

# Tutorial 11: Digital Libraries - Advanced

## DIGITAL LIBRARIES '2000

(c) 2000 Edward A. Fox, all rights reserved

Department of Computer Science  
Virginia Tech, Blacksburg, VA 24060 USA  
fox@vt.edu - <http://fox.cs.vt.edu>  
June 7, 2000

---

## Table of Contents

- 1 DLI Overview for BASIS - in [PDF](#)
  - 2 ETD Genre and Examples - in [PDF](#)
  - 3 DL'99 paper on NDLTD - in [PDF](#)
  - 4 Selections from Online Courseware - [Intro in PDF \(2.4M, 196 pages\)](#), [Advanced \(this\) in PDF](#), [WWW pages](#)
-

# The Digital Libraries Initiative: Update and Discussion

by Edward A. Fox  
Guest Editor

This special section of the *Bulletin of the American Society for Information Science* on the Digital Libraries Initiative begins with an article by the guest editor that provides an overview of the initiative to-date. In the two subsequent articles Michael Lesk gives perspectives on the field, while Stephen Griffin provides important data, including abstracts, of a number of recently funded digital library research projects. Drs. Lesk and Griffin are with the National Science Foundation's Information and Intelligent Systems (IIS) Division, in which Lesk serves as director (on rotation) and Griffin as program officer. The Lesk and Griffin articles are reprinted from *D-Lib Magazine*, v. 25, no. 7/8 (July/August 1999) with the permission of the authors and the Corporation for National Research Initiatives.

## Digital Libraries Initiative (DLI) Projects 1994-1999

by Edward A. Fox

Edward Fox is professor in the Department of Computer Science and Director of the Digital Research Laboratory at Virginia Tech. He directs the Networked Digital Library of Theses and Dissertations (<http://www.ndltd.org>). He also directs the Internet Technology Innovation Center at Virginia Tech (<http://fox.cs.vt.edu/itic/>). He can be reached there by mail at 660 McBryde Hall, M/C 0106, Blacksburg, VA 24061; by phone at 540/231-5113; on the Web at <http://fox.cs.vt.edu>; or by e-mail at [fox@vt.edu](mailto:fox@vt.edu)

Since 1993, the National Science Foundation (NSF) has played a lead role in an interagency federal program called the Digital Libraries Initiative (DLI). DLI emerged after several years of discussion in which a number of researchers, such as Michael Lesk (then at Bellcore), made recommendations through the reports of a series of NSF-sponsored planning workshops (see summary at <http://fox.cs.vt.edu/DLSB.html>). Thus, throughout the 1990s NSF support has been a critical factor in establishing the digital libraries field as an important area for research, development, application and practice. Though total investment around the globe – involving such institutions as libraries, universities, associations, corporations, foundations and other governments – amounts to hundreds of millions of dollars, the single most visible effort is the DLI program, which is the focus of all of the articles in this special section.

### DLI Funding

In the United States, over \$68 million in federal research awards were made through DLI over the period 1994-1999. \$24 million was awarded in 1994 by NSF, DARPA and NASA, split evenly among six "DLI-1 teams." Three were in California: two went to campuses of the University of

California (one to Berkeley and one to Santa Barbara) and the third to Stanford University. Two were in the middle of the country, to the University of Illinois at Urbana-Champaign (UIUC) and the University of Michigan. Carnegie-Mellon University (CMU) received the only East Coast award, leveraging prior work on text, image and speech processing.

Roughly \$44 million, allocated in somewhat different fashion, has already been awarded by NSF, DARPA, National Library of Medicine, Library of Congress, National Endowment for the Humanities, NASA and the FBI (in partnership with National Archives and Records Administration, Smithsonian Institution and Institute of Museum and Library Sciences) in a second phase, the "DLI-2" program (<http://www.dli2.nsf.gov>). A terse summary of these awards is shown in Table 1. Recent commitments to the three California groups in DLI-1, including sub-awards involving other partners in California (University of California, Irvine; University of California, Los Angeles; University of California, San Diego; California Digital Library) and at the University of Georgia, plus an undergraduate education award to Berkeley, account for over \$15 million. CMU also received \$4 million further support, as

well as a separate but related \$450,000 grant. The six other large grants (each for \$1 million or more) went to Columbia University, Cornell University, Harvard University, Michigan State University, Tufts University and the University of South Carolina.

Over \$500,000 was allocated to three awards from 1988 with an undergraduate emphasis (see top section of Table 1). There were six awards focused on international collaboration (see bottom section of Table 1), for a total of about \$2.3

million. The main DLI-2 program (see middle section of Table 1) involved over \$41 million through 21 awards. Of these 21, 10 were large, accounting for over \$35 million, while the remaining 11 account for about \$5.5 million. Please see the accompanying article by Stephen Griffin that provides short summaries of DLI-2 projects announced through August 1999. Other details and newer information can be found at the DLI-2 Web site or set in a broader context as part of the self-study course materials on digital libraries at Virginia Tech (see specifically <http://ei.cs.vt.edu/~dlib/projects.htm>).

**Table 1. Details of DLI-2 Awards by September 1999**

AWARD ID	PI NAME	INSTITUTION	Mos.	\$K
<b>DLI-2 Undergraduate Emphasis</b>				
9817406	Agogino, Alice	UC-Berkeley	12	200
9816026	Maly, Kurt	Old Dominion Univ.	12	80
9816644	Kappelman, John	UT-Austin	24	287
<b>Subtotal</b>				<b>567</b>
<b>DLI-2</b>				
9817485	Kornbluh, Mark	Michigan State	60	3,600
9817484	Crane, Gregory	Tufts	60	2,758
9817434	McKeown, Kathleen	Columbia University	60	5,002
9817496	Wactlar, Howard D.	CMU	48	4,000
9817432	Smith, Terrence	UC-Santa Barbara	60	5,800
9817799	Garcia-Molina, Hector	Stanford University	60	4,300
9817353	Wilensky, Robert	UC-Berkeley	60	5,000
9874747	Verba, Sidney	Harvard University	36	1,800
9817416	Lagoze, Carl	Cornell University	48	2,268
9874759	Etzioni, Oren	Univ. of Washington	36	598
9817492	Gorman, Paul	Oregon Health Sciences	36	650
9817511	Weiderhold, Gio	Stanford University	36	520
9817430	Choudhury, Sayeed	Johns Hopkins	36	530
9874771	Armistead, Samuel G.	UC-Davis	36	497
9817483	Seales, W. Brent	Univ. of Kentucky	36	500
9817444	Buneman, Peter	Univ. of Pennsylvania	36	505
9874781	Rowe, Timothy	UT-Austin	36	500
9817527	Myers, Brad	CMU	36	450
9817473	Chen, HC	Univ. of Arizona	36	501
9817572	Palakal, M.	Indiana Univ.	36	316
9817518	Willer, D.	Univ. of South Carolina	48	1,199
<b>Subtotal</b>				<b>41,294</b>
<b>DL International</b>				
9975164	Larson, Ray	UC-Berkeley	36	305
9905842	Byrd, Donald	Univ. of Mass	36	494
9905935	Hedstrom, Margaret	Univ. of Michigan	36	488
9906025	Calcari, Susan	UW-Madison	36	480
9907892	Lagoze, Carl	Cornell Univ./ePrint	36	292
9905955	Lagoze, Carl	Cornell Univ./ILRT	36	240
<b>Subtotal</b>				<b>2,299</b>
<b>Grand Total</b>				<b>44,160</b>

## Research Coverage of DLI

DLI-1 focused on research, and the six projects were led by individuals with strong backgrounds in technical fields, largely computer and information sciences. An inspection of the available information shows that DLI-2 has greatly expanded the support of different disciplines working in the digital libraries field. Table 2 lists in alphabetical order many of the home departments of investigators funded through DLI-2.

Another illustration of the breadth of coverage in DLI-2 can be seen in Table 3, which deals with the types of content, media or formats being studied. To aid the reader interested in particular topics, universities focusing on them also are listed.

Even with respect to technologies considered, DLI-2 is considerably broader than DLI-1. Table 4 summarizes the technical areas studied along with universities involved in each. The reader is invited to make up a list independently of areas closely related to digital libraries and compare that list with the one given. Alternatively, one might look at lists in other introductions to the field, like that in the April 1995 special section of *Communications of the ACM*. There are areas likely to be on many people's lists that were not much of a focus in DLI-2, such as abstracting, browsing, ethnography, hypertext, indexing, interaction, sociology, storage and virtual reality.

Furthermore, though there are some projects dealing with key issues of information retrieval (IR) (e.g., the Berkeley international effort) or human-computer interaction (HCI) (e.g., the CMU separate project on video editing), these topics seem to play a relatively minor role in the overall initiative. But extensive experimentation in these areas is necessary for the field to mature. Such work on IR and HCI will require readily available test-beds, usability tests involving large numbers

**Table 2. Discipline Coverage of DLI-2**  
(selected home departments of investigators)

Anthropology	Biomedical Information	Classics
Computer Science	Economics	English
Fine Arts	Geography	Geological Sciences
Government	Electrical Engineering	Environmental Science
History	Information Management	Information Studies
Language Technology	Library & Information Science	Linguistics
Management Info. Systems	Medical Informatics	Political Science
Psychology	Religious Studies	Robotics
Sociology	Spanish	Teacher Education

of users, careful comparative experiments and other related studies.

Following along these lines, and possibly of particular interest to ASIS members, is consideration of the ties to information science that are visible in DLI-2. Geographical information and medical informatics are the focus of several efforts. Christine Borgman of UCLA's Graduate School of Education and Information Studies is a co-principal investigator playing a role in the University of California, Santa Barbara project, while Javed Mostafa of the School of Library and Information Science at Indiana is a co-principal investigator in their project. Librarians are co-investigators on several projects. In the international program, two of the projects are run from schools of information (i.e., at Berkeley, Michigan). But overall, few funded DLI-2 projects are run out of library or information science departments or schools. In general most project direction is by computer rather than information scientists.

### Continuing DLI-2

It is clear from the funding for DLI-2 that reviewers and agencies involved largely felt that DLI-1 activities should be continued. While UIUC was not supported, its key partner in DLI-1, University of Arizona, is supported in DLI-2, continuing in particular the work on automatic classification, aiming to consolidate results by scaling up and comparing algorithms. Though the University of Michigan did not receive a follow-on award per se, Margaret Hedstrom in their School of Information is leading a project funded at almost \$500,000 on the topic of preservation (using emulation). Further, work on agents that is rather similar to that at University of Michigan (but somewhat more focused) is being supported at Indiana, Bloomington (for personalized information filtering) and at Washington (to aid retrieval from the WWW). One successful supplement to the project at Michigan was the Joint NSF-European Union (EU) Working Groups on Future Directions of Digital Libraries Research (<http://www.dli2.nsf.gov/workgroups.html>) that stimulated extensive international discussion. Also, the JSTOR effort

(<http://www.jstor.org/>) launched at Michigan has become a serious commercial venture involving digitization of important old journals.

All of the other DLI-1 projects are continuing earlier work with a relatively high level of funding. Consolidation is in evidence too, with coordination of the three California efforts. All three will develop testbeds and foster interoperability, a strong point of the prior work at Stanford. Each will carry out evaluations. All three have efforts on user interfaces, regarding presentation, and on analysis of collection data. In addition, the San Diego Supercomputer Center will act as collection clearinghouse and the California Digital Library will facilitate statewide collaborative knowledge creation and dissemination.

The Santa Barbara effort is focused on building the Alexandria Digital Earth Prototype as a digital earth modeling system made up of Information Landscapes. That effort extends prior work through a broader vision, with many goals for further technical development and with user testing involving UCLA and other partners.

The Berkeley proposal discusses a very large number of

**Table 3. Types of Content and DLI-2 Sites Where They Are Studied**

Types	Universities
Bibliographic Records	Arizona
Engineering Education	UC-Berkeley
EPrints	Cornell (intl ePrint)
Folk Literature	UC-Davis
Geo-referenced Info.	UC-Santa Barbara
Health Care	Oregon Health Sciences
Humanities	Tufts; Kentucky
Library Reference	Washington
Medical Images	Stanford
Mixtures of Media	UC-Berkeley (intl); Cornell (intl ILRT)
Patient Records	Columbia
Sheet Music	Johns Hopkins; UM-Amherst (intl)
Skeletons	UT-Austin
Simulations	South Carolina
Social Science Data	Harvard
Speech	Michigan State
Video	Carnegie Mellon
Web	Arizona; Pennsylvania; Washington
X-ray CT Scans	UT-Austin

research topics around the theme “Re-inventing Scholarly Information, Dissemination and Use.” But the proposal body does not appear to connect this motivating theme to the lively self-publishing efforts expanding around the globe (e.g., e-prints, reports, dissertations, courseware, biomedical

information). Rather, in the tradition of Berkeley UNIX they propose to build general tools to help digital library users do more on their own and also to study models and conduct user studies on dissemination and use.

The Stanford proposal adopts a different approach, emphasizing a comprehensive problem analysis of four barriers to effective digital libraries. One barrier is that contents and systems are highly diverse and heterogeneous. The other barriers are needs for which no solution now exists: filtering mechanisms, portable interfaces and an economic infrastructure that guarantees privacy. Like at Berkeley, the Stanford team will develop software. It will be for value filtering, for portable devices, for extending their earlier InfoBus into the InterServ suite of models and protocols and for economic modeling.

A smaller project at Berkeley (run in connection with the engineering education coalition, NEEDS) is part of the DLI-2 undergraduate emphasis (<http://www.dli2.nsf.gov/addendum.html>), leading toward a national digital library for Science, Mathematics, Engineering and Technology Education (SMETE-lib). Expansion of this effort in upcoming years is likely to go beyond planning and pilot grants to large-scale efforts. Thus it is important that there be closer coordination with other DLI efforts than has occurred to-date.

### Outside Activities

As Michael Lesk indicates in the following article, a great deal of work on digital libraries has proceeded quite independently from DLI. For example, OCLC, the Online Computer Library Center in Dublin Ohio (<http://www.oclc.org>), has led the way on the Dublin Core ([http://purl.org/metadata/dublin\\_core](http://purl.org/metadata/dublin_core)) workshop series, the most important metadata standards activity for the field (though there are others emerging from IMS and IEEE, focused on education). OCLC also has helped run W3C-sponsored work on the Resource Description Framework (RDF) and coordinates CORC (Cooperative Online Resource Catalog), the worldwide cooperative library venture to catalog the WWW, that benefits from a variety of tools developed at OCLC. Another important tool from OCLC is the SiteSearch retrieval system (essentially the same as that used for FirstSearch), recently converted to Java. On the production side of things, OCLC owns one subsidiary (Forest Press) responsible for work on the Dewey Decimal Classification and so is exploring its use in digital libraries and knowledge management. Another OCLC subsidiary handles preservation and digitization; internally there is support as well for electronic journals and their permanent availability.

Commercially, there are many digital library efforts. IBM sells a shrink-wrapped software system called Digital Library. In Japan, several companies involved in library automation sell and adapt digital library software to leading universities. Internationally, thanks to significant funding and other support, digital libraries are under development in many countries, especially in Europe and Asia (see April

**Table 4. Technical Areas and DLI-2 Sites Where They Are Studied**

Types	Universities
3-D Modeling	UC-Santa Barbara; UT-Austin
Access Control	UC-Berkeley
Agents	Indiana-Bloomington; Washington
Archiving/Preservation	South Carolina; Univ. of Michigan (intl)
Audio Retrieval	Johns Hopkins; Michigan State; UM-Amherst (intl)
Classification, Clustering	Arizona
Data (Access) Services	Harvard
Digital Video	CMU
Economic Models	UC-Berkeley; Stanford
Electronic Notebooks	UC-Berkeley
Federation	UC-Berkeley (intl); Cornell; UW-Madison (intl)
Geographic Info. Systems	UC-Santa Barbara
Images	UC-Berkeley; UC-Santa Barbara; Kentucky; Stanford; UT-Austin
Information Filtering	Indiana; Stanford
Information Visualization	CMU
Learning Contexts	UC-Santa Barbara
Linking	Cornell (intl – ePrint)
Log (Trace) Analysis	Oregon Health Sciences
Mobile Computing	Stanford
Multimedia Fusion	CMU; Columbia
Natural Language Processing	Columbia
OCR	UC-Berkeley; Johns Hopkins
Parallel Processing	Arizona
Protocols	Stanford
Personalization	Columbia
Provenance	Penn.
Restoring Manuscripts	Kentucky
Speech Processing	UC-Davis; Michigan State
Summarization	CMU; Columbia
Text Analysis	Tufts
Video Editing	CMU

1998 special section of *Communications of the ACM*). ACM has run international conferences for the field since 1996. Other conferences and workshops have occurred or are planned in Australia, Croatia, France, Germany, Hong Kong, India, Japan, Portugal, Singapore, Taiwan, United Kingdom, etc. Many include reports on or are closely connected with DLI (see <http://www.dli2.nsf.gov/workshops.html>). Two workshops have focused on international cooperation for the field of digital libraries (see <http://www.ks.com/idla/>).

Two of the many other related efforts are especially notable. One is the ongoing series of TREC (Text REtrieval Conference) meetings and competitions. Covering information retrieval and filtering, this National Institute of Standards and Technology (NIST) effort has expanded to handle multiple languages, to deal with interactive sessions and to start to cover media beyond text. The other is the D-Lib activity (<http://www.dlib.org>). Most visible in that category is *D-Lib Magazine*, but also important are the working groups. One has dealt with the Networked Computer Science Technical Reference Library (NCSTRL, <http://www.ncstrl.org>). Another has dealt with metrics. It is likely that others will emerge.

