DIGITAL LIBRARIES '97

Second ACM International Conference on Digital Libraries

First ACM International Conference on Digital Libraries

REGISTRATION IS CLOSED !!!

DL'96 Final Program (text version)

DL'96 Registration Form (text version)

Hypertext'96 immediately proceeds DL'96, at the same location, so some may wish to register and attend both. See the HT'96 Advance Program (text version) or home page

The conference is in Bethesda, MD.

Attendees are encouraged to read the working group materials at http://www.dlib.org to prepare them for the D-Lib Working Sessions.

Related Information:

- **ACM and the 2 main sponsors of DL'96: SIGIR and SIGLINK**
- **DL'94**

- DL'95
- Digital Library Source Book, 1993, ed. E. Fox, etc.
- 37th Allerton Institute 1995: How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum

Digital Libraries '95

The Second Annual Conference on the Theory and Practice of Digital Libraries

June 11-13, 1995 - Austin, Texas, USA

Full Papers

Short Papers

Frank Shipman / shipman@cs.tamu.edu / 95-06-08

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

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 librariesand 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 specturm 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 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.

ALLERTON 1996

38th Allerton Institute
October 27-29, 1996
Allerton Park and Conference Center
Monticello, Illinois

Libraries, People, and Change: A Research Forum on Digital Libraries

Sponsored By:

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

and

The National Science Foundation 

Co-Chairs:

Ann Peterson Bishop, Graduate School of Library and Information Science, University of Illinois at Urbana-Champaign
David M. Levy, Xerox PARC, Palo Alto, California

Allerton '96: Goals

Final program, session descriptions and comments

List of Participants

Discussion Documents

Mapping Us!: The participants mapped

General Information

Registration Form

The 1996 Allerton site is maintained by the GSLIS Publications Office

Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

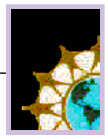
Last Updated: Jan. 27, 1997

37th Allerton Institute 1995

Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign

How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum

This conference was sponsored by the National Science Foundation



The 1995 Allerton site is available via the EDFU Electronic Library

The Publications Office
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign

Last updated: 17 January 1996

Social Aspects of Digital Libraries

A workshop hosted by:
The Department of Library and Information Science
Graduate School of Education & Information Studies (GSE&IS)
University of California, Los Angeles

February 16-17, 1996

Sponsored By:
Information Technology and Organizations Program
Information, Robotics, and Intelligent Systems Division
Computer and Information Science and Engineering Directorate
The National Science Foundation

Contents

Description

This workshop brought together 32 scholars, researchers, and practitioners from the emerging community concerned with social aspects of digital libraries, plus the 8 UCLA investigators (Marcia J. Bates, Christine L. Borgman, Michele V. Cloonan, Efthimis N. Efthimiadis, Anne J. Gilliland-Swetland, Yasmin B. Kafai, Gregory H. Leazer, and Anthony B. Maddox). Our goals were to assess existing knowledge that might inform research in this area and to propose a research agenda that would pose new questions.

We organized the workshop content and selected the participants around two social aspects of digital libraries: information needs, and end-user searching and filtering. In their position papers and in on-site discussions, workshop participants quickly expanded the topical boundaries in several directions. Rather than focusing solely on the individual user who interacts with a digital library, we considered also the group, organization, and community activities and concerns which give rise to information-related behavior. We expanded our interest in information storage and retrieval to include preceding and succeeding phases, incorporating the processes of creating, using, and disposing of information.

Based on the wide-ranging discussions in the workshop, the final report proposes a definition of digital libraries that encompasses two complementary ideas, one emphasizing that they extend and enhance existing information storage and retrieval systems, incorporating digital data and metadata in any form; the other emphasizing that design, policy, and practice should reflect the social context in which they exist. We propose an information life cycle model to illustrate the flow of human activities in creating, searching, and using information and the stages through which information artifacts may pass: activity, inactivity, and disposal.

Research issues raised in the workshop were organized into three foci: human-centered, artifact-centered, and systems-centered. We recommend that research be conducted on these themes, that scholars from multiple disciplines be encouraged to develop joint projects, that scholars and practitioners work together, and that digital libraries be developed and evaluated in operational, as well as experimental, work environments. Only in this way can we build digital libraries to support diverse communities of users in their professional, educational, and recreational activities.

The UCLA-NSF Social Aspects of Digital Libraries Workshop web page includes the final report, the list of attendees, position papers, the UCLA background paper, and links to other sites and materials.

Participant Discussion Papers

Philip E. Agre
Tora K. Bikson
Ann Peterson Bishop
Joseph A. Busch
Donald O. Case
Elfreda A. Chatman
Su-Shing Chen
Paul Conway
Raymond D'Amore
Brenda Dervin
Andrew Dillon
Aimee Dorr
Karen M. Drabenstott and David M. Levy
Susan T. Dumais
Raya Fidel
Edward A. Fox
Rob Kling
Joseph S. Krajcik
Carol C. Kuhlthau
Thomas K. Landauer
Ray R. Larson
Clifford A. Lynch
Gary Marchionini
Daniel V. Pitti
Edie Rasmussen
Vicky Reich
Ronald E. Rice
Philip J. Smith
Velimir Srica
Susan Leigh Star
Nancy Van House

Contributed participant biographies are also available.

Organizers and Managers of the Workshop

Principal Investigator
Christine L. Borgman

Co-Investigators:
Marcia J. Bates
Michele Valerie Cloonan
Efthimis N. Efthimiadis
Anne Gilliland-Swetland
Yasmin Kafai
Gregory H. Leazer

Advising Committee:
Daniel E. Atkins, University of Michigan
Christine L. Borgman, University of California, Los Angeles
Edward Fox, Virginia Polytechnic Institute and State University
Michael Lesk, Bell Communications Research
David Levy, Xerox Palo Alto Research Center
Clifford Lynch, University of California, Division of Library Automation
Gary Marchionini, University of Maryland, College Park

Workshop Coordinator:
Anthony B. Maddox
1-310-206-5865 (voice)
1-310-206-4460 (facsimile)
digital-libraries@gslis.ucla.edu (Internet)

GSE&IS Staff Coordinators:
Keri Botello
Lydia Doplemore
John C. Houser
Mary King
Renée Kneer

GSE&IS Student Assistants:
Nadia Caidi
Venkatachallam Maithili
Marlene Martin
John Schacter
Susan Schreiner
Claude Zachary

The Workshop Site

The Workshop is hosted by the Department of Library and Information Science at the Graduate School of Education & Information Studies Building on the campus of the University of California, Los Angeles which is bordered by the Brentwood, Westwood and Bel-Air sections of Los Angeles.

Other Digital Libraries Sites

- Interoperability, Scaling, and the Digital Libraries Research Agenda
  A Report on the May 18-19, 1995 IITA Digital Libraries Workshop, August 22, 1995, by Clifford Lynch (clifford.lynch@ucop.edu) and Hector Garcia-Molina (hector@db.stanford.edu).

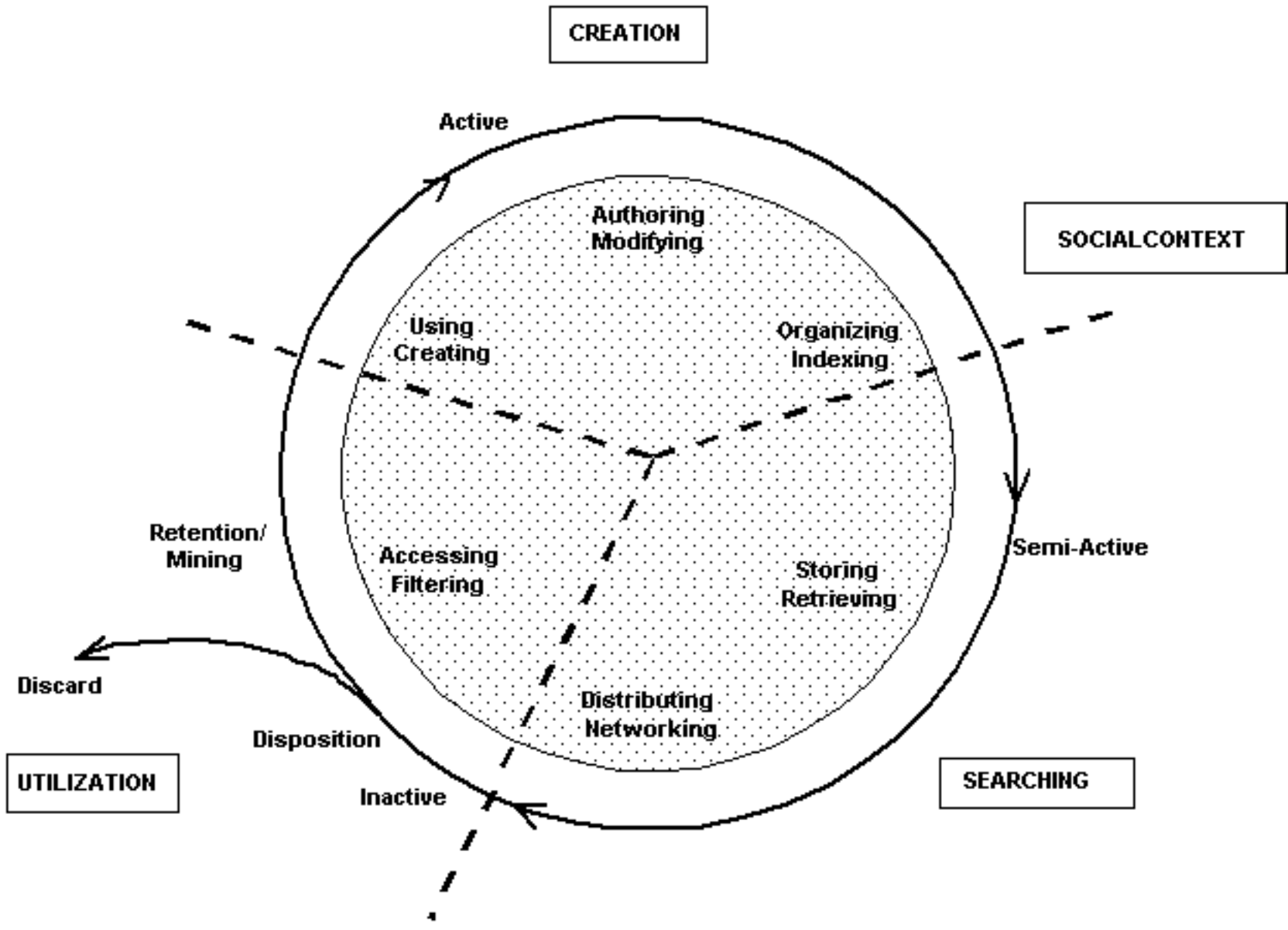
Canned Searches

- InfoSeek Search for "Digital Libraries"
- Lycos Search for "Digital Libraries"
- Alta Vista Search for "Digital Libraries"
- Yahoo - Reference:Libraries:Information Science:Digital Libraries

This page is located at: http://www-lis.gseis.ucla.edu/DL/

Questions regarding this page should be addressed to Jay Baker, hbaker@ucla.edu. Updated January 3, 1996.

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.

D-lib magazine

The Magazine of Digital Library Research

Welcome to D-Lib Magazine, a single site with monthly stories, commentary, and briefings and a collection of resources for digital library research. Visit D-Lib Magazine This Month for new material; visit D-Lib Ready Reference for pointers to projects and collections concerning digital library research.

D-Lib Magazine This Month: May 1997

Search the contents of the monthly magazine and reference pages; or

Browse back issues of **D-Lib Magazine**.
or **D-Lib Magazine's** stories, editorials, and briefings by title or author.

D-Lib Ready Reference

Working Groups: Meetings, Conferences, and Workshops
The Technology Spotlight
Digital Library Research Projects
Elsewhere on the Net

- Clearinghouses for Digital Library Research
- Calendars of Events
- Technical Reports and Papers

About D-Lib
D-Lib's Response Page
Access Terms and Conditions

Mirror sites for D-Lib Magazine are graciously maintained by:

UKOLN: The UK Office for Library and Information Networking: http://hosted.ukoln.ac.uk/mirrored/lis-journals/dlib/, and

The Australian National University Sunsite: http://sunsite.anu.edu.au/mirrors/dlib

---

These activities are coordinated by the Corporation for National Research Initiatives and are sponsored by the Defense Advanced Research Projects Agency (DARPA) on behalf of the NSF/DARPA/NASA Digital Libraries Initiative.

Graphics on D-Lib's pages have been created by Robert Kinneary.
wya/af
Last revised: May 15, 1996

Digital Library Research

From here, you can follow pointers to some of the major cooperative projects, funding and coordinating agencies, and associated activities in digital library research. Many of the research centers maintain lists of projects and project descriptions as well as small collections of technical papers, typically authored by staff.

- Coordinating and funding bodies
- U.S. federally funded cooperative projects
- Centers for research on digital libraries in the U.S.A.
- Programs and projects outside the U.S.A.

For collections of technical papers and a selection of individual items, see:

- **Technical Reports and Papers**

Coordinating and funding bodies

NASA's Digital Library Technology Project. A project that supports the development of new technologies to facilitate public access to NASA data via computer networks.

The Coalition for Networked Information. A joint project of the Association of Research Libraries, CAUSE, and EDUCOM to promote information resources in networked environments.

The Internet Engineering Task Force. The protocol engineering and development arm of the Internet.

The World Wide Web Consortium. The W3 Consortium exists to develop common standards for the evolution of the World Wide Web.

European Research Consortium for Informatics and Mathematics. A consortium of leading research establishments in Europe that encourages collaborative work among researchers and with industry.

European Commision

- Science, Research, and Development: The 4th Framework Programme
- I*M-EUROPE
- General Information

Research Libraries Group (RLG): Strategy for 2000, Supporting Projects. A not-for-profit membership organization for improving access to information to support research and education. Among other services and activities is a research program that facilitates collaboration among institutions in relevant areas of research (e.g., library services, information access and delivery).

Federally funded cooperative projects

The NSF/DARPA/NASA Digital Library Initiative (DLI). Six federally funded projects in digital library research, with partnerships led by universities. The individual projects are listed below.

University of California, Berkeley: An Electronic Environmental Library Project. (A DLI project.)

University of California, Santa Barbara: The Alexandria Project: Towards a Distributed Digital Library with Comprehensive Services for Images and Spatially Referenced Information. (A DLI project.)

Carnegie Mellon University: Informedia: Integrated Speech, Image and Language Understanding for Creation and Exploration of Digital Video Libraries. (A DLI project.)

University of Illinois at Urbana- Champaign: Building the Interspace: Digital Library Infrastructure for a University Engineering Community. (A DLI project.)

University of Michigan: The University of Michigan Digital Library Project. (A DLI project.)

Stanford University: Stanford University Digital Libraries Project. (A DLI project.)

The Computer Science Technical Reports Project (CSTR). A collaboration involving CNRI, five universities, and the Library of Congress.

D-Lib. A forum for researchers and developers of advanced digital libraries.

Centers for research on digital libraries in the U.S.A.

- Center for Intelligent Information Retrieval
- Center for the Study of Digital Libraries
- Center for Electronic Texts in the Humanities
- Center for Research on Information Access
- The Electronic Text Center at the University of Virginia
- The Information Infrastructure Project
- OCLC Online Computer Library Center, Inc.: Office of Research and Special Projects
- Rutgers Center for Information Management, Connectivity, and Integration

Programs and projects outside the U.S.A.

eLib: Electronic Libraries Program (UK): A broad program of projects addressing a wide range of digital library issues.

European Research Consortium for Informatics and Mathematics (ERCIM): A consortium of research organizations from thirteen European countries, which provides a framework for collaboration.

Nordinfo: Coordinating framework for collaboration among three centers devoted to digital libraries and electronic publishing: National Digital Library Centre (NDLC), Nordic Net Centre (NNC), and NordEP.

Distributed Systems Technology Centre (DSTC): A joint venture by the Australian Government's Cooperative Research Centres (CRC) Program and participating organizations to develop the technological infrastructure for global distributed systems.

New Zealand Digital Library: An interactive system with collections of computer science technical reports and literary materials. The site includes pointers to the relevant project descriptions, technology, and related information.

Digital Library Network (DLNet): This page describes digital library work at the University of Library and Information Science, Tsukuba Science City, Japan, in English. From here, follow links to discussions of

D-lib magazine
The Magazine of Digital Library Research

ISSN 1082-9873

Access to Stories, Briefings, and Editorials by Title

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

You may also browse by author.

- SuperJournal, David Pullinger, Christine Baldwin
- TURNIP: The URN Interoperability Project, Renato Iannella
- UK Electronic Libraries Programme, Chris Rusbridge
- Z39.50 and the World Wide Web, Sebastian Hammer, John Favaro

C

Caching for Large Scale Systems, Lessons from the WWW, Robert E. McGrath
Coming Soon to Your Favorite Library: Decision Support on Demand, Hemant Bhargava, Bob Norris
Content Ratings and Other Third-Party Value- Added Information: Defining an Enabling Platform, Martin Röscheisen, Terry Winograd, Andreas Paepcke
Creating a Networked Computer Science Technical Report Library, James R. Davis

D

Digital Libraries and Corporate Technology Reuse, Jonathan T. Hujsak
Digital Libraries: Searching Is Not Enough; What We Learned On-Site, Andreas Paepcke

E

Economic Framework for Pricing and Charging in Digital Libraries, J. Sairamesh, C. Nikolaou, D. Ferguson, Y. Yemini

Editorials:

- All Things in Good Time
- Double-Edged Sword of Access
- Future is a Complex Place
- Getting Used to Technology and Revolutions in the Making
- In This Issue [July/August 1996]
- Levels of Abstraction
- Taking the Measure of the Net
- Telling Time
- Text Is More Than Just Words on a Page, Susan Hockey
- Tragedy of the Commons, Revisited (Again)
- When is Honesty the Best Policy?
- Where Are We in Space?
- Word (or two) of welcome

F

Federating Repositories of Scientific Literature: An Update on the Digital Library Initiative at the University of Illinois at Urbana-Champaign, Susan L. Harum, William H. Mischo, and Bruce R. Schatz
French Minitel: Is There Digital Life Outside of the "US ASCII" Internet? A Challenge or Convergence? Jack Kessler

G

Global Change Data and Information System-Assisted Search for Knowledge (GC-ASK) Project, Roberta Y. Rand

H

Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress, *(Part I)*, Caroline R. Arms
Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress *(Part*

Home Magazine

hdl://cnri.dlib/september96-title.index

Working Groups

One of D-Lib's principal activities is stimulating and supporting working groups that address aspects of Digital Library research. Some of these groups are created by D-Lib; some are affiliated with the Digital Library Initiative, or other federally funded projects; and some are independent groups.

