Database Systems, to appear.

tse88-rulemgr [\[PS\]](#) [\[PDF\]](#) [\[image PDF\]](#)

Stonebraker, M.; Hanson, E.N.; Potamianos, S.

The POSTGRES rule manager.

IEEE Transactions on Software Engineering, July 1988, vol.14, (no.7):897-907.

vis95-tioga2 [\[PS\]](#) [\[PDF\]](#)

Aiken, A.; Chen, J.; Lin, M.; Spalding, M.; Stonebraker, M.; Woodruff, A.

The Tioga-2 database visualization environment.

Database Issues for Data Visualization. IEEE Visualization '95 Workshop. Proceedings. (Database Issues for Data Visualization. IEEE Visualization '95 Workshop. ProceedingsData Issues for Data Visualization. IEEE Visualization '95 Workshop, Atlanta, GA, USA, 28 Oct. 1995). Edited by: Wierse, A.; Grinstein, G.G.; Lang, U. Berlin, Germany: Springer-Verlag, 1996. p. 181-207.

vldb95-tert [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.

Query processing in tertiary memory databases.

Proceedings of the 21st International Conference on Very Large Data Bases. (VLDB '95.

Proceedings of the 21st International Conference on Very Large Data Bases Proceedings of VLDB '95. 21st International Conference on Very Large Data Bases, Zurich, Switzerland, 11-15 Sept. 1995). Edited by: Dayal, U.; Gray, P.M.D.; Nishio, S. San Francisco, CA, USA: Morgan Kaufmann, 1995. p. 585-96.

vldb96-reord [\[PS\]](#) [\[PDF\]](#)

Sarawagi, S.; Stonebraker, M.

Reordering execution in tertiary memory databases.

Proc. 1996 VLDB Conference.

vldb96-cube [\[PS\]](#) [\[PDF\]](#)

Agarwal, S.; Agrawal, R.; Deshpande, P.; Gupta, A.; Naughton, J.; Ramakrishnan, R.; Sarawagi, S.

On the computation of multidimensional aggregates.

Proc. 1996 VLDB Conference.

www5-docs [[PS](#)] [[PDF](#)]

Woodruff, A.; Aoki, P.M.; Brewer, E.; Gauthier, P.; Rowe, L.A.

An investigation of documents on the world world web.

Computer Networks and ISDN Systems 28 (Proceedings of the Fifth International Conference on the World Wide Web, Paris, France, May 1996), 963-980.

[\[Top\]](#) [\[CS\]](#) [\[ERL\]](#) [\[LBL\]](#) [\[S2K\]](#) [\[Theses\]](#) [\[Wisc\]](#) [\[Papers\]](#) [\[Related\]](#)

Related Papers from Other Groups on Campus

Some of the older papers here were produced by groups that no longer exist. The newer papers have mostly been written by the Berkeley [Digital Library](#) project and the Berkeley [Multimedia Research Center](#).

debull96-chabot [[PS](#)] [[PDF](#)]

Carson, C; Ogle, V.E.

Storage and retrieval of feature data for a very large online image collection.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Dec. 1996, vol. 19 (no.4):19-27.

icip96-vods [[PS](#)]

Rowe, L.A.; Boreczky, J.S.; Berger, D.A.; Brubeck, D.W.; Baldeschwieler, J.E.

A distributed hierarchical video-on-demand system.

Proceedings. International Conference on Image Processing (Cat. No.95CB35819). (Proceedings. International Conference on Image Processing (Cat. No.95CB35819) Proceedings International Conference on Image Processing, Washington, DC, USA, 23-26 Oct. 1995). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1995. vol.2.

mm96-vods [[PS](#)]

Brubeck, D.W.; Rowe, L.A.

Hierarchical storage management in a distributed video-on-demand system.

IEEE Multimedia, Fall 1996, vol.3, (no.3):37-47.

nec96-vods [[HTML](#)]

Rowe, L.A.; Berger, D.A.; Baldeschwieler, J.E.

The Berkeley distributed video-on-demand system.

NEC Research Symposium 1995 : Tokyo, Japan. Multimedia computing : proceedings of the sixth NEC Research Symposium / [edited by T. Ishiguro]. Philadelphia : Society for Industrial and Applied Mathematics, c1997.

sosp93-sfi [[image PDF](#)]

Wahbe, R.; Lucco, S.; Anderson, T.E.; Graham, S.L.

Efficient software-based fault isolation.

(14th ACM Symposium on Operating Systems Principles, Ashville, NC, USA, 5-8 Dec. 1993). Operating Systems Review, Dec. 1993, vol.27, (no.5):203-16.

spie94-vods [[PS](#)]

Federighi, C.; Rowe, L.A.

A distributed hierarchical storage manager for a video-on-demand system.

(Storage and Retrieval for Image and Video Databases II, San Jose, CA, USA, 7-8 Feb. 1994).
Proceedings of the SPIE - The International Society for Optical Engineering, 1994,
vol.2185:185-97.

CSD-83-124 [[CS-TR](#)]

Hagmann, Robert Brian.

Performance analysis of several backend database architectures.

Ph.D. thesis. (Prof. D. Ferrari)

CSD-86-258 [[CS-TR](#)]

Gottlob, G.; Paolini, P.; Zicari, R.

Properties and update semantics of consistent views.

Appeared in: ACM Transactions on Database Systems, Dec. 1988, vol.13, (no.4):486-524.

CSD-86-266 [[CS-TR](#)]

Katz, R.H.; Anwarudin, M.; Chang, E.

A version server for computer-aided design data.

Appeared in: 23rd ACM/IEEE Design Automation Conference. Proceedings 1986 (Cat. No.86CH2288-9). (23rd ACM/IEEE Design Automation Conference. Proceedings 1986 (Cat. No.86CH2288-9), Las Vegas, NV, USA, 29 June-2 July 1986). Washington, DC, USA: IEEE Comput. Soc. Press, 1986. p. 27-33.

CSD-86-270 [[CS-TR](#)]

Katz, R.H.; Chang, E.; Bhateja, R.

Version modeling concepts for computer-aided design databases.

Appeared in: (Proceedings of ACM SIGMOD '86. International Conference on Management of Data, Washington, DC, USA, 28-30 May 1986). SIGMOD Record, June 1986, vol.15, (no.2):379-86.

CSD-86-296 [[CS-TR](#)]

Alonso, Rafael.

Query optimization in distributed database systems through load balancing.

Ph.D. thesis. (Prof. D. Ferrari)

CSD-87-341 [[CS-TR](#)]

Katz, R.H.; Chang, E.

Managing change in a computer-aided design database.

Appeared in: Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB. (Proceedings of the Thirteenth International Conference on Very Large Data Bases: 1987 13th VLDB, Brighton, UK, 1-4 Sept. 1987). Edited by: Stocker, P.M.; Kent, W.; Hammersley, P. Los Altos, CA, USA: Morgan Kaufmann, 1987. p. 455-62.

CSD-88-473 [[CS-TR](#)]

Chang, E.E.; Katz, R.H.

Exploiting inheritance and structural semantics for effective clustering and buffering in an object-orientated DBMS.

Appeared in: (1989 ACM SIGMOD International Conference on Management of Data, Portland, OR, USA, 31 May-2 June 1989). SIGMOD Record, June 1989, vol.18, (no.2):348-57.

CSD-89-15 [[CS-TR](#)]

Chang, E.E.

Effective clustering and buffering in an object-oriented DBMS.

Ph.D. thesis. (Prof. R. H. Katz)

CSD-94-796 [[PS](#)]

Rowe, L.A.; Boreczky, J.S.; Eads, C.A.

Indexes for user access to large video databases.

Appeared in: (Storage and Retrieval for Image and Video Databases II, San Jose, CA, USA, 7-8

Feb. 1994). Proceedings of the SPIE - The International Society for Optical Engineering, 1994, vol.2185:150-61.

CSD-94-801 [[CS-TR](#)]

Singhal, V.; Smith, A.J.

Characterization of contention in real relational databases.

Appeared as: Analysis of locking behavior in three real database systems.

VLDB Journal, Feb. 1997, vol.6, (no.1):40-52.

CSD-96-905 [[CS-TR](#)]

Forsyth, D.; Malik, J.; Fleck, M.; Greenspan, H.; Leung, T.; Belongie, S.; Carson, C.; Bregler, C.

Finding pictures of objects in large collections of images.

Appeared in: Object Representation in Computer Vision II. ECCV '96 International Workshop.

Proceedings. (Object Representation in Computer Vision II. ECCV '96 International Workshop.

Proceedings Object Representation in Computer Vision II. ECCV '96 International Workshop.

Proceedings, Cambridge, UK, 13-14 April 1996). Edited by: Ponce, J.; Zisserman, A.; Hebert, M.

Berlin, Germany: Springer-Verlag, 1996. p. 335-60.

CSD-96-913 [[CS-TR](#)]

Zivkov, B. T.; Smith, A.J.

Appeared in: Disk caching in large databases and timeshared systems.

Proceeding. Fifth International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems (Cat. No.97TB100096). (Proceeding. Fifth International

Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication

Systems (Cat. No.97TB100096)Proceedings Fifth International Symposium on Modeling,

Analysis, and Simulation of Computer and Telecommunication Systems, Haifa, Israel, 12-15 Jan.

1997). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1997. p. 184-95.

CSD-97-939 [[CS-TR](#)]

Belongie, S.; Carson, C.; Greenspan, H.; Malik, J.

Recognition of images in large databases using a learning framework.

CSD-97-941 [[CS-TR](#)]

Belongie, S.; Carson, C.; Greenspan, H.; Malik, J.

Region-based image querying.

Appeared in: Proceedings. IEEE Workshop on Content-Based Access of Image and Video

Libraries (Cat. No.97TB100175). (Proceedings. IEEE Workshop on Content-Based Access of

Image and Video Libraries (Cat. No.97TB100175)Proceedings IEEE Workshop on Content-Based

Access of Image and Video Libraries, San Juan, Puerto Rico, 20 June 1997). Los Alamitos, CA,

USA: IEEE Comput. Soc, 1997. p. 42-9.

ERL-M86-40 [[PS](#)] [[PDF](#)]

Rowe, L.A.

A shared object hierarchy.

Appeared in: Proceedings of the 1986 International Workshop on Object-Oriented Database

Systems (Cat. No.86TH0161-0). (Proceedings of the 1986 International Workshop on

Object-Oriented Database Systems (Cat. No.86TH0161-0), Pacific Grove, CA, USA, 23-26 Sept.

1986). Edited by: Dittrich, K.; Dayal, U. Washington, DC, USA: IEEE Comput. Soc. Press, 1986.

p. 160-70.

ERL-M90-12 [[CS-TR](#)]

Bell, J.E.; Rowe, L.A.

Human factors evaluation of textual, graphical, and natural language query interfaces.

Appeared as: An exploratory study of ad hoc query languages to databases.

Eighth International Conference on Data Engineering (Cat. No.92CH3097-3). (Eighth

International Conference on Data Engineering (Cat. No.92CH3097-3), Tempe, AZ, USA, 2-3 Feb.

1992). Los Alamitos, CA, USA: IEEE Comput. Soc. Press, 1992. p. 606-13.

[Paul M. Aoki](#) // graduate student // aoki@CS.Berkeley.EDU

Modified: \$Date: 1998/04/20 22:26:29 \$ by \$Author: aoki \$

Ontologies and Agents in Digital Libraries

Key topics about *Ontology* adapted from *AI Magazine*, Fall 1997, 18(3), include:

- Defn
- Comparison criteria
- Top level categories, taxonomy. categories, realtions, axioms
- Comparison chart

URLs related include:

- [Ontologies](#)
 - [Indented list diagrams of important ontologies](#)
 - [CYC Home Page](#) and [ontology](#) and [table of contents](#)
 - [WordNet Home Page](#) and [online demo](#)
 - Generalized Upper Model: [model](#), [overall organization](#), [concept hierarchy](#), [relational hierarchy](#)
 - [UMLS Home Page](#) and [fact sheets](#), [MeSH](#), [Grateful Med](#) and [demo](#)
 - [TOVE - Toronto Virtual Enterprise](#)
 - [KIF](#) - Knowledge Interchange Format and [brief intro](#)
 - [Stanford KSL Network Services](#) and [Ontology Editor](#)
 - [Guided Tour to Developing Ontologies Using Ontolingua](#); Precursor pages: [Ontolingua](#) with its [Library of Ontologies](#)
 - [EUROKNOWLEDGE Glossary etc.](#)
 - [Stanford DLI](#) and [agents](#), especially for Web browsing
 - Recommend web pages using [Fab](#) (no demos now but several [papers](#))
 - [LIRA](#), a much older system showing key principles (series of screen dumps)
 - A newer [test system](#) without much help (but see [message explaining](#)).
 - [InterPay](#) : [Shopping Models](#), [Secure Electronic Marketplace for Europe](#)
 - [ILU](#) and [Stanford testbed use](#)
 - [Agents '97 Conf.](#)
 - [CHI '97 Software Agents Tutorial](#) by Pattie Maes and her [Software Agents Group](#)
 - [Firefly for music filtering](#) (successor to HOMR from MIT)
 - [My Yahoo](#) (successor to Webdoggie from MIT)
 - [IBM Agent Building Environment \(ABE\): A toolkit for building intelligent agent applications](#)
 - [NEWS WEEDER](#) - naive Bayes classifier - see *AI Magazine* Fall 1997 p. 18
 - [IBM DL: QBIC](#), [agents](#), [WBI tour](#), [watermarking](#)
 - Hal Berghel: [CACM Nov. 1997 40\(11\): Watermarking Cyberspace](#), and [IEEE Computer 29:7 article](#)
 - [DigiCash](#)
 - First Virtual Holdings Inc.: [Green Commerce Model](#)
 - [SSL protocol](#)
-
- Agents: people and places
 - iimam@site.gmu.edu adaptatation, intelligence
 - yves.Kodratoff@Iri.Iri.fr
 - Brian Gaines, U. Calgary: society of agents
 - Haynes, Sen : U. Tulsa: cases
 - Rus, Dartmouth: gather info
 - Decker, Sycara, Williamson: CMU: multiagent society, planning, matchmaker info agent

Questions:

- Try WordNet on "library" and look for coordinate terms on senses 1,2,3
- Try Grateful Med and find MeSH / Meta Terms for "diabetes"

What is an Ontology?

[Tom Gruber <gruber@ksl.stanford.edu>](mailto:gruber@ksl.stanford.edu)

Short answer:

An ontology is a specification of a conceptualization.

The word "ontology" seems to generate a lot of controversy in discussions about AI. It has a long history in philosophy, in which it refers to the subject of existence. It is also often confused with epistemology, which is about knowledge and knowing.

In the context of knowledge sharing, I use the term ontology to mean a *specification of a conceptualization*. That is, an ontology is a description (like a formal specification of a program) of the concepts and relationships that can exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set-of-concept-definitions, but more general. And it is certainly a different sense of the word than its use in philosophy.

What is important is what an ontology is *for*. My colleagues and I have been designing ontologies for the purpose of enabling knowledge sharing and reuse. In that context, an ontology is a specification used for making ontological commitments. The formal definition of ontological commitment is given below. For pragmatic reasons, we choose to write an ontology as a set of definitions of formal vocabulary. Although this isn't the only way to specify a conceptualization, it has some nice properties for knowledge sharing among AI software (e.g., semantics independent of reader and context). Practically, an ontological commitment is an agreement to use a vocabulary (i.e., ask queries and make assertions) in a way that is consistent (but not complete) with respect to the theory specified by an ontology. We build agents that commit to ontologies. We design ontologies so we can share knowledge with and among these agents.

This definition is given in the article:

T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993. [Available on line](#).

A more detailed description is given in

T. R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. Presented at the Padua workshop on Formal Ontology, March 1993, to appear in an edited collection by Nicola Guarino. [Available online](#).

With an excerpt attached.

Ontologies as a specification mechanism

A body of formally represented knowledge is based on a *conceptualization*: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them (Genesereth & Nilsson, 1987). A conceptualization is an abstract, simplified view of the world that we wish to represent for some purpose. Every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization, explicitly or implicitly.

An **ontology** is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an Ontology is a systematic account of Existence. For AI systems, what "exists" is that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, in the context of AI, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names mean, and formal axioms that constrain the interpretation and well-formed use of these terms. Formally, an ontology is the statement of a logical theory.[\[1\]](#)

We use common ontologies to describe *ontological commitments* for a set of agents so that they can communicate about a domain of discourse without necessarily operating on a globally shared theory. We say that an agent **commits** to an ontology if its observable actions are consistent with the definitions in the ontology. The idea of ontological commitments is based on the Knowledge-Level perspective (Newell, 1982) . The Knowledge Level is a level of description of the knowledge of an agent that is independent of the symbol-level representation used internally by the agent. Knowledge is attributed to agents by observing their actions; an agent "knows" something if it acts *as if* it had the information and is acting rationally to achieve its goals. The "actions" of agents---including knowledge base servers and knowledge-based systems--- can be seen through a tell and ask functional interface (Levesque, 1984) , where a client interacts with an agent by making logical assertions (tell), and posing queries (ask).

Pragmatically, a common ontology defines the vocabulary with which queries and assertions are exchanged among agents. Ontological commitments are agreements to use the shared vocabulary in a coherent and consistent manner. The agents sharing a vocabulary need not share a knowledge base; each knows things the other does not, and an agent that commits to an ontology is not required to answer all queries that can be formulated in the shared vocabulary.

In short, a commitment to a common ontology is a guarantee of consistency, but not completeness, with respect to queries and assertions using the vocabulary defined in the ontology.