### Assessment and Conclusion

With work on DLI since 1994, and a new round of funding allowing a broad range of projects to proceed, it seems timely to assess the progress and promise of the Digital Libraries Initiative. That is a difficult task, requiring a book or books, since there have been many hundreds of publications that should be covered (<http://www.dli2.nsf.gov/publications.html>). Furthermore, it is difficult to gauge how many related studies were motivated by DLI efforts or simply parallel the DLI efforts. The comments below reflect this larger scene and provide one person's viewpoint of overall progress.

First, we see ongoing progress and adoption of the work in the information retrieval field. TREC has shown that methods studied before the 1990s scale up to larger collections. A number of projects have demonstrated success with broadening to diverse languages and media forms. While more work is needed, there has been quite a lot done already on image retrieval, and a growing effort on retrieval from speech, music and video. It is time for controlled experimentation and comparative studies, as well as trials with wavelets and other technologies. Much work is needed regarding information visualization, which is really just in its infancy. In that case, as well as in clustering and classification, we are only in the early days of applying the storage and processing capabilities now readily available – to make a significant difference for common users.

Second, we see widespread acceptance of the broadening of the digital libraries field; not only libraries but also museums and archives are within scope. Data collections, Web pages, educational materials, experimental data, simulations and the whole province of electronic publishing are also

under consideration. Collection development is proceeding in novel ways, whether from digitization, the work of dedicated curators, feeds from publishers, user annotations, traces of expert users or through self-archiving. Users are not only scholars and researchers, but also teachers and students, as well as special groups devoted to particularly interesting collections. We are just beginning to see some commercialization, aided by the realization that digital libraries are the high-end of information systems.

Third, we see a coupling of this initiative with attempts to organize the WWW. There will continue to be interplay between work on digital libraries and efforts such as those involving OCLC, in particular Dublin Core, RDF and CORC, that will have Web-wide impact. There will be related advances in searching using various languages as well as media forms. More and more objects will have metadata associated, and (semi) automatic systems will aid in cataloging as well as browsing. As large numbers of collections emerge and are called “digital libraries,” advances will occur to search many together, leading to a second-generation federated search system that allows users to “slice and dice” whatever is available into any convenient organization desired.

Fourth, there is support of undergraduate education that depends on collections or repositories of curriculum or courseware resources. For example, NSF's Division of Undergraduate Education (DUE) funded 16 projects during the period 1993-1998 on collection building within specific disciplines/curricula. One is the Computer Science Teaching Center, available at <http://www.cstc.org>. A number of 1999 awards related to digital libraries are expected to be funded by DUE, in several cases also supported by other parts of NSF.

That leads to the final key point, regarding users. Personalization will indeed become feasible, starting with pilot efforts but ultimately becoming more common. Tailored systems are being built at the level of the organizational library (e.g., through virtual libraries devised by staff for a university community, or with technologies like SFX from University of Ghent and the Los Alamos National Laboratory – see <http://lib-www.lanl.gov/~hvds/sfx/htmls/sfxhome.html>). Content will be aggregated in various ways, community ratings will be considered and user actions will be analyzed, either by client or agent software. Though DLI efforts in this regard continue to operate at the exploratory level and are not a major focus in the initiative, several projects espouse personalization and enough others are working in this area that we can expect some real progress within a few years.

In conclusion, readers are urged to study the next two articles about DLI and to be in touch with the staff of projects that are of interest. The information science field needs closer ties with the important emerging area of “digital libraries” in which the next generation of high-end information systems is gestating and in which a large number of well-supported interesting collections are developing.

# A Digital Library for Authors: Recent Progress of the Networked Digital Library of Theses and Dissertations

Constantinos Phanouriou, Neill A. Kipp, Ohm Sornil, Paul Mather, and Edward A. Fox  
Department of Computer Science  
Virginia Polytechnic Institute and State University  
Blacksburg, VA 24061  
<http://www.ndltd.org/>  
{phanouri,nkipp,osornil,paul,fox}@cs.vt.edu

## ABSTRACT

The Networked Digital Library of Theses and Dissertations (NDLTD) is more than an online collection of Electronic Theses and Dissertations (ETDs). It is a scalable project that has impact on thousands of graduate students in many countries as well as diverse researchers worldwide. Its 59 official members represent 13 countries and integrate some of the world's newest research works, including ETD collections at Virginia Tech and West Virginia University, where ETD submission is now required. The number of ETDs in Virginia Tech's collection has nearly tripled in the last year, while the number of accesses to it has grown by more than half. NDLTD is committed to authors, aiming to improve graduate education for the over 100,000 students that prepare a thesis or dissertation each year. It encourage them to be more expressive by making incorporation of multimedia components into their theses easier. NDLTD activities include: applying automation methods to simplify submission of ETDs over the WWW; specifying the application of the Dublin Core to guarantee that metadata can satisfy needs of searching and browsing; selecting open standards and procedures to facilitate interoperability and preservation; and demonstrating a variety of interfaces, both 2D and 3D, along with exploring their usability.

**Keywords** digital library, user interfaces, information retrieval, usability engineering

## INTRODUCTION

The Networked Digital Library of Theses and Dissertations (NDLTD) is an international effort that seeks to improve graduate education by encouraging all uni-

versities to require submission of Electronic Theses and Dissertations (ETDs). In the process of preparing and submitting their ETDs, student authors learn about the richness of expression that a digital medium makes possible and how to use online resources (i.e., digital libraries). It is through this process that universities can make available immediately and cost-effectively the research results of their graduate students as a contribution to the advancement of education and humanity [Fox, *et al.*, 1996, 1997, 1998].

NDLTD is a digital library in the richest of definitional senses [Borgman, 1999; Lesk, 1997; Fox, *et al.*, 1995]. It has a growing collection of ETDs that it makes available on the Internet; it is concerned with acquisition, preservation, and cataloging of ETDs; it provides useful and usable visualizations of the entire distributed collection. NDLTD is organizing universities and spreading new ideas about scholarly publishing through collaboration and sharing. As each member university joins the NDLTD, a local ETD submission process is planned—be it dictated by university governance, decided by faculty working group, or demanded by the graduate students themselves. Libraries renew their commitments to serve ever-widening scholarly communities, graduate schools sponsor training and workshops, and students and faculty become electronic document authors and publishers. With NDLTD, universities can evolve and share their own systems for collecting and making ETDs available and thereby contribute to the global educational process in exciting ways. As a result, graduate education and scholarly publishing will permanently change, with digital libraries playing a dominant role.

## NDLTD ACTIVITY

“We certainly want to be thorough and we absolutely must get it right, but this is not the sort of thing which will profit from passive study, and we have arrived at the point where we must begin to implement the project.” [WVU, 1998]

## Membership

As of May 1999, NDLTD has 59 members from 13 countries. Fifty-three (53) members are universities; the remainder are coalitions, non-profit organizations, or corporations.

## Governance

The NDLTD steering committee meets in September and April of each year, and is chaired by the initiative's director, Edward A. Fox. Membership includes representatives from Virginia Tech and other NDLTD member universities, Adobe, Association of Research Libraries, Coalition for Networked Information, Council of Graduate Schools, Dissertations Online (Germany), IBM, National Library of Canada, OCLC, UMI, and UNESCO. Topics discussed in the Fall of 1998 and Spring of 1999 included: membership, outreach, expansion programs, archiving, preservation, metadata, ETD submission and workflow processing, workshops, Web sites, ongoing evaluation, results reporting, particular implementations, development and plans, and future opportunities for funding.

## ETDs Required

While most universities in the NDLTD are implementing pilot programs, Virginia Tech and West Virginia University have made ETD submission a requirement for graduate students on their campuses.

**Virginia Tech.** Virginia Tech has required electronic submissions since January 1, 1997, and does not accept paper thesis and dissertation submissions. The Graduate School and University Library have collected more than 1700 ETDs. Of these, 1225 are available worldwide; the remainder are not available beyond the campus at the request of the submitting student. Most documents are in PDF, augmented by various multimedia formats (e.g., JPEG, GIF, TIFF, MPEG, WAV, HTML, VRML, QuickTime, Java applets). Most were created in Word and Word Perfect, but some were created in TeX, LaTeX, and SGML (using the ETD-ML document type definition). The Virginia Tech ETD library uses OpenText for indexing the full text of the collection.

**West Virginia University.** In August 1998, West Virginia University began to require students to submit theses and dissertations electronically [Mendels, 1998]. WVU no longer accepts paper theses and dissertations; exceptions must be approved by the Office of the Provost. WVU requires its documents to be submitted in PDF format. The West Virginia ETD collection contained 210 documents as of April 1999. The local committee for ETD implementation consists of members from its faculty, library, research centers, graduate school, and the Office of Academic Affairs.

## Other Collections

**Australian Digital Theses Project.** Seven institutions in Australia (led by the University of New South Wales, and centered in its library) are collaborating to begin accepting electronic theses from postgraduate students. They have standardized on SGML and PDF as document formats. The collection's oldest work is dated 1968.

**Dissertation.com.** Dissertation.com is part of Amazon.com and functions as a publishing agent for students. It offers electronic dissertations in PDF or paper formats for 20 to 40 US dollars. Abstracts are freely available.

**Dissertations Online.** A national project in Germany involves 4 universities, 2 large libraries, a large computing center, and 4 scholarly societies (chemistry, mathematics, physics, sociology, and education). The focus is on SGML and XML, and helps train students in their disciplines, e.g., to use the markup language for chemistry.

**Encyclopaedia Diplomica.** Encyclopaedia Diplomica is a German company acting as a selling agent for students who prepare scholarly works. Papers are in one of the following formats: Word, PDF, or PostScript. Abstracts and full tables of contents are available for free. Prices for the full documents are 150 to 300 US dollars. The collection offers approximately 20 titles. Most of the documents are in German; the rest are in English or French.

**North Carolina State University.** NCSU has about 30 ETDs in its online collection, which is sponsored by the NCSU Libraries, Graduate School, and Information Technology division. At NCSU, ETD submission is not yet required. Submissions are in PDF format. The Graduate School holds monthly thesis preparation workshops for its students.

**Rhodes University of South Africa.** The Rhodes University of South Africa has begun an ETD pilot project. They request both paper and digital submissions.

**University of Tennessee, Memphis.** The University of Tennessee, Memphis has three documents in its collection. Of these, all are in PDF, but one is also in HTML.

**University of Michigan.** While not an official member of NDLTD, the University of Michigan has begun a thesis pilot program. Instead of PDF, they have four ETDs in SGML, conforming to the Text Encoding Initiative Document Type Definition [Sperberg-McQueen and Burnard, 1994].

**University of Virginia.** The University of Virginia has adopted an ETD pilot; it accepts electronic theses from Engineering bachelor's students. The university plans

to require Master's and PhD's at a later time.

**University of Waterloo, Ontario.** The University of Waterloo in Ontario, Canada is the center of a three-institution cooperative and has sixteen documents online, in PDF with paper and PostScript sources, including one dated 1964. The site is sponsored by the Electronic Thesis Project Team and the University of Waterloo Library. They provide documents for free, but request the name, affiliation, and "reason-why" from the patron before permitting the thesis to be downloaded. The site uses OpenText for searching the full text of the collection.

#### **VIRGINIA TECH INITIATIVE (VT-ETD)**

The Virginia Tech ETD (VT-ETD) initiative has developed software and practices adopted by a number of other NDLTD members.

#### **Authoring and Training**

ETD authors are typically graduate students with above-average knowledge about computers. In a survey of graduate students after their submission, it was found that almost all of them used the Web to find information while doing their research. This makes them aware of what can be published electronically. Keeping with the goal of NDLTD that students should be able to author, submit, and maintain (with annotations and reviews) their work electronically, VT-ETD educates them on how to contribute their own work to the online community with workshops and a comprehensive and informative Web site (<http://etd.vt.edu/>).

Creating a document electronically is simple; enriching it with multimedia, aligning with standards, and making it interactive can be challenging. VT-ETD attempts to make interactive multimedia easier for students by providing the necessary tools and help on how to use them. Usage of multimedia in ETDs is increasing, perhaps due to regular training workshops sponsored by the Graduate School. For example, a dissertation from Chemistry contains 3D VRML models of molecules, a thesis from Animal Science contains audio clips of parrot sounds, and a thesis from Architecture contains video clips from a Turkish coffeehouse.

#### **Acquisition and Collection Management**

To improve upon ETD submission scripts that were written as prototype software, VT-ETD recently developed customized, database-driven ETD management software. As a result, students have more control over their ETD during the entire authoring and submission process. Furthermore, storing ETD metadata in a database enables the VT-ETD to provide better acquisition, searching, and browsing services. It also enables the development of multiple user interfaces to the collection.

VT-ETD encourages students to treat their theses or dissertations as electronic documents from the begin-

ning. With the new submission software students can register their documents early, set up their metadata, and upload their ETD drafts in pieces. This benefits the students in two ways. Primarily, students' work is stored in a secure location, with frequent and reliable backups; thus they are less likely to lose their work if their home or office computer fails. Furthermore, drafts are available for their committee to see and review electronically. Students can restrict public access to their work until they defend. Todd Miller's honor thesis work has developed and tested software to provide online annotation services to the ETDs.

VT-ETD provides students with four options on making their work accessible. The first option is *unrestricted*: release the entire work immediately for access worldwide. The second option is *restricted*: release the entire work for Virginia Tech access only. The third option is *withheld*: secure the entire work for patent and/or proprietary purposes for a period of one year. With the new submission software, VT-ETD now provides students with a fourth option, *mixed*, where they can break down their work and use any of the above options for each part individually. For example, they can make their abstract and introduction worldwide accessible, but restrict the main body of their work for Virginia Tech access only. Table 1 shows the distribution of their choices.

#### **Cataloging**

To facilitate collection sharing, NDLTD members are asked to freely share any MARC records available. Also, due to the variety of heterogeneous DL implementations it became obvious that a standard set of metadata elements should be identified for ETDs (which itself would aid the development of a canonical representation for ETDs). The metadata should be broad enough to permit crosswalks between many popular metadata standards and frameworks such as MARC and Dublin Core, but focussed on the domain of ETDs. VT-ETD began with the standard Dublin Core elements, and then enhanced these to tailor them to the ETD domain.

VT-ETD metadata is designed for practical application across a broad set of information storage and retrieval applications and settings, such as Web, IBM Digital Library, OCLC SiteSearch, and OpenText LiveLink. It is intended to provide a functional medium between static resource description and periodic record-keeping (low-level authority information). As such, part of the metadata is tied to (and derived from) ETD workflow and graduate school policies. Most, however, is static information that is supplied by the ETD author or by a cataloger.

The metadata framework is not rigid. The elements are guidelines, with varying degrees of recommendation for

employment in a local institutional ETD project setting. All NDLTD members are expected to collect all metadata elements marked as “mandatory,” so that there may be a minimal basis for searching across all NDLTD collections.

Each ETD has metadata describing the ETD as a whole, and then each separate part (file) of the ETD has its own metadata. URN identifiers within the metadata parts are used to tie together the metadata parts of an ETD in a parent/child structure. Implicit inheritance is used to minimize the repetition in elements for child items of an ETD.

### Preservation

The Virginia Tech Graduate School requires a specific form for the submission of ETDs to maintain the consistency of these complex documents. The formal statement of these guidelines serves graduate students submitting ETDs, professors with whom they work, and scholars who study the submitted ETDs. VT-ETD defined ETD-ML, a Document Type Definition (DTD) in both SGML and the Extensible Markup Language (XML) for the representation of ETDs. XML is a logical choice for encoding and archiving complex electronic documents [Bray, *et al.*, 1998]. To build ETD-ML, VT-ETD analyzed constructs in existing theses and dissertations and studied the rules for their submission. Software is available to NDLTD members that converts ETD-ML ETDs into HTML for Web accessibility.

### Search and Retrieval

**IBM DL.** VT-ETD developed a preliminary interface to the ETD collection using the IBM Digital Library (IBM DL) product. The Net.Data dynamic page builder component of IBM DL provides a Web front-end to the contents inside the IBM DL and allows users to search the VT-ETD collection. The full-text of the ETD PDF files, as well as abstracts, are indexed by the IBM DL text search server.

The current IBM DL search interface allows users to search the VT-ETD collection in two ways: through the metadata or the full-text index. A user can perform either type of search separately or use both types simultaneously in the same search. Thus, for example, a user can search the collection for each time the phrase “digital library” appears in an ETD, meanwhile specifying that each retrieved ETD must be a thesis from computer science submitted prior to 1997.

**SIFT.** The SIFT filtering software from Stanford has been adapted by Zhambo Sun to work for ETDs. This allows interested parties to specify information needs through email or a WWW interface. When integrated into the rest of the workflow, this should lead to an email notification whenever a new submission matches

any stored user profile.

### NEW DIRECTIONS IN VISUALIZATION

VT-ETD is trying to enhance the information retrieval process for DLs by developing richer browsing interfaces to its ETD collection. In particular, VT-ETD is experimenting with 3D interfaces on desktop machines and on immersive virtual reality devices.

#### 3DL

3DL presents the ETD collection as a 3D VRML model through which users can navigate. It mimics a traditional library: including lobbies, elevators, floors, signs, displays, windows, artworks, doors, rooms, bookcases, and books. Doors are hyperlinks to rooms and books are labeled hyperlinks to items in the ETD collection [Kipp, 1997]. In addition to the usual library components, 3DL uses images extracted from the collection and presents them as hyperlinks in a “virtual art gallery” [Bayraktar, *et al.*, 1998]. VT-ETD developed 3DL as an alternative interface to the ETD collection.