The following working groups in Digital Library research are currently associated with D-Lib:

- Metadata to Describe Information in Digital Libraries
- User Needs Assessment and Evaluation
- Social Aspects of Digital Libraries
- Repository Interfaces
- Digitization and Conversion
- Naming Objects in the Digital Library
- Networked Computer Science Technical Report Library (NCSTRL)
- Task Force on Archiving Digital Information

D-Lib is also a sponsor of:

- The First ACM International Conference on Digital Libraries: Program and Proceedings

D-Lib is coordinated by CNRI and is sponsored by the Defense Advanced Research Projects Agency (DARPA) on behalf of the Information Infrastructure Technology and Applications (IITA) Working Group of the High Performance Computing and Communications (HPCC) program.

Home T-spotlight DL Research

wya/af/reb-a
Last revised: June 14, 1996

WORKING GROUPS

D-Lib Working Group on Metadata to Describe Information in Digital Libraries

Joint Chairs: Michael F. Goodchild, Terence R. Smith, University of California, Santa Barbara

If sense is to be made of the flood of information that will be available through digital libraries, it must be described effectively, so that it can be found, its value assessed, and its acquisition handled efficiently. Metadata is the term most often used to refer to the description of information objects to support these three functions of digital libraries. Digital library technology is capable of both supporting major augmentations to traditional metadata activities and providing a basis for catalog interoperability.

D-Lib is associated with two activities in this field. Both focus on the process by which creators of digital information can add metadata to their work at the time of creation. This metadata is then available for computer programs to use in building indexes and other access tools. It is also available as a basis for subsequent cataloguing or the creation of secondary information services.

The first of these activities comes out of the Alexandria Digital Library project at the University of California, Santa Barbara. This project concentrates on geospatial information, such as maps, but its studies of metadata are broad based and applicable to all types of on-line data. Alexandria is one of the projects in the ARPA/NSF/NASA Digital Library Initiative (DLI) and its metadata studies involve members of several of the other DLI projects.

The second activity is the Metadata I and Metadata II invited workshop series. The first of these was sponsored by OCLC and NCSA in March 1995, chaired by Stuart Weibel of OCLC. Its major contribution was the "Dublin Core" metadata elements. D-Lib has agreed to be a sponsor of subsequent workshops.

These two activities are inter-related. In particular, Alexandria is using the Dublin Core as a building block for its own developments.

wya/reb-a
Last revised: March 17, 1996

WORKING GROUPS

D-Lib Group on Naming in Digital Libraries

The D-Lib Group on Naming in Digital Libraries covers all aspects of naming of digital resources. This topic, which appears simple on the surface, proves to be remarkably subtle when applied to the complex world of digital libraries.

For several years, the Internet Engineering Task Force (IETF) was a focus for efforts to develop Uniform Resource Names (URNs). These are globally- unique, persistent, location-independent names that can be applied to any network resource. This work is being continued by an informal group of URN implementors. The focus of the D-Lib group is on the next stage, how to use names in large scale libraries.

User groups that wish to assign names to objects in a digital library are faced with a variety of issues. One type of question is the relationship of names to semantic concepts such as uniqueness, mutability, etc. Are these managed by the naming system or by an external system? In a large library, rules and conventions for assigning names can be very complicated. If users are to see the names, it is helpful if they have some structure to help them be remembered or recognized, but there are real dangers in attempting to embed semantic information into names.

There will be many naming schemes. Some, already exist and must be merged into the digital library. The integration of naming schemes is a technical challenge and an organizational one, requiring decisions about the registration of naming schemes, and the allocation of top-level names.

Few digital objects exist by themselves. They are parts of larger groups or made up of many components. Naming such complex and compound items proves to be intimately connected to questions of what metadata to keep for each component and how to represent the relationships between them. Proposed solutions include composite objects (which contain several separate objects) and meta-objects (which provide links to other digital objects).

Finally, all questions of naming must consider scale. Processes for naming and organizing small numbers of objects may be totally inadequate for large collections.

Home W Groups

wya
Last revised: February 5, 1996

WORKING GROUPS

D-Lib Working Group on Repository Interfaces

Chair: William L. Scherlis, Carnegie Mellon University

This working group focuses on technical issues associated with repository interoperation. As digital libraries proliferate, many approaches to managing digital assets and associated meta-data are emerging. There are important differences among these approaches, and these differences have technical, legal, social, economic, and political dimensions. How can multiple repositories coexist and interact effectively?

The working group is motivated by several important trends: The complexity and semantic richness of objects and meta-data managed by repositories is increasing. Information objects of greater value are now being managed more routinely, raising issues of security, access control, and support for commerce. Performance demands are increasing, as is the quantity and size of information objects, particularly in multimedia applications. Digital libraries are interacting more often with personal, group, and wide area information services. Finally, the distinction is blurring between digital libraries and other institutional information resources such as databases and corporate webs.

The starting points for the working group are technologies that support management of information objects, their names, and associated meta-data-databases, distributed file systems, object bases, and the Web. Several digital library research groups have started to develop concepts that could provide a basis for repository interoperation, including the CS-TR architectural work of Kahn and Wilensky, the Stanford Infobus project of Garcia-Molina and Winograd, and the agent architecture of the Michigan DLI project. In addition to the need to reconcile these various approaches, there is a broader need to put them in the context of standards efforts in the wider community, including Web-associated standards, CORBA, OLE, z39.50, and SQL and its successors. All of these deal with resolving names to objects, and all deal in some measure with meta-data.

The initial effort of the working group is (1) to identify the dimensions of the space of repository interaction and interoperability, and the issues associated with achieving some transparency for users of the digital libraries, and (2) to assess current research and development efforts to understand the differences among them.

Home W Groups

wya/reb-a
Last revised: February 5, 1996

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.

Home W Groups

clb/wya
Last revised: March 18, 1996

d-lib magazine

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

d-Lib forum d-Lib magazine

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").

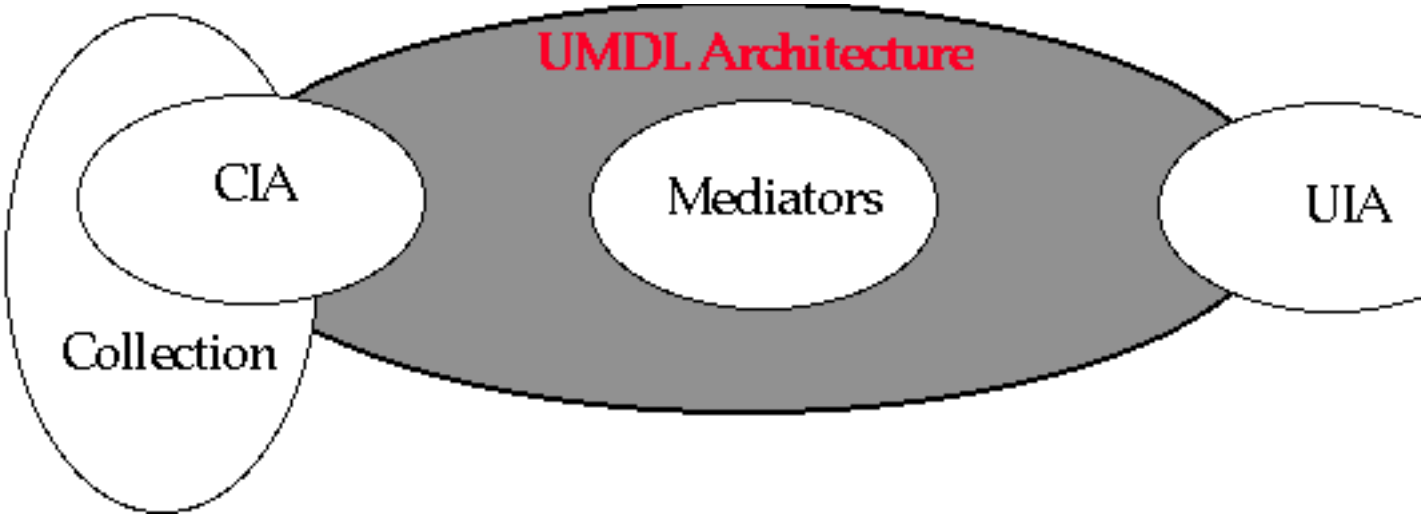


Figure 1: UMDL agent types

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.

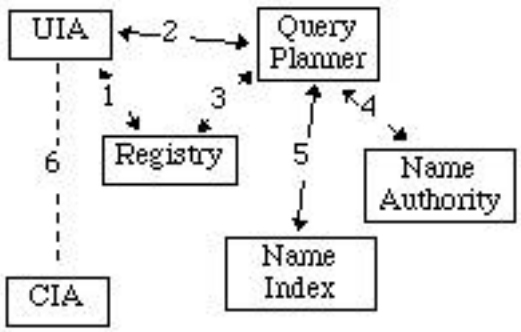


Figure 2: Example search by author

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.

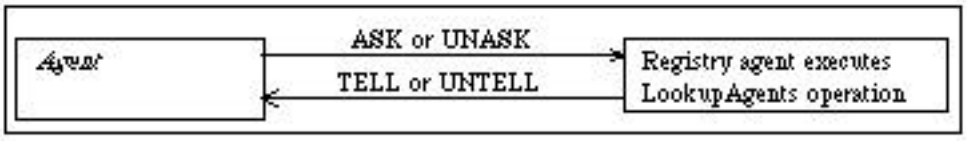


Figure 3: Agent-identification protocol.

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 scaleability, 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. Fritzon, 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.

d-lib magazine

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

d-Lib forum d-Lib magazine

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 responsibility. 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 a 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.

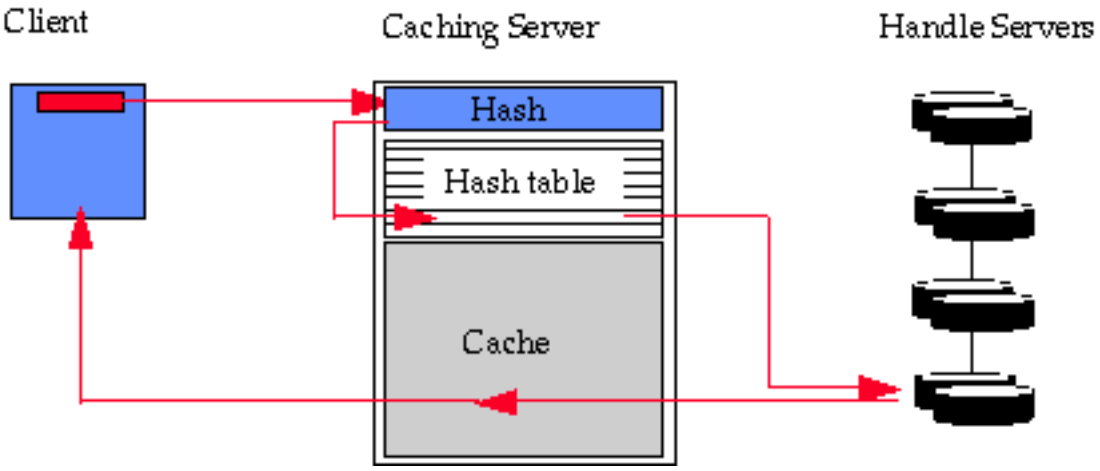


Figure 1. The CNRI handle system

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".

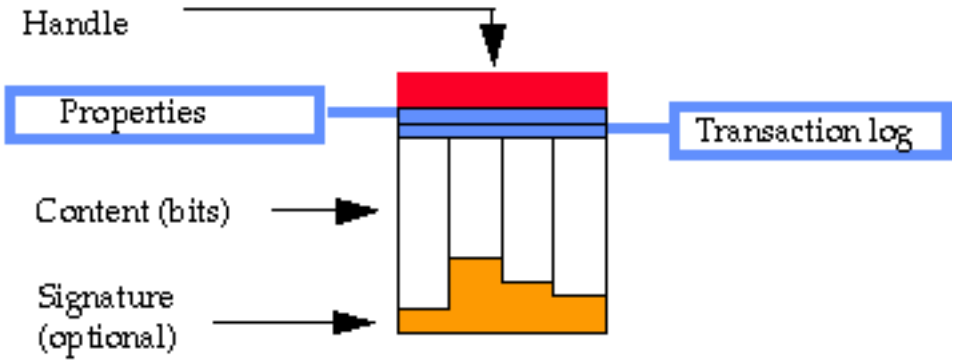


Figure 2. Parts of a digital object

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.

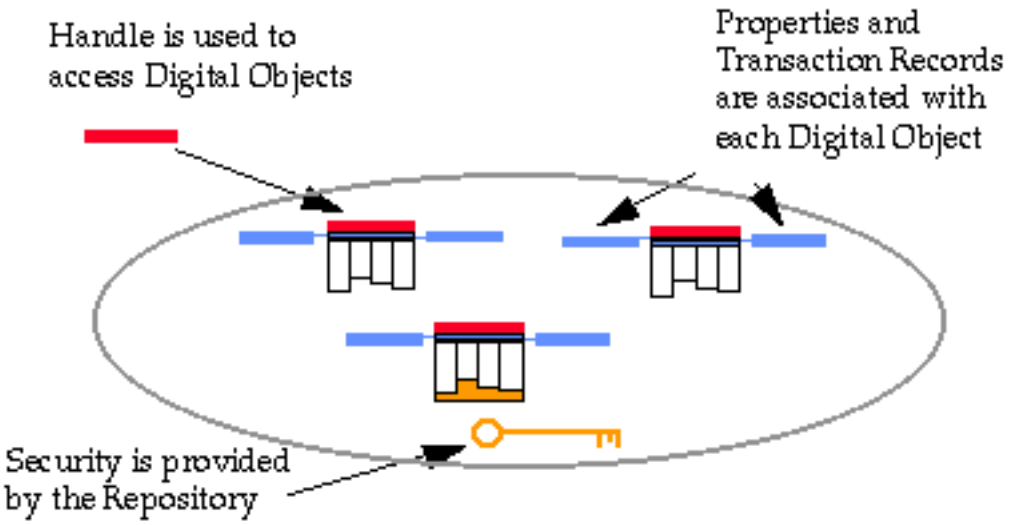


Figure 3. A repository

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.

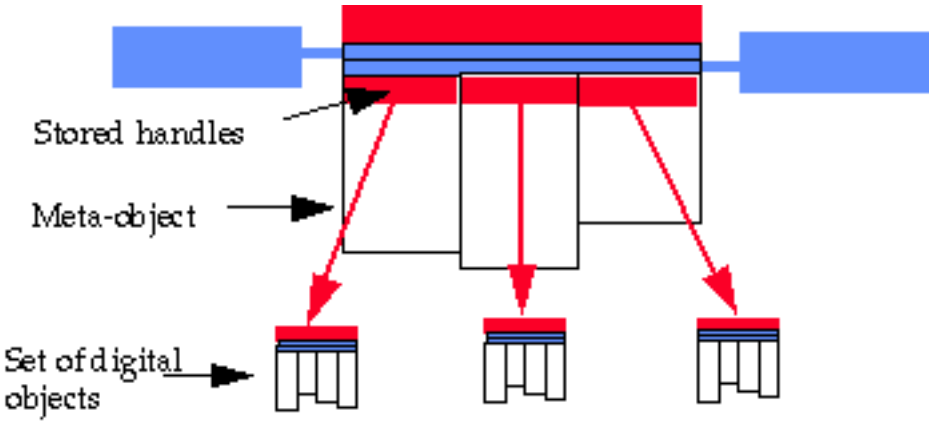


Figure 4. A digital object used as a catalog record

Reference

hdl:cnri.dlib/tn95-01 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)

---

d-Lib forum d-Lib magazine

---

hdl:cnri.dlib/july95-arms.html

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

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/] , 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 Virgina 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/iiir
- 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).

---

d-Lib forum d-Lib magazine

---

hdl:cnri.dlib/july95-weibel

MAGAZINE

Uniform Resource Names

A Progress Report

The URN Implementors

D-Lib Magazine, February 1996

ISSN 1082-9873

Introduction

The development of networked information requires reliable ways to name resources on networks. The Internet community has adopted the term, "Uniform Resource Name (URN)", for a name that identifies a resource or unit of information independent of its location. URNs are globally unique, persistent, and accessible over the network.

The concept of universal names has been warmly embraced by the networking and library communities, but convergence on the details proved difficult until recently. During fall 1995, however, members of the principal groups that are actively working in the field reached outline agreement on most of the major topics. The main characteristics of this agreement are described in this paper.

The catalyst for the recent progress was a meeting in October 1995 hosted by Keith Moore at the University of Tennessee. Invitations were sent to every group that had a current Internet draft on this subject. The URN groups represented are listed at the end of this report. This meeting was followed by a series of discussions including informal sessions at the December meeting in Dallas, Texas, of the Internet Engineering Task Force (IETF).