Notes

[1] Ontologies are often equated with taxonomic hierarchies of classes, but class definitions, and the subsumption relation, but ontologies need not be limited to these forms. Ontologies are also not limited to conservative definitions, that is, definitions in the traditional logic sense that only introduce terminology and do not add any knowledge about the world (Enderton, 1972) . To specify a conceptualization one needs to state axioms that do constrain the possible interpretations for the defined terms.

Ontologies as Indented Lists

- CYC (general ontology for commonsense knowledge, 10K concept types): Thing
 - IndividualObject
 - Event
 - Stuff (parent too of IntangibleStuff)
 - Process (child of Event too)
 - SomethingExisting
 - Intelligence
 - CompositeTangible&IntangibleObject
 - TangibleObject
 - TangibleStuff
 - Intangible
 - IntangibleObject
 - IntangibleStuff (also child of Stuff)
 - InternalMachineThing
 - AttributeValue
 - Relationship (also child of RepresentedThing)
 - Slot
 - Attribute
 - Collection (also child of RepresentedThing)
 - RepresentedThing (parent too of Collection, Relationship)
- WordNet (lexical reference system, 70K synsets): thing, entity
 - living thing, organism
 - plant, flora
 - person, human being
 - animal, fauna
 - non-living thing, object
 - natural object
 - artifact
 - food
 - substance
- Generalized Upper Model (250 concepts, for NLP): Um-thing
 - Configuration
 - Doing&Happening
 - Saying&Sensing
 - Being&Having
 - Element
 - Simple-Quality
 - Simple-Thing
 - Participant
 - Circumstance
 - Process
 - Sequence
 - Expanding-Sequence
 - Projecting-Sequence
- Sowa (based on distinctions, combinations, constraints): T
 - Concrete
 - Object

- Process
- Abstract
- (Level 2)
 - PhysicalObject (child of Concrete, Object)
 - PhysicalProcess (child of Concrete, Process)
 - InformationObject (child of Object, Abstract)
 - InformationProcess (child of Process, Abstract)
- UMLS (153 medical concepts): Entity
 - Physical Object
 - Organism
 - Substance
 - Anatomical Structure
 - Manufactured Object
 - Conceptual Entity
 - Language
 - Occupation or Discipline
 - Organization
 - Group Attribute
 - Group
 - Idea or Concept
 - Finding
 - Organism Attribute
 - Intellectual Product
- TOVE (enterprise modeling): Organization-Entity
 - Organization-Individual
 - Employee
 - Contractor
 - Organization-Group
 - Board of Directors
 - Department
 - Division
- GENSIM (generic simulation): Thing
 - Bacteria
 - E.coli
 - Experiments
 - Nucleic Acid Segments
 - DNA segments
 - Genes
 - RNA segments
 - RNA Genes
 - Protein
 - Active Sites
 - Media



[Company Overview](#)

[The CYC® Technology](#)

[Applications](#)

[Platforms](#)

[Availability](#)

[The Upper CYC® Ontology](#)

[Documentation & Publications](#)

[Staff](#)

★ [Jobs](#) ★

[Where's Cycorp?](#)

[Links](#)

[What's New?](#)

★ [Cyc is Year 2000 compliant.](#) ★

Cycorp has just received an [Advanced Technology Program \(ATP\)](#) award by [NIST](#) (US Commerce Dept) for [1997-2000!](#)

★ [Click here for details.](#) ★

Cycorp is a major participant in the DARPA [High Performance Knowledge Base \(HPKB\) Project](#).

Click [here](#) for more information.

Click [here](#) to download a subset of the constant terms and assertions from the HPKB IKB (Integrated Knowledge Base).

Cycorp, Inc., based in Austin, Texas, is the world leader in commercializing software with common sense.

Cycorp, Inc.
3721 Executive Center Drive, Suite 100
Austin, TX 78731
Internet: info@cyc.com
Telephone: +1 (512) 342-4000
Fax: +1 (512) 342-4040

Copyright © 1995, 1996, 1997, 1998 Cycorp. All rights reserved.
Last updated: December 8, 1997.

WordNet - a Lexical Database for English

Cognitive Science Laboratory

Princeton University

221 Nassau St.

Princeton, NJ 08542

WordNet® is an on-line lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

WordNet was developed by the [Cognitive Science Laboratory](#) at [Princeton University](#) under the direction of [Professor George A. Miller](#) (Principal Investigator). Ongoing development of WordNet is supported by DARPA/ITO (Information Technology Office).

Many people have contributed to the success of WordNet. At the present time, the following individuals at Princeton work on the development of WordNet and research using it:

- Dr. Martin Chodorow
- Dr. Christiane Fellbaum
- Dr. Patricia Gildea
- [Professor Philip Johnson-Laird](#)
- [Shari Landes](#)
- [Professor George A. Miller](#)
- [Randee Teng](#)
- Pamela Wakefield
- Joshua Schechter
- [David Slomin](#)



Unified Medical Language S (UMLS)



NLM's Unified Medical Language System (UMLS) project develops and distributes multi-purpose, electronic "Knowledge Sources" and associated lexical programs. System developers can use the UMLS products to enhance their applications -- in systems focused on patient data, digital libraries, Web and bibliographic retrieval, natural language processing, and decision support. Researchers will find the UMLS products useful in investigating knowledge representation and retrieval questions.

- [UMLS Knowledge Source Server](#) is available to those who have signed the UMLS license agreement.
- [License Agreement for use of the UMLS Knowledge Sources](#) includes a list of vocabularies in the UMLS Metathesaurus. The UMLS products are available free of charge to U.S. and international users. Use of the UMLS Metathesaurus may require additional agreements (which may involve fees) with producers of the individual vocabularies it contains.
- [Obtaining Access to UMLS Resources](#)
- [UMLS Applications](#)
- [Unified Medical Language System Fact Sheet](#)
 - [UMLS Metathesaurus Fact Sheet](#)
 - [SPECIALIST Lexicon Fact Sheet](#)
 - [UMLS Semantic Network Fact Sheet](#)
 - [UMLS Information Sources Map Fact Sheet](#)
- [UMLS Documentation](#) contains complete description of the Knowledge Sources and their distribution formats.
- [Comprehensive Bibliography 1986-96](#) For more recent articles search Unified Medical Language System in MEDLINE.

Send questions, comments about the UMLS project to: custserv@nlm.nih.gov or call 1-888-FINDNLM.

[U.S. National Library of Medicine \(NLM\)](#)

<http://www.nlm.nih.gov/>

Last updated: 12 February 1998

TOVE Manual

Table Of Contents

Chapter 1

[A Common-Sense Model of the Enterprise](#)

- [1.1 Introduction](#)
 - [1.2 Enterprise Modeling Efforts](#)
 - [1.3 Evaluation Criteria](#)
 - [1.4 The TOVE Project](#)
 - [1.5 Acknowledgments](#)
 - [1.6 References](#)
-

Chapter 2

[Ontologies for Enterprise Modelling: Preliminaries](#)

- [2.1 Ontologies and Microtheories](#)
 - [2.2 Time and Action](#)
 - [2.3 References](#)
-

Chapter 3

[An Activity Ontology for Enterprise Modelling](#)

- [3.1 Introduction](#)
 - [3.2 Activities and States](#)
 - 3.2.1 Activity Classes and Instances
 - 3.2.2 Successor Axioms for Status of Terminal States
 - 3.2.3 Status of Non-Terminal States
 - 3.2.4 Status of Activities
 - 3.2.5 Duration
 - [3.3 Aggregation of Activities](#)
 - [3.4 Symbology](#) A
 - [3.5 Summary](#)
 - [3.6 References](#)
-

Chapter 4

The Treatment of Time in the TOVE Project

[4.1 Time Overview](#)

[4.2 Time Taxonomy](#)

[4.3 Time Point Representation](#)

[4.4 Time Point Relations](#)

[4.5 Time Period Representation](#)

[4.5.1 Time Period Relations](#)

[4.6 Time Domain Representation](#)

[4.7 Axioms for Temporal Relations](#)

Chapter 5

Resource Ontology

[5.2 Resource Ontology](#)

[5.3 Relation of the resource ontology with that of the activity-state](#)

[5.4 Conclusion](#)

[5.5 References](#)

CHAPTER 6

An Integrated Product Ontology for TOVE

[6.1 Introduction\(1\)](#)

[6.2 Part Ontology](#)

[6.3 Feature Ontology](#)

[6.4 Parameter Ontology](#)

[6.5 Constraint Ontology](#)

[6.6 Requirements Ontology](#)

[6.7 Symbology](#)

[6.8 Conclusions](#)

[6.9 Selected Bibliography](#)

Chapter 7

An Organisation Ontology for Enterprise Modelling

[7.1 Organisation-Entity](#)

[7.1.1 Organisation-Group](#)

[7.1.2 Organisation-Individual](#)

[7.1.3 Axioms](#)

[7.2 Organisation-Role](#)

[7.2.1 Axioms](#)

[7.3 Organisation Position](#)

[7.4 Organisation Goals](#)

[7.4.1 Objects](#)

[7.5 Communication-Link](#)

[7.5.1 Objects](#)

[7.5.2 Axioms](#)

[7.6 Empowerment](#)

[7.6.1 Objects](#)

[7.6.2 Axioms](#)

[7.7 Authority](#)

[7.7.1 Objects](#)

[7.7.2 Axioms](#)

[7.8 Coordination Speech Acts](#)

[Chapter 8](#)

[8.1 Orders](#)

[8.1.1 Order Overview](#)

[8.1.2 Introduction](#)

[8.1.3 Order Representation](#)

[8.1.3.1 Customer Frame](#)

[8.1.3.2 Order Frame](#)

[8.1.3.3 Line Item Frame](#)

[8.1.4 Order Cluster](#)

[8.2 Symbology](#)

[Chapter 9](#)

[An Ontology for Activity-Based Cost Management](#)

[9.1 Introduction](#)

[9.2 The TOVE Testbed and the Formalization of ABC](#)

[9.3 Competency of the Cost Ontology](#)

[9.4 Cost Ontology for TOVE](#)

[9.5 Resource Cost Point of Activity, a, for Resource, r, at Time point, t: \$cpr\(a,c,t,r\)\$](#)

[9.6 Cost point of Activity, a, at Time point, t: \$cpa\(a,c,t\)\$](#)

[9.7 Taxonomy of Resource Cost Units](#)

[9.8 Taxonomy and Axioms for Cost Orders](#)

[9.9 Activity Cost Taxonomy and Axioms](#)

[9.10 Activity Costs for Cost Orders](#)

[9.11 Successor Cost Axioms for Cost Computations of an Activity for Status Intervals of a Resource](#)

[9.12 Applying \$cpr\(a,c,t,r\)\$, \$cpa\(a,c,t\)\$ and \$cpo\(c,x,t\)\$ for Cost Management](#)

[9.13 Computing cost point of subClass activity, \$ai\$](#)

[9.14 Computing cost point of Class Activity, \$a'ix\$, required to satisfy cost order, \$x\$](#)

[9.15 Computing cost point of cost order, \$x\$](#)

[9.16 Computing cost point of cost order class, \$xc\$](#)

[9.17 Application of the TOVE Cost Ontology towards Activity Based Costing \(ABC\) System.](#)

[9.18 Conventional \(Traditional\) Cost Accounting Systems versus ABC Systems \[Cooper 90\]](#)

[9.19 Mapping the Conceptualization of ABC with the Cost Ontology \[refer figure 7\]](#)

[9.20 Cost Terminology and Semantics 9.20.1 Resource Cost Units](#)

[9.20.1.1 committed res cost unit \(a, r, q, v1\)](#)

[9.20.1.2 enabled res cost unit \(a, r, q, v2\)](#)

[9.20.1.3 disenabled res cost unit \(a, r, q, v3\)](#)

[9.20.1.4 reenabled res cost unit \(a, r, q, v4\)](#)

[9.20.2 Resource Cost Point of Activity, a, for Resource, r, at Time point, t: cpr\(a,c,t,r\)](#)

[9.20.3 Cost Orders](#)

[9.20.4 Activity Costs at time point, t](#)

[9.20.5 Cost Point of an Activity, a, at time point, t: cpa \(a, c, t\)](#)

[9.20.6 Cost Point of an Order, x, at time point, t: cpo\(c, x, t\)](#)

[9.21 Symbology](#)

[9.21.1 Resource Cost Units](#)

[9.21.2 Resource Cost Point of an Activity](#)



[EIL Homepage](#)



UMBC AgentWeb

[UMBC LAIT](#) | [AgentWeb](#) | [AgentNews](#) | agents@cs.umbc.edu | [KQML](#) | [Knowledge Sharing](#) | [NEW!](#) | [Search](#) | [Help](#)

KIF Knowledge Interchange Format

Knowledge Interchange Format (KIF) is a computer-oriented language for the interchange of knowledge among disparate programs. It has declarative semantics (i.e. the meaning of expressions in the representation can be understood without appeal to an interpreter for manipulating those expressions); it is logically comprehensive (i.e. it provides for the expression of arbitrary sentences in the first-order predicate calculus); it provides for the representation of knowledge about the representation of knowledge; it provides for the representation of nonmonotonic reasoning rules; and it provides for the definition of objects, functions, and relations.

KIF 101

[KIF101 - a brief non-technical introduction to KIF](#)

ANSI KIF

- [KIF Specification](#)
- [Model Theoretic Semantics](#) in TeX
- [Standard Ontologies](#)
- [Open Issues](#), [Tabled Issues](#), and [Decisions](#)
- [KIF Electronic Forum](#) and [ANSI KIF Ad Hoc Group](#)

KIF Version 3

The [Manual for Version 3](#) ([postscript version](#))

KIF related software

- [Prologic](#) a common lisp knowledge representation and reasoning system compatible with KIF
- [EPILOG](#) a common lisp inference system compatible with KIF
- [JKP](#) -- a Java Kif Parser which can parse ascii strings representing sentences in a subset of KIF into a Java representation which encodes the logical structure and is ready for further manipulation
- a C [parser](#) for the Knowledge Interchange Format
- IBM [Agent Building Environment](#) -- A toolkit for building intelligent agent applications

AgentWeb is maintained at the UMBC [Lab for Advanced Information Technology](#) by [Tim Finin](#) (finin@umbc.edu).



A Guided Tour to Developing Ontologies Using Ontolingua

This tour is for people who would like an introduction both to developing ontologies and to using the Ontolingua Ontology Editor provided by the [Stanford KSL Network Services](http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/index.html) for creating and modifying ontologies. We suggest that people who have never created an ontology complete this entire tour before using the editor.

Please note: This tour contains a number of screen snapshots. This will mean that the tour is rather more bandwidth intensive than the ontology editor in normal operation, so things will be abnormally slow if you have a low bandwidth connection.

This tour will provide guidance on:

- [How to create an ontology](#)
- [How to create a class](#)
- [How to create a slot](#)
- [How to add a slot to a class](#)
- [How to add a facet to a slot](#)
- [How to create a function](#)
- [How to create an instance](#)
- [How to add an axiom to a class](#)
- [How to create a named axiom](#)

A [glossary of terms](#) is also available.

[Start the tour](#)

ontolingua-librarian@ksl.stanford.edu



Agent Technology Projects in the Stanford Digital Library

Agents are a helpful programming paradigm for object-oriented systems, especially where the task is ill-defined and complex or the agents are making decisions designed to accommodate the user's preferences. The agent paradigm has been used in the Stanford Digital Library Project for information finding (collaborative filtering and distributed gathering) and for economic matters (such as payments).

FAB

FAB is an adaptive multi-agent information retrieval system which finds interesting pages on the web.

"An Adaptive Agent for Automated Web Browsing"

- [Marko Balabanovic](#)
-

InterPay

Designed to permit interoperation between different payment services, InterPay provides levels of abstraction which allow applications to be independent of payment mechanism-specific details.

"InterPay: Managing Multiple Payment Mechanisms in Digital Libraries"

- [Steve Cousins](#)
 - [Prof. Hector Garcia-Molina](#)
 - [Scott Hassan](#)
 - [Steven Ketchpel](#)
 - [Andreas Paepcke](#)
 - [Martin Röscheisen](#)
-

Distributed Commerce Transactions

Gathering information from multiple, self-interested sources that distrust each other requires certain types of structuring to ensure that a multi-stage exchange has the atomicity property.

"A Sound and Complete Distributed Algorithm for Distributed Commerce Transactions"

- [Prof. Hector Garcia-Molina](#)
 - [Steven Ketchpel](#)
-

A Demonstration of the LIRA System

[Marko Balabanovic](#)

Department of Computer Science, Stanford University

February 1995

1. Overview

The LIRA system was designed to help users keep abreast of new and interesting information appearing on the World-Wide Web. Rather than supporting the *searching* task, where the user has a good idea of what they are looking for and can formulate a search query, we are supporting the *browsing* task, often referred to as *surfing*.

Every day the system presents a selection of interesting web pages. The user evaluates each page, and given this feedback the system adapts and attempts to produce better pages the following day. The system starts with completely random pages, and over time attempts to build a profile of the users interests.

2. This Experiment: Finding Music Pages

In order to demonstrate the effectiveness of the LIRA system to people other than the actual user, an objective "interestingness" criterion is required. In this experiment I gave maximum +5 scores to pages related to music, minimum -5 scores to pages unrelated to music, and a +3 score to pages which looked like they might lead to information about music.

The following extracts show the output of LIRA on various days. The first day shows the completely random links from which we start. By day 5 we are starting to get music-related pages. By days 12 and 13 every page suggested is music-related.