#### CAVE-ETD

CAVE-ETD extends the 3DL project from the desktop to an immersive virtual reality environment. CAVE-ETD runs in the Cave Automated Virtual Environment (CAVE), a 10x10x10-foot room, with stereoscopic projections on three walls and the floor, wherein the user may interact with the world through tracking devices, eyeglasses, and a wand. In CAVE-ETD, “Books” are organized on “shelves,” shelves are laid out in “aisles” in a “room,” and rooms are labeled and arranged in a logical sequence. Books can be browsed on the shelves by navigating through the room and reading the titles on the book spines. Real-time clustering methods are being investigated to determine their utility. Although it is unlikely that we will all have a 3D CAVE in our office, it is more likely that we will have a miniature version on our desk.

Both the CAVE-ETD and the 3DL rely on a user’s prior knowledge and experience in a traditional library. This conforms to the usability principle of familiarity [Hix and Hartson, 1993] which aligns with the results of the usability trials of both the 3DL and the CAVE-ETD.

### QUANTITATIVE EVALUATION

**Collection Size.** VT-ETD began collecting ETDs in 1995. By the end of 1998, Virginia Tech had 1546 electronic documents (theses, dissertations, and other documents) in its ETD collection (Table 2).

ETD authors are encouraged to include various multimedia components. Most PDF files include color images or figures. Ninety-three percent (93%) of the files in the collection are PDF and text files, while nearly 7% of the files are supplemental images, sounds, and movies (Table 3).

**Access Statistics.** As the collection grows and gains popularity and more institutions join NDLTD, the number of accesses to the system goes up (Table 4).

The monthly access graph is shown in Figure 1. We can see that number of accesses tends to increase each year. However, there were fewer accesses during the summer break when universities are not in session.

Among US domestic domains, educational institutions contributed to the largest number of requests. Half of these accesses were from users at Virginia Tech, affirming that local researchers and authors are using their own collection. Commercial interest is next, followed by other organizations, while government domains continued to show high interest (Table 5).

Each of the top-five accessing countries has increasing number of accesses every year (Table 6). The United Kingdom and Germany dominated the accesses from outside the US. This trend corresponds to the advancements in network facilities in those countries.

## USABILITY EVALUATION

**3DL.** Human interfaces to VRML browsers are notoriously bad [Carey and Bell, 1997]. VT-ETD usability trials support this conclusion. Even with high-resolution, high-speed desktop displays, refresh rate was poor and navigation was clumsy. Users said that looking at the rooms interface was “nice” but that a plain list of titles would be more useful. VT-ETD is working to improve the speed and usability of the 3DL interface before further evaluation.

**CAVE-ETD.** In trial runs in the CAVE-ETD, we noticed that new users have difficulty adjusting to the interface. Although the interface corresponds most closely with that used in most 3-D computer games, using the CAVE “wand” input device is not natural. “Sideways stepping” would also make browsing through books on shelves more useful, say users. Providing a useful amount of text (e.g., author, title, year) on the book spines substantially slows the CAVE display. The usability study produced many qualitative hints for designers of three-dimensional interactive library interfaces.

## CONCLUSIONS

Digital libraries are more than organizations of information. They are systems by which societies cope with their information problems and through which societies provide information services to users. NDLTD contains a document repository, indeed, but it also consists of the system and society by which that document repository is grown, accessed, maintained, and preserved.

NDLTD is a live test of a new economic model for digital libraries, whereby automation and federation, plus coupling to normal practices and use of standards, lower the

costs sufficiently so that in the normal course of work by authors, graduate schools, and university libraries, a sustainable worldwide digital library can be built, leading to unprecedented sharing of research results. Ongoing research and development work, at Virginia Tech and by other NDLTD members, should expand and improve the services and benefits of this initiative.

## ACKNOWLEDGMENTS

The authors acknowledge the efforts of the NDLTD team, particularly John L. Eaton and Gail McMillan. We also thank Bharadwaja Vadapalli, Prashant Choudhary, Jay Rathi, Nirav Kamdar, Murat Bayraktar, Chang Zhang, for their work on 3DL. For ongoing development of VT-ETD and his careful collection of statistics, we thank Anthony Atkins. For their contribution to the ETD CAVE we thank Kevin Curry, Fernando Das Neves, and Hussein Suleman.

**Funding.** As of September 1, 1996, the U.S. Department of Education Fund for the Improvement of Post-secondary Education (FIPSE) provided grant support for a three-year project, “Improving Graduate Education with the National Digital Library of Theses and Dissertations (NDLTD).” This follows earlier support from the Southeastern Universities Research Association (SURA) for the “Development and Beta Testing of the Monticello Electronic Library Thesis and Dissertation Program.”

## REFERENCES

- Bayraktar, Murat, Chang Zhang, Bharadwaj Vadapalli, Neill A. Kipp, Edward A. Fox, “A Web Art Gallery,” Proceedings of Digital Libraries '98, the Third ACM Conference on Digital Libraries, Pittsburgh, June 1998.
- Borgman, Christine L., “What are Digital Libraries: Competing Visions,” *Information Processing and Management, Special Issue for Digital Libraries*, Gary Marchionini and Edward A. Fox, issue editors, 1999.
- Bray, Tim, Jean Paoli, and C. M. Sperberg-McQueen, “Extensible Markup Language (XML) 1.0,” W3C Recommendation, <http://www.w3.org/TR/REC-xml>, February, 1998.
- Carey, Rikk, and Gavin Bell, *The Annotated VRML 2.0 Reference Manual*. Addison-Wesley, 1997.
- Fox, Edward A., and James Powell, “Multilingual Federated Searching Across Heterogeneous Collections,” *D-lib Magazine*, September, 1998.
- Fox, Edward A., Brian DeVane, John L. Eaton, Neill A. Kipp, Paul Mather, Tim McGonigle, Gail McMillan, William Schweiker, “Networked Digital Library of Theses and Dissertations: An International Effort Unlocking University Resources,” *D-lib Magazine*, September,

**Table 1: Accessibility for first 1729 VT ETDs**

Accessibility type	Number of Documents	Percent
Withheld	338	19.6
Unrestricted	820	47.4
Restricted	542	31.3
Mixed	29	1.7
Total	1729	100.0

**Table 2: ETD collection size through 1998.**

ETD Types	% of ETDs	Pre-1996	1996	1997	1998	Total	%96-97	%97-98
Dissertations	46.0	4	35	167	505	711	377	202
Theses	52.8	14	49	232	522	817	373	125
Others	1.2		1	4	13	18	300	225
Totals		18	85	406	1040	1546	374	158
% of all ETDs		1.16	5.5	26.1	67.3			

**Table 3: Separate multimedia files in first 1454 VT ETDs**

File type	Number of Documents	Percent
PDF, text	5334	93.3
Image	322	5.6
Movie	45	0.8
Sound	18	0.3
Total	5719	100.0

**Table 4: Access Statistics through 1998**

	1996	1997	1998	%96-97	%97-98
Total successful HTTP requests	37,171	247,573	379,742	566	53
Average successful requests per day	102	678	1040	665	153
Distinct hosts served	9015	22,725	36,724	152	62
Total data transferred (Gb)	3.229	25.9	50.0	704	93
Average data transferred per day (kb)	9.038	73.6	136.9	814	186

**Table 5: Accesses from domestic domains**

Domain	96	97	98	%96-97	%97-98
US Education (.edu)	15,314	112,876	254,268	637	125
US Commercial (.com)	5,309	48,540	88,169	814	82
Networks (.net)	2,522	14,026	27,972	456	99
Other Organizations (.org)	375	3,132	1,434	735	-54
US Government (.gov)	282	1,362	6,885	383	406

**Table 6: International Accesses (selected)**

Countries	1996	1997	1998	%96-97	%97-98
United Kingdom	850	2922	8170	244	180
Germany	346	2378	7373	587	210
Australia	608	2501	4223	311	69
France	463	1161	4431	151	282
Canada	713	2367	3970	232	68

1997.

Fox, Edward A., John L. Eaton, Gail McMillan, Neill A. Kipp, Laura Weiss, Emilio Arce, Scott Guyer, "National Digital Library of Theses and Dissertations: A Scalable and Sustainable Approach to Unlock University Resources" *D-lib Magazine*, September, 1996.

Fox, Edward. A, Robert M. Akscyn, Richard K. Furuta, and John J. Leggett, "Digital Libraries," *Communications of the ACM*, 38(4), pp. 22-28, April, 1995.

Hix, Deborah, and H. Rex Hartson, *Developing User Interfaces*, John Wiley and Sons, 1993.

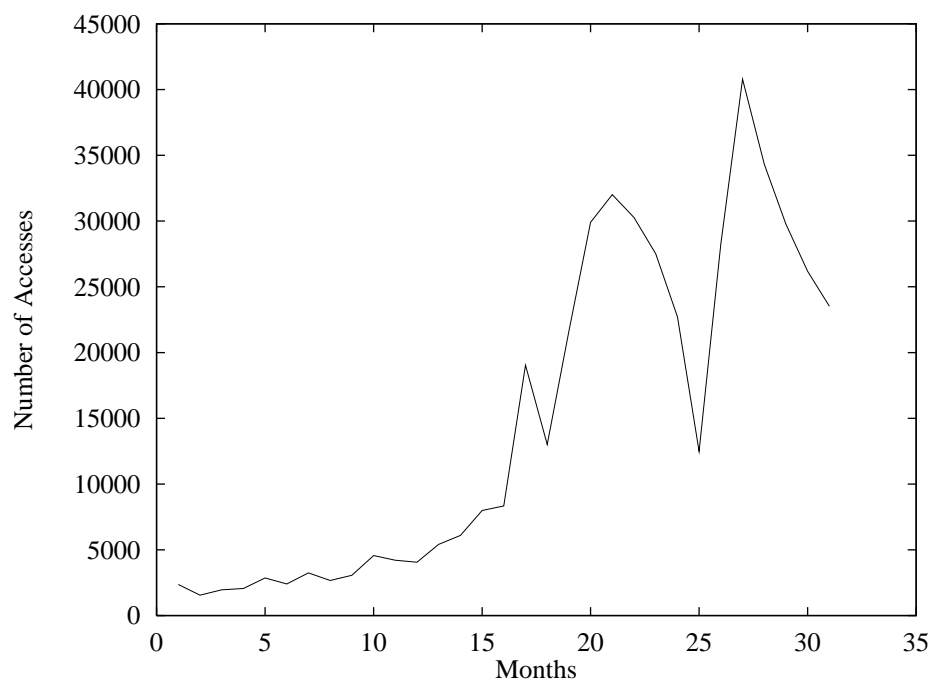
Kipp, Neill A., "Case Study: Digital Libraries with a Spatial Metaphor." SGML/XML '97 Conference Proceedings. Graphic Communications Association, Alexandria, VA, December, 1997.

Lesk, Michael. *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan-Kaufmann, 1997.

Mendels, Pamela. "Paper-Bound Thesis Dusted Off, Digitally," *New York Times*, September 5, 1998.

Sperberg-McQueen, C. M., and Lou Burnard, editors, *Guidelines for Electronic Text Encoding and Interchange: TEI P3*, Text Encoding Initiative, Chicago, 1994.

West Virginia University, "Frequently Asked Questions on ETDs," <http://www.wvu.edu/~thesis/etd-faq.html>, July, 1998.



**Figure 1: Monthly accesses (January 1996—July 1998)**

# Digital Libraries

## Contents

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

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

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

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

- [Hypernews](#)
- 

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

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

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

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

# Contents :

---

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

## Pedagogy:

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

---

[\[Main\]](#)

---

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

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

# Resources:

---

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

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

---

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

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

# Projects:

---

## DLI-2

- [DLI-2 home page at NSF](#)
- [DLI-2 projects funded from 1998-1999 submissions](#)
- [Index to NSF 1-page DLI-2 Award Summaries](#) - with all data available by 9/8/99
- D-Lib Magazine articles on DLI-2 by NSF etc.:
  - [FY 1999 Awards - S. Griffin](#)
  - [Commentary on DLI-2 - M. Lesk](#)
  - [NSF/JISC Int'l Initiative - N. Wiseman, C. Rusbridge, S. Griffin](#)
- [Selected abstracts of IIS awards \(including some DLI-2\)](#)
- Calls:
  - [NSF9863 - Digital Libraries Initiative - Phase 2 \(February 20, 1998\)](#)
  - [Addendum - Special Emphasis: Planning Testbeds and Applications for Undergraduate Education within the Digital Libraries Initiative - Phase 2](#)
  - [NSF996 - International Digital Libraries Collaborative Research \(November 9, 1998\)](#)

## DLI-1

- DLI-1 home page at [NSF](#) and older one at [U. Illinois](#)
- [DLI-1 information & resources](#)
- [DLI-1 publications](#)
- [Carnegie Mellon University](#)
- [Stanford University](#)
- [University of California at Berkeley](#)
- [University of California at Santa Barbara](#)
- [University of Illinois](#)
- [University of Michigan](#)

[Library of Congress](#) and its [American Memory Project](#)

Los Alamos and U. Ghent, SFX: [paper](#) and articles in D-Lib Magazine: parts [1](#), [2](#), [3](#)

[NARA](#) - National Archives and Records Administration

NASA [Digital Library Technology Projects](#)

---

# **NSDL (National Science, Mathematics, Engineering, and Technology Education Digital Library)**

## **DLI-2 Planning Testbeds and Applications for Undergraduate Education**

### **SMETE-Lib Study - NSF Science Mathematics, Engineering and Technology Education Digital Library reports**

#### **Related Projects:**

- **Funded Projects**
  - **SMETE Information Portal:** <http://www.smete.org>
  - **NEEDS - National Engineering Delivery System**
  - **Project Kaleidoscope**
  - **Geoscience:** **Call**; **DLESE** (Digital Library for Earth System Education); **Windows to the Universe**
  - **ODU project** (including buckets)
  - **U. Texas Austin:** **Technology for Education 2000**; **Virtual Multimedia Exams in Physical Anthropology**; **High Res X-ray CT (Computed Tomography) Facility**
  - **Computer Science Teaching Center (CSTC)**
- 

## **Selected International Efforts**

**Australia:** [\*\*National Library DL Initiatives\*\*](#)

[\*\*Bibliotheca universalis\*\*](#): (G7)

[\*\*British Library DL Programme\*\*](#)

[\*\*CIDL\*\*](#) - Canadian Initiative on Digital Libraries

**Electronic Theses and Dissertations Initiative:** [\*\*NDLTD project\*\*](#), [\*\*Collection\*\*](#), [\*\*Submission Instructions\*\*](#)

[\*\*ERCIM\*\*](#): [\*\*DL initiative\*\*](#) (DELOS)

**International Digital Libraries Association:** [\*\*IDLA home page\*\*](#)

**International Fed. of Library Associations and Institutions -** [\*\*IFLA\*\*](#): [page pointing to DL info](#)

## Japan:

- [Workshops - DLnet](#)
- National Museum of Ethnology - [MINPAKU: Virtual Tour](#)
- [Kobe U.: Digital Library Search](#), [TITAN Search using WWW](#)
- [Tokyo Inst. of Technology: Library](#)
- [Kyoto U.: Digital Library](#)
- [NAIST: Digital Library](#)
- [ULIS: Digital Library](#), [Multilingual HTML](#), [Multilingual folk tales](#)
- [University of Tsukuba: Digital Library](#)

**MeDOC**: (German Online Computer Science Library)

**NSF-EU Working Groups and Meetings**: [home page](#)

**Singapore Network**: [SINGAREN](#)

**UK Electronic Library Programme** including a project on preservation: **New Cedars Project: CURL Exemplars in Digital Archives** and a 13M record searchable OPAC called **COPAC**; **Centre for DL Research** (U. Southampton); **DL Group** (De Montfort U., and its **International Institute for Electronic Library Research**)

---

## Selected Publisher / Information-Distributor Projects:

- [ACM DL](#)
  - [UMI](#) and its [Digital Dissertations](#)
  - [Elsevier Electronic Services](#)
  - [IDEAL](#) (INTERNATIONAL DIGITAL ELECTRONIC ACCESS LIBRARY)
  - [IEEE-CS DL](#)
  - [OCLC](#) Electronic Collections Online
  - [Springer's Forum for Science](#) (The LINK Online Libraries)
- 

## Virginia Tech Projects:

- **Interactive Courseware on Digital Libraries** (this site itself is a part of it)
- **Interactive Learning with a Digital Library in CS** <http://ei.cs.vt.edu/>
  - Interactive Learning with a Digital Library in CS arch

<http://ei.cs.vt.edu/~cs5604/Adv/Adv-ILDLCS.html>

- Courseware <http://ei.cs.vt.edu/courses.html>
  - [Project Overview](#) (for FIE'96, in PDF)
  - [Project Interim Report](#), Oct. 1996
  - [Project Report for NSF EI PI Meeting](#), Nov. 1996
  - **Envision (CS literature)** <http://ei.cs.vt.edu/~cs5604/Adv/Adv-Envision.html>
    - Envision report <http://ei.cs.vt.edu/papers/ENVreport/final.html>
  - **CODER** <http://ei.cs.vt.edu/~cs5604/Adv/Adv-CODER.html>
  - **MARIAN**
    - [home page](#)
    - system <http://opac3.cc.vt.edu/htbin/marian>
    - old overview <http://ei.cs.vt.edu/~cs5604/Adv/Adv-MARIAN.html>
  - [CSTC - Computer Science Teaching Center](#) and related effort
  - [CRIM - Curriculum Resources Interactive Multimedia](#)
  - [W3C Web Characterization Repository](#) (of logs, traces, tools, papers)
  - Virginia Tech DL Superstorage Research, using [VT-PetaPlex-1](#), a [PetaPlex](#) system from [Knowledge Systems Inc.](#) with at least 100 processors and 2.5 terabytes
- 