Convergence is important because many people who manage large collections of on-line information have been reluctant to commit to using any form of URN during a period of flux. The present consensus has two major results:

- Users who wish to give permanent names to on-line resources can now plan to incorporate URNs from existing naming schemes in documents, indexes, and on-line systems. They can be reasonably confident that future developments of the URN framework will not force them to reformat or otherwise modify existing URNs.

- The implementation of this framework will remove the concern that using a particular name scheme might affect longevity or the future usefulness of assigned URNs. The framework allows continued support for existing URNs, through other resolution systems, if one name scheme ceases to be supported in its original form. Thus users who assign names within any of the agreed-upon schemes are assured against obsolescence.

This report summarizes the emerging consensus. A strength of the framework is that it allows different approaches to be pursued, and the framework has the ability to evolve over the long term. Naming is a complex issue and the groups are interested in URNs for a variety of different reasons. They bring different philosophies and different technical approaches. Their implementations range in scope and complexity. It is therefore encouraging for the community that they have reached general agreement and are working together to find technical solutions to the outstanding questions.

Background

A good introduction to URNs is Internet RFC 1737, "Functional Requirements for Uniform Resource Names", by Karen Sollins and Larry Masinter, December 1994. The following is an extract from their introduction. It describes the function of URNs and, in particular, how they differ from the Uniform Resource Locators (URL) used by the World Wide Web.

> "A URN identifies a resource or unit of information. It may identify, for example, intellectual content, a particular presentation of intellectual content, or whatever a name assignment authority determines is a distinctly namable entity. A URL identifies the location or a container for an instance of a resource identified by a URN. The resource identified by a URN may reside in one or more locations at any given time, may move, or may not be available at all. Of course, not all resources will move during their lifetimes, and not all resources, although identifiable and identified by a URN will be instantiated at any given time. As such a URL is identifying a place where a resource may reside, or a container, as distinct from the resource itself identified by the URN."

The RFC concentrates on the relationship between a locator (URL) and a persistent name (URN), but naming questions arise in many other contexts. For example, the Resource Cataloging and Distribution System (RCDS), developed in the Computer Science department of the University of Tennessee, uses URNs to support cataloging, replication and caching (for high availability and fault-tolerance), and authenticity and integrity assurances using digital signatures. The paper "A Framework for Distributed Digital Object Services" by Robert Kahn and Robert Wilensky, May 1995, also identifies persistent names assigned to objects in repositories as a key component of a framework to manage intellectual property on networks.

A class of names with some characteristics similar to URNs are the domain names (such as "andrew.cmu.edu"), used to identify computer systems on the Internet. Domain names are supported by a well-tuned computer system, the Domain Name System (DNS). Several URN implementations build on domain names and DNS.

URN Requirements

RFC 1737 lays out functional requirements for URNs. It also makes recommendations about the form that such names might take. An updated version of RFC 1737 is under discussion, but, with some important clarifications, the following list of requirements has been widely accepted.

> "**Global scope**: A URN is a name with global scope which does not imply a location. It has the same meaning everywhere.
>
> "**Global uniqueness**: The same URN will never be assigned to two different resources.
>
> "**Persistence**: It is intended that the lifetime of a URN be permanent. That is, the URN will be globally unique forever, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.
>
> "**Scalability**: URNs can be assigned to any resource that might conceivably be available on the network, for hundreds of years.
>
> "**Legacy support**: The scheme must permit the support of existing legacy naming systems, insofar as they satisfy the other requirements described here. ...
>
> "**Extensibility**: Any scheme for URNs must permit future extensions to the scheme.
>
> "**Independence**: It is solely the responsibility of a name issuing authority to determine the

conditions under which it will issue a name."

Notice that these requirements focus on the URN, but make no assertions about the resource that it identifies. A URN may be globally unique and last for ever without any guarantee that the resource identified by the URN is unique or permanent.

Resolution

To use a URN, there must be a network-accessible service that can map the name onto the corresponding resource. This process is called **resolution**.

Frequently, the resolution system will return the current location of the resource or a list of locations. RFC 1737 concentrates on the case of a URN that resolves to a URL, but a URN can resolve to any network resource or service. For example, in RCDS, a URN may resolve to one or more location-independent file names (LIFNs), which can themselves be considered a specific type of URN. In the Kahn/Wilensky model a URN, known as a "handle", resolves to the name of the repository that holds the resource. In other contexts, a URN may resolve to a data structure containing meta-information about the resource.

The URN Framework

This section describes the URN framework that has emerged from the discussions of the past few months. Although many details remain, the level of agreement is promising.

General Principles

- **Naming schemes and resolution systems.** The framework distinguishes between naming schemes and resolution systems. A naming scheme is a procedure for creating and assigning unique URNs that conform to a specified syntax. A resolution system is a network-accessible service that stores URNs and resolves them.

- **Independence between naming schemes and resolution systems.** A naming scheme is not tied to a specific resolution system. Any resolution systems is potentially capable of resolving a URN from any given name scheme.

- **URN registries.** Since naming schemes and resolution systems are conceptually independent, mechanisms must be created so that the user of a URN can discover what resolution systems are able to resolve the URN. This is called a URN registry or simply a registry.

Multiple independent naming schemes and resolution systems are anticipated. Although the maintainer of a particular URN resolution system may also wish to maintain a registry, it is important to realize that registries and URN schemes are conceptually independent of one another. Any registry is capable of registering resolution services for any URN scheme, and a client may wish to consult multiple registries when attempting to resolve a name.

Syntax

The URN implementors have agreed on the following syntax, with one outstanding difference of opinion; opinions differ whether the leading characters "urn:" should be part of the name. This syntax is acceptable in all proposed naming schemes and resolution systems. There are many details that need to be discussed (for example the precise character sets allowed in URNs).

The following are examples of URNs:

urn:hdl:cnri.dlib/august95
urn:lifn:some.domain:anything-goes-here

urn:path:/A/B/C/doc.html
urn:inet:library.bigstate.edu:aj17-mcc *{Correction to this entry made with permission from the authors. Ed., 2/19/96.}*

Notice that the syntax of a URN explicitly indicates the naming scheme, by including a naming scheme identifier, "hdl", "lifn", "path", "inet", etc. This is followed by a colon and a string that has a syntax defined by the specific naming scheme.

As can be seen from the examples, the different naming schemes use different formats. Some naming schemes divide the name into two parts, a naming authority followed by a unique string, which is assigned by the naming authority. Thus the handle "cnri.dlib/august95" consists of a naming authority, "cnri.dlib" followed by a unique string, "august95". The path URN "/A/B/C/doc.html" consists of a naming authority (or path), "/A/B/C", and a unique string, "doc.html".

The Internet community is developing a general framework of uniform resource identification (URIs), of which URNs are a component. The URI framework was originally outlined in RFC 1630. Under the proposed framework, each participating naming schemes is a URI as defined in the RFC.

Management of Naming Schemes

The long term value of URNs requires the naming schemes to be well managed. Initially, a small number of schemes are under development. Hopefully, a small number of high quality naming schemes will be added in the future.

The criteria for an acceptable URN scheme will be outlined more formally as the URN framework is defined. They are likely to include a requirement that each naming scheme must have a verifiable management system to ensure the integrity of the naming scheme and of the URNs within it. This includes the process for assigning unique URNs within the naming scheme. It must also make sure that there is at least one resolution system able to resolve the names.

Those URN schemes that include naming authorities (e.g., handles, paths) will determine the names of the authority names themselves. Thus, it is possible that different organizations may get the same naming authority string under different naming schemes.

URN Registries

A URN registry is a network service that stores data about URN naming schemes, naming authorities, and resolution systems. A registry provides two types of service. It may provide rules for extracting the naming authority from URNs in a particular naming scheme. In this case, the first step of the URN resolution service may be to provide information on how to find the naming authority in the URN string. The second function is to know which resolution systems are capable of resolving a given URN, from the name scheme and, when appropriate, the naming authority.

The concepts of URN registries and resolution systems are not tied to any specific computing system or set of software. This is important since URNs are intended to be valid for long periods of time, much longer than any computer system can be expected to last. The format of data to be stored in a registry is currently under development. It has been given the working name NAPTR ("Naming Authority PoinTeR"). In practice, it is probable that several URN resolution systems will include URN registries, but every registry need not hold full information for all naming schemes. One proposed implementation is a modified version of DNS. Another uses the handle system.

Flexibility within the URN Naming Schemes and Resolution Systems

This report emphasizes the areas where the various URN developments are converging on a common framework. In a number of key areas, the URN implementors have carefully agreed to support flexibility

rather than to enforce unnecessary conformity.

The value of a naming scheme or a resolution system depends upon a number of assertions. Are the names unique? Can a resource have many names? Can it change? Is it guaranteed to exist? What is the retention scheme? Does a URN resolve to untyped data, typed data, entity-attribute pairs, a URL, the address of a repository, etc.? Within the general URN framework, such assertions about names, semantic decisions, and management issues may be enforced by the naming scheme or the resolution system, or they may be left to external systems. Variations in these important areas will give the schemes their distinctive features and will determine which are most suitable for specific application areas. The objective of the URN framework is to encourage wide flexibility within a stable system of naming and resolution.

URN Implementors

The following projects were represented at the University of Tennessee meeting in October 1995 and have continued to work together to reach agreement on the URN framework.

Resource Cataloging and Distribution Service (RCDS)
: This work is led by Keith Moore, Shirley Browne, Stan Green and Reed Wade of the University of Tennessee. Its aim is to provide transparent replication along with integrity/authenticity assurances, and alleviate the problem of huge demand for some random network resource.

The Handle System
: This work is led by David Ely and William Arms of the Corporation of National Research Initiatives. It is based on the ideas in the Kahn/Wilensky framework.

x-dns-2
: This is a scheme developed by Paul E. Hoffman of Proper Publishing and Ron Daniel, Jr. of Los Alamos National Laboratory. As the name implies it is based on the Internet domain name system (DNS).

URN Services
: This is a proposal by Keith E. Shafer, Eric J. Miller, Vincent M. Tkac, and Stuart L. Weibel of OCLC. It focuses on the syntax and functions of URNs.

Path URN
: This is another scheme that make use of DNS. It has been developed by Dan LaLiberte and Michael Shapiro at the National Center for Supercomputing Applications.

Whois++
: Several groups are working towards using Whois++ as an Internet Directory Service. Work done by Michael Mealling of Georgia Tech and Patrik Faltstrom and Leslie Daigle of Bunyip Information Systems, Inc., focuses on the distribution of URN resolution data and maintenance responsibility in a global publishing environment.

Contributors to this report

The following URN implementors contributed to this report: William Arms (CNRI), Leslie Daigle (Bunyip), Ron Daniel (Los Alamos National Laboratory), Dan LaLiberte (NCSA), Michael Mealling (Georgia Institute of Technology), Keith Moore (University of Tennessee), and Stuart Weibel (OCLC).

Home Magazine Comments
NEXT ▶

d-lib magazine

Working Towards an Understanding of Digital Library Use

A Report on the User Research Efforts of the NSF/ARPA/NASA DLI Projects

Ann Peterson Bishop
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
218 LIS Building, 501 East Daniel Street
Champaign, IL 61820
abishop@uiuc.edu

D-Lib Magazine, October 1995

Introduction

The Digital Library Initiative (DLI) projects, funded jointly by the National Science Foundation (NSF), the Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA) began about a year ago. Their user study teams have already produced some valuable findings and described some provocative theoretical and methodological challenges. From my vantage point as coordinator of the University of Illinois DLI Social Science team, I will highlight the efforts of the six projects to communicate with each other about user research. Links to the DLI project home pages and to some of the papers published by project members have been included for more in-depth coverage of some of the issues summarized. In this article, I will also discuss the upcoming Allerton Institute at the University of Illinois, a methodological forum on digital library use that will provide another means for researchers in a variety of disciplines and settings to share their ideas and concerns about the conduct of social science research related to digital library use.

The Growth of Digital Libraries and the Challenge of Understanding Their Use

Improvements in information technologies and increased support directed towards our national information infrastructure have led to the development of a wide range of digital library collections and services. Academic, special, and public libraries are implementing on-line systems that provide their patrons with electronic access to library catalogs and a variety of other information resources. NASA is developing on-line collections of images and data for scientists and engineers. Museums are digitizing their collections and making them available on the Internet. Members of scientific communities are building collaboratories to support their work and communication. Publishers are experimenting with the creation of digital archives of their journals and books. And individuals and groups from all walks of life are using community-based networks to provide local and global access to information resources they have created. In addition to this array of existing networked information tools and resources--all of which can be thought of as variations on theme of the digital library or as pieces and layers of the digital information infrastructure--research and development projects related to building the next generation of digital library systems are also flourishing.

Digital libraries pose fascinating socio-technical challenges for understanding their use. Those supporting the construction of digital libraries are naturally concerned that their investments pay off in terms of attracting users and making information services more effective and efficient. The design and evaluation of digital libraries,

however, are complicated by the newness of the systems, their ability to integrate a range of functions that were previously designed and evaluated separately, the heterogeneity of their user population, the physically distributed nature of usage, the ability to fragment and rearrange previously integrated documents and images, and the rapid versioning of digital objects. Appropriate user-centered research objectives, measures, and methods for the digital library are just beginning to emerge.

Results from user studies can help digital library designers and policymakers formulate appropriate goals, arrive at a more complete understanding of costs and benefits, design and allocate resources to both technologies and programs that offer the best means of achieving goals, and assess the degree to which network policies and programs have achieved their stated goals. Granted, determining (let alone predicting) impacts from information technology at the individual, organizational, and societal levels is notoriously difficult. But without such investigations, the views, needs, and experiences of individual information creators and consumers will be lost in the push and pull of constituencies with more powerful and direct voices in the process of building digital libraries, a process in which users themselves are all too apt to be considered mere passive consumers in the technology-implementation chain.

How can we learn more about the use and users of digital libraries? How can people involved in user-centered studies associated with the vastly different kinds of digital library initiatives described above share their ideas, concerns, methodological approaches, and findings? I would like to turn now to describing digital library user research, and mechanisms for sharing that research, that I have become involved with as a participant in the NSF/ARPA/NASA Digital Library Initiative.

Synchronizing Work Across the Six DLI Projects: The Role of the User Research Working Group

With encouragement from project sponsors, we have established an informal DLI-wide User Research Working Group. The motivation for the working group stems from our sense that the six DLI projects are similar enough that we can learn from each others' experiences. In addition, we have found that each user research group has different strengths. While our group at University of Illinois, for example, is especially strong in ethnographic approaches to studying system use, other groups have had more experience with conducting usability tests and designing system instrumentation. Common problems include the need to develop new methods for tracking distributed "virtual" users, difficulties in integrating and making sense of data vrom various quantitative and qualitative sources, and dealing with a large and heterogeneous user population.

For this first year, our interactions have been somewhat limited in scope and informal in nature. We get together twice a year at the DLI synchronization meetings and have set up a mailing list for working group members. At the Spring 1995 meeting, we discussed our basic approaches to user evaluation and the recognition that evaluation can proceed at different levels, to reach different goals. Summarizing our discussion, Karen Drabenstott of the University of Michigan suggested the following schema for evaluation levels:

- Adequacy of the collection, functionality, interface, usability.
- Search and retrieval performance and behavior.
- Effect on work of users, fundamental changes in processes
- Public policy implications

It was clear that the six projects are devoting varying amounts of attention to each of these evaluation levels.

We also realized that there were a number of unresolved issues confronting virtually all the user research groups. We discussed the way, for example, in which the new phenomena engendered by digital library technology lead to exploring unfamiliar methods and conceptual realms. Another major dilemma we all faced was figuring out how best to produce useful results for our system designers. Problems arose in juggling conflicting schedules, maintaining effective communication, and knowing how to make our findings operational. We concluded that there were new pulls on both system designers (new ways of thinking about

use and users) and social scientists (new approaches to studying systems), so that it was important to try to keep dialogue open among users, sponsors, social scientists, and computer scientists.

We agreed that members of our cross-project working group would present a status report on goals, methods, results, and problems for each synchronization meeting. We also agreed that we would try to facilitate cross-team sharing through posting our working papers, including instruments, on our project home pages, and that members from each project would complete brief user research "templates" (see Figure 1) to describe their work.

Figure 1: DLI User Research Template

- 1) Capsule summary (3-5 sentences) of the DLI project itself (i.e., what kind of DL your project team is building)
- 2) Age and experience levels of the users you're studying
- 3) Evaluation "levels" you're addressing in your user research:
    - a. Adequacy of the collection, functionality, interface, usability
    - b. Search and retrieval performance and behavior
    - c. Effect on iwork of users, fundamental changes in processes
    - d. Implications for public policy
- 4) Use settings you're studying (e.g., laboratories, public libraries, high schools)
- 5) Methods you're using
- 6) Theoretical considerations of particular import in your work

---

Based on information provided by members of each DLI project, I have prepared capsule descriptions of each project's user research efforts (http://anshar.grainger.uiuc.edu/dlisoc/home_page.html/user_research_wg).

Building Understanding Beyond the DLI Projects: The Allerton Institute

One recommendation that arose from the User Research Working Group at the Fall 1994 DLI synchronization meeting was to find a way that we could get together with other interested researchers to explore methodological approaches associated with the use of digital libraries. This recommendation has been realized in the 1995 Allerton Institute conducted by the Graduate School of Library and Information Science at the University of Illinois, titled "How We Do User-Centered Design and Evaluation of Digital Libraries: A Methodological Forum." The Institute, sponsored by the National Science Foundation (NSF), will be held on October 29-31.