- [Day 0](#)
- [Day 5](#)
- [Day 7](#)
- [Day 10](#)
- [Day 13](#)
- [Summary Graph](#)

3. Further Information

N.B. LIRA has been superseded by



The LIRA system was created by [Marko Balabanovic](#), [Yoav Shoham](#) and [Yeogirl Yun](#). Here are some papers which describe the system:

Learning Information Retrieval Agents: Experiments with Automated Web Browsing

(Extended Abstract)

Marko Balabanovic and Yoav Shoham

To appear in [*AAAI-95 Spring Symposium on Information Gathering from Heterogenous, Distributed Environments*](#)

- [Abstract](#)
- [Full paper \(163K postscript\)](#)

An Adaptive Agent for Automated Web Browsing

Marko Balabanovic, Yoav Shoham and Yeogirl Yun

- [Stanford Digital Library Project Working Paper SIDL-WP-1995-0023](#)

For more information please [email Marko](#) at marko@cs.stanford.edu.



InterPay: A Project in the Stanford Digital Library

Public libraries have set many expectations by providing free access to high quality information. However, as budgets are slashed and prices on books and journals continue to rise, this expectation is harder and harder to support. Many libraries now charge a fee for extra services, and of course, the for-profit services and sources need to charge as well. The need for electronic currency was clear, and many rose to the challenge of providing the technical mechanisms to transfer money over the network.

However, none of the newly established vendors was able to become the dominant market player, and several competing standards co-exist, each jockeying for a position in the customer's electronic wallet. InterPay (developed in late 1994 -- early 1995) introduced three layers of abstraction which were designed to insulate the application programmer from the details of a payment mechanism. At the *application layer*, the only difference between a for-pay application and a for-free one was an additional parameter, the *payment agent* that was passed from customer to merchant. The merchant made the transition to the *payment policy layer*, asking an object known as a *Collection Agent* to collect the amount of the invoice. The collection agent dealt with the customer's payment agent, which implemented his payment policy--e.g., small amounts should be automatically approved, while larger ones required explicit user approval. The payment agent would then select one of its *payment capabilities*, such as DigiCash, First Virtual, or NetCheque. At this lowest level, called the *payment mechanism layer*, the payment capability interacted with the *collection capability* to effect the transfer and notify higher levels of the outcome.

The implementation of the InterPay architecture showed payments made by the First Virtual system co-existing with payments made through DigiCash and account-based mechanisms.

Improvements to the InterPay architecture led to UPAI, a *Universal Payment Application Interface*. In addition to a cleaner separation of the payment process from the rest of the application, UPAI specifies an asynchronous process, so multiple payments may proceed in parallel, or an initiated but not yet completed payment may be canceled.

However payment is only one part of the shopping experience, and therefore, we inaugurated a project on "**shopping models**" which broadened the scope to increase its coverage. The basic architecture seeks two objectives:

1. Interoperation of existing mechanisms for payment and delivery, and
2. Flexible specification of customer/merchant interaction without requiring significant software development for each new interaction model (like subscription, pay-per-view, gift certificates, auctions, etc)

The first goal was addressed by using UPAI and a related protocol for delivery called U-DEL to interface with existing payment and delivery mechanisms.

The second goal resulted in the specification of a shopping model, which is divided into three parts, for

handling

- Orders
- Payment
- Delivery

Proxies to the customer and merchant application are similarly divided into those three parts. An interface (in ISL, an interface specification language similar to CORBA's IDL) describes the functions that each part of the proxies must support. The shopping model acts as a "director" deciding what step to call next based on the current state of the transaction and the most recent message.

The Shopping Models project (initially called InterPay II) started in February of 1996, and is ongoing.

7/29/96: Check out the SEMPER (Secure Electronic MarketPlace for Europe) site at <http://www.semper.org>

3/17/96: Also see the eCo project from Commerce.Net at http://www.commerce.net/meetings/dec_96/eco_general

Note: We have nothing to do with InterPay, Inc. (aside from a few phone calls from their lawyers....). If you're looking for help with automated payroll processing, see their [home page](#).

Publications

[Shopping Models: A Flexible Architecture for Information Commerce](#)

by **S. Ketchpel, H. Garcia-Molina, and A. Paepcke**

To appear in DL '97

[Working Paper version]

[InterPay: Managing Multiple Payment Mechanisms in Digital Libraries](#)

by **S. Cousins, S. Ketchpel, A. Paepcke, H. Garcia-Molina, S. Hassan, and M. Röscheisen**

Digital Libraries 95

[postscript, 2.6 MB] added Mar. 20, 1995

[UPAI: A Universal Payment Application Interface](#)

by **S. Ketchpel, H. Garcia-Molina, A. Paepcke, S. Hassan, and S. Cousins**

USENIX 2nd Workshop on E-Commerce

[postscript, 210K] added July 16, 1996

[The Mailing Archive](#)

Restricted to project participants.

Project Members

- [Ali Bahreman](#) (Verifone)

- [Steve Cousins](#)
- [Prof. Hector Garcia-Molina](#)
- [Scott Hassan](#)
- [Steven Ketchpel](#)
- [Andreas Paepcke](#)
- [Martin Röscheisen](#)



ketchpel@cs.stanford.edu



Stanford Digital Library Project



SIDL-WP-1996-0052

Shopping Models: A Flexible Architecture for Information Commerce

Steven P. Ketchpel, Hector Garcia-Molina, Andreas Paepcke

ketchpel@cs.stanford.edu

Abstract: In a digital library, there are many different interaction models between customers and information providers or merchants. Subscriptions, sessions, pay-per-view, shareware, and pre-paid vouchers are different models that each have different properties. A single merchant may use several of them. Yet if a merchant wants to support multiple models, there is a substantial amount of work to implement each one. In this paper, we formalize the shopping models which represent these different modes of consumer to merchant interaction. In addition to developing the overall architecture, we define the application program interfaces (API) to interact with the models. We show how a small number of primitives can be used to construct a wide range of shopping models that a digital library can support, and provide examples of the shopping models in operation, demonstrating their flexibility.

Note: Papers in this series are in development and are not in a final form for publication or general dissemination. They are subject to change. Please do not quote or further distribute them without explicit permission from the authors.

This paper was created on: 11/09/96 and last revised on: 6/15/1997

Author's Comments: The final version that appears in the published proceedings.

Status: PUBLIC

[Click here to see the full text of SIDL-WP-1996-0052](#) (PS)

[Click here for the full text of SIDL-WP-1996-0052](#) (PDF)

Revision History

Version	Format	Date	Comments
4	PS	1/14/1997	The version submitted to DL '97. A major revision of the earlier work.
3	PS	1/14/1997	The version submitted to DL '97. A major revision of the earlier work.
2	PS	11/23/1996	Adds some pictures, plus a better explanation of "roles".
1	PS	11/9/1996	A first, very-preliminary discussion of how InterPay2 enables the support of different shopping models.



Secure Electronic Marketplace for Europe

ACTS Project AC026

SEMPER is a European R&D project in the area of secure electronic commerce over open networks, especially the Internet. It is executed by an interdisciplinary [consortium](#), combining experts from social sciences, finance, retail, publishing, IT and telecommunications, and has established [liaisons](#) with several related efforts.

SEMPER is part of the [European Commission's ACTS Programme](#) (Advanced Communications Technologies and Services), executing [Task 503](#). Funding is provided by the partner organisations, the European Union and the Swiss Federal Department for Education and Science.

For more information, see

- **Project Synopsis in English** ([PDF](#)) and in **Français** ([PDF](#))
- [Public Project Reports and Deliverables](#)
- [Mailing Lists](#)
- [... or contact us directly!](#)

For pointers to information on secure electronic commerce outside *SEMPER* click [here](#) (collected by *SIRENE*).

History

29 May 98

New publications on [risks in implementation of digital signatures](#) and [dispute handling in payment systems](#).

31 Mar 98

Press release: ["SEMPER" Security on the Internet: Advanced European E-Business Prototype Goes Online](#) (also in [German](#))

22 Mar 98

[Poster on SEMPER](#)

5 Mar 98

[Slideshow](#) presented at TEN-Telecom Workshop, Brussels, 4 March 1998

10 Dec 97

Public Deliverable on [New Payment Instruments Prototype](#)

24 Sep 97

[Slideshow](#) presented at

- 21st Century: The Communications Age, European Parliament, Brussels, June 1997; and
- Ministerial Conference "Global Information Networks: Realising the Potential," Bonn, July 1997.

21 Sep 97

Several new reports added:

- Birgit Baum-Waidner: [Liability Cover Service for Digital Signatures in Electronic Commerce](#). *Technical Report*.
- Gerard Lacoste: [SEMPER: A Security Framework for the Global Electronic Marketplace](#). *Overview*.
- Matthias Schunter, Michael Waidner: [Architecture and Design of a Secure Electronic Marketplace](#). *Overview*.
- Michael Waidner: [Secure Electronic Marketplace for Europe; Presentation at "The Open Group -- European Public Forum, Venice, 27 October 1997."](#) *Overview*.
- Arnd Weber: [The Necessity for Secure Implementation of Digital Signatures / Zur Notwendigkeit sicherer Implementation digitaler Signaturen](#). *Technical Report*.
- Dale Whinnett: [End-User Acceptance of Security Technology for Electronic Commerce](#). *Technical Report*.

25 Jul 97

Updated [JAVA bindings of SEMPER APIs](#)

19 Dec 96

Public Deliverable on [Survey Findings, Trial Requirements, and Legal Framework](#)

18 Dec 96

Slides presented at [CommerceNet Global Summit 1996](#)

10 Dec 96

Reports on [Optimistic Protocols for Fair Exchange](#)

4 Dec 96

Revised Overview of [Payment Manager](#)

3 Dec 96

Public Deliverable on [Architecture of Payment Gateway](#)

4 Oct 96

Slides of the [SEMPER presentations](#) at PKS'96, Oct. 2nd 1996, Zürich

4 Oct 96

Paper and slides presented at [ESORICS 96, Sept. 1996, Rome](#)

3 Oct 96

Public Deliverable on [Basic Services: Architecture and Design](#)

16 July 96

[Public Mailing Lists](#)

7 May 96

Presentation on SEMPER (1. Deutscher Internet Kongress, Karlsruhe, 7 May 1996) --
Superseded by ESORICS presentation.

1 May 96

Updated report on [Electronic Payment Systems](#)

21 Mar 96

Summary of [Payment Manager](#)

19 Feb 96

Summary of **Objectives and Initial Architecture of SEMPER** -- *Superseded by ESORICS presentation.*

2 Nov 95

Press release: [European Commission funds first open solution for secure commerce over the internet](#) (also in [German](#))

[[Contact](#) | [EU ACTS Home Page](#) | [SEMPER internal only](#)]

Last modified: Tue Feb 10 08:31:28 MET 1998

[SEMPER Server Administrators](#) semper-admins@www.sempet.org

Stanford Digital Library Testbed Development

Department of Computer Science
Stanford University
Stanford, CA



Inter-Language Unification or ILU:

- [The ILU Project](#)
- [More ILU Information](#)
- [ILU 2.0 alpha 7 Reference Manual](#) **NEW**
- [ILU 2.0 Reference Manual](#)
- [Postscript ILU 2.0 Release Reference Manual](#) **NEW**
- [Postscript ILU 1.8 Release Reference Manual](#)
- [Postscript ILU 1.7 Release Reference Manual](#)
- [ILU Mailing list Archives](#)

Stanford ILU Installation:

ILU currently supports c, c++, lisp, modula 3, and [python](#).

Configuration Script:

In order for us to update and change the ilu installation, please add the following lines to your .cshrc script:

```
setenv OS_SYSTEM `uname -s | tr -d ' / ' \  
setenv OS_RELEASE `uname -r | tr -d ' / ' | sed 's/\\...*//'
```

```

setenv OSNAME ${OS_SYSTEM}
setenv OSTYPE      `uname -s | tr "[A-Z]" "[a-z]"`

if($OS_SYSTEM =~ SunOS) then
    setenv OSNAME ${OS_SYSTEM}${OS_RELEASE}
    setenv OSTYPE ${OSTYPE}${OS_RELEASE}
endif

if(-r /usr/local/ilu20a8/ilu.cshrc) then
    source /usr/local/ilu20a8/ilu.cshrc
endif

setenv DLHOME ~/dldev
setenv DLPATH ${DLHOME}/lib/c/${OSNAME}/${DLHOME}/lib/python
if ($?PYTHONPATH) then
    setenv PYTHONPATH ${PYTHONPATH}:${DLPATH}
else
    setenv PYTHONPATH ${DLPATH}
endif

if ($?ILUPATH) then
    setenv ILUPATH ${ILUPATH}:${DLHOME}/interfaces
else
    setenv ILUPATH ${DLHOME}/interfaces
endif

setenv CVSRROOT /u/testbed/CVSRROOT

```

with these configure options:

```
# ./configure --enable-udp-transport --enable-python-support --enable-corba-iiop --
```

We currently have ILU compiled, installed, and available on:

HP-UX platform: Version 2.0alpha8

distribution location:

```

/usr/local/ilu
/usr/local/ilu20a8

```

HP-UX platform: Version 1.8 RELEASE

distribution location:

```
/usr/local/ilu18
```

supported languages:

```
c, c++, python
```

supported hosts:

```

adrian.stanford.edu
walrus.stanford.edu
sealion.stanford.edu

```

eric.stanford.edu
davidg.stanford.edu
dan.stanford.edu

AIX platform: Version 2.0alpha

distribution location:

/usr/local/ilu
/usr/local/ilu20

AIX platform: Version 1.8 RELEASE

distribution location:

/usr/local/ilu18

supported languages:

c, python

supported hosts:

thames.stanford.edu
oi.stanford.edu
xingu.stanford.edu
mjosa.stanford.edu
db2.stanford.edu

Installation Notes:

Had to statically link the ilu libraries into python.
Cannot get the g++ compiler to produce good code.

OSF1 platform: Version 2.0 alpha

distribution location:

/usr/local/ilu
/usr/local/ilu20

OSF1 platform: Version 1.8 RELEASE

distribution location:

/usr/local/ilu18

supported languages:

c, python,
c++

supported hosts:

Whale.Stanford.EDU
Abalone.Stanford.EDU
Anemone.Stanford.EDU
Porpoise.Stanford.EDU
Starfish.Stanford.EDU
Octopus.Stanford.EDU
Marlin.Stanford.EDU

Barracuda.Stanford.EDU
Clam.Stanford.EDU
Coral.Stanford.EDU
Cowry.Stanford.EDU
Flounder.Stanford.EDU
Halibut.Stanford.EDU
Mako.Stanford.EDU
Mussel.Stanford.EDU
Remora.Stanford.EDU
Scallop.Stanford.EDU
Seal.Stanford.EDU
Jordan.Stanford.EDU
Khatanga.Stanford.EDU

Installation Notes:

need to use the dec linker and pass the -taso flag.

SUN platform: Version 2.0 alpha 8**distribution location:**

/usr/local/ilu
/usr/local/ilu20a8

supported languages:

c, python,
c++

supported hosts:

grunion sole cero bigeye amberjack blenny cunner durgon hake mullet coke

Installation Notes:

Linux platform: Version 2.0 alpha 8**distribution location:**

/usr/local/ilu
/usr/local/ilu20a8

supported languages:

c, c++, python

supported hosts:

coho snapper batray trout congo limpopo rhine grand huron

Linux platform: Version 1.8 RELEASE**distribution location:**

/usr/local/ilu
/usr/local/ilu18

supported languages:

c, c++, python

supported hosts:

all INTEL denoted pc's.

ILU Demonstration Programs:

[pyhello](#) - a very simple python hello client/server object.

[c++hello](#) - a very simple C++ hello client/server object.

[pyinherit](#) - two object servers that implement a type of implementation/interface inheritance (Delegation)

[pyhelloselect](#) - an object server that implements a mainloop in conjunction with ILU's mainloop using select.

There are many examples programs in: '\$(ILUHOME)/examples' like:

rwho2

is an example module implemented in C++ and Python. This example also contains some simple C++ and Python clients that use the rwho2 module.

test1 is a rambling, random example, which serves as a basic regression test. It just calls some sample routines in a pre-determined order. It is rendered in C, C++, Modula-3, and Python; the servers and clients in those languages should all interoperate in all 16 combinations.

timeit

performs timing tests between a server and a client, both implemented in ANSI C. The client performs 3 tests, a cardinal ping, a real ping, and a 10-char string ping (where ping means sending a value and receiving that value back again).

foo lots of stuff.

foogen

This example exercises some Modula-3 capabilities of ILU.

fs is an example of a service written in CommonLisp. It implements a file server protocol, where clients contact the file server with filenames, and receive file objects in return. Methods defined on the file objects allow the users to manipulate the files.

pyhello

a very simple python hello client/server object.

c++hello

a very simple c++ hello client/server object.