## Approaches to DL:

- Build upon existing electronic materials
  - Netlib (numerical analysis) <http://www.netlib.org/> and its search: [http://www.netlib.org/utk/misc/netlib\\_query.html](http://www.netlib.org/utk/misc/netlib_query.html)
- Build upon publishers collections
  - AAAS - Science Online <http://www.aaas.org/>
  - ACM DL <http://www.acm.org/dl/>
  - ACS (Chemistry) - Online <http://www.acs.org/>
    - CORE Overview <http://ei.cs.vt.edu/~cs5604/DL/DL2.html>
    - D-Lib Magazine, Dec. 1995, Making a Digital Library, Chemistry Online Retrieval Experiment <http://www.dlib.org/dlib/december95/briefings/12core.html>
    - CORE at OCLC <http://www.oclc.org:5047/oclc/research/projects/core/>
  - Elsevier
    - Science Direct <http://www.elsevier.nl/>
    - TULIP (material science & engineering) homepage <http://www.elsevier.nl/inca/homepage/about/resproj/tulip.shtml>

- With universities + OCLC
  - [Highwire Press](#)
  - [IEEE](#)
  - [IEEE-CS DL](#)
  - [JSTOR](#)
- Commercial services and systems
  - IBM <http://www.software.ibm.com/is/dig-lib/>
    - Version 2 <http://www.software.ibm.com/is/dig-lib/v2factsheet/>
    - collection treasury <http://www.software.ibm.com/is/dig-lib/treasury/>
    - images - QBIC <http://www.qbic.almaden.ibm.com/>
    - news archive <http://www.software.ibm.com/is/dig-lib/newsarchive/>
- Enhance WWW (hypertext):
  - HyperWave <http://www.hyperwave.de/>
  - HyperWave [information server](#)
  - HyperWave author <http://www2.iicm.edu/hyperwave/author>
  - HyperWave author features <http://www2.iicm.edu/hyperwave/author/features.html>
  - HyperWave author specs <http://www2.iicm.edu/hyperwave/author/specifications.html>
  - Harmony <http://www2.iicm.edu/harmony>
  - Harmony screens <http://ei.cs.vt.edu/~cs5604/Adv/Adv-Harmony.html>
  - Amsterdam model <http://ei.cs.vt.edu/~mm/gifs/Amsterdam-hm.html>
- Community network multimedia history
  - BEV <http://www.bev.net>
  - BEV History <http://history.bev.net/bevhist/>
    - Timeline <http://history.bev.net/bevhist/historyBase/mainTimeline.html>
    - [Screen for Spring 1992](#)
    - [Screen for Article](#)
- Discipline - Greek Literature <http://www.perseus.tufts.edu/>
  - Evaluation - [article in TOIS](#)
- Discipline - Computer Science
  - Technical reports
    - [WATERS](#) - through 1995
    - CSTR <http://WWW.CNRI.Reston.VA.US/home/cstr.html>
    - NCSTRL <http://www.ncstrl.org/>
      - Search results, Search results abstract

- Doc. thumbnails, Doc. page 1
- CoRR: <http://xxx.lanl.gov/archive/cs/intro.html>
- Ptrs
  - DLs for CS <http://fox.cs.vt.edu/DLCS.html>
  - Results page, document page from search
- Genre - ETDs - electronic theses and dissertations
  - Virginia Tech <http://etd.vt.edu/>
    - Submission form <http://scholar.lib.vt.edu/ETD-db/ETD-submit/login>
    - Approval form <http://etd.vt.edu/submit/approval.htm>
    - Letter to students <http://etd.vt.edu/submit/letter.htm>
    - Standards <http://etd.vt.edu/submit/mm.htm>
  - Collection <http://www.theses.org>
  - Project - Networked Digital Library of Theses and Dissertations <http://www.ndltd.org>
    - Brief description <http://www.ndltd.org/info/descr.htm>
    - D-Lib Magazine Overview September 1996  
<http://www.dlib.org/dlib/september96/theses/09fox.html>
    - D-Lib Magazine Update September 1997  
<http://www.dlib.org/dlib/september97/theses/09fox.html>
    - D-Lib Magazine Federated Search September 1998  
<http://www.dlib.org/dlib/september98/powell/09powell.html>
    - FIPSE (US Dept. of Education) funding of 1996-1999 project
      - proposal abstract <http://www.ndltd.org/support/fipseabs.htm>
      - proposal full-text <http://www.ndltd.org/support/fipse10.pdf>
      - project final report ([PDF](#))

---

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

---

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

**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**



# DIGITAL LIBRARIES INITIATIVE

a community of  
researchers and  
agencies working  
together to bring the  
world's knowledge  
to your desktop

Digital Libraries Initiative  
Phase 2 HOME

Digital Libraries Initiative  
Phase 1 (1994-1998)

Search

[Register Now for  
DLI2 All-Projects  
Meeting 2000](#)

## Program Announcements

[Digital Libraries Initiative  
Phase 2 \(NSF 98-63\)](#)  
(closed)

[Planning Testbeds for  
Undergraduate Education](#)  
(closed)

[International Digital  
Libraries Collaborative  
Research \(NSF 99-6\)](#)  
(next target date Jan.15  
'01)

## Related Program Announcements

[National Science,  
Mathematics,  
Engineering, and  
Technology Education  
Digital Library \(NSDL\)](#)  
(next deadline April 14,  
'00)

[Geoscience Education](#)  
(next deadline April 10,  
'00)

## Feature

[DLI2 Funded Projects](#)  
[DLI2 Undergraduate  
Emphasis](#)

## Sponsoring Agencies and Programs

National Science Foundation ([NSF](#))  
[Digital Libraries Initiative](#)

Defense Advanced Research Projects Agency ([DARPA](#))  
[Information Technology Office](#)

National Library of Medicine ([NLM](#))  
[Extramural Programs](#)

Library of Congress ([LOC](#))  
[Digital Library Initiatives](#)

National Endowment for the Humanities ([NEH](#))  
[Digital Library Initiative](#)

National Aeronautics & Space Administration ([NASA](#))  
Federal Bureau of Investigation ([FBI](#))

## In Partnership with

[National Archives and Records Administration](#) (NARA)  
[Smithsonian Institution](#) (SI)  
[Institute of Museum and Library Services](#) (IMLS)

## [NSF Contact](#)

## [Agency Contacts](#)

**Digital Libraries Initiative Phase Two** is a multiagency initiative which seeks to provide leadership in research fundamental to the development of the next generation of digital libraries, to advance the use and usability of globally distributed, networked information resources, and to encourage existing and new communities to focus on innovative applications areas.

Since digital libraries can serve as intellectual infrastructure, this Initiative looks to stimulate partnering arrangements necessary to create next-generation operational systems in such areas as education, engineering and design, earth and space sciences, biosciences, geography, economics, and the arts and humanities. It will address the digital libraries life cycle from information creation, access and use, to archiving and preservation.

Research to gain a better understanding of the long term social, behavioral and economic implications of and effects of new digital libraries capabilities in such areas of human activity as research, education, commerce, defense, health services and recreation is an important part of this initiative.

[DLI2 International Projects](#)

[Special Projects](#)

[Funded Workshops](#)

[NSF-EU Working Groups](#)

[Publications](#)

[D-Lib Magazine](#)

## Related Information

[Glossary](#)

[News](#)

[Events](#)

[Recent Articles](#)

[Reports](#)

[DL Resources](#)

[iMP Magazine](#)

[DLI2 Material Submission Guidelines](#)

## Quick Search:

Submit comments and suggestions for digital library activities to [dli2 coordinators](#)

4.26.2000

This web site is maintained for the community by the Special Projects Program in the Information and Intelligent Systems ([IIS](#)) Division of the Directorate for Computer and Information Science Engineering ([CISE](#)). For official NSF documents, please visit the [NSF](#) web site.



# FUNDED PROJECTS

[DLI2 HOME](#)

[DLI1 \(1994-1998\)](#)

[SEARCH](#)

The following projects do not constitute a complete list of awardees from the Digital Libraries Initiative-Phase 2. Announcements of additional grant recipients will be made as they become official.

Projects are ordered alphabetically by institution

## [University of Arizona](#)

Project Web Site: [High-Performance Digital Library Classification Systems: From Information Retrieval to Knowledge Management](#)

Project Start Date: May 1, 1999

Project End Date: April 30, 2002

Expected Total Amt. \$499,998 (Estimated)

[NSF Awards Abstract](#)

[Hsinchun Chen](#) Principal Investigator

[Robin Sewell](#), Co-Principal Investigator

[Artificial Intelligence Lab](#), [Department of Management of Information Systems](#)

[Project Summary](#) (pdf)

Related Links:

["Beyond Geography: Mapping Unknowns of Cyberspace"](#) (Digital Library Research in the New York Times (9/30/1999))

[OOHAY Project for Digital Libraries](#)

[Spiders are Us](#)

[Information Analysis and Visualization](#)

[Medical Informatics](#)

## [University of California Berkeley](#)

Project Web Site: [Re-inventing Scholarly Information Dissemination and Use](#)

Project Start Date: April 1, 1999

Project End Date: March 31, 2004

Expected Total Amt. \$5,000,000 (Estimated)

[NSF Award Abstract](#)

[Robert Wilensky](#), Principal Investigator

[David Forsyth](#), Co-Principal Investigator

[Computer Science Division](#), [School of Information Management and Systems](#)

[Project Summary](#) (pdf)

Related links (html)

[Information about the Digital Library Project](#)

[University of California Davis](#)

Project Web Site: [A Multimedia Digital Library of Folk Literature](#)

Project Start Date: July 1, 1999

Project End Date: June 30, 2002

Expected Total Amt. \$495,317 (Estimated)

[NSF Award Abstract](#)

[Samuel Armistead](#), Principal Investigator

[Department of Spanish](#)

[Bruce Rosenstock](#), Co-Principal Investigator

[Classics](#), [Religious Studies](#)

[Project Summary](#) (html)

[University of California Santa Barbara](#)

Project Web Site: [Alexandria Digital Earth Prototype](#)

Project Start Date: Sep 01, 1999

Project End Date: Aug 31, 2004

Expected Total Amt. \$5,400,000 (Estimated)

[NSF Award Abstract](#)

[Terence Smith](#), Principal Investigator

[Computer Science Department](#), [Geography Department](#) , [University of California Santa Barbara](#)

[Christine Borgman](#), Principal Investigator

[Department of Information Studies](#), [University of California at Los Angeles](#)

[Nick Faust](#), Principal Investigator

[Georgia Tech Research Institute](#), [Georgia Tech](#)

[Reagan Moore](#), Principal Investigator

[San Diego Supercomputer Center](#)

[Amit Sheth](#), Principal Investigator

[Department of Computer Science](#), [University of Georgia](#)

[Mike Goodchild](#), Co-Principal Investigator

[Geography Department](#)

[Anurag Acharya](#), [Divyakant Agrawal](#), Co-Principal Investigators

[Computer Science Department](#)

[James Frew](#), Co-Principal Investigator

[Donald Bren School of Environmental Science and Manangement](#)

[Bangalore Manjunath](#), Co-Principal Investigator

[Electrical and Computer Engineering Department](#)

[Richard Mayer](#), Co-Principal Investigator

[Psychology Department](#)

[Project Overview](#) (pdf)

[Project Proposal](#) (pdf)

Related links (html)

[Alexandria Digital Library](#)

[Carnegie Mellon University](#)

Project Web Site: [\*\*Informedia-II: Auto-Summarization and Visualization Over Multiple Video Documents and Libraries\*\*](#)

Project Start Date: May 1, 1999

Project End Date: April 30, 2003

Expected Total Amt. \$4,000,000 (Estimated)

[NSF Award Abstract](#)

[Howard D Wactlar](#), Principal Investigator

[Takeo Kanade](#), [Christos Faloutsos](#), [Alexander Hauptmann](#), [Michael Christel](#),

[John Lafferty](#), [Yiming Yang](#), Co-Principal Investigators

[School of Computer Science](#)

[Project Description](#) (pdf)

[Project Summary - slides](#) (pdf)

[Carnegie Mellon University](#)

Project Web Site: [\*\*Simplifying Interactive Layout and Video Editing and Reuse\*\*](#)

Project Start Date: Sep 01, 1999

Project End Date: Aug 31, 2002

[Brad Myers](#), Principal Investigator

[Albert Corbett](#), [Scott Stevens](#), Co-Principal Investigators

[Human Computer Interaction Institute](#), [School of Computer Science](#)

[Project Summary](#) (html)

[Related links](#) (html)

[Informedia-II: Auto-Summarization and Visualization Over Multiple Video Documents and Libraries](#)

[Columbia University](#)

Project Web Site: [\*\*A Patient Care Digital Library: Personalized Search and Summarization over Multimedia Information\*\*](#)

Project Start Date: Sep 01, 1999

Project End Date: Aug 31, 2004

Expected Total Amt. \$5,002,375 (Estimated)

[NSF Award Abstract](#)

[Kathy McKeown](#), Principal Investigator

[Computer Science Department](#)

[Shih-Fu Chang](#), Co-Principal Investigator

[Department of Electrical Engineering](#)

[James J. Cimino](#), [George Hripcsak](#), Co-Principal Investigators

[Department of Medical Informatics](#)

[Judith L. Klavans](#), Co-Principal Investigator

[Center for Research on Information Access](#)

[Project Overview](#) (html)

[Related links](#) (html)

[Natural Language Processing Group](#)

[Medical Informatics](#)

[Computer Graphics & User Interfaces Lab](#)

[On-Line Demos of Image and Video Search Systems](#)

## [Cornell University](#)

Project Web Site: [\*\*Project Prism at Cornell University: Information Integrity in Digital Libraries\*\*](#)

Project Start Date: May 01, 1999

Project End Date: Apr 30, 2003

Expected Total Amt. \$2,268,608 (Estimated)

[NSF Award Abstract](#)

[Carl Lagoze](#), Principal Investigator

[Kenneth P. Birman](#), [Fred B. Schneider](#), Co-Principal Investigators

[Computer Science Department](#)

[Anne Kenney](#), [Sarah Thomas](#), Co-Principal Investigators

[Cornell University Library](#)

[Project Summary](#) (html)

## [Harvard University](#)

Project Web Site: [\*\*An Operational Social Science Digital Data Library\*\*](#)

Project Start Date: July 1, 1999

Project End Date: June 30, 2002

Expected Total Amt. \$1,800,000 (Estimated)

[NSF Award Abstract](#)

[Gary King](#), Principal Investigator

[Department of Government](#)

[Sidney Verba](#), Principal Investigator

[Dale Flecker](#), [Nancy M. Cline](#), Co-Principal Investigators

[University Library](#)

[Micah Altman](#), Director and Co-Principal Investigator

[Department of Government](#), [University Library](#)

[Project Summary](#) (html)

[Project Summary](#) (pdf)

[Related links](#) (html)

[Library Digital Initiative](#)

[Harvard-MIT Data Center](#)

## [Indiana University Indianapolis/Bloomington](#)

Project Web Site: [\*\*A Distributed Information Filtering System for Digital Libraries\*\*](#)

Project Start Date: June 15, 1999

Project End Date: May 31, 2002

Expected Total Amt. \$315,387 (Estimated)

[NSF Award Abstract](#)

[Mathew J Palakal](#), Principal Investigator

[Rajeev R. Raje](#), [Snehasis Mukhopadhyay](#), Co-Principal Investigators

[Department of Computer and Information Science](#)

[Javed Mostafa](#), Co-Principal Investigator

[School of Library and Information Science](#)

[Project Summary](#) (pdf)

[Project Summary](#) (ps)

[Johns Hopkins University](#)

Project Web Site: [\*\*Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music, Phase Two\*\*](#)

Project Start Date: April 15, 1999

Project End Date: March 31, 2002

Expected Total Amt. \$529,951 (Estimated)

[NSF Award Abstract](#)

[Sayeed Choudhury](#), Principal Investigator

[Cynthia Requardt](#), Co-Principal Investigator

[Digital Knowledge Center](#)

[University of Kentucky](#)

Project Web Site: [\*\*The Digital Atheneum: New Techniques for Restoring, Searching, and Editing Humanities Collections\*\*](#)

Project Start Date: March 15, 1999

Project End Date: February 28, 2002

Expected Total Amt. \$499,924 (Estimated)

[NSF Awards Abstract](#)

[William Brent Seales](#), Principal Investigator

[James N Griffioen](#), Co-Principal Investigator

[Department of Computer Science](#)

[Kevin S Kiernan](#), Co-Principal Investigator

[Department of English](#)

[Project Summary](#) (pdf)

Related links (html)

[Electronic Beowulf](#): a study of the digitization, representation, archival and access of library manuscripts and artifacts.