As chairperson, my goal is to bring together an interdisciplinary group of researchers and practitioners involved in the design and study of information systems, in user-centered research in traditional libraries, and in a wide range of digital library projects. The purpose of the forum is to present both the range of user-centered methods available for studying digital libraries and rationales for choosing amongst them; we also want to look ahead to new methods and developments and map out the challenges that lie ahead. This methodological forum will give the 60 invited participants an opportunity to share their expertise, experiences, and ideas with their peers in a relaxed environment. Forum activities will be devoted to issues such as:

- What are appropriate measures for gauging digital library outcomes at the individual, group, institutional, and global levels?
- How can we best incorporate knowledge of user needs and behavior in designing digital library interactions and interfaces?
- What do we need to know about how people use electronic texts and how can we gain this knowledge and apply it to the development of digital libraries?
- What can we learn from studies of traditional library use?
- How can we develop an understanding of the computerization of library work that will help as digital

systems are incorporated into current institutional practices?
- How can we deal with the ethical, practical, and conceptual issues that arise in the remote observation of on-line (and off-line) behavior on a very large scale?
- How do we foster effective communication among digital library designers, users, and social science researchers?

Each participant submitted a brief discussion document outlining their work and the issues they were most eager to explore. These papers were used to develop the five major Institute sessions, which will focus on co-design approaches, work practice and institutional change, migrating foundational approaches to virtual library environment, electronic information seeking behavior, and understanding diversity and change. Participants include researchers from the fields of Computer Science, Sociology, Library and Information Science, Education, and Psychology who are involved in digital library projects in a wide range of settings. Presentations will be given by a number of participants, including Michael Twidale, Annelise Mark Pejtersen, William L. Anderson and Susan L. Anderson, F. W. Lancaster, Andrew Dillon, John M. Carroll, Brenda Dervin, Rob Kling, Chip Bruce, and Gary Marchionini. Discussion documents from participants, plus perhaps other related material from the Institute, will be made publicly available at some point after the Institute.

I hope that the user research efforts of the DLI projects, along with the ideas arising from the Allerton Institute, will help in building a framework for understanding the use and implications of digital information infrastructure, as our research methods, systems, and expectations of systems continually evolve. By situating the study of DLI testbed use within the broader context of professional work and social practices, I believe we will gain more robust insights into the functions and features that will make digital libraries more effective. In addition, we will get a sense of large-scale changes in work and cognition that occur as the nation's entire information infrastructure begins to change.

References

- Bishop, A. P., & Bishop, C. M. (1995). The policy role of user studies. **Serials Review**, 21(1), 17-25.
- Buttenfield, B. P. (1995, draft). User evaluation for the Alexandria Digital Library Project. [Discussion document submitted for the 1995 Allerton Institute: "How We Do User- Based Design and Evaluation for Digital Libraries: A Methodological Forum"].
- Digital libraries [Special issue]. (1995). **Communications of the ACM**, 38(4).
- Finholt, Thomas A. (1995, draft). Understanding digital libraries: Possible lessons from the analysis of collaboratories. [Discussion document submitted for the 1995 Allerton Institute: "How We Do User-Based Design and Evaluation for Digital Libraries: A Methodological Forum"].
- Gaston, B. (1994, Sept. 27). NSF announces awards for digital libraries research. Washington, DC: National Science Foundation. (Available: http://www.grainger.uiuc.edu/dli/release.htm)
- Lasher, R. (1994, Oct. 11). Library issues for the Joint Initiative Digital Library Projects. Unpublished manuscript. (Available: http://www-diglib.stanford.edu/diglib/pub/dllibrary/library-issues.html)
- Reich, V. (1995, draft). Allerton discussion document. [Discussion document submitted for the 1995 Allerton Institute: "How We Do User-Based Design and Evaluation for Digital Libraries: A Methodological Forum"].
- Reich, V., & Weiser, M. (1994). Libraries are more than information: Situational aspects of electronic libraries. Serials Review, 20(3), 31-37. (Available: http://www.ubiq.com/hypertext/weiser/SituationalAspectsofElectronicLibraries.html)
- Van House, N. A. (1995). User needs assessment and evaluation for the UC Berkeley Electronic Environmental Library project. In F. M. Shipman, III, Richard Furuta, & David Levy (Eds.). **Proceedings of Digital Libraries '95: The second annual conference on the theory and practice of digital libraries** (pp. 71-76). College Station, TX: Texas A&M University. (Available: http://csdl.tamu.edu/DL95/papers/vanhouse/vanhouse.html)
- Van House, N. A. (1995, draft). Current project: UC Berkeley's NSF/ARPA/NASA Digital Libraries Project. [Discussion document submitted for the 1995 Allerton Institute: "How We Do User-Based Design and Evaluation for Digital Libraries: A Methodological Forum"].

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 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.

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 that 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

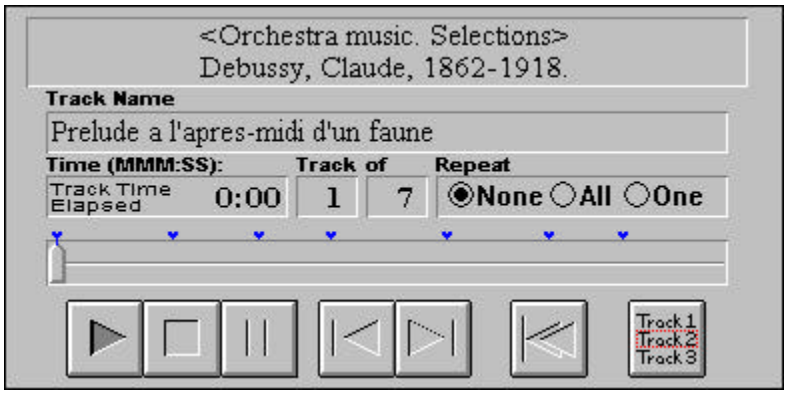
Library patrons access sound recordings in the system from 45 IBM Pentium computers located throughout the library and in a teaching

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 files 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

Home Magazine Comments
NEXT ▶

- Electronic Document Collections at Virginia Tech
- Imagebase (Uses PURLs and Dublin Core Metadata standard)
- Meeting Schedule
- Pointers to information about Digital Libraries
- *Project Reports*
- References: Research Department Virginia Tech Computing Center

---

References

- ***General***
  - Digital Libraries Research and Development Forum (D-Lib)
  - Florida Center for Library Automation's Digital Library Project
  - IBM Digital Library
- ***Specific Topics***
  - **Agents**
    - UMBC Intelligent Software Agent Resources
    - Survey of Intelligent Software Agents
    - More agent links
  - **Metadata**
    - Metadata: the Foundations of Resource Description
    - OCLC/NCSA Metadata Workshop Report
    - RFC-1807
    - TEI
  - **Naming**
    - Handles
    - PURL
  - **Z39.50**
    - Isite Software
    - Library of Congress WWW/Z39.50 Gateway/Info
    - Prise 1.0 Software
    - Willow

Some of these documents are in Adobe's Portable Document Format (PDF). In order to view them, you will need a PDF viewer

---

University Libraries, Virginia Tech
Send Suggestions or Comments to webmaster@scholar.lib.vt.edu
Last updated: October 1, 1996

URL: *http://scholar.lib.vt.edu/digilib/*

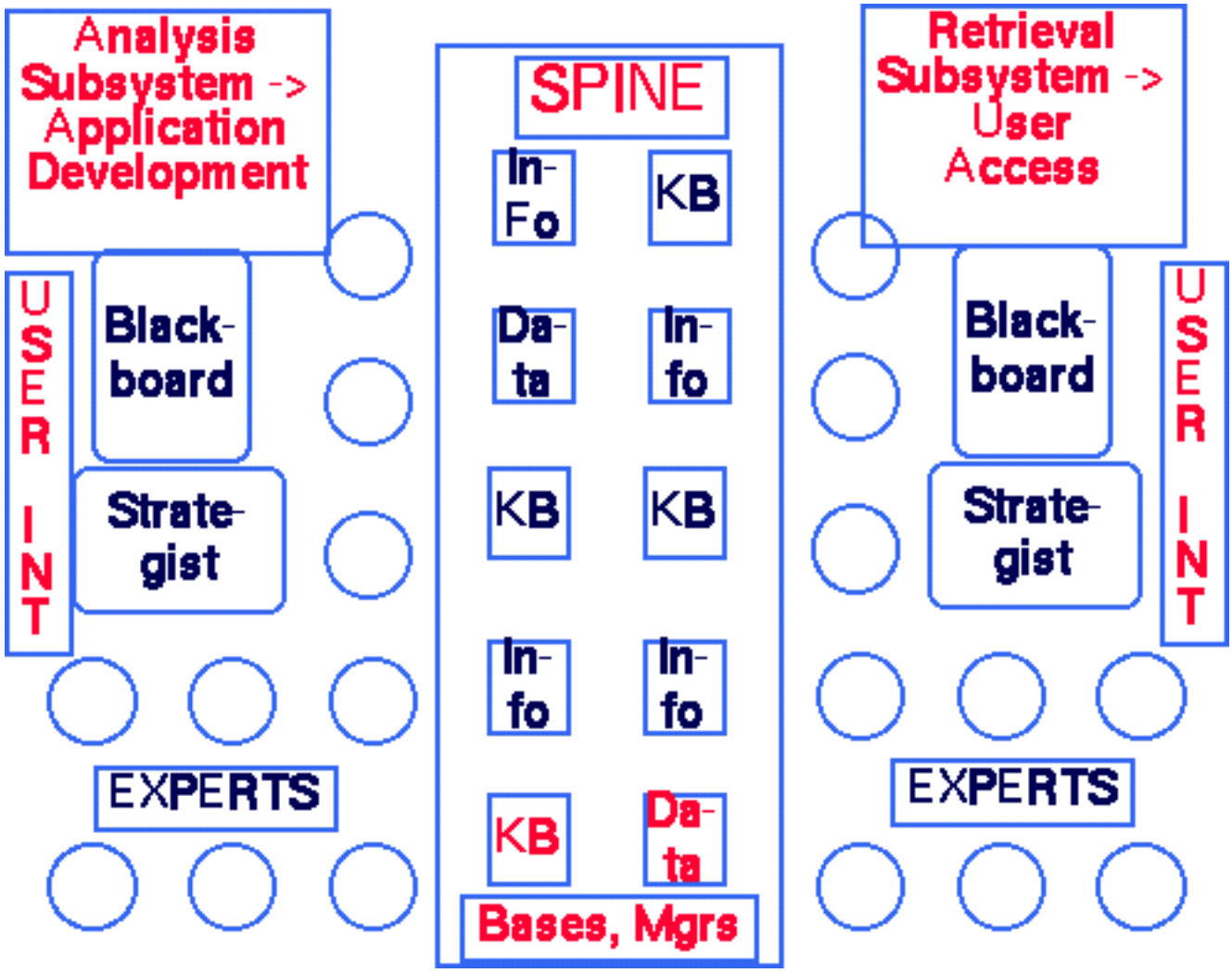
Digital Library Source Book, 1993, ed. E. Fox

To order a paper copy, or find out background information please look at the README file. To use an Adobe Acrobat Reader or Exchange to work with the book, look at the PDF version. Otherwise, use the PostScript version that appears below in sections.

See also more information of interest:

- April 1995 *Communications of the ACM*
- *Gladney et al. report on DL requirements and architecture (PostScript)*
- *PowerPoint presentation by Fox for 1994 Digital Libraries Workshop at Rutgers (to be decoded by binhex)*
- *PowerPoint presentation by Fox for 1994 Digital Libraries Workshop at Texas A&M (printable, in black and white, to be decoded by binhex)*
- *PowerPoint presentation by Fox for DL Keynote at EG-MM'94 in Graz (to be decoded by binhex)*
- *PowerPoint presentation by Fox for DL Keynote at ISMIS'94 in Charlotte (to be decoded by binhex)*
- *WWW Pages for CS2984 Course Notes on Digital Libraries*

COmposite Document Expert/extended/effective Retrieval (NSF: 1985-9)



Application Domains: 1985 -

- Electronic mail (AIList Digest, Collins English Dictionary)
- Navy messages (Naval Message Analyzer)
- Medical information on cardiology (400M)
- Campus Catalog - MARIAN (1G - underway)

Environment

- Communications: TCP/IP (supporting our own language for interprocess data/information/knowledge transfer)
- Operating Systems: versions of UNIX
- Programming Languages: C, C++, Prolog

MARIAN Mosaic Interface

To try out the MARIAN system you can use the MARIAN Mosaic interface which provides the top-ranked 30 items after searching against the Dec. 1993 version of the Virginia Tech catalog.

Overview

MARIAN (Multiple Access Retrieval of Information with ANnotations) is a system developed by the Virginia Tech Computing Center starting in 1991. It runs on a collection of NeXT computers, using one or more threads for each user session. A similarity value is computed for each field (e.g., title, author, subject). The combiner module computes an overall similarity that is used to rank the documents, so users see the top-ranked items only.

Using the NeXT interface, one can call for successive sets of 30 items. Also, one can request circulation data that is obtained from the VTLS computer by way of an expert system analysis module.

In addition there is a Gopher+ interface, one using the curses interface package and access using telnet.

In 1995 the MARIAN system should be opened for wider use, after the data is brought into synchronization with the current contents of the VTLS computer.

History

One precursor of MARIAN is the CODER system. Many of the ideas from it, the interprocess communication approach, and the English lexicon developed for it, are used in MARIAN.

The other precursor of MARIAN is the REVTOLC study --- Retrieval Experiment, Virginia Tech On Line Catalog. A pilot study was done with 300,000 records and 52 users. The full study is reported at length in the 1993 Ph.D. dissertation of Amjad Daoud. Students preferred our approach to VTLS. They also preferred ranked retrieval to standard Boolean retrieval. Details follow:

- Experiment with 500,000 records
- 4 methods compared
    - Boolean queries
    - Extended Boolean interpretation
    - Vector queries
    - Vector queries + probabilistic feedback
- Each person tried 2 methods, 2 queries/method
- 12 groups: 4 choices for method 1, 3 left for method 2
- 18 queries: taken from ones recorded in library
- Tested with 216 users, each spending 60-90 minutes

Welcome to

ARIAN

Searching the Virginia Tech Library Collection

Query

Quit MARIAN | Help

Preferences

*** NEWS ***

New format for document close-ups.

MARIAN is now using a new display format for individual items.
Use the "Comment" feature to tell us what you think of it.

Some more data is shown for each work in the libraray collection. More importantly, the information has been set up to be easier on the eye. The order has also been changed, for instance to bring all the authors (including illustrators, performers,and so forth) into a single place. We hope you find this organization sensible.

The short descriptions in the list of results from a search have been slightly cleaned up too.

MARIAN is a search system for library catalog data. The system you have reached searches the Virginia Tech Library collection, updated regularly.

This page and the WWW gateway to MARIAN are both under construction.

It is always better to use the buttons on MARIAN pages than to move back through your Web browser history pages and start from a page there. That's because the MARIAN gateway does not always interact well with pages cached in your Web browser. This page should be an exception, but even it may not be. If the gateway (or your browser) gets confused, reconnect to this URL and start again.

The MARIAN system is in the final phase of development. Things are pretty trustworthy, but the occassional failure is to be expected. As a rule of thumb, if your results look anything like what you expected, they are probably completely correct. If they are completely off-the-wall, or if nothing comes back, the system has probably dropped a ball. Try again later.

Envision

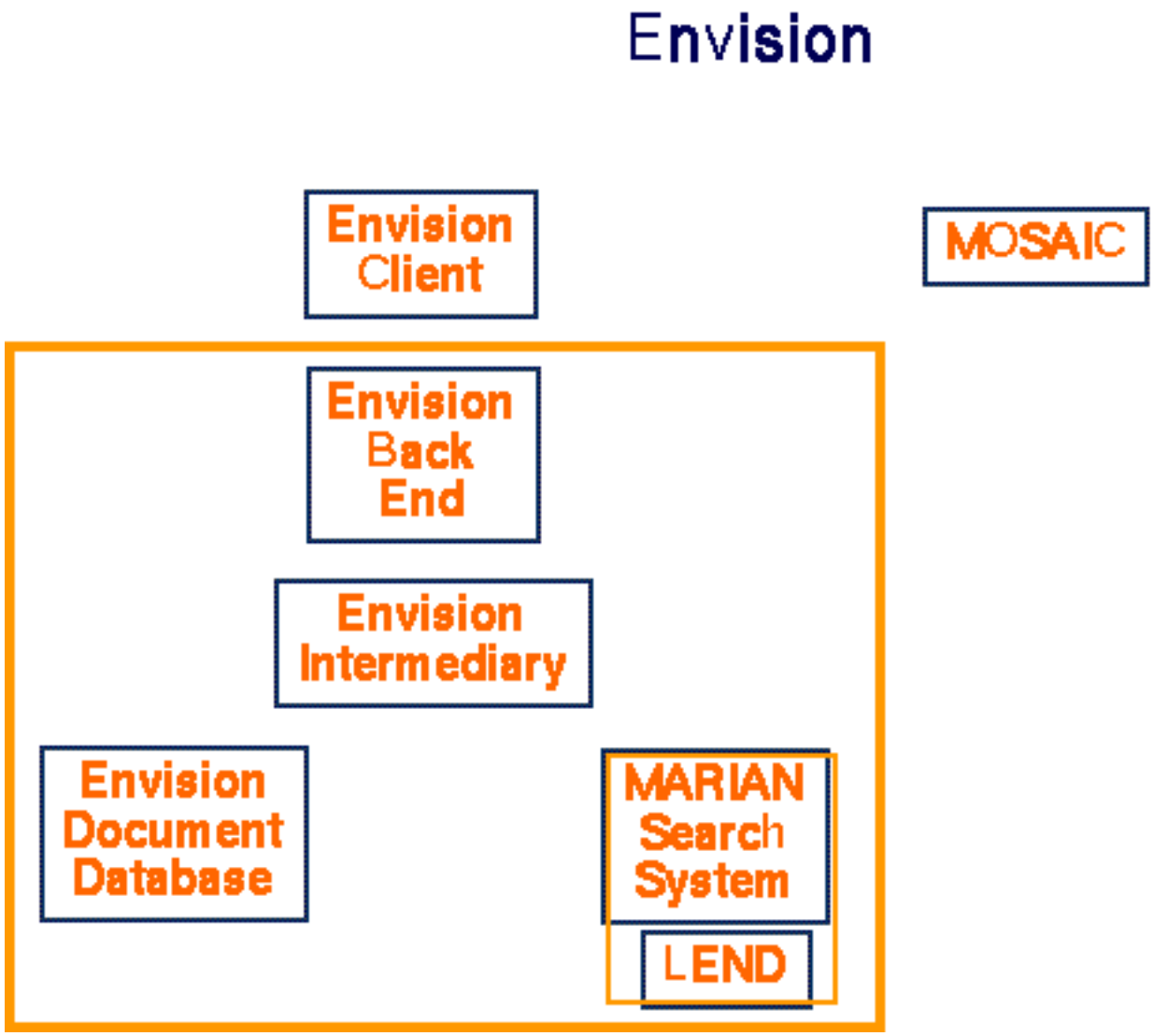
The Envision Project was funded as **A User Centered Database from the Computer Science Literature** by NSF for 1991-95. ACM has provided free access to their publications.