Related Information:



Digital Libraries Webmaster

Webmaster@diglib.stanford.edu



The First International Conference on

Autonomous Agents

Marina Beach Marriott Hotel, Marina del Rey, California

February 5-8, 1997

- [Order Agents '97 Conference Materials](#)
- [Read About What Happened at Agents '97](#)
- [Get Ready for Agents '98](#)
- [About the Conference](#)
- [Visit our European Mirror Site](#)
- [Our Sponsors](#)
- [Conference Program](#)
 - [Tutorials](#)
 - [Invited Talks](#)
 - [Panels](#)
 - [Papers](#)
 - [Posters](#)
 - [Robot Demonstrations](#)
 - [Software Demonstrations](#)
 - [Videos](#)
 - [Banquet](#)

CHI97 Software Agents Tutorial

4/2/97

[Click here to start](#)

Table of Contents

[Software Agents](#)

[Agenda](#)

[Agent Visionaries](#)

[Video Knowledge Navigator](#)

[What is an Agent?](#)

[Common Issues Studied](#)

[Types of Agents](#)

[What is a Software Agent?](#)

[How is an Agent different from other Software?](#)

[Why do we need Software Agents?](#)

[Why do we need Software Agents?](#)

[Direct Manipulation](#)

[Indirect Management/Agents](#)

[Criticisms of Software Agents \(Lanier, Schneiderman\)](#)

[Software Agent =? Expert System](#)

[Agents vs Expert Systems \(cont.\)](#)

[Types of Software Agents](#)

[Types of Software Agents](#)

Author: Pattie Maes

Email: pattie@media.mit.edu

Home Page:

<http://www.media.mit.edu/~pattie>

Other information:

Software Agents Group MIT Media Laboratory

[Types of Software Agents \(cont.\)](#)

[User-Programmed Agents](#)

[Nature of “Intelligence” \(cont.\)](#)

[Knowledge-Based Agents](#)

[Nature of “Intelligence” \(cont.\)](#)

[Learning from the User](#)

[Learning from other Agents](#)

[Example: Email Agent](#)

[Second Example: News Agent](#)

[Real example of User-Programmed Agent:
OVAL \(Malone, MIT Sloan\)](#)

[OVAL \(cont.\)](#)

[PPT Slide](#)

[Other examples of User Programmed Agents](#)

[Real Example of Knowledge-Based Agent:
CHORIS \(Tyler & Sullivan, Lockheed\)](#)

[PPT Slide](#)

[Other examples of Knowledge Based Agents](#)

[Real Examples of Learning Agent: Maxims
\(MIT Media Lab\)](#)

[Maxims Learning Agent \(cont\)](#)

[Learning from Peers: Peer-Peer model](#)

[Other examples of Learning Agents](#)

[Pros/Cons of Approaches](#)

[Pros/Cons of Approaches](#)

[Pros/Cons of Approaches](#)

[Which approach is best?](#)

[Types of Software Agents](#)

[Location of Agents](#)

[Location of Agents \(cont.\)](#)

[Mobility of Agents](#)

[Mobility of Agents \(cont.\)](#)

[Exercise](#)

[Technical challenges](#)

[Software Agent <-> Application](#)

[Technical Challenges](#)

[Interface Agent <-> Other Agent](#)

[Common Language, Common Ontology](#)

[Negotiation & Commitment methods](#)

[Modeling other Agents](#)

[Authentication & Identification](#)

[Technical challenges](#)

[Interface Agent <-> User](#)

[Issue 1: Understanding](#)

[Issue 1: Understanding](#)

[Issue 2: Control](#)

[Issue 2: Control](#)

[Issue 3: Distraction](#)

[Issue 3: Distraction](#)

[Issue 4: Ease of Use](#)

[Issue 4: Ease of Use](#)

[Issue 5: Personification](#)

[Video: Persona Project \(Microsoft\)](#)

[Issue 5: Personification](#)

[Personified Agents Research](#)

[Roles for Software Agents](#)

[Agents as Eager Assistants](#)

[Eager \(Cypher\)](#)

[Video Eager](#)

[PPT Slide](#)

[Video Meeting Scheduler \(MIT Media Lab\)](#)

[Maxims Email Agent \(MIT Media Lab\)](#)

[Eager Assistant Agent \(cont.\)](#)

[Details of one Example](#)

[Prediction & Confidence Level Computation](#)

[Using the Prediction](#)

[Multi-agent Collaboration](#)

[Agent Interface](#)

[Explanation](#)

[Video Maxims](#)

[Does it work? Results:](#)

[Discussion \(cont.\)](#)

[Video Schlimmer's Note Taking Agents](#)

[Eager Assistants: Dimensions](#)

[Agents as Guides](#)

[Video Guides \(Apple\)](#)

[Letizia \(Lieberman\)](#)

[Video Letizia \(Lieberman\)](#)

[Feature-Based Filtering: Analyzing Documents for Relevant Keywords](#)

[Feature-Based Filtering: Updating User Profile Based on one More Datapoint Document](#)

[PPT Slide](#)

[Feature-Based Filtering: Filtering Based on User Profile](#)

[Agents as Memory Aids](#)

[Remembrance Agent \(Rhodes/Starnner\)](#)

[PPT Slide](#)

[Remembrance Agent URL](#)

[MIT Media Lab Wearables Group](#)

[Agents as Filters/Critics](#)

[Video NewsTaylor](#)

[Technology Underlying Critics: Feature-based Filtering \(FBF\)](#)

[PPT Slide](#)

[PPT Slide](#)

[PPT Slide](#)

[Demo Firefly and Webhound](#)

[Complementary Technology Underlying Critics: Automated Collaborative Filtering \(ACF\)](#)

[ACF - Predicting a Rating](#)

[Collaborative vs Feature based](#)

[Comparison \(cont.\)](#)

[Some other Neat Features](#)

[Neat Features \(cont.\)](#)

[Other quantitative Results](#)

[Neat Features \(cont.\)](#)

[Neat Features \(cont.\)](#)

[Quantitative Results – Base case Algorithm:](#)

[Quantitative Results – HOMR/Ringo:](#)

[Agents as Matchmakers](#)

[Yenta \(Foner\)](#)

[Yenta applications](#)

[Yenta Methods](#)

[Yenta Algorithm](#)

[Agents for Buying/Selling](#)

[PPT Slide](#)

[PPT Slide](#)

[Bargain Finder demo](#)

[PPT Slide](#)

[PPT Slide](#)

[Fido shopping Doggie Demo](#)

[Kasbah \(Chavez\)](#)

[Kasbah \(Chavez\)](#)

[Kasbah Example](#)

[The Future of Agent Technology](#)

[Future Direction: Markets for Agents](#)

[Future Direction: Ecologies of Agents](#)

[Implications of Software Agents](#)

[Benefits to Consumer](#)

[Benefits to Marketers/Producers](#)

[Speculations: Effects of Agents on Markets](#)

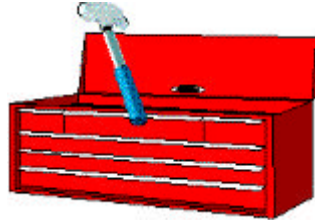
[Open Questions](#)

[Further References](#)



[What's New?](#)
[What are Agents?](#)
[Services](#)
[Technologies](#)
[Download Software](#)
ABE
[WBI](#)
[White Papers](#)
[Resources](#)

Agent Building Environment (ABE)



A Toolkit for Building Intelligent Agent Applications

- [ABE Introduction](#)
- [Technical Overview](#)
- [Latest level information](#)
- [Software requirements](#)
- [Frequently asked questions](#)
- [Download](#)
- [Questions or Problems](#)

ABE Introduction

IBM's Agent Building Environment is a toolkit for software developers that makes it easy to build an application based on intelligent agents or to add agents to an existing application. In this version, the intelligent agent watches for a certain condition, decides what to do based on the rules you've given it, and triggers an action as a result. For example, you can ask your agent to frequently check a stock quote over the internet and if the price drops below a certain point, the agent can alert you by paging you. Or, an agent can check on the inventory of a product and if stock is running low, automatically e-mail a refill order to a supplier. The conditions are based on your areas of interest and the rules for agent behavior under those conditions are based on your preferences.

This developer kit comes with a number of pre-built parts which make it easy for you to add agent technology to applications. The "central intelligence" brain for the agent is based on reasoning engine and adapter technologies from IBM's T.J. Watson Research Lab. "Adapters" or interfaces allow the agent to interact with the rest of the world. For instance, the HTTP adapter provided with ABE interfaces with the world-wide-web. The NNTP adapter interfaces with internet USENET news services, and the timer adapter allows events to be triggered based on time. You can write your own adapters as well and guidelines and a sample adapter are also provided. Custom adapters can be written in either C++ or Java **.

A simple full-screen interface is also provided to allow you to specify the rules for the agent's behavior.

A [Technical Overview](#) with more details is available.

The adapters provided with this version include:

- Time: an "alarm clock" that can trigger events based on time
- NNTP: an adapter that triggers events based on USENET News articles
- HTTP: an adapter that triggers events based on use of the world wide web
- File: an adapter that allows observation and manipulation of files that are located on a local system or any file server to which the agent has access
- SampMail: an adapter that demonstrates how to write an adapter in C++ by providing a basic way to send e-mail.
- Stock: a sample adapter that demonstrates how to write an adapter in Java by providing a basic way to fetch stock quotes from a web site.
- Utility: an adapter that provides utilities for string manipulation and arithmetic operations.

In addition to the above list of adapters, the toolkit also includes:

- A library to hold the facts and rules that the engine uses to intelligently guide the agent. Facts and rules are in a KIF-like format. Guidelines are included on how to write your own rules.
- A generic rule editor to help you write and edit rules.
- Programming interfaces (classes) used for writing C++ or Java agent applications and adapters. Reference documentation for these interfaces is also provided.
- Tutorial and guidance information on designing and writing agent applications and adapters using the toolkit. Some small sample adapters and agent applications are also provided.

This toolkit is being provided for experimentation and without any formal support.

Latest Level Information

The latest level of ABE for Windows 95/NT and AIX is Level 6 (6/29/97).

The latest level of ABE for OS/2 is Level 5 (3/28/97).

The latest level of ABE for OS/390 OpenEdition (Unix System Services) is Level 7 (10/31/97).

Level 7 (10/31/97) of ABE includes the following enhancements:

- ABE has been ported to run in the OpenEdition (Unix System Services) environment of OS/390.

Level 6 (6/29/97) of ABE includes the following enhancements:

- The Time adapter has been extended to support a persistent mode in which alarms are saved and restored across shutdown of the agent. This allows the Time adapter to be used more naturally to generate reminders, follow-ups, expirations, etc.

IMPORTANT: The Level 6 Time adapter provides a different set of effectors and events than the previous versions did. Rule sets that were used in conjunction with the previous versions will require changes to run with the new TIME adapter. See the CHANGES file for details.

- A new AGENT:CONFIG startup-time event has been added to provide an additional opportunity to configure adapters in the agent. The new event makes it easier to do rule-based configuration in situations where some agent-wide configuration effectors have to be called before per-user configuration effector calls are made.
- A new reset() service has been added to the IAAGent class to cause adapters and engines to discard information for a selector (i.e. a user).
- Support for developing ABE components using Microsoft Visual C++ has been extended, and support for using Borland C++ has been added.
- A complete server agent example is now provided (the RemAgent) to illustrate most of the aspects of using the Agent and Library component APIs.
- The Java based components now exploit JDK 1.1.x capabilities, including use of the JNI within ABE's Java interface adapter to improve performance.
- The ABE documentation is now provided in browsable HTML form in addition to Postscript form.

Level 5 (3/28/97) of ABE includes the following enhancements:

- The agent interface (the interface for initializing and controlling the agent) now supports Java as well as C++. A sample Java agent is provided.
- Rule and Atom parsing improvements: more and better messages are provided; escape characters are supported; underscore characters and numbers are permitted in symbol names.
- Several new effectors have been added to the Time Adapter.
- Real numbers are supported as terms for rules and facts.
- The documentation has been reorganized to make material easier to locate and navigate.
- **IMPORTANT :** The format of rules files has changed such that files used with previous versions of ABE will not work with Level 5. A conversion utility (cvtrules) is provided to convert the affected files to the new format. See the READ.ME file for more information.

Software Requirements

For the OS/2 version:

- OS/2 Warp Connect (V3) or OS/2 Warp V4
- TCP/IP for OS/2 Programmer's Toolkit (when using Java adapters or the supplied HTTP, NNTP, or sample stock adapters)
- Java Development Kit 1.0.2 (when running Java agents, Java adapters or the Rule Editor). This JDK is available from IBM at <http://www.ibm.com/Java>.
- The download image is approximately 6MB, and requires approximately 20MB of disk space when installed.

OR

For the Windows version:

- Windows 95 or Windows NT
- Java Development Kit 1.1.2 or later (when running Java agents, Java adapters or the Rule Editor). This JDK is available from Sun Microsystems, Inc. at <http://www.Javasoft.com>.
- The download image is approximately 6MB, and requires approximately 20MB of disk space when installed.

OR

For the AIX version:

- AIX Version 4.1.4
- Java Development Kit 1.1.1 or later (when running Java agents, Java adapters or the Rule Editor). This JDK is available from IBM at <http://www.ibm.com/Java>.
- C SET++ for AIX, version 3.1.4
- The download image is approximately 12MB, and requires approximately 30MB of disk space when installed.

OR

For the OS/390 OpenEdition version:

- OS/390 Version 1 Release 3
- TCP/IP Version 3 (when using the supplied HTTP, NNTP, or sample stock adapters)
- TCP/IP X-Windows Client feature (when using the ABE RuleEditor)
- Java Development Kit 1.1 or later (when running Java agents, Java adapters or the Rule Editor). This JDK is available from IBM at <http://www.ibm.com/Java>.
- The download image is approximately 10MB, and requires approximately 27MB of HFS space when installed.

The following C++ compilers can be used for developing ABE components::

- OS/2 Version
 - IBM Visual Age C++ 3.0
- Windows Version
 - IBM Visual Age C++ for Windows 3.5
 - Microsoft Visual C++ Version 4.2
 - Borland C++ Version 5.02
- AIX Version
 - IBM C SET++ for AIX, Version 3.1.4

Be aware that at this level, the Open Class libraries, ibmcl.a, and ibmcls.a, are not thread safe, and cannot be used for writing adapters or agent controller components.

- OS/390 Unix System Services Version
 - IBM OS/390 C/C++ Version 1 Release 3

Downloadable Software



[Click Here](#) to download.

The download for OS/2 and Windows is a self-extracting EXE file.

After downloading Agent Building Environment into the ABE target directory, to install on OS/2 or Windows, make that directory your current directory (e.g., if you are installing ABE in directory TOOLKIT issue command CD \TOOLKIT) then:

For the OS/2 version issue the command:

IAGTKOS2

For the WIN 95/NT version issue the command:

IAGTKW32

The download for AIX is a compressed tar file.

After downloading Agent Building Environment, to install the AIX version, issue the following commands with the iagtkaix.tar.Z file residing in the directory in which you wish to install ABE:

```
zcat iagtkaix.tar.Z | tar xvf -
```

The download for OS/390 OpenEdition is a compressed tar file.

After downloading Agent Building Environment, to install the OS/390 OpenEdition version, issue the following commands from within the OpenEdition shell, with the iagtkos390.tar.Z file residing in the directory in which you wish to install ABE:

```
pax -rvf iagtkos390.tar.Z
```

Questions or Problems?

This version of the toolkit does not have formal support. Support is AS IS and only on a best-can-do basis by the development team. If you have problems or need assistance, or would like to provide feedback, feel free to [e-mail us](mailto:iagent@us.ibm.com), iagent@us.ibm.com.

Program last modified: October 31, 1997 - Level 7

Return to the [Intelligent Agent Home Page](#)

[[Home](#) | [Order](#) | [Privacy](#) | [Legal](#) | [Contact IBM](#)]

** Java is a trademark of Sun Microsystems, Inc.

Online [Machine Learning](#) software and datasets

Each of the following provide source code and data to accompany examples discussed in the textbook [Machine Learning](#).

- [Neural network learning to recognize faces](#) (example from Chapter 4)
- [Bayesian learning for classifying netnews text articles](#) (example from Chapter 6)
- [Decision tree code](#) (to accompany Chapter 3)

This code and data is made available free of charge for non-commercial use. Please cite this web page in any publications that make use of this data or software.



[What's New?](#)
[What are Agents?](#)
[Services](#)
[Technologies](#)
[Download Software](#)
[White Papers](#)
[Resources](#)



What's New?



[GINKGO Knowledge](#)
[Management Tools](#)

Welcome to IBM's Intelligent Agents Consulting and Programming Services Group. We bring together a broad portfolio of intelligent agent and knowledge management technologies to build solutions that help businesses handle information in smarter ways, become more productive, more innovative and more competitive.

What's New?



"Ginkgo" Knowledge Capture and Virtual Consultation

Are you looking for better ways for the people in your organization to share information, easily learn from experts who are always busy, eliminate re-work and duplication, and lead to new innovations? IBM now offers:

- ***Ginkgo Knowledge Capture*** learns what people know and builds a knowledge base incrementally while people do their normal jobs
- ***Ginkgo Virtual Consultation*** lets people consult the knowledge of others without needing to speak to them in person.

Turnkey solutions built just for your organization are available using these tools, which are based on new intelligent agent learning technology from IBM. Consulting, prototyping, and complete solutions are some of the service offerings available from [IBM's Knowledge Management Services group](#).

Details about the new Ginkgo technology are available in the [Ginkgo White Paper](#).



Consulting and Programming Services

Knowledge Management Solutions for Your Business

IBM offers a complete range of [Consulting and Programming Services](#) for applied intelligent agent applications, including:

- smart web-based applications
- personal and task profiling
- applications that learn and improve over time
- collaborative knowledge sharing and learning in an organization

[Read more](#) about the types of applications that use agent technology.

Some current projects underway are:

- The [NIIP](#) Consortium is deploying agents to help build a "Virtual Enterprise" system, allowing manufacturers and their suppliers to interoperate as if they were part of the same enterprise, thereby increasing efficiency and global competitiveness for each company
- [SHIIP](#) is building an advanced information infrastructure, using agents to improve the shipbuilding design and engineering process.

IBM Knowledge Management and Intelligent Agent Technologies

IBM has a broad portfolio of intelligent technologies that can be applied to solve a wide range of business problems.