[The Digital Library at the British Library](#)

[Michigan State University](#)

Project Web Site: [\*\*Founding a National Gallery of the Spoken Word\*\*](#)

Project Start Date: Sep 01, 1999

Project End Date: Aug 31, 2004

Expected Total Amt. \$3,599,989 (Estimated)

[NSF Award Abstract](#)

[Mark Kornbluh](#), Principal Investigator

[H-Net](#), [MATRIX](#), [History Department](#)

[Jack Deller](#), Co-Principal Investigator

[Department of Electrical and Computer Engineering](#)

[Joyce Grant](#), Co-Principal Investigator

[Department of Teacher Education](#), [College of Education](#)

[Michael Seadle](#), Co-Principal Investigator

[Michigan State University Libraries](#)

[Douglas Greenberg](#), Co-Principal Investigator

[Chicago Historical Society](#)

[John Hansen](#), Co-Principal Investigator

[University of Colorado](#)

[Jerry Goldman](#), Co-Principal Investigator

[Department of Political Science](#), [Northwestern University](#)

[Project Description](#) (html)

[Oregon Health Sciences University](#)

[Oregon Graduate Institute of Science and Technology](#)

Project Web Site: **[Tracking Footprints through an Information Space:  
Leveraging the Document Selections of Expert Problem Solvers](#)**

Project Start Date: January 1, 1999

Project End Date: December 31, 2001

Expected Total Amt. \$649,997 (Estimated)

[NSF Award Abstract](#)

[Paul Gorman](#), Principal Investigator

[Biomedical Information Communication Center](#), [Oregon Health Sciences  
University](#)

[David Maier](#), [Lois Delcambre](#), Co-Principal Investigators

[Department of Computer Science and Engineering](#), [Oregon Graduate  
Institute of Science and Technology](#)

[Project Description](#) (pdf)

[Project Summary](#) (html)

[University of Pennsylvania](#)

Project Web Site: **[Data Provenance](#)**

Project Start Date: June 1, 1999

Project End Date: May 31, 2002

Expected Total Amt. \$504,988 (Estimated)

[NSF Award Abstract](#)

[Peter Buneman](#), Principal Investigator

[Val Tannen](#), [Susan B. Davidson](#), [Chris Overton](#), Co-Principal Investigators

[Department of Computer and Information Science](#)

[Mark Liberman](#), Co-Principal Investigator

[Department of Linguistics](#)

[Project Summary](#) (html)

[Project Summary](#) (pdf)

[Project Summary](#) (ps)

Related links (html)

[The Data That Archiving Fails to Capture](#)

[University of South Carolina](#)

Project Web Site: [A Software and Data Library for Experiments, Simulations, and Archiving](#)

Project Start Date: April 1, 1999

Project End Date: March 31, 2003

Expected Total Amt. \$1,199,215 (Estimated)

[NSF Award Abstract](#)

[David Willer](#), Principal Investigator

[Department of Sociology](#)

[E. Elisabet Rutstrom](#), Co-Principal Investigator

[Department of Economics](#)

[Project Summary](#) (pdf)

[Stanford University](#)

Project Web Site: [Stanford Interlib Technologies](#)

Project Start Date: April 1, 1999

Project End Date: March 31, 2004

Expected Total Amt. \$4,297,585 (Estimated)

[NSF Award Abstract](#)

[Hector Garcia-Molina](#), Principal Investigator

[Terry Winograd](#), [Dan Boneh](#), Co-Principal Investigators

[Department of Computer Science](#)

[Stanford University](#)

Project Web Site: [Image Filtering for Secure Distribution of Medical Information](#)

Project Start Date: January 1, 1999

Project End Date: December 31, 2001

Expected Total Amt. \$519,594 (Estimated)

[NSF Award Abstract](#)

[Gio Wiederhold](#), Principal Investigator

[Department of Computer Science](#)

[Project Description](#) (pdf)

[University of Texas at Austin](#)

Project Web Site: [A Digital Library of Vertebrate Morphology, Using High-Resolution X-ray CT](#)

Project Start Date: June 1, 1999

Project End Date: May 31, 2002

Expected Total Amt. \$499,964 (Estimated)

[NSF Award Abstract](#)

[Timothy Rowe](#), Principal Investigator

[Department of Geological Sciences](#)

[Project Summary](#) (html)

[Tufts University](#)

Project Web Site: [A Digital Library for the Humanities](#)

Project Start Date: June 15, 1999

Project End Date: May 31, 2004

Expected Total Amt. \$2,758,400 (Estimated)

[NSF Awards Abstract](#)

[Gregory Crane](#), Principal Investigator

[Department of Classics](#)

[Robert Jacob](#), Co-Principal Investigator

[Electrical Engineering and Computer Science Department](#)

[Holly Taylor](#), Co-Principal Investigator

[Psychology Department](#)

[Ross Scaife](#), Co-Principal Investigator

[Kentucky Classics](#), [University of Kentucky](#)

[Nancy Allen](#), Co-Principal Investigator

[Museum of Fine Arts, Boston](#)

[Project Summary](#) (html)

[University of Washington](#)

Project Web Site: [Automatic Reference Librarians for the World Wide Web](#)

Project Start Date: January 1, 1999

Project End Date: December 31, 2001

Expected Total Amt. \$598,110 (Estimated)

[NSF Award Abstract](#)

[Oren Etzioni](#), Principal Investigator

[Dan Weld](#), Co-Principal Investigator

[Department of Computer Science](#),

[Project Description](#) (pdf)

[Related links](#) (html)

[Internet Softbot Research](#)

[Ahoy! The Homepage Finder](#)

[Grouper, A Document Clustering Interface for HuskySearch](#)

Undergraduate Emphasis

[University of California Berkeley](#)

Project Web Site: [\*\*Using the National Engineering Education Delivery System as the Foundation for Building a Test-Bed Digital Library for Science, Mathematics, Engineering and Technology Education\*\*](#)

Project Start Date: October 1, 1998

Project End Date: September 30, 1999

Expected Total Amt. \$399,999 (Estimated)

[NSF Award Abstract](#)

[Alice Agogino](#), Principal Investigator

[College of Engineering](#)

[Project Description \(pdf\)](#)

Related links (html)

[SMETE Information Portal](#) - A Digital Library for Science, Mathematics, Engineering and Technology Education

[NSF SMETE-Lib Study](#) - An initiative of the National Science Foundation's Division of Undergraduate Education to examine the potential impact of digital libraries on science, mathematics, engineering, and technology education (SMETE), with emphasis at the undergraduate level.

[Columbia University](#)

Project Web Site: [\*\*Columbia Earthscape: A Model for a Sustainable Online Educational Resource in Earth Sciences\*\*](#)

Project Start Date: December 1, 1999

Project End Date: November 30, 2002

Expected Total Amt. \$581,068 (Estimated)

[NSF Award Abstract](#)

[Kate Wittenberg](#), Principal Investigator

[Columbia University Press](#)

David S Millman, Co-Principal Investigator

[Academic Information Systems](#)

[Lewis E Gilbert](#), Co-Principal Investigator

[Project Summary \(html\)](#)

[Project Summary \(pdf\)](#)

[Georgia State University](#)

Project Web Site: [\*\*Research on a Digital Library for Graphics and Visualization Education\*\*](#)

Project Start Date: October 1, 1999

Project End Date: September 30, 2002

Expected Total Amt. \$330,278 (Estimated)

[NSF Award Abstract](#)

[G. Scott Owen](#), Principal Investigator

[Mathematics and Computer Science Department](#), [Hypermedia and Visualization Laboratory](#)

[Yanqing Zhang](#), Co-Principal Investigator

[Rajshekhar Sunderraman](#), Co-Principal Investigator

[Department of Computer Science](#)

[Project Description](#) (html)

[Project Summary](#) (html)

[Related links](#) (html)

[Hypergraph](#)

[Eckerd College](#)

Project Web Site: **Digital Analysis of Whale Images on a Network (DARWIN)**

Project Start Date: May 1, 2000

Project End Date: March 15, 2002

Expected Total Amt. \$32,870 (Estimated)

[NSF Award Abstract](#)

[Kelly R Debure](#), Principal Investigator

[Computer Science Department](#)

[Project Summary](#)

[Related links](#) (html)

[University of Maryland](#)

Project Web Site: **Digital Libraries for Children: Computational Tools that Support Children as Researchers**

Project Start Date: January 1, 2000

Project End Date: December 31, 2002

Expected Total Amt. \$613,437 (Estimated)

[NSF Award Abstract](#)

[Allison Druin](#), Principal Investigator

[Institute for Advanced Computer Studies \(UMIACS\)](#), [Department of Human Development](#)

[Project Summary](#) (html)

[Related links](#) (html)

[UMD Human-Computer Interaction Lab](#)

[Our approach to partnering with children to develop new technologies](#)

[Old Dominion University](#)

Project Web Site: **Planning Grant for the Use of Digital Libraries in Undergraduate Learning in Science**

Project Start Date: October 1, 1998

Project End Date: September 30, 1999

[Kurt Maly](#), Principal Investigator

[Mohammed Zubair](#), [Stewart Shen](#), [Steven Zeil](#), Co-Principal Investigators

[Department of Computer Science](#)

[Project Description](#) (pdf)

[University of Texas at Austin](#)

Project Web Site: **Virtual Skeletons in Three Dimensions: The Digital Library as a Platform for Studying Anatomical Form and Function**

Project Start Date: October 1, 1998

Project End Date: September 30, 2000

Expected Total Amt. \$287,147 (Estimated)

[NSF Award Abstract](#)

[John Kappelman](#)

[Department of Anthropology](#)

[Project Description](#) (pdf)

[Project Description](#) (html)

[Related links](#) (html)

[Technology for Education 2000](#)

[Virtual Examinations in Physical Anthropology](#)

[High-resolution X-ray CT \(Computed Tomography\) facility](#)

[DLI2 Home](#)

comments to [dli2 coordinators](#)

4.12.2000



## AVAILABLE RESEARCH

### [University of California at Berkeley](#)

Environmental Planning and  
Geographic Information Systems

### [University of California at Santa Barbara](#)

The Alexandria Project:  
Spatially-referenced Map Information

### [Carnegie Mellon University](#)

Infomedia Digital Video Library

### [University of Illinois at Urbana-Champaign](#)

Federating Repositories of Scientific  
Literature

### [University of Michigan](#)

Intelligent Agents for Information  
Location

### [Stanford University](#)

Interoperation Mechanisms Among  
Heterogeneous Services

### [DLI Project \[Contacts\]\(#\)](#)

### [DLI Workshop Series](#)

### [DLI Publications](#)

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

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

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

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

### [NSF Digital Libraries Contact](#)

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

**Foreign Language Versions of this page available:**

**[[Chinese](#)]-[[French](#)]-[[German](#)]-[[Italian](#)]-[[Japanese](#)]-[[Korean](#)]-[[Russian](#)]-[[Spanish](#)]**

comments to [DLI coordinators](#)

4.29.1999

# DLI - Carnegie Mellon:

---

- [Home page - Infromedia](#)
- [IEEE Computer article](#)
- [NetBill](#)

---

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

---

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

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



The Informedia Digital Video Library project is a research initiative at Carnegie Mellon University funded by the NSF, DARPA, NASA and others that studies how multimedia digital libraries can be established and used. The Informedia project has pioneered new approaches for automated video and audio indexing, navigation, visualization, search and retrieval and embedded them in a system for use in education, information and entertainment environments. Intelligent, automatic mechanisms are being developed to populate the library. Research in the areas of speech recognition, image understanding, and natural language processing supports the automatic preparation of diverse media for full-content and knowledge based search and retrieval.

**Informedia-I** - Informedia-I was one of the original NSF-funded Digital Library Initiative (DLI) projects, uniquely combining speech recognition, image understanding and natural language processing technology to automatically transcribe, segment and index linear video.

**Informedia-II** - The Informedia-II Project continues the pursuit of search and discovery in the video medium. This phase will transform the paradigm for accessing digital video libraries through meaningful, manipulable overviews of video document sets, multimodal queries, and adaptive summarizations of very large amounts of video from heterogeneous distributed sources. Video information collages are the key technology in Informedia-II and will be built by advancing information visualization research to effectively deal with multiple video documents.

**Experience-on-Demand** - Informedia Experience-on-Demand (EoD) is a DARPA-sponsored effort, developing tools, techniques, and systems that allow users to capture complete records of personal experience and to share them in collaborative settings.

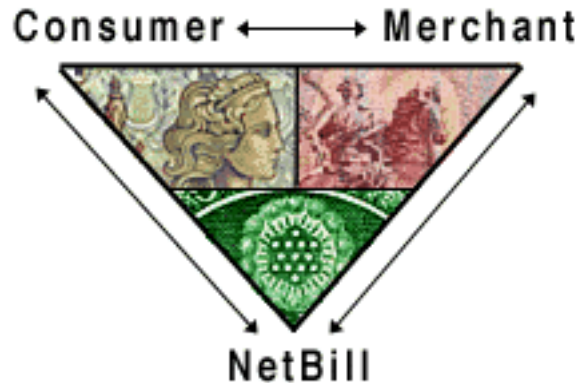




# The NetBill Project

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

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



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

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

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

<http://www.netbill.com>

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

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

If you require further information, please contact us at [support@netbill.com](mailto:support@netbill.com)



All contents copyright © 1995,1996,1997 Carnegie Mellon University.

All rights reserved.

Last revision: Fri Oct 10 11:54:34 EDT 1997

# DLI - Stanford:

---

- [Home Page](#)
- [IEEE Computer article](#)
- [testbed development](#)
- [info finding](#)
- [user interfaces](#)
- [DLITE \(task env\)](#)
- [SDLIP](#) (Simple DL Interop. Protocol) - also see [D-Lib Magazine article](#)
- [mediation infrastructure](#)

---

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

---

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

**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**



# STANFORD DIGITAL LIBRARY TECHNOLOGIES

<a href="#">PROJECTS</a>	<a href="#">DOCUMENTS</a>	<a href="#">PEOPLE</a>
<a href="#">SEMINARS</a>	<a href="#">TESTBED</a>	<a href="#">RESOURCES</a>

## [HOME](#)

## [PROJECTS](#)

[Resource Discovery](#)  
[Retrieving Information](#)  
[Interpreting Information](#)  
[Managing Information](#)  
[Sharing Information](#)

## [DOCUMENTS](#)

[Publications/Working Papers](#)  
[Dissertations](#)  
[Presentations](#)  
[Project Reports](#)

## [PEOPLE](#)

[Stanford DataBase Group](#)  
[Project on People, Computers, and Design](#)  
[Theory Group](#)  
[Stanford Libraries](#)

## [SEMINARS](#)

## [TESTBED](#)

[SDLIP](#)  
[InterBib](#)  
[PalmPilot Infrastructure](#)

## [RESOURCES](#)

[External Resources](#)  
[Seminars](#)

## [SPONSORS/PARTNERS](#)

[Government](#)  
[University Partners](#)  
[Corporate Affiliates](#)

The Stanford Digital Library Technologies Project was initiated in July as part of the Federally funded Digital Library Initiative Phase 2. The goal of this Project is to design and implement the infrastructure and services needed for collaboratively creating, disseminating, sharing and managing information in a digital library context.

The Stanford Digital Library Technologies Project is one participant in the [DLI2](#), Digital Library Initiative Phase II, started in 1999 and supported by the

National Science Foundation [NSF Digital Libraries Initiative](#)

Defense Advanced Research Projects Agency [DARPA Information Technology Office](#)

National Library of Medicine [NLM Extramural Programs](#)

Library of Congress [LOC Digital Library Initiatives](#)

National Endowment for the Humanities [NEH Digital Library Initiative](#)

National Aeronautics and Space Administration [NASA](#)

Federal Bureau of Investigation [FBI](#)

The Stanford Digital Library Technologies Project was funded from three coordinated proposals, from The University of California at Berkeley [UCB](#), the University of California at Santa Barbara [UCSB](#), and Stanford University. One of our major goals is to demonstrate our technologies on the emerging California Digital Library, [CDL](#) and to implement and evaluate these technologies on a testbed system to be built with the help of the San Diego Supercomputer Center, [SDSC](#). All three projects together yield a synergistic and comprehensive digital libraries project.

The Stanford component of this effort will develop the base technologies that are required to overcome the most critical barriers to effective digital libraries. One of these barriers is the heterogeneity of information and services. Another impediment is the lack of powerful filtering mechanisms that let users find truly valuable information. The continuous access to information is restricted by the unavailability of library interfaces and tools that effectively operate on portable devices. A fourth barrier is the lack of a solid economic infrastructure that encourages providers to make information available, and give users privacy guarantees. See the [summary](#) for more information.

In November 1998, we spent some time to look back at our efforts of our DLI1 research. These ruminations led to a [publication](#) and a [presentation](#). Both are entitled: "Building the InfoBus. A Review of Technical Choices in the Stanford Digital Library". We talk about infrastructure decisions, about why USMARC in the end wasn't quite right for us, and about how deeply user traditions impacted the details of our technical designs.

Our collection in DLI1 was primarily computing literature. However, we also had a strong focus on networked information sources, meaning that the vast array of topics found on the World Wide Web are accessible through our project as well. At the heart of the DLI1 project is the [testbed](#) running [the "InfoBus" protocol](#), which provides a uniform way to access a variety of services and information sources through "proxies" acting as interpreters between the InfoBus protocol and the native protocol. The InfoBus is implemented on top of a [CORBA-based](#) architecture using [Inprise's Visibroker](#) and [Xerox's ILU](#).