Efforts have concentrated on building an archive based upon SGML, developing an object-oriented database, applying the MARIAN retrieval system and WWW, and constructing a special search interface based upon user wishes.

The interface includes:

- a query screen
- a results list screen
- a results visualization screen
- Mosaic display of retrieved documents

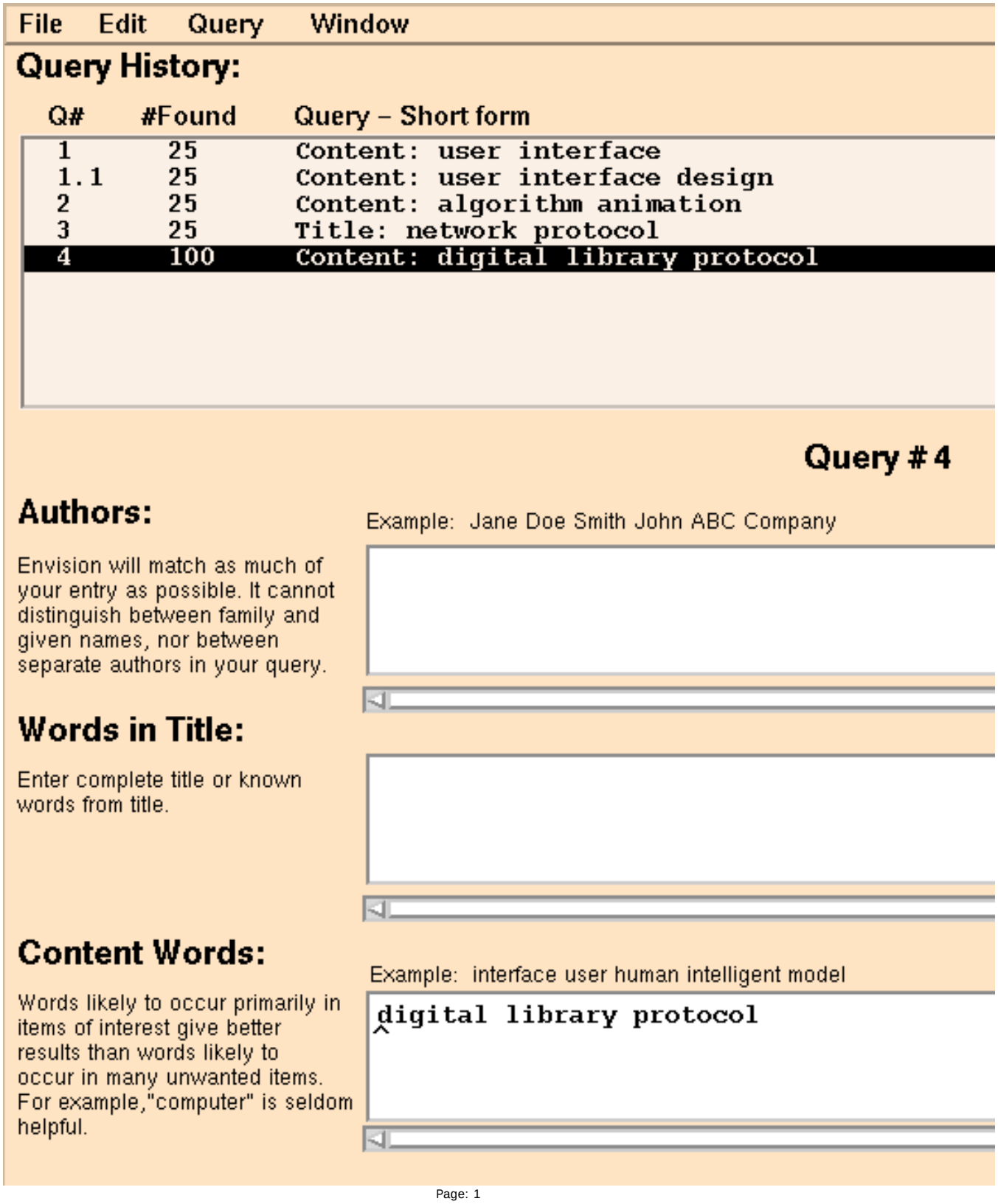
The system architecture is a combination of various elements:

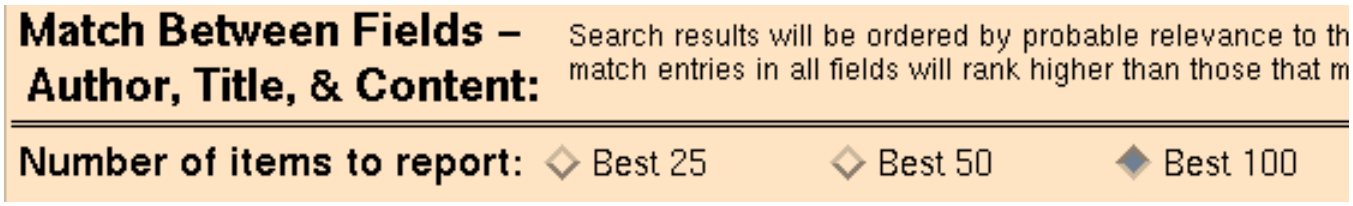


Envision - Query Screens

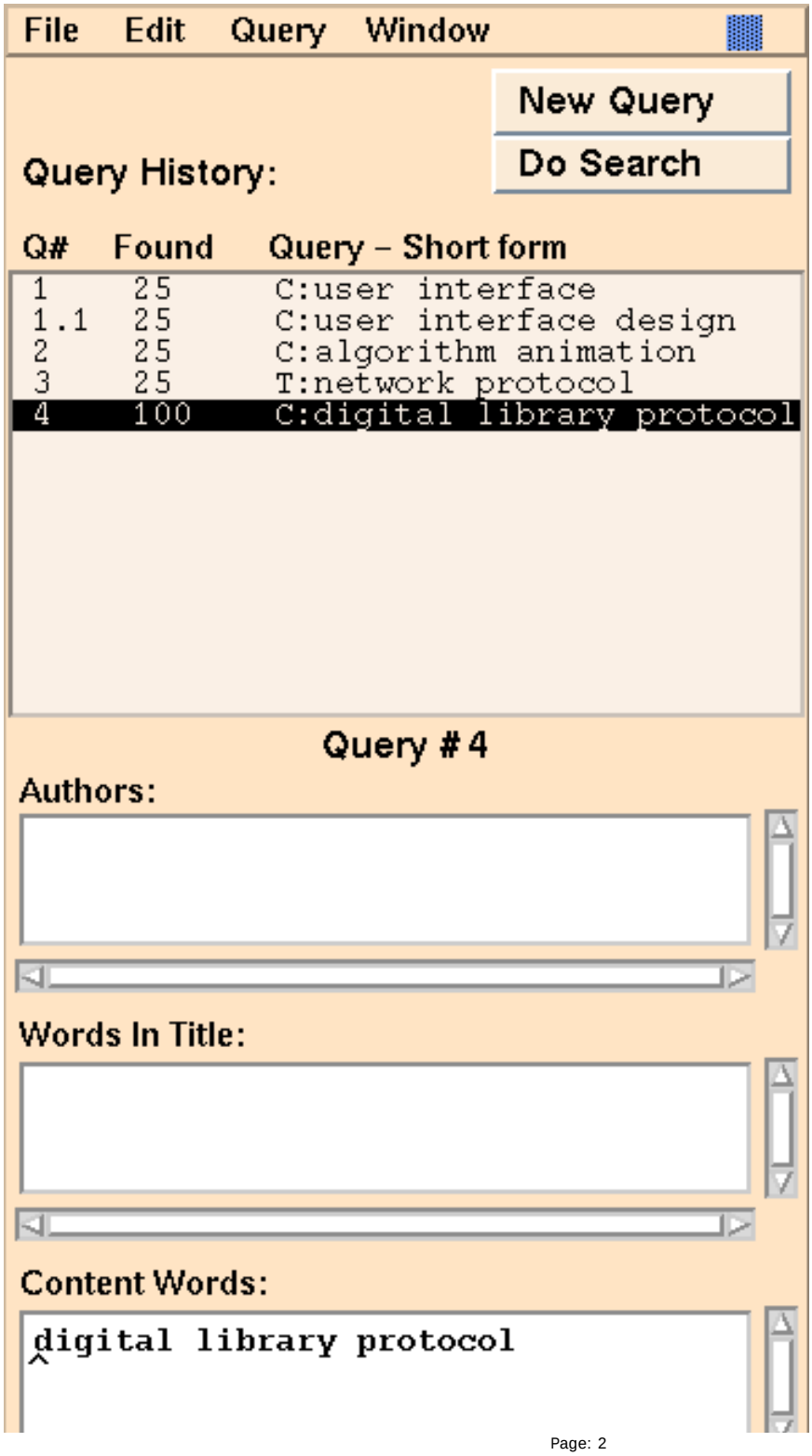
The interface includes:

- Big query:





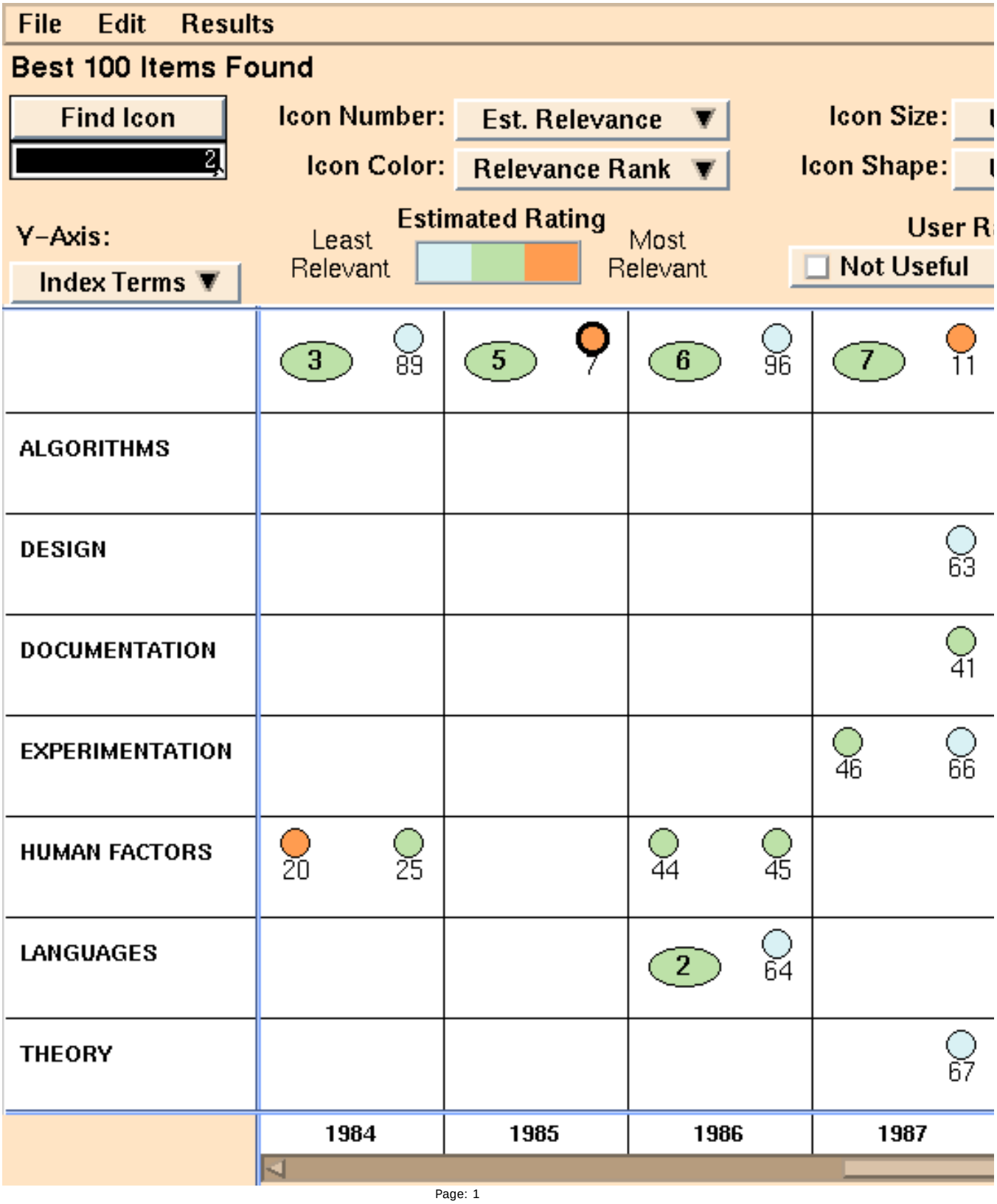
- Small query 1:

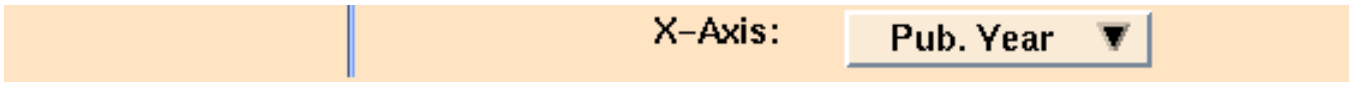


Envision - Results Screens

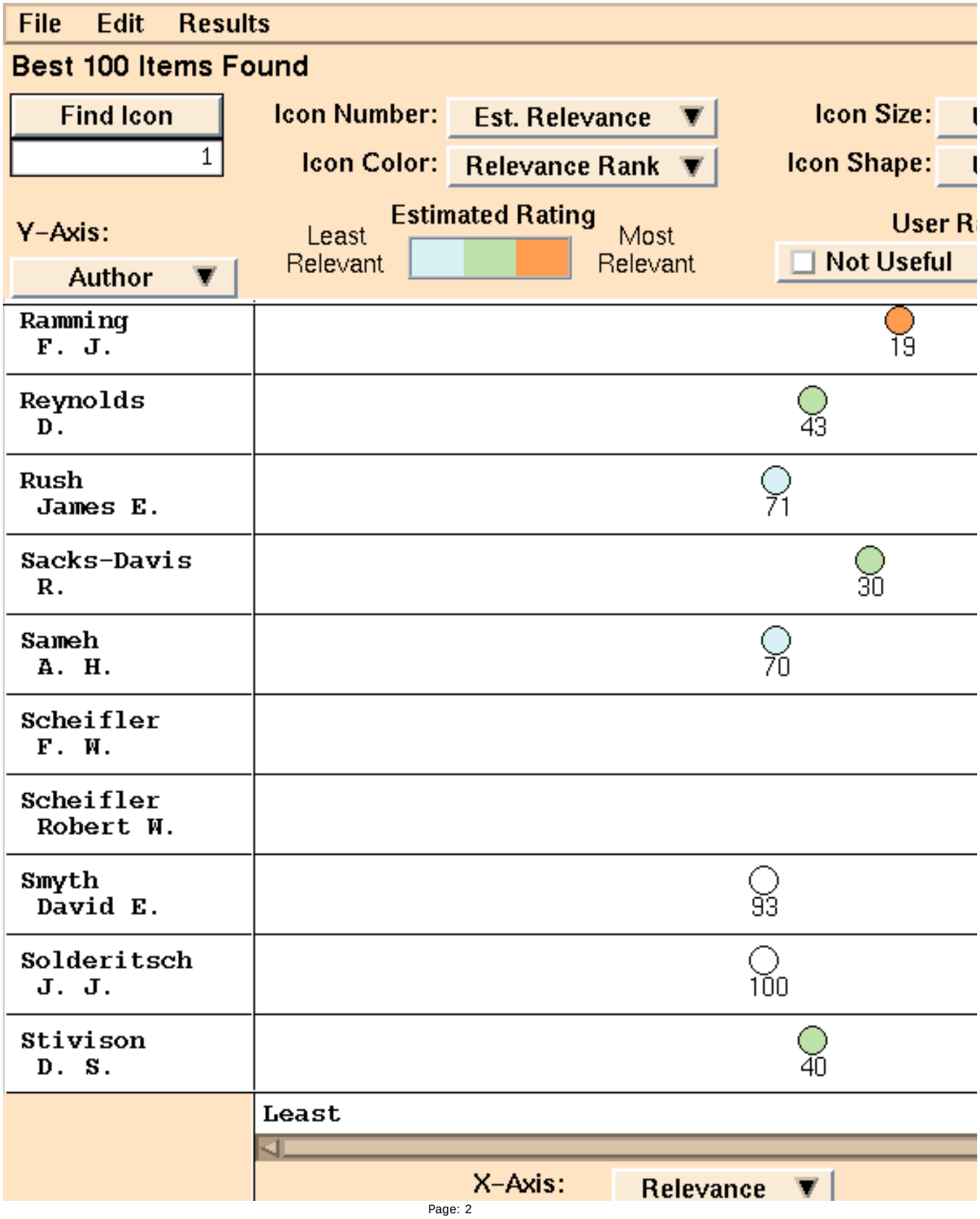
The interface includes:

- Graphic View 1:





- Graphic View 2:



PROJECT ENVISION FINAL REPORT
A User-Centered Database from the Computer Science Literature
NSF Grant IRI-9116991

Edward A. Fox, Lenwood S. Heath, Deborah Hix
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061-0106

Converted to HTML Wed Jul 5 17:41:14 EDT 1995

Summary of Completed Project

With the support of the National Science Foundation and the Association for Computing Machinery (ACM), the Envision project has developed a prototype digital library of computer science literature that is highly usable (from user-centered design), highly structured (from SGML and an object database), and highly integrated (from hypertext links among objects). The result is a representation of part of the computer science literature as a cohesive body of knowledge that can be searched and viewed in innovative ways. The user interface was designed with careful attention to user needs and desires (through interviews with potential users), to graphic detail (through involvement of an artist and attention to the research literature on graphical perception and psychophysics), and to usability (through an iterative process of usability evaluation). Recognizing the need to translate enormous quantities of documents in an unlimited variety of input formats into a single standard format, the project developed a flexible system for analyzing the structures (e.g., titles, authors, paragraphs, and references) within a document and translating that structure into any standard markup scheme. The Envision distributed server supports simultaneous access to the library by a number of users and in a variety of ways. The Envision software is soon to be installed at ACM headquarters and made available to ACM members. The Envision system will continue in use at Virginia Tech and Norfolk State University to support the work of a related NSF Educational Infrastructure grant.