- [Aglets](#) are mobile java agents that can roam the internet. The Aglets Workbench lets you develop your own java agents.
- [ABE \(Agent Building Environment\)](#) is a rules-based system consisting of a reasoning engine, a library for rule sets, and adapters to allow the rules to be applied to outside applications, databases, the internet, and other systems. ABE's inference engine is from the [T.J.Watson Research Lab](#), which conducts research on highly reusable intelligent agent technologies.
- [Ginkgo](#) is an agent-based learning system which learns what people do or prefer, and anticipates what they may need; it also learns complex processes and can make improvements. Ginkgo uses an associative memory technique for personal and task profiling, and for multi-agent collaboration
- [Java-based Moderator Templates](#) (JMT), is a framework for collaborative work of multiple mobile agents. JMT allows developers to build a complex plan by simply combining them so that multiple agents can work together toward a common goal.
- [KnU \(Knowledge Utility\)](#) is a knowledge management engine which enables software to reflect a user's preferences.
- [MediaMiner](#) is a set of tools for multimedia information-retrieval and mining applications. MediaMiner includes text search and retrieval, text mining and image mining components, and companion tools.
- [WBI \(Web Browser Intelligence\)](#) can personalize a web user's experience, and give a web server the ability to add or modify web content on the fly. WBI is a component-based development environment that makes it easy to manipulate web data streams, inserting helpful agents at desirable points in the web server-web browser interaction.

What are Intelligent Agents?

Intelligent Agents are software entities that assist people and act on their behalf. They make your life easier, save you time, and simplify your complex world. They're like your private secretary, assistant, or personal advisor, who learns what you like and can anticipate what you want or need. Intelligent agents are being employed today in business applications in the manufacturing, health, financial, travel, retailing and many other industries, in electronic commerce, and on the internet.

Agents can also automate complex processes, such as in manufacturing complex products, and take action to improve the process or eliminate costly errors.

Agents can also communicate with each other, allowing collaborative sharing of information, learning, and smarter organizational systems.

See our [Consulting and Programming Services](#) page and the [Ginkgo White Paper](#) for some examples of real-world applications using agents today.



Feel free to drop us an [E-mail Note](#).

Linda Guyer, IBM Corporation

lguyer@us.ibm.com

[[Home](#) | [Order](#) | [Privacy](#) | [Legal](#) | [Contact IBM](#)]

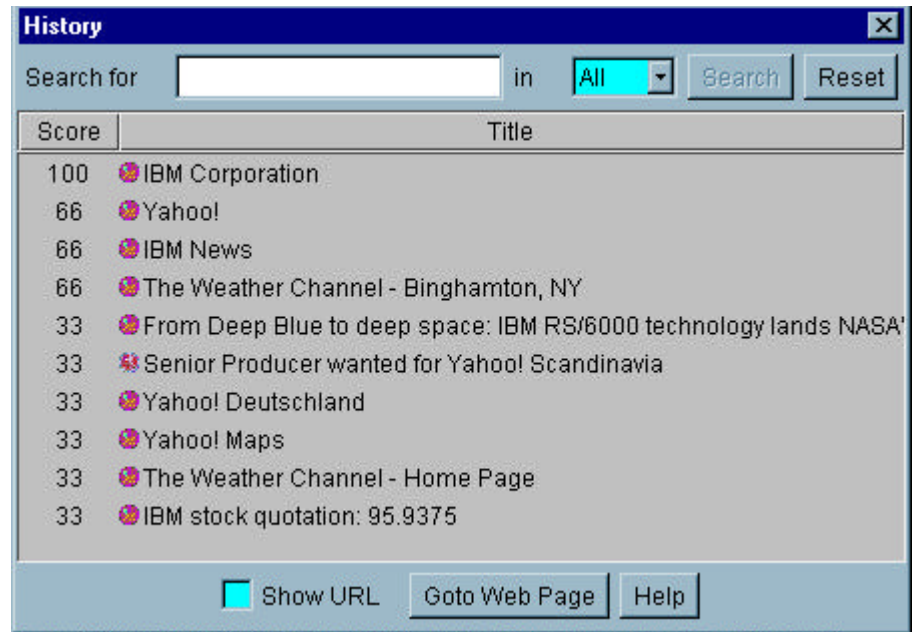


WBI Personal Web Agent Quick Tour

Take a quick tour through some of the services your WBI agent can provide: Personal History, Watch, Path, Shortcuts, and Offline Browsing.

Personal History

- Where have I been on the web?
- Where was that word I saw?
- Your WBI agent keeps track of web page titles, content, and url's
- Your personal history is fully searchable by keyword
- Searches return a list ranked by relevancy (how applicable is this hit to my search?)
- Save time by searching only *your* browsed pages, not the whole web



[Next Feature](#)

Return to the [Intelligent Agent Home Page](#)

[[Home](#) | [Order](#) | [Privacy](#) | [Legal](#) | [Contact IBM](#)]

Commerce, Economics, Publishers:

NetBill

- [Home Page](#)
- [Demo](#)
- [Overview article on payment systems from IEEE Spectrum](#)
- [Set of PowerPoint slides discussing Internet Commerce, Payment Systems and NetBill](#)
- Questions: How would this work with ETDs? What are the advantages and disadvantages relative to other approaches?

Commerce part of CS6604 lecture

- [Workshop on Tech. of Terms and Conditions](#) and [Final Report to NSF](#) - including Breakout Group Reports
- [EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce](#), Hamburg, Germany, June 4-5, 1998
- [Stanford U. work on electronic commerce, legal pointers](#)

[Projections for Making Money on the Web](#) (Michael Lesk, Harvard Infrastructure Conference, 23-25 January 1997)

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta



**What's
New**



**About
NetBill**



**Consumer
Services**



Shopping



**Merchant
Info**



Index



Help

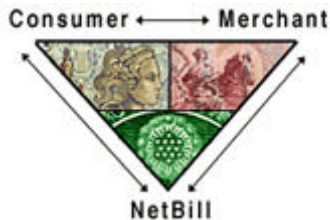


**Send
Feedback**



Search

...a dependable, secure and economical payment method for purchasing digital goods and services through the Internet.



NetBill enables consumers and merchants to communicate directly with each other, using NetBill to confirm and ensure security for all transactions.

NetBill lets you use the Internet to order, pay for and receive information goods easily and securely. And NetBill makes it possible for merchants to sell images, articles (even a paragraph of information), and other goods at low cost.

NetBill is only a trial system, however we invite you to explore this site, learn more about NetBill, and give it a try.

[Demo](#) | [Introduction](#) | [Using this site](#) | [FAQs](#) | [Open an Account](#)



Money Tool

NetBill is a service of:



Mellon



CyberCash



All contents copyright © 1995, 1996, 1997 Carnegie Mellon University.

All rights reserved.

Last revision: Tue Oct 28 14:17:12 EST 1997

Version: R1.4.10

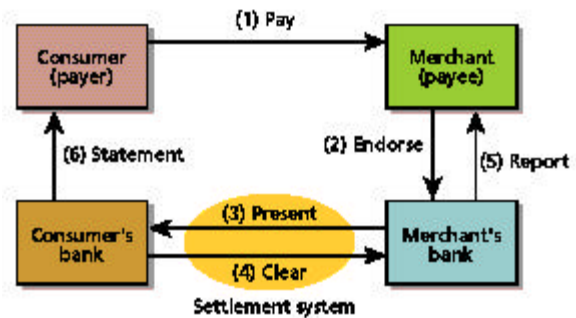
electronic paymentsby Marvin A. Sirbu,
Carnegie Mellon University

CREDITS AND DEBITS ON THE INTERNET

A plethora of technologies and business models are in development to enable electronic payments

Since the advent of banking in the Middle Ages, bank customers have used paper-based instruments to move money between accounts. In the past 25 years, electronic messages moving through private networks have replaced paper for most of the value exchanged among banks each day. With the arrival of the Internet as a mass market data network, new technologies and business models are being developed to facilitate electronic credit and debit transfers by ordinary consumers.

These new systems include CyberCash (which is a gateway between the Internet and the authorization networks of the major credit cards) and the Secure Electronic Transactions protocol (a standard for presenting credit card transactions on the Internet), as well as First Virtual (a way of using e-mail to secure approval for credit card purchases of information), GC Tech (a payment system that can use credit or debit via an intermediation server), and NetBill (a public-private-key encryption system for purchasing information).



Conventional checking

In today's banking world, money consists of ledger entries on the books of banks or other financial institutions. A checking account, also known as a demand deposit account (DDA), records deposits by the consumer and can be used, via the consumer's instructions in the form of a check, to make payments to third parties. Typically, a check is written by a consumer, authenticated by signature, and presented to a merchant, who may endorse it with a signature before presenting it to a bank for payment. If the merchant's bank and the consumer's bank are the same, it can simply transfer the funds on its ledgers from the consumer's account to the merchant's. If the payer and the payee keep accounts at different banks, the payee bank presents the check for settlement to the payer's bank and receives the funds in return through a settlement system. Several private check clearinghouse systems, as well as the Federal Reserve system, provide settlement services in the United States [Fig. 1].

When checks are sent to banks for deposit, merchants do not yet know if consumers have adequate funds and therefore need to find out whether the checks cleared. Similarly, consumers receive statements from their banks showing which checks have been paid. Any discrepancy between bank records and those of the payers may indicate that forged checks were presented against consumers' accounts.

This model works equally well when there is a negative balance in consumers' accounts, at least if the consumers' banks are willing to extend credit—that is, to lend the consumers funds needed to pay off the checks. Many banks in the United States and Europe provide such credit facilities, sometimes referred to

as "overdraft protection." A credit card is another example of an account that lends money to the consumer.

The simple model below illustrates the major issues that must be addressed in designing an electronic credit or debit system.

- Naming: there must be an unambiguous way of identifying the payers' bank accounts and the payees' bank accounts.
- Signatures: it must be possible for the payers' banks to verify that payment instructions were generated by people authorized to use accounts.
- Integrity: electronic checks should be difficult to alter.
- Confirmation: payees must have confirmation that transfers took place; payers must have notification of transfers out of their accounts.
- Confidentiality: third parties should not be able to monitor such payments.
- Settlement: separate banking institutions must have a way of settling their accounts.

Such a system does exist for paper checks. In the United States and Canada, a bank identification code and account numbers are encoded in magnetic ink on the check. But the naming of accounts is not standardized internationally. Payees provide their account numbers when endorsing checks. The payers' banks match the signatures on checks with customers' signatures on file at banks. Integrity is ensured by the use of special paper and the practice of writing checks in ink with no alterations. The U.S. Federal Reserve system provides a vehicle for settlement, and confirmation takes the form of periodic statements or special notices for bounced checks. If checks are presented in person or mailed in sealed envelopes, they are generally protected from observation by third parties.

From a business perspective, payment systems differ in the warranties the different parties make and in the liabilities they assume. For example, the payers' banks are responsible for verifying signatures on checks. If this fails to happen, the payers are not liable for forged checks drawn on their accounts. It is possible to cut the cost of the entire process if payment messages can be readily tied to the parties' accounting systems--for instance, by including purchase order numbers or a consumer's account number with a merchant on all checks. It may also be desirable to link payment to some proof that merchandise has been delivered. These links to other processes are among the principal benefits of electronic payments.

In a payment processing system, the cost of normal operations is frequently outweighed by the costs associated with exception handling. If a typical transaction costs US 5 cents to process, and the manual labor associated with handling errors and exceptions comes to an average of \$25, even with an error rate of only two per thousand, exception costs will equal normal processing costs. As electronic processing drives down the cost of normal transactions, exception handling becomes relatively more significant. Payment systems must therefore be implemented to the highest standards of reliability, with automated procedures for recovering from errors whenever possible.

The case of credit cards

The credit card system was designed to provide immediate gratification of the wants of consumers by allowing them to purchase goods or services on credit. A credit card is a token of trust that transfers the risk of granting credit from a merchant to the card-issuing bank. Once a merchant has had a purchase authorized by the card issuer over the private authorization network, the merchant is assured of payment and the card issuer assumes responsibility for billing the consumer and collecting the money. Settlement

takes place later, when the merchant periodically submits a batch of authorized transactions to the merchant's (acquiring) bank for settlement with the card issuer. But the issuer's assumption of risk is limited, however, to "card-present" transactions, such as those taking place in retail stores. When a merchant accepts a credit card by mail or phone ("card not present"), the card issuer accepts only the risk of nonpayment; the merchant bears the risk of fraudulent card usage. Merchants pay the costs of credit card use because selling on credit expands their business. Under U.S. law, a consumer's liability if someone else fraudulently uses the consumer's card is limited to \$50 [Fig. 2].

In a card-present transaction, the merchant validates the payer's signature by matching the one on the back of the card against the one on the charge slip. Integrity is protected by the device of giving the consumer a carbon copy of the slip. The consumer's account number is verified by the embossed number on the credit card. Settlement is handled by card associations (such as Visa and MasterCard). The merchant receives immediate confirmation of a transaction while submitting it for authorization by way of the card association's private data network.

When a catalog sale takes place by mail or phone, the merchant has no way of verifying the consumer's right to use the card number proffered. At best, the merchant can request the consumer's billing address and receive an address verification. In effect, a credit card purchase requires only that the card number be conveyed from buyer to seller. For this reason, consumers are asked to protect their credit card numbers.

While conventional checking and credit card systems may seem quite similar, the legal meaning of credit card and check payment differ significantly. Credit card companies warrant their merchants; a person can challenge a credit card charge if dissatisfied with the goods. Checks provide no such recourse. If a person buys a plane ticket on an airline with a check, and the airline goes bankrupt before the ticket can be used, the unlucky purchaser becomes an unsecured creditor, behind many other claimants. By contrast, someone who pays for a ticket with a credit card may claim restitution from the card-issuing bank, and the card issuer in turn is entitled to redress from the airline's bank, which must stand behind the airline.

Payment systems vary significantly in their allocation of liability and in the warranties made by the different parties. Technical mechanisms have a strong influence on the willingness of parties to assume liability. If only the payer's bank can verify a signature on a check, the merchant or payee bank will not assume any liability for fraudulent signatures. But if public-key-based "signatures" make it possible for a merchant to verify them on an electronic check, merchants can be expected to undertake verification as they now do in card-present transactions.

Transactions on the Internet

Translating checks or credit card transactions to the Internet requires finding electronic and business model equivalents for the functions described above.

Signatures and confidentiality are the two biggest problems in creating digital payment instruments. These issues are typically handled with some form of cryptography. The use of public-private-key pairs allows a message to be "signed" digitally and verified by anyone who has the public key. Some form of public-key infrastructure, such as certificates, must be employed to associate a named user or an account unambiguously with a particular public key. Message digests provide integrity.

Most payment systems require special consumer and merchant software to prepare and process electronic payment messages. Although the consumer software is often described as an "electronic

wallet," that term is misleading; funds are never kept in the wallet, which acts rather as an electronic checkbook for signing payment orders--managing keys, performing cryptographic operations, and formatting messages, as well as acting as a check register for keeping track of transactions.

The use of credit cards over the phone for catalog shopping is well established. Some of the first Internet systems propose to extend that model to shopping from Web-based catalogs.

CyberCash's gateway

CyberCash Inc., Reston, Va., implemented a system for protecting credit card presentation on the Internet in April 1995. The system was one of the first of its kind. The company, which provides software to both consumers and merchants, operates a gateway between the Internet and the authorization networks of the major credit card brands. As Nathaniel Borenstein, chief technical officer for First Virtual Holdings Inc., San Diego, Calif., noted, "Debugging obscure problems with incompatible implementations of Internet protocols is not a core competence of most financial institutions"--hence, the role for a gateway service.

The consumer begins by downloading the wallet software, which supports encryption and transaction record keeping. Like a physical wallet that may hold a number of credit cards, the software wallet can be used by the consumer to register several credit cards. Another software package provides similar services to the merchant. Messages are encrypted using a random symmetric key, which in turn is included in the message encrypted under the recipient's public key. The CyberCash public key is built into the wallet and merchant software. Consumers generate a public-private-key pair when they register credit cards with the wallet software, and the public key is sent to CyberCash, where it is maintained in a database. While consumers, merchants, and CyberCash all have public-private-key pairs, only CyberCash knows for certain everyone's public key. As a result, the company can exchange information securely with consumers or merchants, but they communicate with one another in the clear, relying on CyberCash to authenticate all signatures [Fig. 3].

When the time comes to make a purchase, the consumer requests the item desired by selecting it with a Web browser. The merchant's server sends the wallet software a cleartext, signed payment-request message that describes the purchase and indicates which credit cards the merchant accepts. The wallet software thereupon displays a window that lets the consumer select which credit card to use, and approve the purchase and the amount.

A credit card payment message, including a signed and encrypted description of the transaction, along with the consumer's credit card number, is sent back to the merchant, which forwards the payment message, along with the merchant's own signed and encrypted description of the transaction, to the CyberCash gateway. There, CyberCash decrypts and compares the two messages and their signatures. If they match, it submits a conventional authorization request and returns the charge response to the merchant, whose software confirms the purchase to the consumer's wallet software (credit card response). Additional messages cover refunds, voiding transactions, capture, and status inquiries.

CyberCash operates its gateway as an agent of the merchant's (acquiring) bank. Thus it must be trusted to decrypt the information for resending over conventional authorization networks.

Since the information is encrypted under CyberCash's public key, the merchant does not actually see the consumer's credit card number--a procedure that in theory cuts the risk that customer credit card numbers will be abused. In practice, so many catalog companies organize their customer marketing

records by credit card numbers that an acquirer usually authorizes CyberCash to provide them to merchants on request.

Secure electronic transactions

In February 1996, Visa and MasterCard announced their joint support of a standard protocol, dubbed Secure Electronic Transactions (SET), for presenting credit card transactions on the Internet. SET is designed to operate both in real time, as on the World Wide Web, and in a store-and-forward environment, such as e-mail. As an open standard, it is also designed to permit consumer, merchant, and banking software companies to develop software for their respective clienteles independently and to have them interoperate successfully.

In the CyberCash protocol, only CyberCash knows everyone's public key. SET, however, assumes the existence of a hierarchy of certificate authorities that vouch for the binding between a user and a public key. Consumers, merchants, and acquirers must exchange certificates before a party can know what public key to employ to encrypt a message for a particular correspondent [Fig. 4].