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

Questions or Comments? Send email to  
[dlwebmaster@db.stanford.edu](mailto:dlwebmaster@db.stanford.edu)

[PROJECTS](#) [DOCUMENTS](#) [PEOPLE](#) [SEMINARS](#) [TESTBED](#) [RESOURCES](#) [SPONSORS/PARTNERS](#)



# STANFORD DIGITAL LIBRARY TECHNOLOGIES

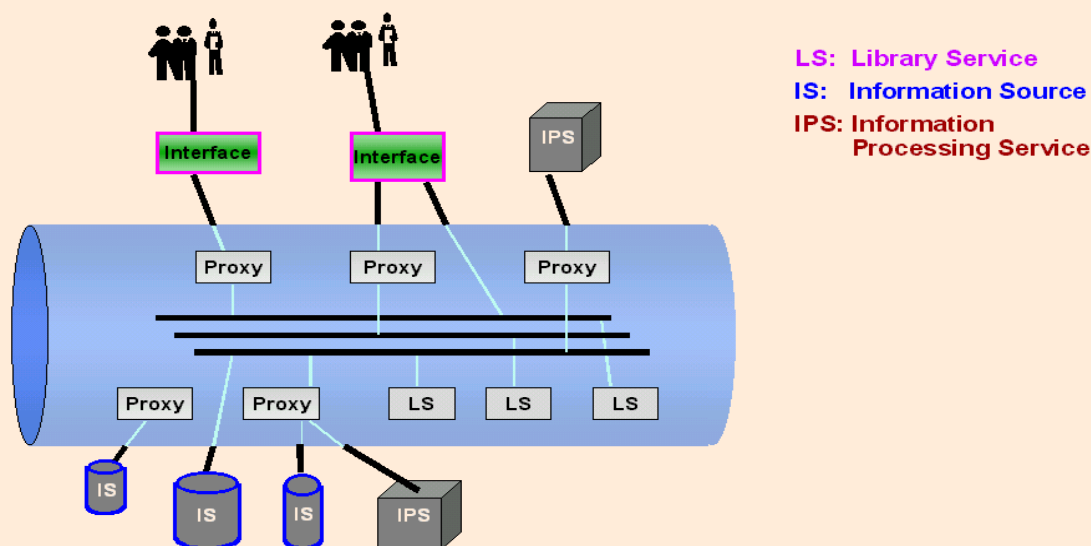
<a href="#">PROJECTS</a>	<a href="#">DOCUMENTS</a>	<a href="#">PEOPLE</a>
<a href="#">SEMINARS</a>	<a href="#">TESTBED</a>	<a href="#">RESOURCES</a>

## Testbed Highlights

[HOME](#)
[TESTBED](#)
[SDLIP](#)
[InterBib](#)
[PalmPilot Infrastructure](#)

## The Stanford Digital Library Testbed

The Stanford Digital Library testbed is our platform for experimentation with interoperation among online services. Our basic approach is to use **distributed objects** to allow integrated access to heterogenous services across networks. We call this system the InfoBus. The distributed approach allows the interaction of processes on different machines, with different architectures, implemented in different languages. We use **CORBA** to provide communication between remote processes. In particular, we use Xerox PARC's **ILU**, a free implementation of a CORBA superset, **MICO**, a free CORBA implementation under the Gnu license, and **Visigenic**, a commercial provider. We use Java, C++, and the interpreted, object-oriented language Python for our development work. Our computing platforms include Sun, PC-based architectures, and 3COM Palm Pilots.



For more information on the underlying technologies, see:  
CORBA

- Information from the [OMG](#), including a [Beginners' page](#)

ILU

- [Xerox PARC's ILU Home Page](#)

MICO

- [MICO's Home Page](#)

Visibroker

- [Visibroker Home Page](#)

## What Protocol does the Testbed Use?

We have developed the [Simple Digital Library Interoperation Protocol \(SDLIP\)](#) (pronounced S-D-Lip) for information access and retrieval. It supports both synchronous and asynchronous operation, providing robustness in the face of network or server outages. Moreover, it also gives the programmer a high degree of control over where and when information objects are materialized, affecting tradeoffs of space and cost vs. time. Protocol bindings are defined for both CORBA and HTTP. SDLIP is carefully designed so that it can be implemented even on very small footprint PDAs, but that it can scale up to serve interactions with complex information sources.

## Mobile Access to Digital Libraries



One portion of our testbed is devoted to making digital library resources available everywhere a user travels. We are developing proxies that prepare information for transmission over low bandwidths to portable digital assistants (PDAs) with very small screen real-estate. A part of this effort includes support for secure transactions between PDAs and online services.

We have developed a [software library](#) that supports our work on the 3COM Palm Pilot. It includes facilities for memory management, event handling, TCP/IP communication, and XML parsing. We are also working on DietORB, a scaled-down CORBA ORB for the Pilot.

## Publicly Available Software Services

- [InterBib](#). A bibliography tool for converting bibliographies among various formats. The tool also processes RTF and Framemaker files, including bibliographies when given a BibTeX bibliography source. This extends LaTeX's BibTeX capability to MS-Word and Framemaker documents.

## Various Operating Instructions

- [DL PowerPoint Presentation Template](#) (position mouse over this link, right click the mouse button, and choose "Save Link As...", and save in "Program Files/Microsoft Office/Templates")
- [Emacs Support for Entering BibTex Records](#)
- [Visigenic .cshrc setup](#)
- [How to use CVS on our SUN and PC machines](#)
- [How to keep services running on the InfoBus](#)
- [How to call C functions from Java](#)
- [Palm Pilot infrastructure](#)
- [SDLIP protocol](#)
- [Examples of how to use Visigenic on the InfoBus](#)

Questions or Comments? Send email to  
[dlwebmaster@db.stanford.edu](mailto:dlwebmaster@db.stanford.edu)

[PROJECTS](#) [DOCUMENTS](#) [PEOPLE](#) [SEMINARS](#) [TESTBED](#) [RESOURCES](#) [SPONSORS/PARTNERS](#)



# Information Finding Projects in the Stanford Digital Library

---

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

---

## FAB

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

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

- [Marko Balabanovic](#)
- 

## GLOSS

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

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

- [Luis Gravano](#)
- 

## [Query Translator](#)

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

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

## [SenseMaker](#)

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

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

- [Michelle Q Wang Baldonado](#)

## Grassroots

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

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

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

## The Stanford Digital Library Metadata Architecture

Services need to provide

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

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

[Metadata for Digital Libraries: Architecture and Design Rationale](#)

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

## STARTS: Stanford Protocol Proposal for Internet Retrieval and Search

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

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

## **BackRub**

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

- **Larry Page**

## [ComMentor](#)

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

["Content Ratings and Other Third-Party Value-Added Information: Defining an Enabling Platform"](#)

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

## [InterOp Protocol](#)

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

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

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

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

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

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

## [InterBib](#)

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

[Information on InterBib](#)

- [Andreas Paepcke](#)
- 

[Stanford]

[DigLib]

[Write  
Webmaster]

---



# User Interface Projects in the Stanford Digital Library

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

---

## [SenseMaker](#)

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

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

- [Michelle Q Wang Baldonado](#)
- 

## [DLITE: A Digital Library Interface](#)

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

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

- [Steve Cousins](#)
- 

## [Grassroots](#)

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

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

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

## [ComMentor](#)

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

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

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

## Audio Interfaces to HyperText

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

[AHA: Audio HTML Access](#)

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

## WebWriter

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

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

- [Arturo Crespo](#)
- 

## [RManage/FIRM](#)

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

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

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

[\[Stanford\]](#) [\[DigLib\]](#)

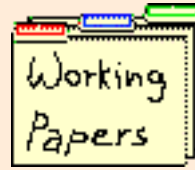
---

[dlwebmaster@db.stanford.edu](mailto:dlwebmaster@db.stanford.edu)



# Stanford Digital Library

## Technologies



### SIDL-WP-1996-0049

#### The Digital Library Integrated Task Environment (DLITE)

Steve B. Cousins, Andreas Paepcke, Terry Winograd, Eric A. Bier, Ken Pier

[cousins@cs.stanford.edu](mailto:cousins@cs.stanford.edu)

**Abstract:** We describe a case study in the design of a user interface to a digital library. Our design stems from a vision of a library as a channel to the vast array of digital information and document services that are becoming available. Based on published studies of library use and on scenarios, we developed a metaphor called workcenters, which are customized for users' tasks. Due to our scenarios and to prior work in the CHI community, we chose a direct-manipulation realization of the metaphor. Our system, called DLITE, is designed to make it easy for users to interact with many different services while focusing on a task. Users have reacted favorably to the interface design in pilot testing, but a problem surfaced: we need a mechanism to teach new users about the metaphor and interface. We conclude by describing our approaches to this problem.

---

**Note:** Papers in this series are in development and are not in a final form for publication or general dissemination. They are subject to change. Please do not quote or further distribute them without explicit permission from the authors.

---

This paper was created on: 9/20/96 and last revised on: 1/14/1997

**Author's Comments:** Submitted to DL'97

**Status:** PUBLIC

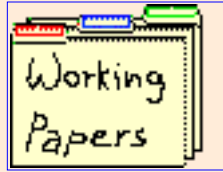
[Click here to see the full text of SIDL-WP-1996-0049](#) (PS)

[Click here for the full text of SIDL-WP-1996-0049](#) (PDF)

## Revision History

Version	Format	Date	Comments
<a href="#">4</a>	PS	1/12/1997	Updated related work section. Pre-DL'97 draft.

<a href="#">3</a>	PS	1/9/1997	Draft to be submitted to DL'97.
<a href="#">2</a>	PS	9/30/1996	Added a figure which was left off in the previously-submitted version.
<a href="#">1</a>	PS	9/27/1996	Submitted to CHI'97



[dlwebmaster@db.stanford.edu](mailto:dlwebmaster@db.stanford.edu)

# The Digital Library Integrated Task Environment (DLITE)

Steve B. Cousins, Andreas Paepcke, Terry Winograd

Stanford University  
Computer Science Department  
Stanford, CA 94305 USA  
+1 415 723 7784

{cousins,paepcke,winograd}@cs.stanford.edu

Eric A. Bier, Ken Pier

Xerox PARC  
3333 Coyote Hill Rd.  
Palo Alto, CA 94304 USA  
+1 415 812 4000  
{bier,pier}@parc.xerox.com

## ABSTRACT

We describe a case study in the design of a user interface to a digital library. Our design stems from a vision of a library as a channel to the vast array of digital information and document services that are becoming available. Based on published studies of library use and on scenarios, we developed a metaphor called workcenters, which are customized for users' tasks. Due to our scenarios and to prior work in the CHI community, we chose a direct-manipulation realization of the metaphor. Our system, called DLITE, is designed to make it easy for users to interact with many different services while focusing on a task. Users have reacted favorably to the interface design in pilot testing. We conclude by describing our approaches to this problem.

**Keywords:** Digital library, user interface, direct-manipulation, world-wide web, holophrasting

## 1. INTRODUCTION

The Stanford Digital Library project is focused on creating technology that will allow a user to access digital resources, from static document citations to dynamic information exploration and management services, without having to know the details of the format of each document and the mechanics of each service. This technology, loosely called the "InfoBus" [Paep96a], provides a unifying framework that can bring together services now provided on the world-wide web, as well as traditional information retrieval and document services, and new kinds of information services that are being developed.

The InfoBus technology makes access to the resources possible, and must be accompanied by an interface that provides a consistent model for users to deal with the plethora of offerings. Now that the web has brought us consumer-level distributed systems, we need powerful metaphors and systems to help us handle the new capabilities. DLITE is the result of a task-oriented, user-centered design process. It is a system that a person can use to interact with many services in pursuit of a complex goal.

A spectrum of simplicity versus power is being played out with the development of the world-wide web. The web began as hypertext: the ontology consisted of pages and links, and the user interface was very simple. Pages contained formatted text and images. Programs could be invoked, but their parameters were limited to those that could be encoded in a link. User interface designers could

only give users access to program parameters by setting up a choice between links.

HTML forms changed the web from a tool for browsing to a tool for building distributed user interfaces of the kind found in simple interactive systems. User interface programmers could use a variety of widgets to let users specify how programs would be invoked. The user's conceptual model was more complex, but the ontology still consisted of pages and links, where the notion of a page now incorporated fill-out forms.

The current generation of web browsers gives user interface designers even more control over what users can see and do. Using Java applets, designers can implement the direct-manipulation interfaces that the CHI community has been talking about and using for years. These interfaces allow users to point to objects without having to name them explicitly.

Applet-based systems like Java provide the means to build direct-manipulation interfaces that have the same level of user-interface technology as single-user systems. However, the design of previous GUI systems has been shaped by an environment (the PC or workstation) which has different characteristics from the heterogeneous and distributed environment of the net. Although there is much experience building direct-manipulation interfaces, building direct-manipulation *distributed* interfaces raises new challenges. For example, in our domain we can draw an icon representing a remote service, but in addition to whatever "normal" manipulations the user can perform, we also need to show the state of the network and of the remote service. In this work, we explore how direct-manipulation can be used to build a user interface for a distributed information system.

DLITE is a digital library interface that

- gives the user control in a task context,
- provides smooth integration of services,
- supports services that run at very different speeds, and
- supports sharing, re-use, and persistence.

We are developing the DLITE user interface assuming that eventually we will be able to widely deploy a direct-manipulation, drag-and-drop style interface. There are two versions of the interface: one is implemented using Python and Tk on X Windows, and one in Java's simple windowing toolkit AWT. Both implementations use the Netscape Navi-

gator web browser as an auxiliary input and display mechanism. We use CORBA distributed objects [OMG93] as our distribution framework, as implemented by Xerox PARC's Inter-Language Unification (ILU) system [Cutt93].

In the next section, we describe the workcenter metaphor. Then we briefly discuss the objects that can exist in a workcenter, as background for an extended example of the system in use. We then summarize the interaction modalities of the system, and discuss user reactions from a pilot study. Finally, we discuss some of the issues that surfaced and how DLITE relates to other systems.

## 2. WORKCENTERS

We introduce the notion of a workcenter, that is inspired by the work on Rooms [Hend86]. A workcenter is a place where the tools for a task are ready-to-hand. A kitchen is a good real-world example of a workcenter: the tools for cooking are handy. You could do woodworking in a kitchen, but a kitchen is not set up for woodworking.

Workcenters suggest activities that can be done in them, just as a kitchen is a subtle reminder of what you might cook for breakfast. In DLITE, the tools are called "components." For example, a DLITE workcenter might contain a component for automated summarization of documents. If such a component were present, it would be because the designer of the workcenter put it there for a reason, or because a previous user of the workcenter had found it useful.

We expect that DLITE workcenters will be created by people with expertise in information exploration and management, such as librarians and editors, and that these expert users will tailor the workcenters to the needs of their colleagues and patrons. For example, a workcenter at a walk-up terminal in a public library might be designed for anonymous users to access free or advertising-supported services. A similar search-oriented workcenter at a research firm might have a component for accessing Knight-Ridder's commercial Dialog information service in a prominent location. A reference librarian's workcenter could have components that are powerful and efficient, but that require training to use effectively. Each user would have access to any number of workcenters, each appropriate for a specific task.

The digital library should support publishing tasks as well as retrieval tasks. A company might have a workcenter for publishing papers, that could include components for routing documents for intellectual property approval and automatically adding approved works to the company's list of publications. As one experiment, we have been building a workcenter to help with the task of producing a high-quality digital document from a printed color document.

Workcenters and their contents persist over time, allowing users to come back and continue a task at a later date. This is consistent with our reading of the library use literature [Marc95, Nard96, Oday93]. Persistence also reinforces the notion that the workcenter is a place, and supports sharing. DLITE is designed so that multiple users can interact with the same workcenter from different physical locations at the

same time. This collaboration mechanism will be presented in a later paper.

## 3. COMPONENTS

All objects in a workcenter are components. DLITE components fall into five basic categories: documents, collections, queries, services, and representations of people.



Document components may be anything from simple citations to complex entities with hundreds of meta-information fields and multiple content representations. In our environment, common document types are results from searches of Dialog databases, library bibliographic holdings, or the world-wide web, and documents that users upload from their local disks for processing (such as Microsoft Word documents, text files, or scanned images).



Collections are containers for other components. Basic collections are used to build up sets of interesting information that can be displayed or processed by services. A common sub-type of collection is the result set, which is filled in by search services in the course of processing. Result sets are interesting because they can contain partial results. They begin empty, and a search will put some initial results in them. Users can ask result sets for more results, at which point the underlying InfoBus protocols do the necessary work to reconnect and continue the search.



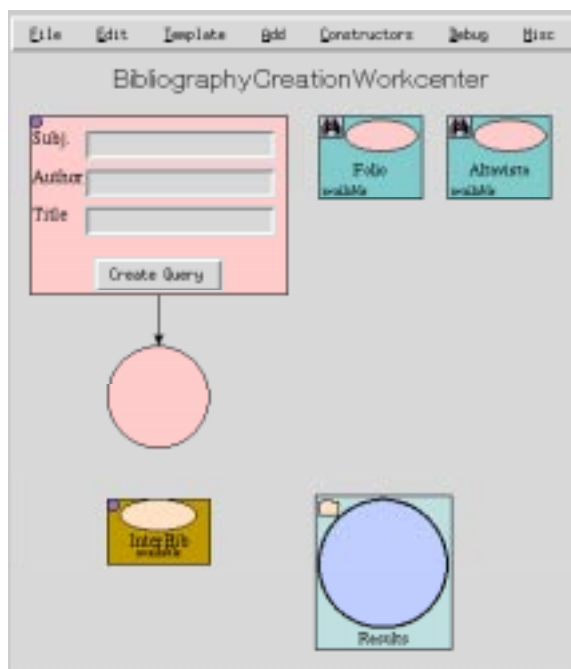
Queries are expressions of a user's information need, and range from simple lists of keywords to complex boolean expressions. The InfoBus has a query translation system built in, so that the same query can be sent to many different search services [Chan96]. From the HCI perspective, the challenge is to help users construct queries that will make sense for the sources in the current workcenter. Queries represent encapsulated retrieval expertise. Since they are first-class objects in DLITE, they can easily be re-used or shared with other people.