Technical Information

The list of publications resulting from Envision research appears in the References section. The data collected during this project include electronic versions of computer science literature (Section 2.1). A great deal of software was created or adapted during this project (Section 2.2). A number of people have contributed to the success of the Envision project. These are listed in Appendix A. We are particularly proud of the number of undergraduate students who were able to obtain research experience on the Envision project.

Computer Science Literature

The library contains bibliographic records, full-text articles, and scanned page images. The bulk of the approximately 100,000 bibliographic records are from ACM's *Computing Archive.* We have also incorporated publicly available bibliographies from Ohio State University, the University of Arizona, and the University of Melbourne. We have approximately 700 full-text articles from *Communications of the ACM* and several of the

ACM *Transactions.* Finally, we have about 13,000 scanned page images, from various ACM publications and the technical report series of the Virginia Tech Department of Computer Science.

Envision Software

The major software components of the Envision system are the following.

1. **The Envision Client.** This component interacts with a user to accomplish the tasks of querying the Envision library and visualizing result sets in the Envision graphical display. This client interface is a major innovation of the Envision project and required the greatest amount of effort in interaction design and evaluation, in software design, and in software development.
2. **A WWW Viewer.** Envision employs a WWW browser as its presentation front end. Currently we use Mosaic running on a UNIX workstation.
3. **The Envision Intermediary.** This component communicates with the Envision client over the network to maintain session information, packages queries for the MARIAN search system, and packages result sets to pass back to the Envision client.
4. **The MARIAN Search System.** This component, developed in a separate research effort to access a library catalog, searches the Envision library for documents relevant to the user's query. The search can be based on a combination of title, author, and content words. Result sets are ranked by estimated relevance.
5. **Enhanced WWW Server.** Envision documents are viewed via a WWW interface that accesses a WWW server enhanced by CGI scripts that retrieve Envision objects from the object database and package them into HTML for presentation.
6. **The Object Database.** The Envision object database maintains our view of the structure of the library in terms of classes such as document, person (author), institution, publication, and keywords. Objects in this database refer to related objects, providing a rich hypermedia structure.
7. **The DELTO System.** The DELTO (Document Analysis and Translation) system addresses the need to convert documents in many ill-defined input formats that are received for inclusion in the Envision library into the standard SGML structural representation needed by the Envision object database and MARIAN searchers. This system emphasizes flexibility and automation. DELTO is a major innovation of the Envision project.

Components 1 and 2 run under the X Window System; these have been tested on Sun, DECstation, and DEC Alpha workstations. Components 3 and 4 run on a NextStation. Components 5, 6, and 7 run on a DEC Alpha and should port easily to other UNIX systems.

A public release of the Envision software is due during the summer of 1995. The Envision client will be freely available over the Internet by anonymous ftp from Virginia Tech. Initially, the server components (3, 4, 5, 6, and 7) and the actual library of electronic documents will be released to the ACM, as well as used in a related NSF Educational Infrastructure project at Virginia Tech and Norfolk State University.

References

1 G. A. Averboch. A system for document analysis, translation, and automatic hypertext linking. Master's thesis, Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1995.

2 S. Betrabet, E. A. Fox, and Q. Chen. A query language for information graphs. Technical Report TR 93-03, Department of Computer Science, Virginia Polytechnic Institute and State University, 1993.

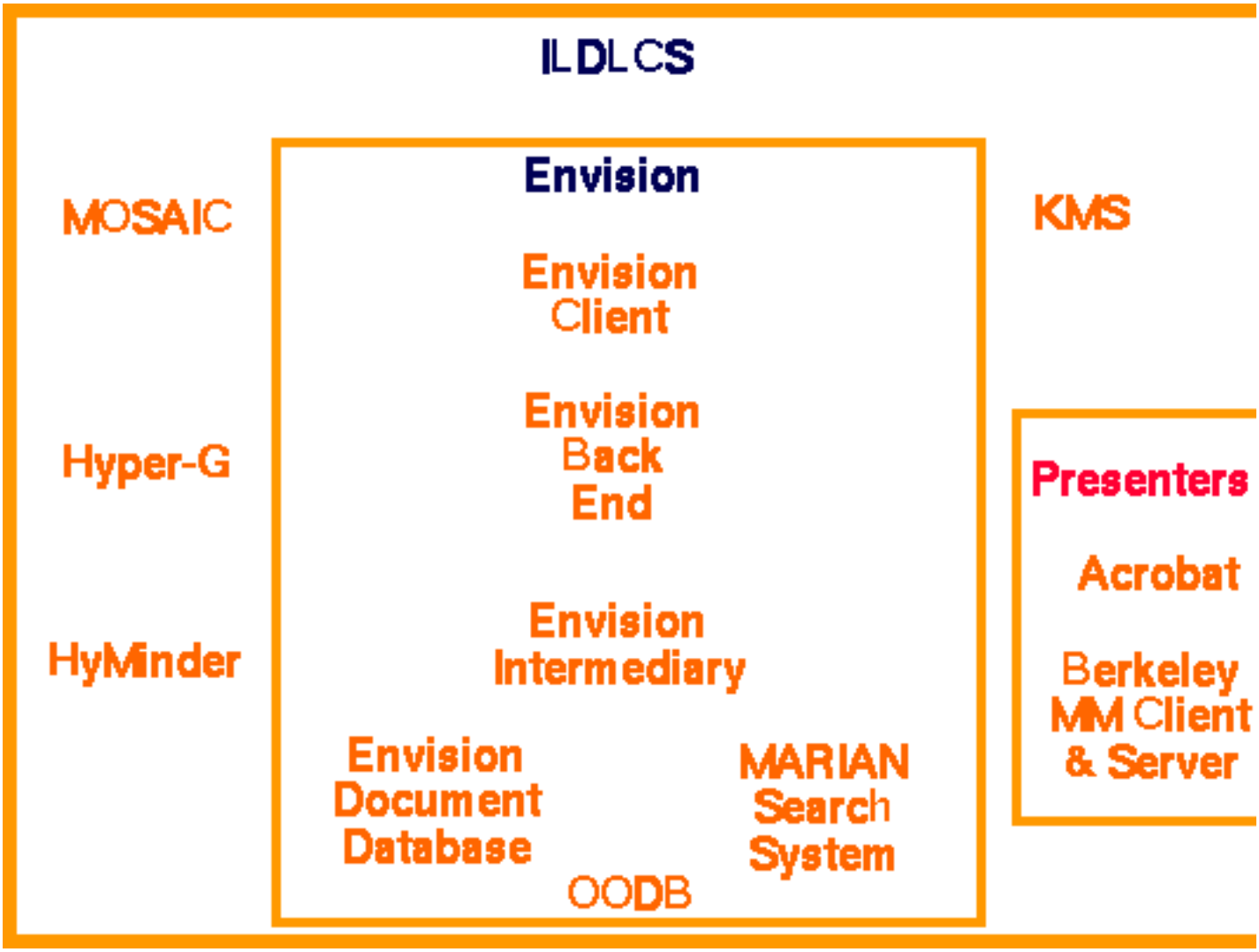
3 D. J. Brueni, B. Cross, E. A. Fox, L. S. Heath, D. Hix, L. T. Nowell, and W. C. Wake. What if there were desktop access to the computer science literature? In *Proceedings of the 21st Annual ACM Computer Science Conference*, pages 15-22, 1993. Also available as Tech. Report TR 92-42,

ILDLCS

The ILDLCS Project was funded as **Interactive Learning with a Digital Library in Computer Science** by NSF for 1993-96. ACM has provided free access to their publications, as have several other publishers. Norfolk State University is a partner in this effort, which building upon the Envision Project. More details are given online.

Efforts have concentrated on developing courseware for 4 courses that have been redone in paperless manner, constructing tools to help with algorithm visualization, and extending the Envision efforts to help with as many CS courses as possible.

The system architecture is a combination of various elements:

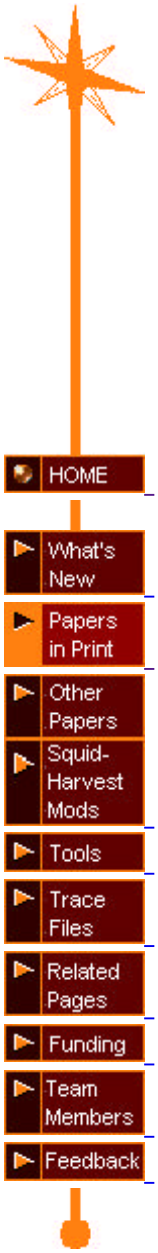


Computer Science Department
Virginia Tech
Blacksburg, VA 24061-0106

Research Group Mission

- To collect and make available to other researchers a collection of Web traffic traces from a variety of networks
- To work to make the use of proxy caches more effective through performance evaluation of different proposed cache designs
- To produce tools to assist in the collection and analysis of Web traffic and in the evaluation of cache designs

Last modified: March 27, 1997



Published Papers

- Roland P. Rooster and Marc Abrams, *Proxy Caching that Estimates Page Load Delays* , to appear in WWW6, April 1997.

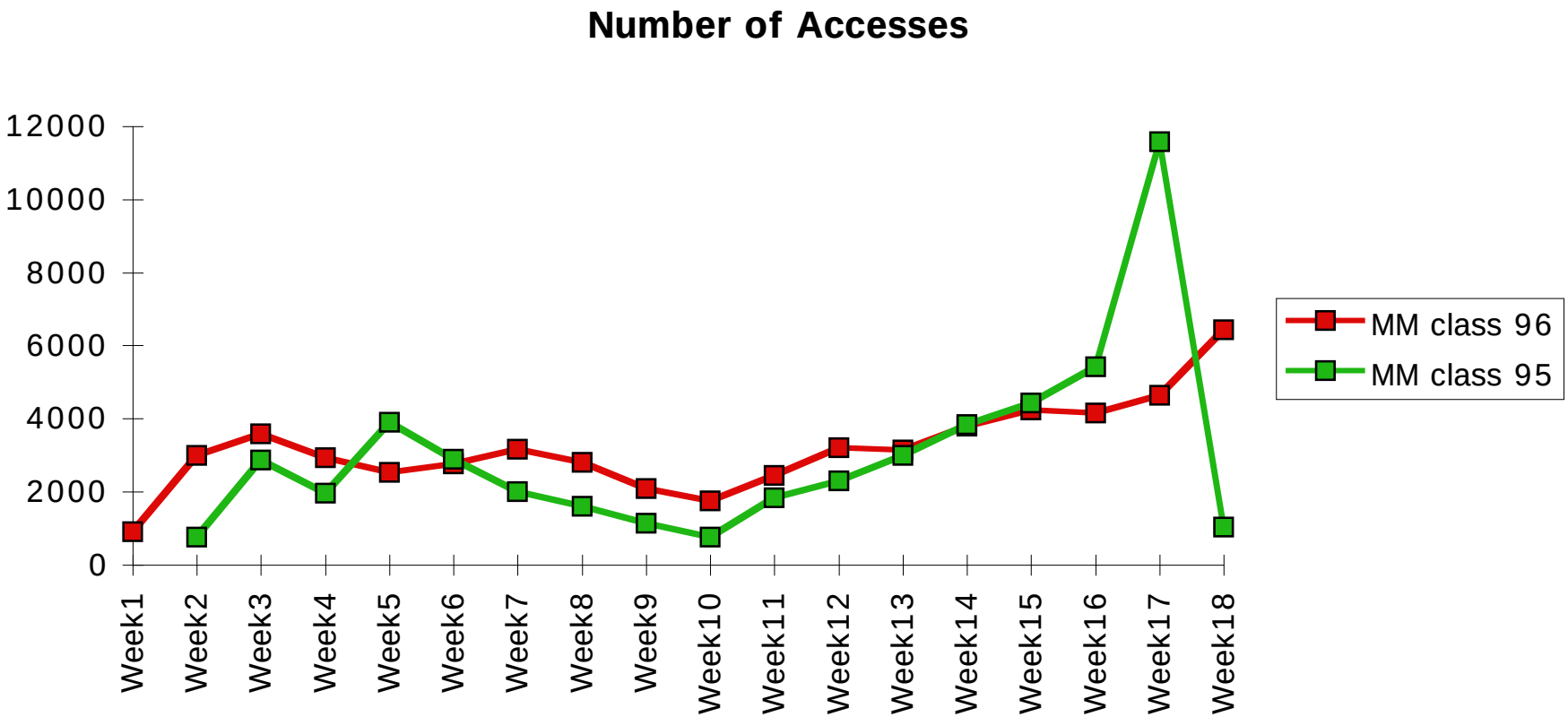
- Marc Abrams, Stephen Williams, "Complementing Surveying and Demographics with Automated Network Monitoring," *World Wide Web Journal*, No. 3, Vol. 1., June 1996. Slides from a presentation on this paper [also in Adobe pdf or postscript format].

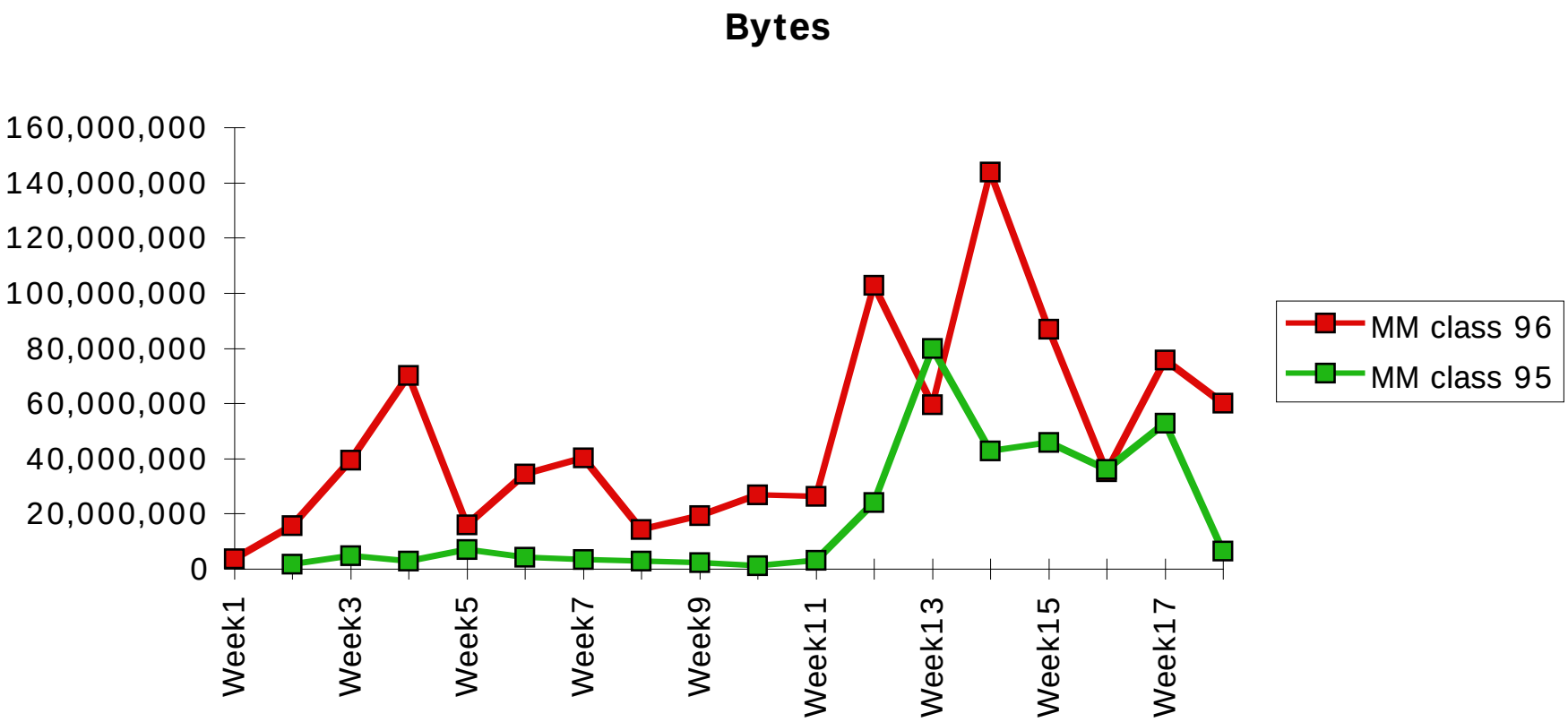
- Stephen Williams, Marc Abrams, Charles R. Standridge, Ghaleb Abdulla, Edward A. Fox, "Removal Policies in Network Caches for World-Wide Web Documents," Proceedings, *ACM SIGCOMM*, Stanford, CA, revised August 1997, pp. 293-305. Slides [gzip'd slides] from the conference presentation.