Although the software industry is moving rapidly to implement SET, the protocol poses significant problems for banks. Card issuers must invest considerable sums to have public key pairs and certificates issued to their card holders. Yet the benefits to the SET card issuers are not clear. A standard protocol may reduce software costs to merchants and consumers, as well as inhibit merchant fraud, but the cost of such dishonesty is borne by the acquirers, not the card issuers. What is more, it is not clear that SET will generate significant new credit card volume, as opposed to merely displacing mail and telephone orders. The card associations suggest that SET transactions, like card-present ones, should involve lower payments to card issuers. Thus a shift from telephone orders to SET could actually reduce the revenues of card issuers, while increasing costs by requiring them to issue certificates. Aligning benefits with costs will require a reallocation of the merchant discount between issuers and acquirers, a politically difficult task for card associations.

First Virtual: no hide and seek

First Virtual provides a mechanism that lets information providers accept credit cards for Internet purchases without resorting to cryptography. Consumers establish account IDs with First Virtual and fax or telephone their credit card numbers to it. To buy information, consumers present those account IDs to merchants, who then connect to the First Virtual server to verify that IDs are valid; if so, the information is sent directly to the consumers. The server then sends them an e-mail message asking if they are willing to pay for the information. Consumers e-mail a reply indicating "yes," "no," or "fraud." If the answer is yes, First Virtual submits the user's credit card number through its acquirer, and the consumer's card is charged. After holding the funds for 90 days, the company transfers them to the merchant by means of an automated clearing house.

The First Virtual model has several key premises. First, consumers do not really know if they want a piece of information until they have looked at it. Second, the cost of sending information electronically "on approval" is negligible, so a merchant has lost very little if a consumer's answer is "no." Third, most consumers are honest: they will not systematically order goods and then answer "no" even when they are satisfied. (As an added deterrent to dishonest behavior, First Virtual will cancel a consumer's account if the pattern of usage suggests abuse.) By not charging consumers until they are satisfied, the system eliminates the cost of reversing charges for information that was not delivered as a result of network or computer problems.

Since the request for payment approval comes by e-mail, while goods are typically delivered over the Web, First Virtual believes that its model is so hard for an attacker to abuse that the risks are justified. Moreover, because the company delays payment to merchants for 90 days, consumers have plenty of time to discover fraudulent charges on their credit card statements, in which case First Virtual can easily reimburse the credit card with the funds it is holding.

In the First Virtual model, naming is provided by the account ID. In lieu of signatures, the company relies on the integrity of the Internet's e-mail infrastructure to ensure that a real consumer is answering yes or no. There is no message confidentiality, except that the account IDs may be viewed as pseudonyms. Confirmation is provided by e-mail and credit card statements. Settlement is handled first by the credit-card provider transferring payment to First Virtual and then First Virtual transferring payment to the merchants.

The company has been in operation since October 1994. It claims more than 180 000 consumer accounts.

Electronic checks

Beginning in the early 1970s, banks began searching for ways to reduce the costs of check processing (6.5¢-13¢ per item) by handling payments electronically. In direct payroll deposit, an employer sends a list of payroll payments to its bank, which then transfers funds to the employees' accounts at their banks through one of several automated clearinghouses (ACH). Consumers use direct payment to deal with recurring bills, such as utility, mortgage, and auto loan payments. In 1995, four ACH operators--the Federal Reserve, the New York Clearinghouse, the Arizona Clearinghouse, and VisaNet ACH Services--handled 2.9 billion transactions worth \$13 trillion on their private electronic networks. The cost to banks was only half of what they would have spent processing checks manually. Payers and payees saved even more.

For both direct payroll deposit (used today by more than 45 percent of the U.S. workforce) and for direct payments, transactions begin when a large organization sends a batch file or tape to its bank with a list of payments or requests for payment. Because this is a batch system, it can take as many as three days for a payee to receive confirmation that a payment has cleared. The existence of these ACH systems for settlement between banks provides a strong base on which to build consumer-oriented electronic payment systems that can accept individual electronic requests for payment originating with consumers.

On the Internet, a paper check can readily be replaced by a digitally signed message--that is, an electronic check. A consortium of banks working through the Financial Services Technology Consortium (FSTC) Inc. has demonstrated a prototype electronic check system that maps directly into the model described above for conventional checks. The payer uses a secure processor, in the form of a PC card, to generate a digitally signed payment instruction, or "check," that is transmitted to the premises of the merchant where it is "endorsed" digitally before it is sent on to the merchant's bank. There, the check can be settled through an existing ACH [Fig. 5]. Other scenarios are also supported; for example, payers can send electronic checks to their own banks, which would then transfer funds directly to the payees' banks.

Standards for conveying invoice and remittance information so that payments can be readily linked into accounts payable and accounts receivable processing systems are an important component of the electronic check concept.

The FSTC model assumes that public keys and certificates are widely available, with banks vouching for their customers and associations of banks, such as an ACH, vouching for one another. The insistence on a hardware token for protecting a private key is designed to provide a high level of protection against such threats as Trojan horse software.

Instant debit systems

To the extent that FSTC's electronic checks rely on the conventional ACH system for clearing, they cannot give the merchant immediate payment confirmation of the sort provided by credit card authorization. CyberCash, Carnegie Mellon University, and GC Tech have introduced, or are developing, low-cost debit payment systems that give a merchant an immediate assurance that the payment will go through.

These systems provide a service model based on the concept of an on-line bank account, with immediate posting of transactions so that payees can get real-time confirmation that funds are available. In addition, they offer an interface to existing electronic funds-transfer mechanisms, including both ACHs and credit cards, so that consumers can easily transfer funds between their primary banks or credit accounts and their Internet payment accounts. Furthermore, these systems aggregate many on-line transactions for batch settlement over traditional settlement networks. They differ in the order of steps required for a transaction, in the consumer protection they provide in the event that goods are not delivered, and in the balance they strike between computationally expensive public-key cryptography and the use of shared-key cryptography.

GC Tech's turnkey offering

GC Tech SA, headquartered in Paris, France, bases its business model on turnkey payment systems software for banks and other financial institutions [Fig. 6]. The intermediation server in the GC Tech model maintains a "ledger" of consumer funds on account in the payment system. These funds may actually be on deposit at the consumer's bank, but their disposition is accounted for on the intermediation server's books. Account funding may take the form of a charge against the consumer's credit card or a transfer from the consumer's checking account to the payment system account. A consumer opens an account by downloading the wallet software and specifying a credit card used to fund the account.

When the consumer has selected a product for purchase, the merchant responds with a digitally signed payment-request message that is sent to the consumer's electronic wallet, which verifies the terms of the transaction and forwards the message to the intermediation server. The server then issues an authentication challenge to the consumer's wallet software. Upon receiving a correct response, the server debits the consumer's account and credits the merchant. Accumulated merchant credits will be settled in a single periodic batch transaction. If the consumer has sufficient funds, the server returns a digitally signed proof of payment (PPT) to the consumer's wallet software, which forwards it to the merchant. Assured of payment, the merchant can now deliver the goods.

The GC Tech cryptographic model assumes that the intermediation server and the merchant have public-private-key pairs, while consumers have only a PIN number. When the consumer forwards the proof of payment to the server, it proposes a session key encrypted under the server's public one. This session key is used to encrypt the authentication challenge and response, as well as to protect the PIN from disclosure. The proof of payment, signed by the server's private key, can be independently verified by both consumers and merchants. This model eliminates the need to issue and manage certificates for

consumers.

Various entities are expected to use the GC Tech system, marketed under the brand name GlobeID. The GlobeID operator in France is Kleline SA, a joint venture operated by Moët Hennessey Luis Vuitton SA and Compagnie Bancaire SA, all three of which are in Paris. U.S. operations are expected to start in early 1997.

NetBill for information delivery

NetBill, a system under development at Carnegie Mellon University (CMU), Pittsburgh, in cooperation with Mellon Bank Corp., also in Pittsburgh, is optimized for delivering such information goods as text, images, and software over the Internet. Its developers, who include the author, have stressed the importance of guaranteeing that consumers receive the information they pay for. To that end, consumers are not charged until the information has actually been delivered to them. Similarly, merchants are guaranteed payment for goods delivered. The basic NetBill protocol has eight steps, beginning with the authentication of identity (using public-key cryptography) and ending with the transmission of a decryption key to the consumer so that the information being purchased can be decrypted and presented [Fig. 7].

In this system, consumers are not charged until the (encrypted) goods reach them. At the same time, if there is not enough money in the consumer's account, the transaction will be rejected and the key never delivered, preventing the consumer from using information that has not been paid for. The merchant's endorsement of the electronic payment order also serves as a warranty that what was received by the consumer is what the merchant intended to deliver. In the unlikely event that the merchant or client machine goes down after the consumer has been charged but before the key is delivered, the consumer can request a copy of the receipt--which contains the key--from the NetBill server.

Note the contrast in message flows between the GC Tech and NetBill systems. GC Tech requires merchants to communicate with the intermediary by way of the consumer's software. In a NetBill microtransaction, only the merchant talks directly to the accounting server.

NetBill will fund its accounts by charging the credit cards of consumers to put spending money in their NetBill accounts. These funds will be held at NetBill's bank. As merchants accumulate credit balances, funds will be transferred via VisaNet to the merchants' banks.

CMU and Mellon Bank expect to launch a commercial trial of the NetBill system in the first half of 1997. Transaction fees, paid by the merchant, are expected to range from 2.5 cents on a 10 cent transaction to 7 cents on a \$1 transaction.

CyberCoin for small deals

In September 1996, CyberCash Inc., Reston, Va., introduced its CyberCoin service, which is designed to support low-cost (25 cent to \$10) transactions for information goods over the World Wide Web. Like the NetBill and GC Tech systems, this one relies on a real-time account database to track Internet transactions. The CyberCash business model assumes that many banks will want to offer a bank-branded payment service that CyberCash would operate on their behalf. This approach would be similar to the recent trend in credit cards: fewer than 25 percent of banks do their own processing; most of them leave it to specialized companies such as First Data Corp., Atlanta, Ga.

A CyberCoin account can be "loaded" either by a charge to a credit card or by a transfer from the consumer's checking account. In the latter case, the transfer is handled in one of several ways: as an ACH transaction, by direct access through a debit or ATM network, or by other means. Depending on the mode of access and the user's level of authorization, funds may become available immediately or held for as many as three days until the transaction clears. While it is less costly to the intermediary to obtain the funds through the clearing house--thus avoiding the credit card discount fee--consumers are likely to prefer credit cards that give them instant access to the funds and 30 days before they have to pay the bill.

In the CyberCoin system, like the NetBill one, merchants deliver the goods encrypted and provide the key only after payment is confirmed. But rather than using RSA digital signatures on every small transaction, the CyberCoin system uses asymmetric cryptography only to load accounts and establish a session key. Individual transactions are then signed with this symmetric key, thus reducing the data processing burden.

CyberCash has established partnerships with a number of important players. Netscape, for one, has agreed to bundle the CyberCash wallet software with its browser software products.

The future is not like the past

Payment systems can be expected to go on proliferating for the next several years, until the market determines the most desirable combinations of functions, price, and performance. The paper world, after all, has many different instruments, which embody different tradeoffs among risk, cost, complexity, responsiveness, and the time until the transaction is final. The same variety should be expected in electronic credit and debit systems.

Yet new technologies uncover new ways to distribute risk, liability, and cost among the parties to a transaction, so that new financial instruments with no comparable paper analog are also to be expected. They will take somewhat longer to develop, however, as they require changes in regulatory assumptions, case law, and participant behavior, all of which evolve much more slowly than technology does .

About the author

Marvin Sirbu holds a joint appointment as professor in the departments of Engineering and Public Policy, the Graduate School of Industrial Administration, and Electrical and Computer Engineering, at Carnegie Mellon University, Pittsburgh. In 1989 he founded the university's Information Networking Institute, which is concerned with interdisciplinary research and education at the intersection of telecommunications, computing, business, and policy studies. Before joining Carnegie Mellon in 1985, he taught in the Sloan School of Management at the Massachusetts Institute of Technology, where he also directed a research program on communications policy.

[To probe further...](#)

Towards A Formalism for Terms and Conditions

Workshop Homepage September 24 - 26, 1996

A major obstacle to the further development of digital libraries, and the national information infrastructure as a whole, is the lack of adequate means of providing digital objects and information on any basis other than free, unrestricted access. Authors are increasingly taking the path of self-publishing using assorted home-grown schemes to seek payment and to impose terms and conditions on use. Publishers wish to specify terms of use and ensure those terms are enforced (optionally collecting payment), before providing valuable materials on the net. While payment and related topics are the subject of much commercial activity, mechanisms for the specification of terms of use seem to have been largely neglected.

Accordingly, a workshop was held on developing a formalism for terms and conditions for the use of digital objects and information. The Workshop organizers were [James R. Davis](#) (Xerox) and Judith L. Klavans (Columbia University, [Center for Research on Information Access](#) and Department of Computer Science).

The workshop took place September 24 - 26, 1996 at the Columbia University Conference Center at Arden Homestead, north of New York City. Now that the workshop is complete, we'll use this page as a reference source for further work on terms and conditions.

- [Workshop schedule](#)
- [List of attendees](#)
- [Readings from the workshop](#)

Reports and presentations about the workshop

- [Workshop Summary: Technology Issues for Terms and Conditions](#) (a brief summary from [D-Lib magazine](#), October 1996)
- [Presentation](#) given at Conference on "Digital Content", Center for Law and Technology, University of California at Berkeley, California, November 8, 1996, by Judith L. Klavans
- Presentation at 1996 NSF Digital Libraries Initiative Meeting, Stanford University, Stanford, California, December 17, 1996. Slides from [David Millman](#), [Vicky Reich](#), and [Judith L. Klavans](#).
- [Final report to the NSF](#) by Judith Klavans. (Added May 27, 1997)

Related links

- [Economics of Digital Information and Intellectual Property](#) (draft papers from a conference held at Harvard, January 23-25, 1997)
- [Bridging Digital Technologies and Regulatory Paradigms](#) Conference June 27-28, 1997, Haas School of Business, University of California, Berkeley.

Projections for Making Money on the Web

Michael Lesk

Harvard Infrastructure Conference, 23-25 January 1997

Abstract

Numerous groups will sell you advice on getting rich on the Web; they discuss online sales of information, retail catalog shopping, advertising, consulting, and connectivity. What will actually pay for the Web? What is the 'killer ap'?

This paper contains a great many conflicting numbers. The different predictors of future revenues differ; even the measures of current success differ. No effort is made to resolve the conflicts, since knowing the spread in estimates may be of value to the reader.

My personal projection for getting rich: connectivity.

1. Introduction.

What is going to pay for the Web? Why should Web site providers continue to do the work of writing, drawing, and coding, plus bear the cost of equipment and communication lines? Justification for Web sites includes many reasons which involve no direct financial gain, such as self-promotion. However, many site builders are hoping to get rich, despite costs that may run over \$1M for a large professionally designed set of corporate web pages.

Among the possible models of finding wealth on the Internet are:

1. Selling objects via the Web. In this dream, the L. L. Bean catalog is replaced by a set of web pages, and calling the 800 number changes into web clicks. Delivery would still be via a parcel carrier.
2. Selling information via the Web. Again, people look at web pages and buy things, but the result can be sent to them directly since it is electronic access. People may pay item by item, or for continuous or regular access to a particular information service.
3. Selling advice. In this case, the basic information is free; and what is being sold is some kind of selection, editing, or consulting related to it.
4. Selling advertising. Again, the information is free, and is supported by advertising in the same way as broadcast TV. For the last year this has been perhaps the most hyped possibility for paying for the web.
5. Selling connectivity. One service that most people do pay for is access to the web itself. Will it be possible to fund the services provided out of such connectivity charges?

This paper discusses some of the estimates for the amount of money that people might pay for each of these services. The traditional business motivations are greed and fear. So far, a great deal of Web activity could be said to be motivated by fear: 'our competitors have a web page.' Is it likely that greed will take over? And will it actually result in riches?

2. Selling objects.

Mail-order catalogs are the model for online sales of objects. In the United States in 1995, catalog sales may have been \$53B out of total retail sales of \$2.2 trillion [Ziegler 1995], or perhaps they were \$60B out of total retail sales of \$1.7 trillion [Sanders 1996], or perhaps \$46B out of \$2.3 trillion total retail sales [Cwiklik 1996]. Online sales, by contrast, are still small: maybe \$324 million sold online in 1995 (Wall Street Journal), or maybe \$575M (Jupiter Communications), with other measures and forecasts ranging down to \$160M. Most distressing is that some of the forecasts were higher than the actual numbers reached; online shopping has not boomed as rapidly as some forecasters thought. 1996 is estimated at either \$518M or \$1B [Stoltz 1996].

There are something like 100,000 retailers on the Web, or maybe only 10,000. The most active shopping seems to be for airline tickets, computer parts, CDs and books. As examples, NECX Direct (computers) sells about \$1M per month; CDNow (audio disks) sold \$2M in 1995 and expects \$9M total sales in 1996; Amazon Books expects to sell \$5M worth of books in 1996. Looking at a more traditional retailer, the Sharper Image sold about \$250K online in 1995 and expects to sell \$3M in 1996.

To consider a typical retailer, one online store is 'She Sails,' which features clothing with a nautical theme or for use on boats. The owner, Ursula Kuehn, described her choice. If she prints and mails a catalog, costing \$20,000 for 12,000 copies, she can expect 1 or 2 percent of those who get it to order something. Her Web page cost about \$3,000 to set up and \$300 a month to keep going; and it has 200 hits a week, producing about 2 orders a week of around \$200 each. If we imagine amortizing the initial cost of the web page over 2 years, the cost per order from either the paper catalog or the Web is about \$100. Clothing is a particularly hard sell on the Web. The HERMES survey of the University of Michigan Business School reports that only 2% of its respondents buy even casual clothes online; nobody buys formal clothes online; by comparison 12% buy computer hardware and 13% buy books.