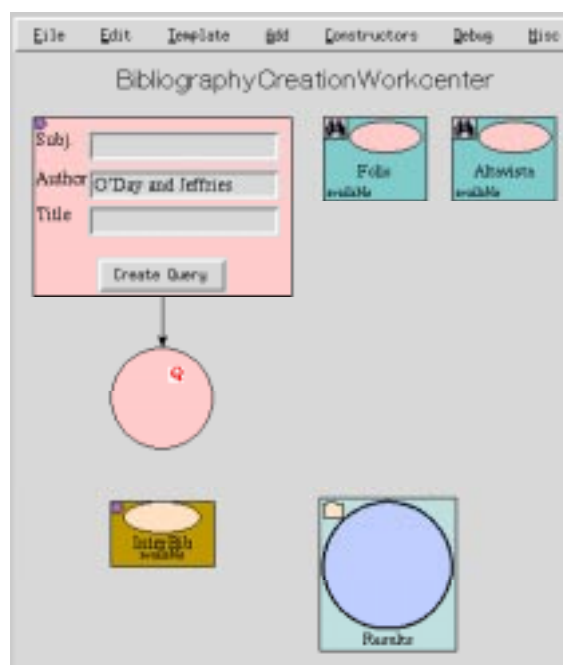
The InfoBus contains many different services. There are now hundreds of collections accessible via search services, and other types of services as well, including summarization, optical character recognition, query expansion, format translation and bibliography processing. These services are commonly implemented by building proxy objects that translate from the native protocol of the service to the InfoBus protocol that we use to integrate service components into the user interface. This architecture provides great flexibility in bringing services on-line, but because it allows transparent integration of independent services, it also provides potential points of failure that must be accounted for in the user interface.



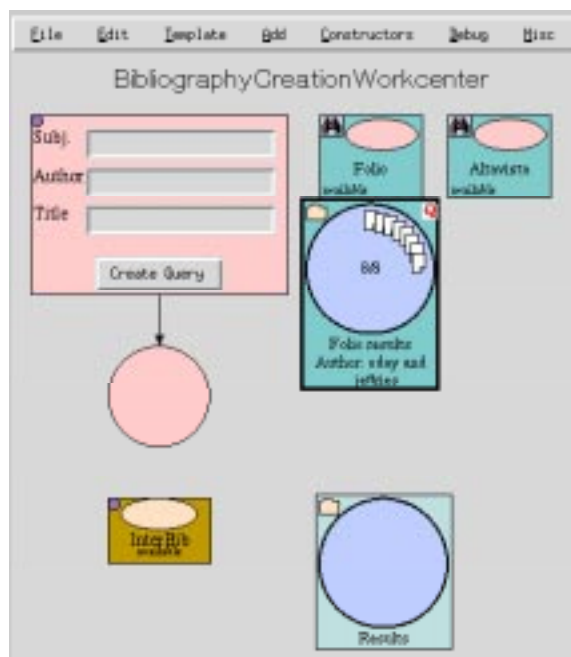
Finally, representations of people in the interface are used to allow the user to manage concerns such as access control, communication, payment, and authorization. For example, a simple icon of a person on the bottom of the workcenter indicates who is currently logged on to the workcenter, and therefore who will be responsible for payment if a query is sent to a fee-based service. The underlying mechanism will automatically send account information when the service is accessed [Cous95].



(a)



(b)

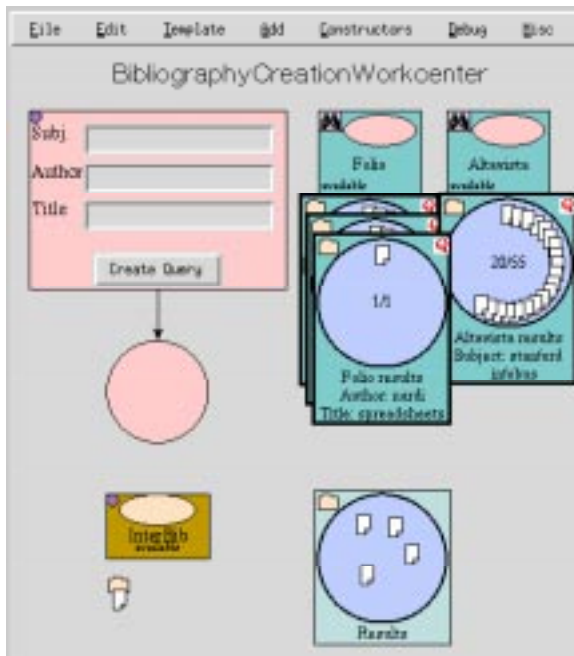


(c)

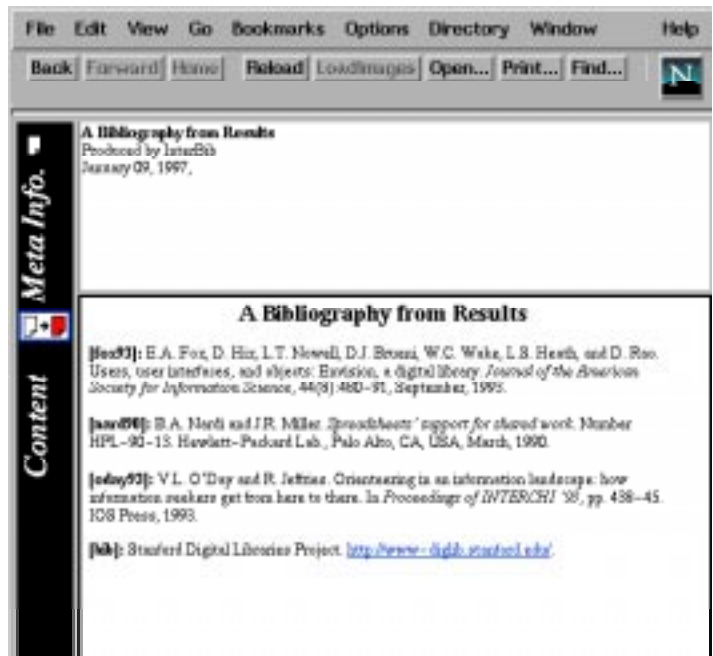


(d)

**FIGURE 1 (a-d).** A DLITE workcenter for bibliography processing. (a) The workcenter template before processing starts. (b) The user creates a query for the authors “O’Day and Jeffries”. (c) The user sends the query to the Folio search service by dropping it onto the service. The service responds by creating a result set object, attaching the query to it, and moving the result set away from the search service. As results come from the search service, they are attached to the result set (small document icons). (d) The user activates the collection to see a summary of it in a web browser. The summary contains buttons that can affect objects in the workcenter.



(e)



(f)

**FIGURE 1 (e-f). (e) The user does more searches, and collects relevant documents in a personal collection. The personal collection is then dropped into InterBib which creates a bibliography object. (f) The resulting bibliography is uniformly formatted in a form suitable for inclusion in a technical paper.**

The ontology for the digital library differs from that of a typical desktop with drag-and-drop. Documents differ from files in a number of ways. For example, documents are not necessarily associated with disk storage, and they may be citations to information objects that do not even exist in digital form.

Collections and services also differ from folders and application programs in similar ways. Collection components are visualizations of InfoBus collection objects, which may be stored at a remote computer and accessed over the network. Folders or file directories are usually (but not always) associated with disk storage. Service components may correspond to application programs, but may also be associated with objects running at remote sites. They also contain additional meta-information, in the form of defaults for how the service is to be invoked within this particular task, while applications programs in conventional interfaces are usually customized on a per-user basis at best.

#### 4. LEVERAGING THE WEB BROWSER

We have taken advantage of the fact that web browsers such as the Netscape Navigator can be controlled remotely by designing DLITE to work alongside the browser. In particular, we use the browser to display documents which may have been retrieved from search services, or were generated by other parts of the system. This view includes meta-information, controls that allow documents to be marked in the workcenter, and access to the document content. The marking mechanism is particularly useful because it allows us to keep the visual representation of documents small in the workcenter, saving valuable screen real estate. This integration of a standard Web browser also allows users to move

smoothly between searching, browsing, and document processing.

In the next section, we show how the components interact in a walk through of a typical use of one workcenter.

#### 5. EXAMPLE

A common practice when writing research papers is to delay formatting the bibliography until the very end of the process. Researchers often leave comments like "[[cite O'Day&Jeffries]]" in the text, intending to replace them in the final step with a marker like "[1]" and add the appropriate reference at the end of the paper. Tools to help with this task have existed for a long time, such as EndNote, BibTex, and Refer, but they have typically assumed that the researcher had access to a large database of citations in the right format. Researchers would either spend time maintaining and extending files of citations, or would manually format the bibliographies at the end of each paper-writing process.

Recently, InterBib, a new bibliography generating tool, was developed by Paepcke [Paep97]. InterBib exists as a free service on the web. To use it, you fill out an HTML form, upload your citation databases and your incomplete document, and wait for the resulting document to be returned. InterBib differs from its predecessors in that it can accept input in many different formats.

InterBib also has an InfoBus interface, and can accept as input InfoBus collections, such as result sets returned from search services. For example, a query to Dialog via the InfoBus proxy will return a result set that can be passed directly on to InterBib. InterBib can either format the entire result

set as a bibliography, or can use the result set as one of several collections to search while processing the citations in a paper.

InterBib, combined with the ever-growing set of collections accessible via the InfoBus, provides a complete tool for finishing research paper references painlessly. These pieces interact more smoothly in DLITE than in existing HTML interfaces. More generally, for tasks that require the use of multiple services, the direct-manipulation approach gives users more power than a forms-based approach would provide. The advantage of DLITE, in short, is that results of one operation can be directly passed to another operation, without the need to save them, copy them to a clipboard, reformat them, or otherwise do manual processing in between.

The workcenter in Figure 1a is designed for the task of completing bibliographies. The component labelled 'Create Query' is a simple query creation tool. Its field labels correspond to the typical search attributes for this kind of task. The components with oval-shaped input places represent two search services: Folio, a Stanford University Library resource serving the INSPEC bibliographic database, and AltaVista, a general-purpose World-Wide Web search service. The component with the round circle is a user-maintained collection of documents. The last component represents the InterBib service.

To complete the bibliographic task, the user begins by filling in one or more of the fields in the query constructor. For example the authors "O'Day and Jeffries" might be used as a query. Figure 1b shows how clicking the button in the constructor creates a query component labelled "Q". This query component may be dragged across the screen. In our collaborative workspaces it can also be shared with other users.

Figure 1c shows what happens when the user drags the query component onto the oval input place of the Folio service. A result set is created by the system, and the query component is attached to the upper corner of that result set. The entire result set is animated away from its source. As this animation happens, the query is sent to the search service proxy object which uses the InfoBus query translation system to translate the query and forward it in native form to the Folio search service.

As result messages are returned from the search service, they cause information on the display to be updated. Initially, the message "0/8" appears in the result set to let the user know that there are 8 results available. Soon the results begin to appear, and the message changes accordingly. DLITE also adds document icons to the result set as the results arrive. With very fast services, all of the results appear almost instantaneously, while slower services add results one at a time. By default, DLITE initially only requests the first 20 documents to be returned from searches.

Dragging the query from the result set to the AltaVista component's input place begins a similar activity on the AltaVista service. Both searches may occur in parallel.

Next, the user clicks on the result collection to have the results summarized in the Web browser (Figure 1d). In the browser window, the user can mark each relevant document, turning the related document icon in the workcenter dark.

The user copies and drags each selected document into the collection of relevant articles (marked 'Results' in Figure 1). When all citations have been found and placed into the collection, that collection is dropped onto InterBib's input place, and after a bit of processing, the collection and the resulting bibliography move out of InterBib (Figure 1e). Finally, the user views the resulting bibliography in the web browser by clicking on it (Figure 1f), and adds the formatted bibliography to the document. (InterBib can add the bibliography directly into the document, but we have simplified the interaction for the purposes of this example).

If the query did not yield proper results, the user can edit or replace the information in the query constructor to create a new query. For services that accept entire documents as queries for relevance feedback, the user could drop relevant document icons onto the service component input place, although we have not yet implemented any proxies to such services.

Imagine what would be required to accomplish this task using existing tools. Even if there were a web page for searching Folio, and even if it gave its results in a machine-readable format, the user would still have to somehow combine the results from Folio and AltaVista into a single file, and then go to a third web page and upload that file for processing. By contrast, this task is completed in DLITE without requiring the user to change focus: all tools are ready-to-hand.

## 6. INTERACTION MODALITIES

DLITE affords three basic operations, "point," "activate," and "drag-and-drop." It uses animation to show the user what is happening, and leverages a web browser as a handy document viewer that we assume users can already use. We will describe each of these concepts, and then look at the "result set" component in more detail to make the discussion concrete.

### 6.1. Pointing

The most basic affordance in DLITE is pointing. When the user moves the mouse over a DLITE object icon, a small yellow box appears with a short description of the object. This is done to reduce the screen real estate required to display objects, and allows us to display up to about 100 objects in a very small area of the screen. For more information about the object, the user can use the activate affordance.

### 6.2. Activate

In our current prototype, activate is invoked by "double-clicking" or by clicking the right mouse button, and is more general than double-clicking on an icon in a standard desktop interface. The component that receives the activate message can do an arbitrary action, and the component gets information about which part of itself was activated, and can take different actions for different parts.

Components in DLITE contain one or more visible graphical elements. Most components have an icon in their upper-left that represents the component as a whole. The default response to an activate message on that icon is to hide the rest of the component so that only the icon is visible. The default response to an activate message on other parts of the component is to display information about the component in the web browser. For example, activating the body of a collection will display the summary of the collection in the web browser, while activating the folder icon associated with the collection will shrink the representation of the collection on the screen.

### 6.3. Drag-and-drop

Drag-and-drop dates back at least to the Xerox Star GUI [John89], but most desktop file system interfaces do not exploit the power of the technique. A document may be dropped onto an application program, but the mechanism ends there: once the application program starts, the icons are covered by application windows and not seen again. Some programs process documents directly, so that drag-and-drop is the only user action necessary. For example, some utilities for uncompressing documents work this way. Unfortunately, users are often left guessing where the results went, and have to resort to other clues, such as menu or application bars to see if the program has even finished yet.

We use animation to indicate that the service has begun processing, is finished processing, or has raised an exception. We allow components to be attached to one another. When a subcomponent is dragged, the semantics may be to remove it, copy it, or drag a link. This is a generalization of the situation (with all of the attendant problems) in desktop file systems -- in those systems the only attachment possible involves objects within folders.

Users can use the icon in the upper-left of components to iconify components in the interest of screen real estate. But they can also drag the icon off of the component to pass the component to a service or to make a copy of the component. This affordance is both powerful and potentially very confusing to users, and we are watching reactions to it closely to find ways to provide the functionality in a non-confusing way.

One problem with conventional uses of drag-and-drop is that it requires recipient components to have single inputs. In our context, the services represented by our components often have more than one input parameter. Because we allow intermediate objects (such as queries) to be first-class objects, we can use an iterative technique to solve this problem, by creating objects that have one or more parameters fixed, and can then be provided as parameters to services.

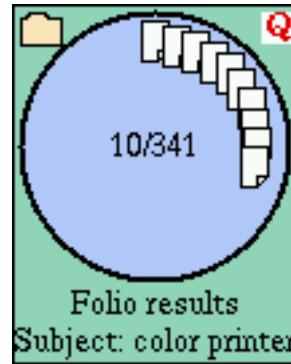
### 6.4. Animation

Animation is used extensively in DLITE to convey to the user that an action is in progress. When drag-and-drop is used, the object receiving the drop normally moves the dropped object into its geographical center during processing, and then moves it to a standard output location when processing is finished. When an input is dropped where it cannot be handled, it “waggles” by moving back-and-forth

quickly, then animates back to the place from which it was dragged.

We have avoided using pop-up dialog boxes since our focus is on conveying as much information as we can in the direct-manipulation style. We are adding more text to the interface on an as-needed basis, as determined by experience with users.

### 6.5. Example: Result Set Component



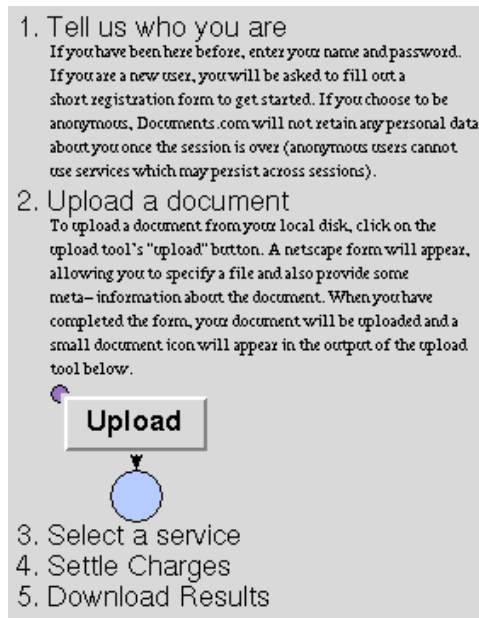
A result set is created when a query is dropped onto a search service. This operation provides a good example to illustrate the interaction techniques we have described. A result set contains three types of icons: its folder icon, the query that was used to create it, and a number of documents, which represent the results. When a user points at the icons, descriptions of

the objects are displayed in yellow boxes. For example, pointing at the query shows the query text and pointing at a result item shows its title.

Five different activations can be invoked on a result set, depending on where the click occurs. If the user activates the folder icon, the collection is iconified so that only the folder icon is visible. If the query is activated, a view of the query is displayed in the web browser. We intend for this view to support query refinement, but have not implemented this feature yet. If a document icon is activated, the document is viewed in the web browser. If the background of the collection is activated, the entire collection is summarized in the browser. Finally, if the numbers in the center of the result set are activated, the result set requests more documents from the search service. This last activation is a hidden affordance, and is particularly problematic because users often invoke it when they intend to summarize the collection in the browser because of its central location. This particular feature will be changed in the next version of the interface.