- Marc Abrams, Stephen Williams, Ghaleb Abdulla, Shashin Patel, Randy Ribler, Edward A. Fox, "Multimedia Traffic Analysis Using Chitra95," *Proceedings: ACM Multimedia '95*, San Francisco CA, November 1995, pp 267-276.

- Marc Abrams, Charles R. Standridge, Ghaleb Abdulla, Stephen Williams, Edward A. Fox, "Caching Proxies: Limitations and Potentials," *Proceedings: 4th Inter. World-Wide Web Conference*, Boston, MA, Dec. 1995, pp 119-133.

Last modified: March 25, 1997


Number of Accesses
12000
10000
8000
6000
4000
2000
0
Week1
Week2
Week3
Week4
Week5
Week6
Week7
Week8
Week9
Week10
Week11
Week12
Week13
Week14
Week15
Week16
Week17
Week18
MM class 96
MM class 95


Bytes
160,000,000
140,000,000
120,000,000
100,000,000
80,000,000
60,000,000
40,000,000
20,000,000
0
Week1
Week3
Week5
Week7
Week9
Week11
Week13
Week15
Week17
MM class 96
MM class 95

SiteSearch

Search for specific topics

Go to OCLC Home Page

- **Demonstration**
- **Descriptions of Products and Services**
- **News**
- **Publications**
- **User Documentation**

Demonstration

- The WebZ Demo Page lets you search OPAC data and several other reference databases (http://tikal.dev.oclc.org:2000)

Descriptions of Products and Services

- Introducing OCLC SiteSearch: To the Next Stage of the Electronic Library
- Elsevier Science/OCLC Electronic Publishing Pilot Program
- WebZ Server Questions & Answers
- Z39.50 Server System Questions and Answers

News

***OCLC Newsletter* Features**

- Georgia's GALILEO Project
- Interview: Merryll Penson, Ralph E. Russell, and William Gray Potter. The directors of three libraries involved in the GALILEO project discuss the creation of the statewide project, its current status and future plans

News Releases

- 'GALILEO' to Use OCLC SiteSearch Software to Deliver Information, FirstSearch to Georgia Libraries--December 1, 1995

See the complete news release list for earlier news releases.

Publications

Reference News

- Winter 1996, No. 29
- Fall 1995, No. 28
- Summer 1995, No. 27
- Spring 1995, No. 26
- December 1994, No. 25

Introducing OCLC SiteSearch

To the Next Stage of the Electronic Library

Welcome to the **OCLC SiteSearch®** family of software products. For librarians, SiteSearch brings you a significant step closer to the dream of a virtual library--seamless integration of local and remote information resources. You enhance access to your local collection, and your library becomes the doorway to the global information environment.

For end-users, SiteSearch means one interface, one access point, and one search process--all from one desktop. Users have the most comprehensive reference system available whether they are in the library, dormitory, home or office. Read on to see how SiteSearch enables your college or university and other groups to build databases and link your local area networks, OCLC reference services, and the World Wide Web into **a customized virtual library that accommodates both print and electronic information.**

What is SiteSearch?

SiteSearch is a complete set of software tools that lets you create an integrated information home for your users with your library. It is based on client server technology, and it works for individual libraries or library groups, such as regional, state, or local consortia.

SiteSearch allows you to:

- load commercial databases locally
- bring remote databases--commercial or public--to your library
- build your own unique, local databases
- index those databases
- provide one user interface to search your databases and any remote Z39.50 compatible databases

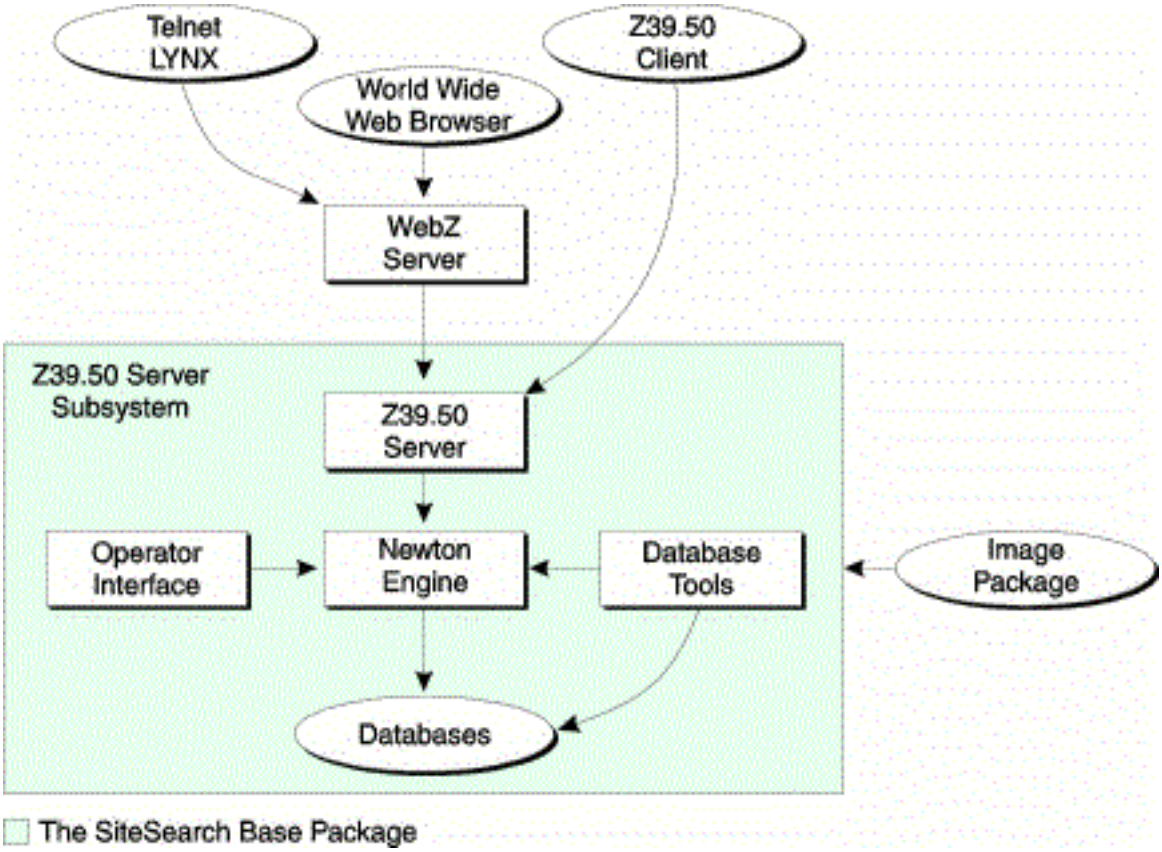
The SiteSearch software tools are:

- **Z39.50 Server System**
  - **Database loading, creation and maintenance software:** Helps you mount commercial databases locally and build and maintain unique, local databases. Lets you define indexing rules and display formats.
  - **Search engine:** Supports searching very large or very small databases of citations, full text, images, and sound with speed and efficiency. The search engine used is Newton, which is used for all of OCLC's online reference systems.
  - **Z39.50 server:** Manages connections to database servers and supports the Z39.50-1992 protocol for communication between the user interface and the search engine.
  - **Image support:** Provides tools for creating electronic image collections for applications

such as photo collections, archives and reserves.

- **Z39.50 client:** Provides a gateway to Z39.50 servers and access to OCLC SiteSearch databases for World Wide Web browsers, such as NetScape, Mosaic and Lynx. The client is called WebZ.

These components can be purchased individually or as an entire package with site licenses based on user population or simultaneous log ons.



Satisfy Your Users Information Demands

By linking resources from your library, remote sites and the World Wide Web, users will have a comprehensive and thorough information search. Their journey could take them to an electronic journal on the Internet, a database halfway around the world, or deep within your library's archives--all transparent to the user thanks to SiteSearch software.

Increase their awareness of your library's value. SiteSearch draws people to your library, both electronically and physically. Your library collection becomes more accessible, and your library becomes the starting point to the global information networks.

Enhance their library experience. Users will see your library as bringing some order to today's chaotic global information environment, where a myriad of systems and a mountain of information are available to them. Their confidence in their research will increase because they are searching and gathering information globally. And they can search at their convenience, wherever they are.

Z39.50 resources - a pointer page

****
The Library of Congress is the official maintanence agency for Z39.50. As such they are the place to go to get the most official current legal information related to Z39.50. This page you are reading may phase out (though not soon) as they develop their page (started July 1995).
****

This page is meant as a reference point for resources related to the Information Retrieval Service and Protocol standard, ANSI / NISO Z39.50. This standard was first successfully balloted in 1988; several companies implemented this standard or variants of this; but it did not develop large scale acceptance. A noteworthy implementation based on this standard is WAIS (Wide Area Information Services). Also see the **Profiles** section for more info on present development of WAIS within Z39.50.

The standard was significantly rewritten for its next version. This is ANSI/NISO Z39.50-1992 (Version 2). One important step in this version of the standard was alignment with ISO 10162/10163, the Search and Retrieval (SR) Service Definition and Protocol Definition. Also beginning with this version, the protocol data units are described in ASN.1 (A "Layman's Guide" to ASN.1 is available from RSA) -- The Version 3 ASN.1 is available as flat ascii as well as in a wonderfully useful HTML format. from Library of Congress's various servers.

The next version (Version 3) of the standard was balloted in December 1994, and officially accepted by ANSI in July 1995. The official version of the standard is available electronically, at the Library of Congress's ftp server (ftp.loc.gov). Note this is a copyrighted document - many thanks to whoever achieved this electronic availability. The official text is available in postScript and wordPerfect, in four parts:
postscript: Part1, Part2, Part3, and Part4.
WordPerfect: Part1, Part2, Part3, and Part4.

The Z39.50 ImplementorsGroup (ZIG) works closely with the standard's maintenance agency, the Library of Congress. This group meets 2 - 3 times a year and has discussions on its listserv Z3950IW@NERVM.NERDC.UFL.EDU. For meeting minutes, more about the LISTSERV, scheduled future meetings, and other related information check out the relevant sub-section at Library of Congress

Freely available implementations of Z39.50 and related code are starting to become available. Those I know of (let me know of others) are:

- CNIDR's Isite, Isearch, FreeWAIS, etc
- Index Data, a software development enterprise operating out of Copenhagen, Denmark has developed a Version 3 API toolkit to aid in the implementation of the ISO SR and Z39.50-1995 protocols. They say: "software is available free of charge, on a liberal license: Commercial re-use is explicitly permitted."
- National Library of Canada has made its client and server code available;
- NIST is making available a Z39.50 client/server package based on the PRISE search engines.
- OCLC has made its Z39.50 Client API available to the public
- University of California - Berkeley demonstration client/server protocol engine
- USGS is making available a freeware implementation of Z39.50 as an OLE add-on to WWW browsers. You can fetch the executable software, README.TXT, and source files by anonymous FTP to host www.usgs.gov, in the directory /gils/ciir/dtic_a02.
- Willow -- the Washington Information Looker-upper Layered Over Windows.

- John Lamp is doing a good job tracking sites with Z39.50 tools and resources.

Electronic documents of interest (let me know of more) are:

- Z39.50 in a Nutshell - (An Introduction to Z39.50) by John A. Kunze & R. P. C. Rodgers. Written at National Library of Medicine, July 1995.
- The ANSI/NISO Z39.50 Protocol: Information Retrieval in the Information Infrastructure by William Moen (added here July 1995)
- IETF NIR document by Mark Needleman
- RFC1729: Using the Z39.50 Information Retrieval Protocol in the Internet Environment, by Clifford Lynch. December 16, 1994
- RFC1625: WAIS over Z39.50-1988, by M. St Pierre et al. June 1994
- Facilitating the Creation of Z39.50 Origins in the UK, by A.M.Addyman
- Z39.50 FAQ - very dated. A new Z39.50 FAQ is being developed.
- A list of Z39.50 available databases. This list is available as a Z39.50 database (at z3950.research.att.com) or via a WWW gateway.

There are several people/organizations working on Z39.50 -> WWW gateways.

- Prentiss Riddle is keeping a good page tracking WWW-to-Z39.50 Gateways. I recommend checking there rather than depending on this list!!
- The implementation being done by CNIDR is available for general use and free; this is part of the CNIDR software described above. The gateway is at cnidr.org.
- The Stanford gateway, with perl source code is supposed to be very useful and flexible.
- The Library of Congress has up a fairly complete set of resources at their gateway.
- The AT&T Library Network is also experimenting with gateways; to see this in action select here.

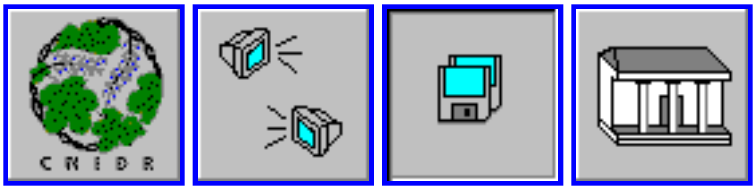
Profiles

Profiles are formal implementation agreements within the context of standard. There are two major profiles being worked on with the Open Systems Environment Implementors Workshop (OIW), the Government Information Locator Service (GILS), and the Wide Area Information Service (WAIS).

- The Office of Management and Budget, in concert with the Information Policy Committee of the Information Infrastructure Task Force, to promote the establishment of an agency-based Government Information Locator Service (GILS). GILS is intended to help the public locate and access public information throughout the U.S. government. Note GILS is based on Z39.50.
  The document is available via anonymous FTP in
  Microsoft Word for WIndows format, Word Perfect 5.0, Rich Text Format, and as ASCII text.
- The WAIS Profile of Z39.50 V2 specifies the required components of Z39.50-1994 for full WAIS functionality. Please send comments on the Profile to oiw-l@mozart.esl.com.
- A draft of the Geospatial Metadata Profile (GEO) is ready for review at the URL listed below. This profile is intended to be a guide to developers to support the attributes defined in the Content Standards for Digital Geospatial Metadata promoted by the U.S. Federal Geographic Data Committee.

Online views of Z39.50 related products/services

- Ameritech Library Services
- BookWhere for Windows - Z39.50 Client software
- Chemical Abstracts Service
- Data Research Associates, Inc
- Geac Computer Corporation, Ltd
- Library of Congress
- Online Computer Library Center, Inc (OCLC)
- The Research Libraries Group, Inc. (RLG)
- SIRSI Corporation
- VTLS, Inc.



CNIDR Isite

CNIDR Isite is an integrated Internet publishing software package including a text indexer, a search engine and Z39.50 communication tools to access databases. Isite includes the CNIDR ZDist, Isearch and Search API distributions.

See what Isite can do for you

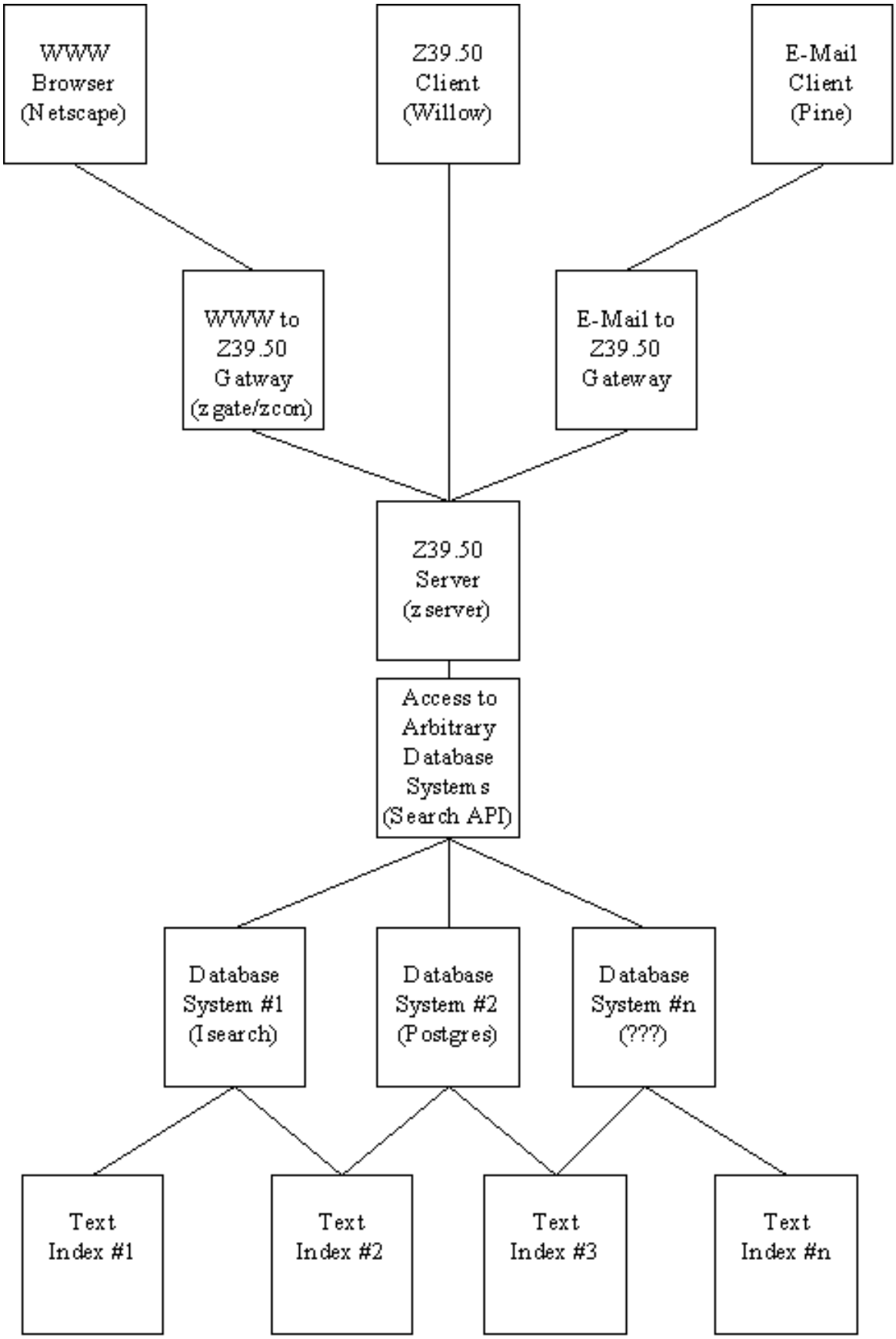
- **Help us to better serve you!**
- Diagram of Overall Architecture - Details available via Administrator's Guide below
- Demo of Stateful http to Z39.50 Gateway - Demonstrates access to various database systems
- Other systems using Isite
    - NASA Global Change Master Directory
        - Z39.50 Ranked Search
        - Z39.50 Boolean Search
        - Distributed Document Search
    - American Astronomical Society: Electronic Astrophysical Journal Letters
    - United Nations International Drug Control Programme
    - University of Tennessee Office of Research Services: Friends and Partners Cookbook
    - Microlytics, Inc.
    - Library of Congress Z39.50 Gateway
    - U.S. Department of Housing and Urban Development GILS Service
    - **YOUR LINK GOES HERE** - *Please* send me pointers to your Isite-based systems!!