More people want to buy books online, and bookstores, particularly Amazon Books, are prominent in Web sales publicity. The specialty store Pandora's Books says 30% of their customers are online. CD sales on the Web are also popular; specialty items that are hard for retail stores to stock can be easily handled by online book or CD operations. The CD vendors often allow you to access samples of the music \- listen to thirty seconds and then buy the record if you like it.

Another online operation is 'Peapod,' which sells groceries to about 13,500 people (starting in Chicago). 80% of their users order on a computer. They charge \$4.95 per month, plus \$6.95/order plus 5% (note that they are delivering groceries, which are not usually sold by mail). In 1995 they grossed \$16M, but their balance sheet showed a \$6M loss. They hope in the long run to change the grocery business by taking advantage of the lower costs of warehouse space as opposed to supermarket space; they predict \$750M of revenue for the year 2000.

Despite the publicity given to sales of books and CDs, Nevertheless, Jupiter Communications says that half of all online sales are airline tickets. Anyone using the Web also knows that a lot of pornography is advertised, and Forrester claims that about 10% of Web sales (\$50M/year) are porno. Only 14% of Web shoppers have actually bought anything; there are a lot of browsers. Other bad precedents are the Prodigy florist, which sold \$4M in 1994 but has not thrived as Prodigy has lost business (in 1996, Prodigy had 31 online merchants selling \$35M). There used to be a set of shopping CD-ROMs produced; but such operations as 'Compuserve CD,' '2 Market,' 'Shopping 2000,' and Spiegel's CD-ROM have either stopped producing their CD-ROMs or retargeted them in some way.

Forecasts for shopping remain optimistic. Here are some forecasted volumes from Forrester:

Year	Shopping
1996	\$518M
1997	1138M
1998	2371M
1999	3990M
2000	6579M

and other numbers include \$46B in 1996 (Activmedia), \$300B in 2000 (Killen & Associates), and \$1 trillion in 2010 (The Times, London).

Real stores have advantages over any kind of Web experience. Looking at any shopping mall, most people are in groups, not alone. Shopping is partly a social experience, and so far Web shopping doesn't replicate that. For some items, such as clothes or tools, the ability to try them on or handle them is important in making a purchase. And retail shopping provides instant satisfaction; no waiting for the delivery to arrive.

3. Selling information

Web shopping for information, rather than objects, avoids the problem with delivery. Not only can the Web deliver information immediately, but for many purposes the digital information may be more useful than some kind of paper bought at a bookstore or newstand. Electronic information sales are of course a substantial industry today, although many of the sales are to professionals (e.g. librarians or stockbrokers). The information industry was about \$13.5B gross in the United States in 1993, with much of this representing credit reports and other financial information [Hillstrom 1994]. It was showing a 25% growth rate and it is an industry in which the United States has an enormous positive balance of payments. The table below shows US imports and exports of data and information services for 1993.

United States Trade in Data and Information

	Whole world	Japan only
Exports	\$735M	\$94M
Imports	\$88M	\$10M

An enormous amount of information, of course, is distributed via the Web. About 200 GB of text is now online, perhaps the equivalent in volume of a 250,000 book library. Walnut Creek distributes about a terabyte a month from its ftp site, equivalent to about 5,000,000 copies of a magazine. As everyone knows, however, Web information is extremely variable in quality, and although copious in some subjects (current sports events, computer parts, and so on) extremely sparse in other areas. Ian Irvine (head of Reed Elsevier) said that Elsevier journals rejected 80% of what was submitted to them, and that those rejected manuscripts were what you found on the Web. "Information you get for nothing," he continued, "is worth nothing" [Milliot 1996]

Publishers have been reluctant to sell information on the Web. One obvious problem is the risk of piracy. It is estimated that something like \$15B worth of United States software is pirated every year, and \$2B worth of movies are sold illegally. A software game industry on PCs flourished briefly in the early 1980s but was destroyed by illegal copies, and book publishers worry that the first digital copy of a book sold might be the last one sold. Another problem is competition with existing channels; bookstores might choose to put books available on the Web down on the bottom shelf. In addition to the loss of money, publishers fear a loss of control; for example, the book 'Le Grand Secret' (a report of Mitterand's

illnesses) was banned in France as an invasion of privacy, but it is widely circulating on the Web. It is estimated that something like one half the global software market is pirate and one fifth of the global music market is private; and this does not count private copies which are not distributed (it is estimated that three times as much music is copied privately as is sold commercially).

Nevertheless the attractiveness of information sales has stimulated many attempts. Companies like First Virtual and OpenMarket systems have tried selling items one at a time on the Web. Others such as Commercenet also aim to support companies making such sales. However, to date various problems seem to be discouraging the use of item-by-item sale of information. Both users and publishers fear the lack of predictability in item-by-item sales. Journal publishers, for example, have traditionally relied on selling subscriptions, so that they collect their income in advance; they know before the issues are printed how many copies will be sold at what price and thus what size the journal can be; they have no problem collecting bad debts; and so on. Users, in their turn, gain a certainty of knowing what their expenses will be. If information moves to a model in which articles, or perhaps even paragraphs, are charged for individually with billing taking place later, both sides worry.

In addition, electronic commerce is not really well enough developed for the small amounts that would have to be charged for individual articles. Most billing on the Web is still via conventional credit cards, and if small amounts of intellectual content are going to be sold separately, some kind of micropayment scheme will need to be adopted. There are many proposals for small payment systems, e.g. Netbill [Sirbu 1995] and Commercenet [Tennenbaum 1995], but none yet that has general acceptance.

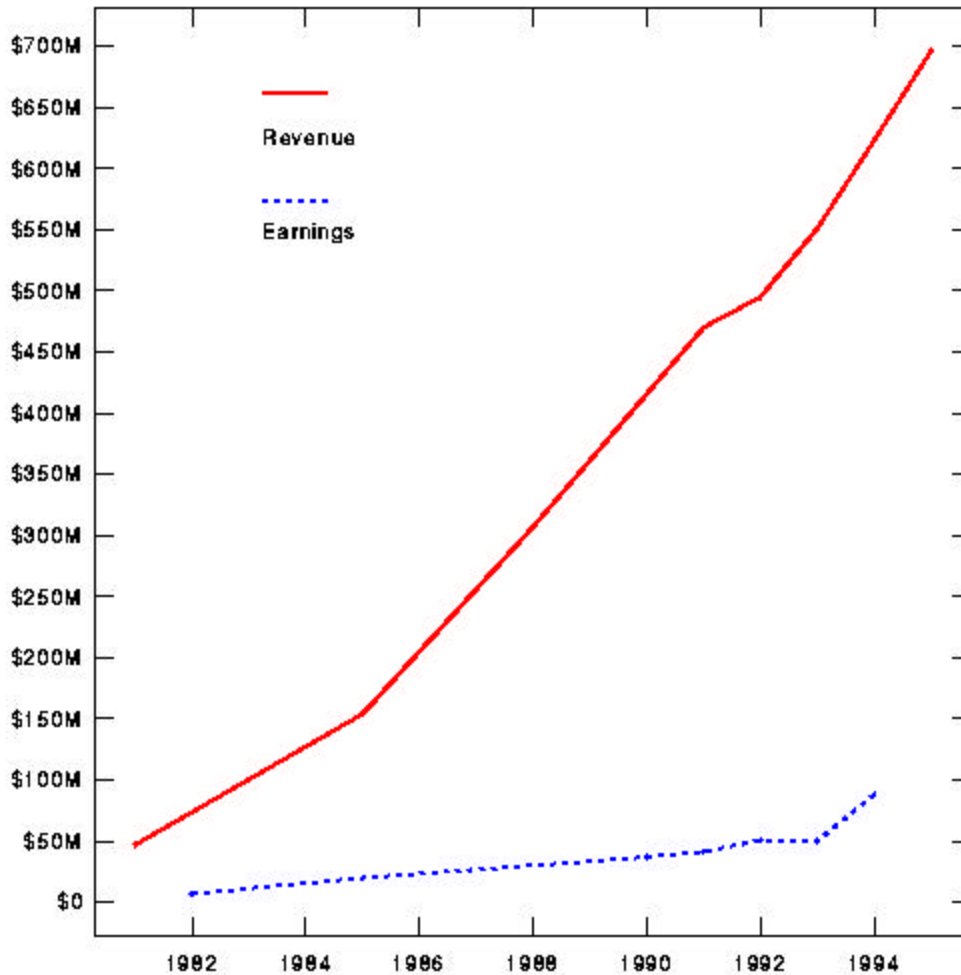
America On Line used to employ a kind of item-by-item payment; when AOL charged its users about 10 cents per minute, AOL would pay a small fraction of this (one or two cents per minute) to the information provider whose information was being read. The rise of the Web, with free information, has discouraged this model, which never paid very large sums anyway. For example, *Time* magazine got several times as much in the way of bounties paid to it by AOL for people who joined by filling out an ad coupon in *Time* than it did for people reading *Time* online through AOL.

For these reasons a subscription model is becoming more common. The 'Electric Library' charges \$10 per month; InfoSeek charges \$5/month; and the *Wall Street Journal* is charging about \$4/month. As yet it is unclear how well these companies are doing. NewsPage, also at \$4/month, went from 55,000 to 15,000 users when it instituted charging, but rebounded to 38,000 users as of December 1995. The *Wall Street Journal* is still providing free access to those who use a particular browser, so it is premature to judge what fraction of its subscribers will actually pay in the long run. As of October 1996 they had 30,000 paid users of their 'interactive edition' while retaining about 100,000 users who get it free [Weber 1996].

Subscription sales of information look on balance as if they are thriving. In fact, they pose a threat to traditional companies like Mead Data Central or Knight-Ridder's Dialog which are accustomed to charging much higher prices to professional librarians; in addition to CD-ROMs, librarians can now get some of the same databases from the cheaper online services.

The figure below, for example, shows the revenues and profits of Mead[Weber 1996] Data Central. As can be seen, although revenue growth has been steady, profits have not kept track. Although some kinds of online information (library catalogs, indexing services) have largely replaced the paper versions, it is not yet as lucrative a business as the sales of paper.

Revenues & profits of Mead Data Central



Many CD-ROM publishers have also had problems turning a profit; the costs of producing multimedia CDs are high and sales have been disappointing for many producers. Some CDs are very successful; CD-ROM encyclopedias, for example, have largely taken over the market for paper encyclopedias. Many games sell well. But most reference CDs do not sell in quantity and profits have been hard to find. The boom of 1993-1994 tailed off at the end of 1995, although a record 2600 CD-ROM titles were released in 1996 [Lasky 1996]. This did not prevent 96 CD-ROM sales from dropping further; in 3Q 1995 43.3M CD-ROMs were bought, but in 3Q 1996 only 38.9M discs were purchased. Forecasts were that 25% fewer discs would be sold in 4Q96 and that this would be 39% less spending per household. The average CD title sells about 28,000 units and brings in \$638K. Problems of declining interest might repeat on the Internet.

Forrester Research has taken a skeptical attitude towards online subscriptions. They predict few will pay more than \$20 per year and that few sites will attract more than 12,000 subscribers, i.e. revenue of \$240,000 per site. This would not be enough to make much difference to a major newspaper or magazine; e.g. the *Wall Street Journal* has 1.8 million circulation of its printed editions. Even projecting to the year 2000, Forrester believes that a site will do well to get \$900K per year from subscriptions.

4. Selling consulting

Esther Dyson is known for saying that on the Internet information should be treated as if it were free, implying that money will be made more by selling advice, comments, ratings, consulting and the like [Ross 1996]. There are various ratings services available on the Internet today, both formal and informal. Several Web searchers, for example, advertise the fact that they review the material they index instead of just grabbing all possible URLs (to provide material of higher quality, or to assure the absence of pornography).

The most interesting form of advice services are the shopping robots. Andersen Consulting, for example, built a program called the BargainFinder. It looks at different CD store web pages, looking for the lowest price for a CD that you wish to buy. The problem with this service is that many CD stores block the robot, since it is doing straightforward price comparison and they wish an opportunity to compete on service, delivery, or something else.

An interesting advice technology is the use of cooperative ratings. Groups at both Bellcore [Hill 1995] and MIT [Maes 1995] have introduced technology in which users are asked to rate something, and the ratings are used to model the users as well as the data. Once you have rated thirty movies, for example, the system compares your ratings to those of other people, and looks for movies that it thinks you will like but that you have not seen. Bellcore has used this for movies and restaurants; MIT, under the name Firefly, has used it for audio CDs. A new company, Agents, Inc. (founded by some of the MIT researchers) is commercializing this technology. Sapien is another company which attempts to deliver health care information using agents.

How to make these systems commercially successful is a problem. There is a great deal of free advice on the Web, and it looks as if finding people willing to pay for better advice may be difficult.

5. Selling advertising

In 1996 the most hyped possibility has been that the Web, like broadcast television, would be supported by advertising. Web users are generally wealthier than the average American (after all, either they or their employer must be able to afford a computer), and they seemed a desirable group to which to advertise some products, particularly high-tech products. In the last quarter of 1995 about \$13M was spent advertising on the Web, with estimates for the whole year running to \$55M or \$64M. Predictions for 1996 made early in the year ranged from \$44M to \$74M to \$110M in advertising revenues. Estimates made late in the year don't agree much better, as shown in the following table.

Web ad spending 1996

Estimator	Estimate
Forrester	\$74M
Simba	\$110M
Alex Brown/McCann-Erickson	\$150-200M
Jupiter	\$312M

One problem has been that the price of advertising on the Web, as more and more sites try to host ads, has been declining rapidly. In 1995 people talked about 10 cents per "view" of an ad, coming down to 6 cents, and by mid-1996 to perhaps 2 cents. Higher rates are charged for a kind of targeted advertising service, appropriate for search services, in which the ad shown depends on the search word typed. As a trivial example, an automobile company might wish to have its ad shown whenever the user typed *car* as

one word in the search. The table below shows some late 1996 advertising rates.

Ad rates in cents per exposure

Host	Bulk	Targeted
DejaNews	2.0	4.0
Excite	2.4	4.0
Infoseek	1.3	5.0
Lycos	2.0	5.0
Yahoo	2.0	3.0

Other services charge per click-through: that is, not per view of an ad but per person who clicks on the ad to reach the advertiser's site. *Playboy*, for example, charges 10 cents per click-through. To give an idea of the balance, one statistic for 1995 shows about 40 views per click-through.

Cheaper total costs are of course available from smaller publications; *American Horseman Online* charges \$40/month. For comparison, the Morris County, New Jersey cable company charges \$65 for 30 seconds on CNN with about 1000 viewers; this is the equivalent of 6 cents per view (although the TV commercial may be considered to have more impact than a web page image).

The next table shows the actual Web advertising revenues for the first quarter of 1996.

Web Advertising

Revenue (,000)	Site
\$3,107	Infoseek
2,622	Lycos
2,190	Yahoo
1,909	Netscape
1,330	CNet
1,244	Excite
1,111	ZDNet
1,100	ESPN SportsZone
989	Pathfinder
925	Webcrawler
833	Hotwired
750	CMP TechWeb
738	NewsPage
625	CNN Interactive
320	Web Review
300	Magellan
274	Wall Street Journal
246	T@P Online
232	Discovery Channel Online

231	PCWorld Online
216	Word
200	Playboy
192	Mercury Center (San Jose Mercury)
181	Jumbo Coolest Shareware
175	GNN Select

In June, for example, \$11.9M total was spent advertising on the Web, with \$8.9M going to the top sites in the table above. The trend, for April to May to June, was from \$10.2M to \$11.7M to \$11.9M, according to Webtrack.

Despite these relatively moderate numbers, the estimates for future advertising are enormous. Here are the estimates from Forrester:

Year Ads on Web

1996 \$80M
 1997 \$200M
 1998 \$1000M
 1999 \$2700M
 2000 \$4800M

and here are a set of estimates for the year 2000:

Web ads in 2000

Source	Estimate
Simba	\$1.9B
Alex Brown	\$2.0B
Forrester (old)	\$2.6B
Jupiter	\$4.5B
Forrester (newer)	\$4.8B
New York Times	\$5.0B

As Web advertising becomes more common, advertisers have started flashing or rotating their ads, adding motion to attract attention. And there are now products coming, such as 'Internet Fast Forward,' which promise to delete ads from the web pages you see.

6. Selling connectivity

The largest amount of money being paid today is being paid by people to be connected to the Internet. How many Web users are there? Here are various measurements and projections. Estimates taken in mid-1996 ranged widely, as seen in the table below [Kantor 1996].

How many Web users?

Estimator	Number
Morgan Stanley	9M
Computer Intelligence Infocorp	15M
Hoffman/Novak	16M
Louis Harris	29M
Intelliquest	35M
Wirthin Worldwide	42M

The *Wall Street Journal* on October 21st reported 14.7 million households. Part of the differences reflect how serious the users are. One study suggests that only 16% of the Web users connect every day; an estimate for September 1996 suggested that per day about 9 million Americans connected to the Web. Here are some forecasts for the future:

Internet Users

Year	Goldman Elect. Sachs-1	Inf.Rep.	Goldman Electronic Sachs-2	Pen
1994	6M			
1995	9M	10M	10M	
1996	15M	30M	25M	95M
1997	22M	60M	40M	190M
1998		95M		380M
1999		130M		760M

Somewhat more concrete are the sales of PCs. Here is a table which compares the number of PCs sold and the number of PCs which have Internet access (world-wide).

Year	PCs with internet access	PCs	PCs
1994	13.6	153	
1995	28.3	184.6	217
1996	52	233.7	240
1997	91.6	266.4	268
1998	133.2	303	303
1999	184.3	326.8	

Today, the majority of Web users are in the United States; Bill Gates says 75%. Forrester thinks that this will remain true until 2000; others see the U. S. dropping to perhaps half the users. About half the PCs today are in the US.

The most extreme forecast of users from Nicholas Negroponte: he predicts 1 billion users in the year 2000. That is comparable to the number of people today with telephone service.

How much do those people pay to get access to the Web? Revenue for ISPs (Internet Service Providers) was \$125M in 1995. Again, prices are declining as the number of users increases. For example, *America Online* pricing is as follows:

\$5.95/mo + 5.00/hr over 1 1992

\$7.95/mo + 6.00/hr over 2 1993

(no change in 1994)

\$9.95/mo + 2.95/hr over 4 1995

\$9.95/mo + 2.95/hr over 5 1996

and AOL has recently announced a change to flat rates. Here is a survey taken in January 1996 of New Mexico ISP providers, showing the number of vendors at each price:

#ISPs Price

2 \$12/mo

2 \$15/mo

6 \$20/mo

7 \$25/mo

3 \$29/mo

3 \$30/mo

1 \$35/mo

1 \$36/mo

25 (total)

By the middle of the year prices seemed to be settling around \$20/month, although some companies had lower prices and AT&T offered a package whereby users could get a free year's connectivity (by signing up for long distance telephony as well). Note that connect time charges in the traditional online database industry were around \$100 per *hour*. So connectivity is much cheaper now, but it is a rapidly growing industry.

In the public eye, the most visible online provider is AOL (America On Line). Here is a chart showing the number of subscribers to various services, including a few of the traditional professionally-oriented online vendors.

Service	Owner	No. subscribers		
		1993	12/31/95	6/30/96
AOL	independent	.53M	4.5M	6M
Compuserve	H&R Block	1.7M	4M	5M
Prodigy	IBM/Sears	2.1M	1.4M	1.4M
MSN	Microsoft		0.6M	1.3M
eWorld	Apple	0.126M		
Delphi	NewsCorp	0.085M	0.05M	0.05M
Genie	General Electric	0.2M	0.075M	
Ziffnet	Ziff-Davis	0.2M	0.3M	0.275M
Pershing	DLJ*	0.1M		
Lexis/Nexis	Reed Elsevier		.744M	.762M
DJ News Ret.	Dow Jones		.233M	.240M
Dialog	Knight-Ridder		.200M	.200M
Reuters	Reuters		.327M	.345M

* Donaldson Lufkin Jenrette

These numbers are subject to enormous turnover. AOL, for example, has seen nearly a quarter of its subscribers quit in one 3 month period. In first quarter 1996, AOL gained 2.3M customers but lost 1.4M. This makes it difficult to judge how successful the marketing effort of AOL will turn out to be.

Whatever the details, however, if we imagine that in the near future there are 50 million people paying \$20/month, or \$240/year, to connect to the Internet, this represents a \$12B industry. This is the largest revenue stream directly predictable for the Internet.

7. Individual connectivity

If people started making voice or video telephone calls on the Web, this would represent a huge boom in Web business. Trying to measure the size of the Web as compared with the U. S. telephone network is difficult. Counting switching decisions, the Web is already larger (since it switches every packet, with a few dozen bytes per packet, while the phone network switches a typical 3-minute voice call only once). If one counts by bytes transmitted, the Internet is something like one-hundredth the size of the long distance phone network. However, the Internet is doubling every year; this would suggest it will catch up by 2004 or so. The growth rate is hard to project: on the one hand, it has been accelerating with the attractiveness of the Web, but on the other hand fairly soon it will become impossible to double the Internet merely by connecting more households, and further growth must represent additional use by people already connected.

One well-hyped possibility is making voice (or fax) calls via the Internet. In principle, Internet telephony can reduce costs by packetizing and compressing the speech. For fax transmission, where it is possible to allow some delays, costs could be reduced even further. Voice transmission can not tolerate delays, although users at the moment seem willing to put up with fairly poor quality (in exchange for free service). There is a complex legal and regulatory issue revolving around the payment of 'access charges' by Internet users. Internet connection by modem, with an average holding time of 20 minutes, is having

an effect on a local phone network designed for 3 minute average call durations, and we do not yet know what mechanisms will eventually dominate Internet connectivity.

Here are the market shares of current Internet telephony vendors (this is for the sale of the software that permits the phone calls):

Market shares in Web telephony

Company	Product	Share
VocalTec	InternetPhone	80%
Quarterdeck	WebTalk	5%
Netspeak	WebPhone	3%
Televox		2%
Camelot	Digiphone	1%
other		9%

And here is another optimistic forecast for how many copies of this software are going to be sold in the future, with an expectation perhaps 10-20M users by the end of the century.

Year	Consumer users	Business users	Total	Software Bought
1995	475K	25K	500K	\$3.5M
1996	1500K	500K	2000K	\$70M
1997	2500K	2500K	5000K	\$175M
1998	4000K	6000K	10000K	\$350M
1999	6000K	10000K	16000K	\$560M

Electronic mail is another popular service which is attracting people to the Internet. Fax numbers became standard on business cards about 10 years ago; now email is being added. The demand for interpersonal communication is probably more certain than the demand for advertising or other services.

If videotelephony on the Internet became common, its bandwidth demands would dwarf anything else. There is little experience for guidance; to judge from AT&T's attempt to sell Picturephone in 1970, it would appear that videotelephony may be less attractive to people than plain voice telephony.

8. Other motivations

There are indirect reasons for having a web page. These represent some kind of cost avoidance or miscellaneous benefits. As examples,

1. Sun Microsystems believes that it saves several million dollars a month in help line costs by having much of its documentation available on the Web.
2. The U. S. cable TV industry has offered to donate Internet connections to schools, presumably expecting to gain public relations benefits if not money.
3. The software used on the Web can be sold for internal use in corporations. Sales of information retrieval and browsing software, for example, are now about \$700M per year, up from \$75M in 1990. Total Internet software sales are expected to be \$1.3B in 1997 and \$4.3B in 1998. However, the idea of giving away Web content to sell software seems the reverse of the usual model in which one provides razors cheap in order to sell blades (John D. Rockefeller was known for giving away kerosene lamps).

There are even stranger motivations; one company has proposed to pay people just to look at ads.

The strongest motivation in early 1996 seemed to be not to sell products, but to sell companies. Netscape reached a market value of over \$5B on 1995 sales of \$80M (Maytag, by contrast has a market value of \$2.5B on sales of \$3.3B). Spyglass was valued at \$242M on 1995 sales of \$10M. For a while there was a rush of new stock offerings for Internet startups. As one would expect, some went up and some went down.

Name	Offering	Top Price	Sep. 96 Price
Excite	\$17	\$21.25	\$5.75
Yahoo	\$13	\$43	\$19.625
Lycos	\$14	\$29	\$18.75
Infoseek	\$15	\$16.50	\$6.75

9. Summary

What is the balance between the different sizes of the markets, as predicted?

Here are the forecasts from Jupiter:

Source	1995	2000	
Access	\$215	\$145	per household
	\$2.1B	\$14B	total
Content	0	\$3.6B	
Ads	\$55M	\$5B	
Shopping	\$300M	\$7.3B	
commission	15%	8%	
Shop.Rev.	\$45M	\$580M	

which emphasizes access and advertising, but with strong components from content sales and shopping. By contrast Forrester expects almost everything to come from ads. They predict that in the year 2000, **90%** of Internet revenue will be from advertising, with **6%** from subscriptions and **2%** from transactions.

Feeling that it would be churlish not to offer my own prediction, I suggest that the answer is 20% from ads, 20% from subscriptions, and 60% from connectivity. What we lack at the moment is an effective way for content providers to benefit from the money paid to connectivity providers. Connectivity does not have the economies of scale of content provision, and so its prices can be more stable. Perhaps in the end some model such as the German tax on blank tape will be applied to the Internet. Alternatively, a technological solution to the piracy problem might make subscription publishing on the Internet more feasible.

On balance, my guess is that just as telephony and the post office are larger enterprises than broadcasting and publishing, we shall also see inter-personal communication and access as more significant than shopping and advertising. Come back and read this paper in the year 2000.

References

Robert Cwiklik, "Revisit the Cyberseers at www.hindsight.1996," *Wall Street Journal*, December 9, 1996.

[Hill 1995] Will Hill, Larry Stead, Mark Rosenstein, and George Furnas, "Recommending and evaluating choices in a virtual community of use," *Proc. CHI Conference on Human Factors in Computing Systems*, Denver, Col., 1995, pages 194-201.

[Hillstrom 1994] *The Encyclopedia of American Industries, Volume Two: Service and Non-Manufacturing*, ed. Kevin Hillstrom, Gale Research, 1994.

[Kantor 1996] Andrew Kantor, Michael Neubarth, "Off the Charts," *Internet World*, vol. 7, no. 12, pp. 44-51 (December 1996).

[Lasky 1996] Michael Lasky, "Best CD-ROMs of 1996," *PC World*, December 1996, page 113.

[Maes 1995] Upendra Shardanand and Pattie Maes, "Social information filtering: algorithms for automating 'word of mouth'," *Proc. CHI Conference on Human Factors in Computing Systems* Denver, Col., 1995, pages 210-217.

[Milliot 1996] Jim Milliot, "Publishers still searching for profits in new media," *Publishers Weekly*, vol. 243, no. 1, page 22 (January 1996).

[Ross 1996] Philip Ross, "Cops vs. Robbers in Cyberspace," *Forbes*, September 9, 1996

[Sanders 1996] Jared Sanders, "At Last, Main Street.com is Opening for Business," *Wall Street Journal*, June 17, 1996.

[Sirbu 1995] M. Sirbu and J. D. Tygar, "Netbill: an Internet commerce system optimized for network-delivered services," *IEEE Personal Communications*, vol. 2, no. 4, pp. 34-39 (August 1995).

[Stoltz 1996] Craig Stoltz and Carolyn Spencer Brown, "wwwwhy./shop.on/line? We spent several weeks doing it -- and still aren't sure we know the answer.." *The Washington Post* April 24, 1996, page R04.

[Tennenbaum 1996] J. M. Tennenbaum, C. Medich, A. M. Schiffman, and W. T. Wong, "Commercenet \- spontaneous electronic commerce on the Internet," *Proc. COMPCON conference*, pages 38-43, (March 1995).

[Weber 1996] Thomas Weber "Interactive Edition Attrats More than 38,000 Paid Users," *Wall Street Journal Interactive Edition*, October 9, 1996

[Ziegler 1995] Bart Ziegler, Associated Press, Dec. 20th, 1995.

Intellectual property rights, copyright laws and legal issues:

(Chapter 10, page 223, "Books, Bucks and Bytes", Michael Lesk)

- [Cyberspace Law for Non-Lawyers](#): This is an electronic course : a "real" course in the "real world" This site includes a discussion function which will allow you, if you are so inclined, to post your own comments and reactions to the individual messages that the instructors have mailed out.
- [Overview of Copyright Laws in the Digital Domain](#) and [References](#) : Check out the references for some very good links and information on copyright laws and related issues.
- [Pamela Samuelson](#) and pointers based on her pages and recommendations
- [Electronic Commerce](#)
- [Workshop on Tech. of Terms and Conditions](#) and [Final Report to NSF](#) - including Breakout Group Reports
- [EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce](#), Hamburg, Germany, June 4-5, 1998
- [Stanford U. work on electronic commerce, legal pointers](#)

Other related references:

- Digital Copyright Protection - Peter Wayner - AP Professional - Boston, 1997
- Scholarly Publishing: The Electronic Frontier - ed. Robin P. Peek and Gregory B. Newby - The MIT Press, Cambridge, MA, 1996
- The Network Nation - Starr Roxanne Hiltz and Murray Turoff - The MIT Press, Cambridge, MA, 1994
- Ubiquitous Email ...

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

Pamela Samuelson Plus Recommendations on Law and Digital Libraries

[Professor Pamela Samuelson](#) is one of the leading authorities on legal issues in the area of intellectual property rights (IPR). A new [MacArthur Fellow](#), a Fellow of the [Electronic Frontier Foundation](#), a Fellow of the [Cyberspace Law Institute](#), she is a Professor at the [University of California at Berkeley](#) with a joint appointment in the [School of Information Management and Systems](#) and the [School of Law](#).

For more information on this and related topics, see

- [Selected Papers by Pamela Samuelson](#)
- [Law 276: Cyberlaw](#) - by Pamela Samuelson, University CA, Berkeley
- [Infosys 296A: Future of the Information Society, Copyright & Community](#) - by Peter Lyman and Pamela Samuelson, University CA, Berkeley
- [Cyberspace Law for Non-Lawyers](#), which attracted over 20,000 subscribers, by [David Post](#), [Temple U. School of Law](#); Lawrence Lessig, [Harvard Law School](#); [Eugene Volokh](#), [UCLA School of Law](#)
- [Crash Course in Copyright](#) from UT system, including the [Digital Library](#)
- [Copyright Management Center](#) of IUPUI, directed by [Kenneth Crews](#)
- [The ILTguide to Copyright](#) at Columbia, for educators
- [Copyright Law Materials](#) at Cornell Legal Info. Institute
- [Copyright & Fair Use](#) site of Stanford University Libraries
- [Copyright Basics Circular from the U.S. Copyright Office](#)
- [Copyright Clearance Center \(CCC\)Online](#)
- [Digital Future Coalition \(DFC\)](#)
- [IIP Policy Gateway, Harvard Information Infrastructure Project](#)
 - [Bibliography](#)
 - [Policy resources in the area of Internet governance](#), supplement to MIT Press [book](#)
 - [The Impact of the Internet on Communications Policy conference](#)
- [ALAWON](#) - ALA (American Library Association) Washington Office Newslines providing urgent and late breaking news
- [ARL Federal Relations and Information Policy Program](#), Prue Adler

Social Issues:

- Social Aspects [D-Lib Working Group](#)
- UCLA Workshop, Social Aspects of Digital Libraries, Feb. 16-17, 1996
<http://www-lis.gseis.ucla.edu/DL/>
 - Life Cycle http://www-lis.gseis.ucla.edu/DL/UCLA_DL_model.gif

[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

Please send comments/suggestions to [Ed Fox](#).

(c) Copyright 1998, Edward A. Fox, Rajat Gupta

WORKING GROUPS

D-Lib Group on Social Aspects of Digital Libraries

I. UCLA-NSF Workshop on Social Aspects of Digital Libraries

An invitational workshop was held at UCLA, February 15-17, 1996; 32 researchers, developers, and practitioners, 9 UCLA faculty facilitators, and 6 UCLA graduate research assistants participated. All materials from the workshop, including schedule and agenda, list of participants, participants' discussion papers and biographical statements, and summary reports presented at the meeting are available on the web site (<http://www.gslis.ucla.edu/DL/>).

We selected two research areas, each with three sub-topics, as focal points for a two-day workshop:

Information Needs: Identifying real information needs and developing digital libraries to meet those needs.

- Social context and culture
- Information needs and information seeking
- Linking user-learner needs and behavior to digital library design

End user searching and filtering: Designing digital libraries in which it is possible to find the right information in a glut of information.

- Organization, description and representation of information
- Search capabilities for users
- Interface design for information retrieval

II. Results of the workshop

While we bounded the scope of the workshop to provide a starting point for discussion and a set of criteria for selecting participants, our participants quickly expanded those boundaries.

The boundaries expanded in several directions:

- Level of analysis: Our scope, as stated in the background paper (see web site), focused on the needs and activities of the individual user. While important, we must recognize that individuals do not work with information resources in isolation from their communities. They perform individual tasks in the context of their work teams, classroom, and other social organizations. Many tasks are performed in group contexts; we must consider CSCW and collaboratory environments as well. Multiple levels of analysis are required.
- Scope of analysis: Our scope addressed information searching and retrieval processes. While important, we must set searching in the context of the cycle of information creation and utilization. People will create information in digitized form that becomes part of digital libraries and need tools and functional capabilities for doing so. They will search for information created by other people, and for purposes other than those intended by the creators, requiring a variety of searching functions. Once located, they will incorporate new information into other products and processes that become part of the life-cycle. We need consistent means to organize, describe,

represent, and dispose of information throughout these activities and processes.

- Content vs. process: Our scope addressed digital libraries as a set of digitized resources and associated technical capabilities for searching for information, which is roughly the scope defined in the digital libraries initiative. This scope statement addresses the digitized content of digital libraries but does not recognize the social processes around digital libraries -- the "library" in digital libraries. We need to address both, hence the distinction made in the second definition stated in the beginning of this report.

III. Research agenda for Social Aspects Of Digital Libraries

We will present the research agenda with respect to the two definitions of digital libraries outlined above. These two definitions converge in a model of the life cycle of information and information processes.

The model covers the sequence from the creation of information (author, artist, memo-writer, data-generation scientist, publisher, etc.), through the searching for information, and the utilization of it, often for very different purposes than it was originally created. An exit from the loop is given to indicate that we do not need to save everything created in digital form -- indeed, we need criteria and mechanisms to decide what to keep and what to destroy. The model addresses the social context for all aspects of the cycle -- people create information for one purpose, search for it for another, and utilize for another. We need to organize, describe, and represent for multiple uses but we must design based on an understanding of what those uses might be. Similarly, we need searching and utilization interfaces that support many perspectives and purposes, with a variety of functional capabilities -- but all must be based on some understanding of the underlying tasks/roles that the information will play in a social context.



clb/wya

Last revised: March 18, 1996