Download a copy

- Stable Version - includes precompiled binaries
- Untested Versions - require a C++ compiler

Read the documentation

- Isite Administrator's Guide - Refers to stable versions
- Untested Isite Administrator's Guide - Refers to untested versions
- Isearch Tutorial - Step-by-step guide on building databases with Isearch
- Z39.50 Maintenance Agency - Everything you always wanted to know about Z39.50 and more!
    - Includes electronic copies of the ANSI/NISO Z39.50 standard
    - Includes implementor agreements
    - Includes various papers written by experts in the field
    - Includes lots of other stuff you will need to get the most out of Isite
- BSn Doctypes - Many of the input files supported by the Isearch indexer are documented here


WWW Browser (Netscape)
Z39.50 Client (Willow)
E-Mail Client (Pine)
WWW to Z39.50 Gatway (zgate/zcon)
E-Mail to Z39.50 Gateway
Z39.50 Server (zserver)
Access to Arbitrary Database Systems (Search API)
Database System #1 (Isearch)
Database System #2 (Postgres)
Database System #n (???)
Text Index #1
Text Index #2
Text Index #3
Text Index #n

[text-only]

The LIBRARY of CONGRESS

AMERICAN MEMORY

Historical Collections for the National Digital Library

SEARCH
American Memory Collections

BROWSE
List of all American Memory Collections

LEARN
Organized help for using the collections

American Memory consists of primary source and archival materials relating to American culture and history. These *historical collections* are the key contribution of the Library of Congress to the National Digital Library. Most of these offerings are from the Library's unparalleled special collections.

Access Collections by Type


Prints & Photos


Documents


Motion Pictures


Sound Recordings

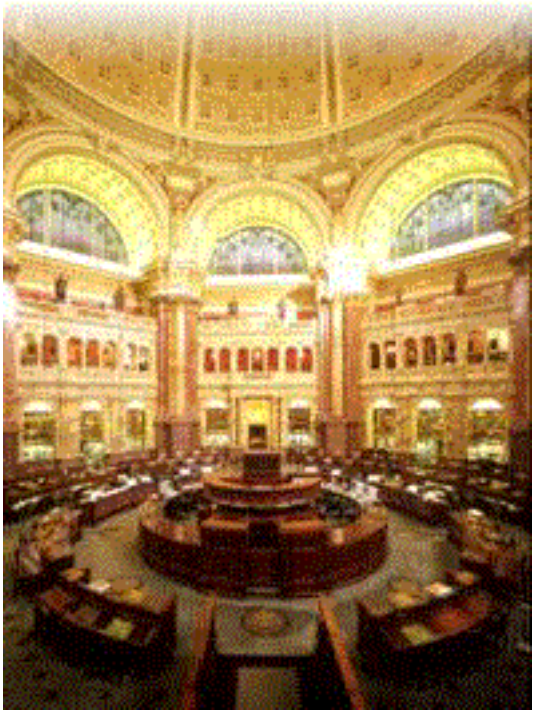
Showcase

Three new collections:

Evolution of the Conservation Movement, 1850-1920 (manuscripts, legal documents, photographs)
Gottscho-Schleisner (photographs)
Horydczak (photographs)

LIBRARY OF CONGRESS ❖ AMERITECH

National Digital Library Competition



***The*Library *of* Congress Home Page**

Library of Congress
Comments:lcweb@loc.gov (07/03/96)

Corporation for National Research Initiatives

Key Architectural Issues in The Digital Library

William Y. Arms

Acknowledgments

- This is work in progress.
- This is a personal interpretation of ideas developed by the CSTR Project.
- CSTR is a joint project of CNRI with Carnegie Mellon, Cornell, MIT, Stanford and UC Berkeley, funded by ARPA.
- For background information, see the CSTR home page.
- The architecture is more fully described in a paper by Robert Kahn and Robert Wilensky.

Key Issues and CSTR Terminology

This set of WWW pages looks at the following six key issues in the architecture of the digital library.

- Items in the library - digital object.
- Identifiers - handle.
- Storage - repository.
- Sets of objects - composite and meta-object.
- Information about objects - properties.
- Semantic layering (schema) - data model.

The architecture under development is an open architecture. In general, it allows these topics to be considered separately.

The CSTR Architecture and the World Wide Web

Many of the concepts in the CSTR architecture can be partially implemented within the framework of the World Wide Web and fit with recent IETF discussions.

Corporation for National Research Initiatives

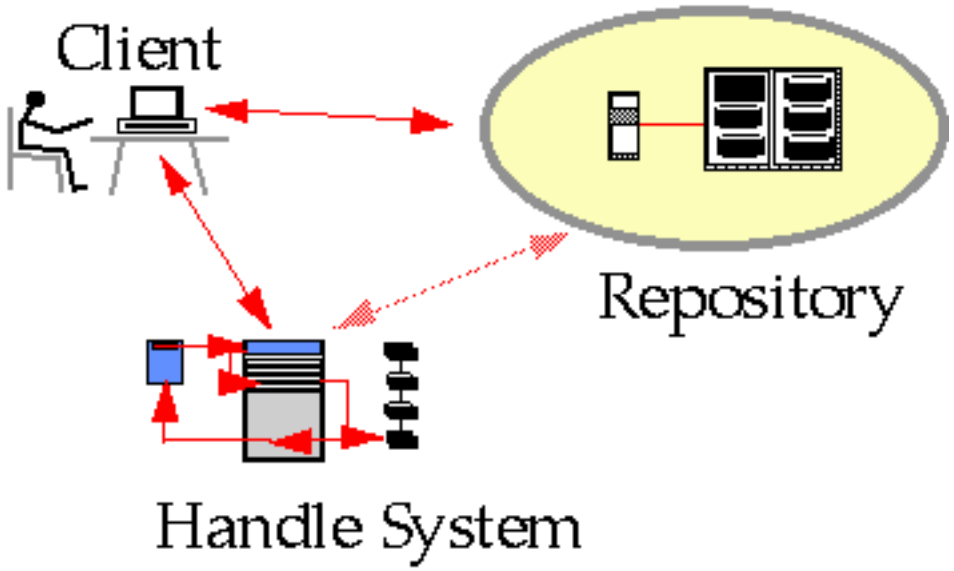
Digital Object Architecture Project

Principal investigators

Robert E. Kahn
William Y. Arms

Summary of the project

This project continues the architectural work of the DARPA-funded Computer Science Technical Reports (CS-TR) project. That project developed a Framework for Distributed Digital Object Services and implemented some key components. This project continues research and development of this framework and two extensive testbeds at the Library of Congress.



The basic entity in the system architecture is the "digital object", which contains copyright material or other material in which other rights and interests are manifest. There may also be rights and interests associated with digital objects themselves. The major components of the system are: (a) repositories of digital objects that allow network based deposit and access, (b) handle servers that record the location of digital objects over long periods of time, (c) registration and recordation mechanisms to keep track of rights and interests associated with digital objects, and (d) client software to enable use of these components over the network. Digital object

fingerprints are used in the registration system to permit validation of the objects at a later time.

The first testbed is with the Copyright Office at the Library of Congress. This is a system to register electronic materials for copyright and recordation of changes in copyright ownership. The second testbed is with the National Digital Library Program at the Library of Congress. This is a very large scale project to convert historic materials from the library's collections to digital form and make them available to the world.

Background 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
- Historical Collections for the National Digital Library: Lessons and Challenges at the Library of Congress Caroline R. Arms, D-Lib Magazine, April 1996. Part 2

Funding

Funding for this work is provided by the Defense Advanced Research Projects Agency (DARPA) and the Library of Congress.

Home

---

wya
6/30/96

A PURL is a **P**ersistent **U**niform **R**esource **L**ocator. Functionally, a PURL is a URL. However, instead of pointing directly to the location of an Internet resource, a PURL points to an intermediate resolution service. The PURL resolution service associates the PURL with the actual URL and returns that URL to the client. The client can then complete the URL transaction in the normal fashion. In Web parlance, this is a standard HTTP *redirect.*

The OCLC PURL Service has been strongly influenced by the active participation of OCLC's Office of Research in the IETF Uniform Resource Identifier working groups. There is nothing incompatible between PURLs and the ongoing URN work. PURLs satisfy many of the requirements of URNs using currently deployed technologies and can be transitioned smoothly into a URN architecture once it is deployed.

Further Information and Resources

- A brief introduction to PURLs
- A longer introduction to PURLs
- Frequently Asked Questions
- Download the PURL software NEW
- PURL-L mailing list
- More info

Interacting with This Resolver

- Create your first PURL
- Register as a user
- Create PURLs, domains, groups
- Modify PURLs, domains, groups, users
- Search this resolver
- Power user's page (all features)

As of *Sat Jul 13 13:26:28 PDT 1996* : PURLs Created = **6768** , PURLs Resolved = **473905** and Unique Client Systems = **13121** (See the complete Database Stats for more details.)

The PURL Team
purl@oclc.org

Corporation for National Research Initiatives

Handles and the Handle System

Forms for handle administration

Forms to add and edit handles, to create naming authorities, and to set up groups of administrators are available through the Handle Administration Page.

Information about the Handle System

Technical information

- An overview of the system.
- Implementation of the Handle Management System.
- FTP server to download documentation and code.
- Browsers that support handles.

Architectural considerations

- The use of handles within a a framework for distributed digital object services.
- Handles as a key concept in the digital library.
- The IETF's work on Universal Reference Names.

Presentations and demonstrations

- An architectural overview.
- The handle system.
- D-Lib Magazine with handles.

A brief introduction to Handles

A **handle** is a unique identifier for a digital object. This object can be stored in a digital library repository, in an ftp archive, in a World Wide Web server, or any other digital store. Handles can also be used for other forms of identification, such as electronic mail addresses. A high performance Handle Management System is publicly available on the Internet. The useful properties of handles include the following.

- Handles are guaranteed to be unique.
- Handles are permanent. Therefore, they can be used to identify objects for purposes of copyright or archiving.
- Handles are location independent. The object may be moved to a different location without changing its handle. This enables handles to be used to refer to an object, for example, in a bibliographic citation.

A handle has the syntax:

naming authority / string
or: hdl://*naming authority / string*

The **naming authority** is a globally unique name. The **string** is unique for that naming authority.

Return to CNRI home page

---

hdl://cnri/handle-intro

wya
Last revised: November 11, 95

UMBC **AgentWeb**

UMBC

An Honors University in Maryland

***L*aboratory for *A*dvanced *I*nformation *T*echnology**

UMBC AgentWeb

Intelligent Software Agents

Information and resources about intelligent information agents, intentional agents, software agents, softbots, knowbots, infobots, etc. Send comments and suggestions to Tim Finin (finin@umbc.edu).

- **About the AgentWeb...**
  - What's new... NEW
  - Current AgentNews webletter NEW
  - About the AgentNews webletter and mailing lists
  - AgentWeb help...
  - AgentWeb salon ... NEW
  - About the UMBC Laboratory for Advanced Information Technology
- **Agent basics ...**
  - Introductory material
  - Agent FAQ
  - Agent theory - philosophy, formalisms, ...
  - Agent technology - systems, tools, languages, standards, ...
  - Mobile agents ... NEW
- **Agent resources ...**
  - Agent papers
  - Agent events, conferences, workshops, ...
  - Agent mailing lists and newsgroups
  - Agent courses and seminars
  - Other agent related web resources
- **Who is doing what ...**
  - Agent-related R&D groups and companies
  - Agent-related projects
  - Example Agents

- Employment opportunities

- **Agents and ...**

  - Agents and the Knowledge Sharing Effort
  - Agents and security
  - Agents and learning
  - Agents and Ontologies.
  - Agents and Robots
  - Agents and artificial life
  - Agents and sex
  - Agent and humor.
  - Agents and virtual environments, muds, ...
  - Agents and other miscellaneous topics (e.g., IR)

- **Agents for ...**

  - Agents for Manufacturing
  - Agents for Commerce
  - Agents for human-computer interfaces
  - Agents for, on, and by the web.

AgentWeb is maintained at the UMBC Lab for Advanced Information Technology by **Tim Finin** (*finin@umbc.edu*).

Modified on Wednesday, 10-Jul-96 12:56:04 EDT -- 02980 hits since June 25, 1996

Preserving Digital Information: Final Report and Recommendations

May 20, 1996

At the end of 1994 the Commission on Preservation and Access (CPA) and RLG created a Task Force on Archiving of Digital Information charged with investigating and recommending means to ensure "continued access indefinitely into the future of records stored in digital electronic form." The 21-member task force, co-chaired with distinction by Donald Waters, Associate University Librarian, Yale University, and John Garrett, Chief Executive Officer of CyberVillages Corporation, recently completed their final report. RLG and CPA are making this widely available online and in print.

Electronic versions are available from RLG's FTP server (ftp.rlg.org) and this Web site:

HTML version
Adobe Acrobat version: /pub/archtf/final-report.pdf
Microsoft Word for Windows 6.0 version: /pub/archtf/final-report.doc
ASCII Rich Text Format version: /pub/archtf/final-report.rtf

Notes:

To download an Adobe® Acrobat® viewer to use as a helper application with your web browser, connect to the Adobe web site.

Copies of the printed, bound report are available for $15.00 (prepayment required) from the Commission on Preservation and Access, 1400 16th Street, N.W., Suite 740, Washington, DC 20036-2217.

RLG will be mailing the printed report to the member representative at each of our member institutions in North America and Europe as well as to each member liaison in our collaborative SHARES (Shared Resources) and PRESERV (Preservation) programs.

The task force's final report benefits from their action last September to make a draft version available online and to open a listserv for comments by the community. Many thanks to all of you who responded. That draft report can still be found on RLG's server and Web site:

Adobe Acrobat version: /pub/ArchTF/Draft-Report.pdf
Microsoft Word for Windows 6.0 version: /pub/ArchTF/Draft-Report.doc
ASCII version: /pub/ArchTF/Draft-Report.txt

RLG has already built into its agenda work on several of the task force's nine recommendations. (Our archival server and digital collections projects are directly related.) We will be following up on other recommendations with other stakeholders.

Please share your comments and advice with us regarding this report and the specific recommendations; you can send them by e-mail to Nancy Elkington, RLG member services officer and member of the task force.

Sincerely,

James Michalko
President

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 help 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.

Problems with this page? Send email to jdavis@parc.xerox.com

TEI Guidelines for Electronic Text Encoding and Interchange (P3)

Made available from the Electronic Text Center at the University of Virginia.

Search the *TEI Guidelines*.

Word or phrase (omit all quotes): Search

1-50

Other types of searches:
You may also combine words or phrases within a specified proximity, or locate segments such as sections where two words or phrases both occur.

Browse the *TEI Guidelines*.

Resources of Related Interest

- The Text Encoding Initiative Home Page
- Other Electronic Versions of the TEI Guidelines
- TEI P3 now available on CD-ROM
- The Electronic Text Center Introduction to TEI and Guide to Document Preparation.
- TEI DTD Browser, courtesy of CETH

Digital Libraries - Implementation Principles

As we build digital libraries, it is important to consider key principles so that these libraries will be easily usable, and have long-term archival value.

1. Declarative representations of documents should be used.
2. Document components should be represented using natural forms, namely objects that can be manipulated by users familiar with those objects.
3. Links should be recorded, preserved, organized and generalized.
4. There should be a separation between the digital library and user interfaces to it.
5. Searching should make use of advanced retrieval methods.
6. Open systems that include the user, and where (some of) the functions of librarians are carried out by the computer, must be developed.
7. Task-oriented access to electronic archives must be supported.
8. A user-centered development approach should be adopted.
9. Users should work with objects at the right level of generality.