All of the icons attached to a result set may be “copied off.” If a user drags a document out of the collection a copy will remain. We tried using “move-off” for the documents, since that is the way that ad hoc collections need to behave, but found that doing so destroyed the integrity of the result collection: once documents were removed from the result set, it no longer represented the results of the query. For the same reason, dropping components onto result sets is not allowed. Using copy-off for the query has proven to be quite helpful. It facilitates query reuse and refinement, when combined with the behavior that queries dropped onto query constructors cause the fields of the constructor to be filled from the query.

Animation is used mainly during result set creation. The current implementation creates a result set on top of the



**FIGURE 2. A holophrastic interface with items 1 and 2 open.**

search service, animates the query to the upper-right of the result set, attaches it, and then animates the entire set away from the service while the search is initiated. As results come in, they are rendered in the result set. The resulting animation, which can only be imagined from a text description, gives the strong impression that the query was accepted and that processing has begun.

## 7. WALK-UP USER TRAINING APPROACHES

We have explored two techniques for helping users find their way into the DLITE interface: “holophrasting” and “web first.”

Holophrasting means the use of a word or phrase to stand for a long text. It was used by the hypertext community in the 1980s to refer to the concept of expanding a link in place (e.g. [Bier92]), but its use in user interfaces dates back to the early 1970’s [Hans71]. We have implemented holophrastic components that progressively disclose instructions and other components. The model we have in mind is that a user begins with a short list of instructions summarizing the steps of a task. For each step, the step name is activated to reveal the details of that part of the task. When the task step is finished, it can be collapsed and the next step expanded. Figure 2 shows a holophrasting component for the task of accessing a for-fee document processing service. As users step through these simple instructions, they become familiar with the environment in which the instructions are embedded, and will soon be able to manipulate the environment without the need of such specific instructions.

Since our system is integrated with a web browser, another approach to walk-up users is to assume that they see web pages first. This approach effectively puts the DLITE environment in a conceptual box: the thing that is accessed by clicking on a specific link. It also allows us to ensure that

documentation comes first, since it can be accessible from the web instead of being yet another affordance in the interface. Of course, this approach can be combined with holophrasting if both are required or desirable.

## 8. PILOT STUDY

The DLITE prototype has been under development for two years, undergoing continuous evolution on the basis of feedback from users and observers, both within the project and from outside. An initial pilot usability study was conducted with six novice users.

Each subject was given a copy of a short research paper in which four citations were in the raw, informal shape described in our example above. The subjects were asked to find bibliographic references relevant to the citations, to replace the informal citations in the document with the automatically generated keys for the references they found, and to cause a bibliography to be attached to the document. We gave the subjects a short explanation of the components in the workcenter, and a short demonstration of the affordances. The goal of this pilot study was to learn whether these general explanations were enough to enable users to complete a very specific task with DLITE. We also wanted to learn which aspects of the interface might be confusing.

Our sample consisted of four women and two men, ranging in age from early 20s to late 40s. They were: an engineer with some programming experience but no Internet experience; an administrative associate with no formal computer training but day-to-day experience with Netscape; a teacher with Netscape and computer programming experience; a librarian with extensive training in the use of traditional information retrieval systems; a computer science graduate student; and a government administrator with limited computer experience.

All of the subjects were able to successfully complete the task in under 30 minutes. Because of the small number of subjects, we cannot make any generalizations, but we noted some interesting common actions:

- Two users took advantage of the fact that multiple searches could be performed in parallel.
- Two users clicked on the number of results (asking for more results), when they intended to summarize the collection in the web browser.
- When documents are displayed in the web browser, a document icon is included to remind the user that the page they are seeing originates in DLITE. One user tried repeatedly to drag this icon into the workcenter.
- Two users dragged documents into InterBib instead of collections, and one dragged a folder icon onto a search service, intending to reuse the attached query. In all cases, after wondering why it didn’t work, the subjects realized what had happened and continued.
- We have both right-click and double-left-click cause an “activate” event. Unfortunately, Netscape uses single left-click on links to essentially activate them. As a

result, two users repeatedly clicked with the wrong button in one application or the other and waited for something to happen.

We did not test whether subjects will remember the novel affordances of DLITE in this study, but plan to test this in the future.

## 9. DISCUSSION

Our preliminary results show DLITE to be particularly useful in providing user-level uniformity in interacting with heterogeneous services, and in allowing intuitively obvious parallel activities when interacting with Digital Library services. The latter is important for user productivity in that it matches the multi-track working style required of many Digital Library users [Paepcke96b]. Parallelism all the way up to the user surface is important because different services take significantly different amounts of time to complete.

Here are some design considerations and lessons from the pilot study pertaining to particular aspects of DLITE.

### 9.1. Workcenters

The notion of workcenter is necessary to partition the space of available resources. There are just too many resources in Digital Libraries that span the Internet to present to the user at one time. Equally important, many of the services will only be useful for particular tasks and would consume valuable screen space and user attention if displayed all the time.

An alternative design for managing the large number of services might be to use a fisheye view technique. The advantage would be that services would at least be in the peripheral view of the user, rather than not being shown at all unless explicitly placed in a workcenter. We decided on our design for two reasons: First, we expect that even for systems that are easy to use, organizations will wish to develop standard collections of tools for given tasks in order to help less sophisticated users, and to ensure that resources are properly used for the organization's tasks. While workcenters could be designed by individual users, they could also be provided by departments such as an organization's research library, or other centers of expertise. This would not be the case with the fisheye view approach.

A second reason for our workcenters is that they make part of the expertise for each task a discrete entity in the system. The collection of tools that are combined in one workcenter may be carefully thought out, and may then be re-used by many users. The fact that a component is included reminds the user of its existence and makes it ready-to-hand. A well-designed workcenter may even suggest an order of activity by the placement of its components.

### 9.2. Components

One big challenge around our components was to find a small enough set to preserve simplicity, yet to introduce enough of them to span the space of services and notions we needed users to gain access to. We arrived at the present set through a straight-forward requirements analysis. While the set presented here has taken us far, we may have to add a few more components carefully, as the Digital Library grows richer in services. But our strong preference is to

keep the number of different component types to a minimum.

The extreme simplicity of our components is both an asset and a liability. We have worked hard to keep the number of inputs to our components to a minimum. In general we do this by allowing users to construct first-class objects (such as queries) out of parts, and then to use those objects as inputs to services. This has proved understandable and useful for the limited examples we have explored, but may need to be revisited when the sophistication of available services (and users' knowledge of them) calls for a less minimalist approach (e.g., in allowing a user to specify more details of the resources to be used in carrying out a particular search, such as time, cost, user context, etc.).

One difficulty for any interface that helps users deal with large numbers of items is the maintenance of display orderliness. One example of this problem is in our decision to have only a few distinct component types. This reduces confusion, but means that there will be many identical-looking objects on the screen. As an alternative, we could introduce more distinctions among the types (e.g., different images for different kinds of documents), increasing information and also increasing visual clutter. Currently we provide simple means for color-marking individual items, and for low-overhead label display (just moving the cursor over the item). Future experiments will determine if these are sufficient.

The one-to-one correspondence between visible components and remote services provides a simple way to distinguish error sources, using the component as an attachment point for the error indicator (which in some cases is a simple as a red X across the component). These indicators can be updated continuously while the user is attending to other processes.

### 9.3. Animation

In animating the components in the DLITE interface we make it possible for users to track the results of actions and have their attention drawn to salient events (such as a document being added to a result set). In order to avoid confusion, we have found it important not to animate components over long distances on the screen. The animation required significant engineering effort [Cous96] to support our design goal of allowing users to simultaneously share a DLITE workcenter on different machines, potentially separated over wide-area networks.

### 9.4. Use of the Web Browser

As exemplified by the mouse button issue in our pilot study findings, differences between the Web browser and DLITE interaction conventions sometimes cause problems. It is difficult to provide completely consistent use of the mechanisms, since other interfaces (such as web browsers) have different kinds of functionality. The same problem arises, for example, in ordinary GUIs, where single clicking on a desktop document will select (rather than opening) it, which is not consistent with single-clicking on a link in a browser. We are exploring alternative conventions for selecting, activating, and dragging DLITE component objects, and in our usability studies we will be examining the ways in which

users understand these as integrating with activities in the browser and operating system.

The use of a commercial Web browser as companion to DLITE has otherwise proven to be very advantageous. The browser, of course, covers navigation needs on the Web, and users are often familiar with its facilities. In our Java implementation we are also using the Web browser to deliver DLITE itself.

### 9.5. Scalability

Scalability of the interface is, as always, an issue. Graphical facilities like DLITE inherently suffer from potential screen real-estate problems. Scrolling is one solution, but it feels more clumsy than when text is scrolled. We have used two mechanisms to deal with scalability strains. One is the use of the Web browser as a surface on which to summarize entire result collections of arbitrary size. While the result collection components are limited as to the number of document icons they can hold, the collection summaries are linearly appended extracts that may be scrolled naturally in the browser. We mostly use document attributes like author, title, abstract and URL for our summaries, but large degrees of sophistication could be used to improve on this simple scheme. The point is that collection summarization does scale.

We are also developing special tools for dealing with large result sets which would be unwieldy in any interface. These tools work through clustering of large result sets which decreases the space needed for display. See [Bald96] for details.

### 10. RELATED WORK

In the last 15 years, there have been other attempts to give users direct-manipulation interfaces to multiple services within an information system. The Xerox Star and its successors allow simple document objects to be dropped onto applications programs or simple services like printers [John89]. In the 1980s, Metaphor Computer developed "capsules" that gave users the ability to link multiple application programs together. Rose has experimented with extending the traditional desktop with a 'pile' metaphor [Rose93].

The Virtual Notebook System [Gorr94], developed in the late 1980s, gave users the ability to combine text and images from disparate sources into a new, shared notebook artifact. Shipman, Marshall, and Moran have experimented extensively with systems for manipulating and understanding information objects arranged on canvases (e.g. VIKI [Ship95]).

Rao's InfoGrid system emphasized document space visualization and management [Rao92], and Mackinlay's Butterfly experimented with handling asynchronous requests to services that could take varying amounts of time to respond [Mack95].

Chang and Ungar experimented with animation techniques in the user interface for the Self programming language [Chan93]. Maloney and Smith also incorporate extensive animation in the Morphic user interface [Malo95].

Recently, Hendry and Harper described a system called SketchTrieve that reifies documents and services [Hend96]. SketchTrieve uses much more of a dataflow flavor in its interactions than DLITE, and does not emphasize the variety of services. We will compare the systems in detail elsewhere.

### 11. CONCLUSIONS

DLITE goes beyond existing direct-manipulation interfaces, bringing distribution of services, asynchronous processing, and multiple representations into the interface in a unified way. Using the metaphor of workcenters containing components, we have experimented with affordances such as drag-and-drop, and feedback mechanisms such as animation, in the context of a functioning digital library prototype. The resulting design supports users in coherent tasks, allowing multiple services to be invoked in parallel and for their results to be smoothly integrated into the user's environment.

### 12. ACKNOWLEDGEMENTS

Thanks to Michelle Baldonado for comments on an early draft of this paper, and to Scott Hassan and Tom Schirmer for help with InfoBus integration issues. Alan Steremberg implemented the Java/AWT interface. This material is based upon work supported by the National Science Foundation under Cooperative Agreement IRI-9411306. Funding for this cooperative agreement is also provided by DARPA, NASA, and the industrial partners of the Stanford Digital Libraries Project. Any opinions, finding, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the other sponsors.

### REFERENCES

- [Bald96] Baldonado, M.Q.W. and Winograd, T. Sense-Maker: An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests. To appear in CHI'97. Available as <http://www.diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1996-0048>.
- [Bier92] Bier, E.A. EmbeddedButtons: supporting buttons in documents. ACM Transactions on Information Systems (Oct. 1992) vol.10, no.4, p. 381-407.
- [Chan93] Chang, B.-W. and Ungar, D. Animation: from cartoons to the user interface. Proceedings of the ACM Symposium on User Interface Software and Technology (UIST'93). ACM Press, 1993. p. 45-55.
- [Chan96] Chang, K.C-C., et al. Boolean Query Mapping Across Heterogeneous Information Sources. IEEE Transactions on Knowledge and Database Engineering, v. 8, n. 4, August, 1996, 515-512.
- [Cous95] Cousins, S.B., et. al. InterPay: managing multiple payment mechanisms in digital libraries. In Proceedings of Digital Library, (Austin, TX, June, 1995) <http://csdl.tamu.edu/DL95/papers/cousins/cousins.html>.

- [Cous96] Cousins, S.B., et. al. A Distributed Interface for the Digital Library. Stanford University, 1996. <http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1996-0037>.
- [Cutt93] Cutting, D., et. al. ILU Manual. Xerox PARC. Dec., 1993. <ftp://ftp.parc.xerox.com/pub/ilu/ilu.html>.
- [Gorr94] Gorry, G. A., et al. Experience with the Virtual Notebook System: Abstraction in Hypertext. In Proceedings of CSCW (Chapel Hill, NC, October, 1994), ACM Press, 133-143.
- [Hans71] Hansen, W.J. User engineering principles for interactive systems. In Proceedings of Fall Joint Computer Conference, (Las Vegas, November, 1971), AFIPS Press, 523-532.
- [Hend86] Henderson, D. A. Rooms: multiple virtual workspaces. In ACM Transactions On Graphics, July, 1986.
- [Hend96] Hendry, D. G., and Harper, D. J. An architecture for implementing extensible-seeking environs. Proceedings of SIGIR (Zurich, 1996), 94-100.
- [John89] Johnson, J., et. al. The Xerox Star: a retrospective. Computer (1989), 22(9):11-26, 28-29.
- [Mack95] Mackinlay, J., et. al. An organic user interface for searching citation links. In Proceedings of SIGCHI, (Denver, CO, May, 1995), Addison Wesley, 67-73.
- [Malo95] Maloney, J.H. and Smith, R.B. Directness and liveness in the Morpheus user interface construction environment. Proceedings of UIST (1995). ACM Press, 21-28.
- [Marc95] Marchionini, G. Information seeking in electronic environments, Cambridge Univ. Press, 1995.
- [Nard96] Nardi, B., and O'Day, V. Intelligent agents: What we learned in the library. In Libri, v. 46, n. 2, 1996.
- [OMG93] Object Management Group. The Common Object Request Broker: Architecture and Specification. December, 1993. <http://www.omg.org>.
- [Oday93] O'Day, V.L. and Jeffries, R. Orienteering in an information landscape: how information seekers get from here to there. In Proceedings of INTERCHI '93, (Amsterdam, NL, May 1993), IOS Press, 438-445.
- [Paep96a] Paepcke, A., et. al. Towards Interoperability in Digital Libraries: Overview and Selected Highlights of the Stanford Digital Library Project. IEEE Computer Magazine, May, 1996. Also: <http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1995-0013>.
- [Paep96b] Paepcke, A. Information needs in technical work settings and their implications for the design of computer tools. Computer Supported Cooperative Work (1996), v. 5, n. 1, 63-92.
- [Paep97] Paepcke, A. InterBib: Bibliography-related services. <http://www-interbib.stanford.edu/~testbed/interbib>. 1997.
- [Rao92] Rao, R., et. al. The Information Grid: a framework for information-retrieval centered applications. In Proceedings of UIST (Nov. 15-18, 1992), 23-32.
- [Rose93] Rose, D.E. Mander, R. Oren, T. Ponceleon, D.B. Salomon, G. Wong, Y.Y. Content awareness in a file system interface: implementing the 'pile' metaphor for organizing information. Proceedings of SIGIR (1993), 260-269.

# DLI - Berkeley:

---

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

## Pedagogy:

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

- MVD
  - How well does [MVD 0.9](#) work for you? Could you get the links on that page to work (use 2 windows of browser, one for the instructions, and one for testing)? What do you like most about it?
  - Did you use it on video or a PC or Mac with Netscape 4?
  - Did you work out Lens overlaying, such as OCR and then Magnify?
  - For the TableSort example, could you under Anno view the note?
  - Could you get the special behaviors to work: Biblio, where you Select a type of format, use the mouse to select an entry, use Edit and Copy to get a version in that format, and then paste elsewhere?
  - Could you get Doublespace in the View menu to work?
- Cheshire
  - Can you find interesting environmental documents using Cheshire II?
- TileBars
  - What happens with TileBar search of "document" and "retrieval"?

- What happens with TileBar search of "fault" and "dam"?
- When is TileBar searching useful on a single document?
- Collections
  - What is the name of the DBMS used?
  - What is a database "schema"? How does it relate to "metadata"?
  - How many documents and how many images are in their collection?
  - How good is the OCRing? What research is underway to improve OCRing beyond that of ScanWorX and how well does it work? What is the main idea behind it?
  - How can you find the dams for a county?
  - How does the database table information for Almond dam relate to the page about it? To the OCR output about that page?
  - What is a VLURL? How do you construct it? Can you build one and show results for getting pictures of California wildflowers that have the string "rose" in their common names?
  - Display a distribution map for your favorite flower in California.
  - Can you tell the direction of flight from the aerial photos?
  - How do layers help with managing GIS information with the [GIS viewer](#)? Can you zoom in and out and pan around?

---

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

---

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

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



## Re-inventing Scholarly Information Dissemination and Use

The UC Berkeley Digital Library Project is developing the tools and technologies to support highly improved models of the "scholarly information life cycle." Our goal is to facilitate the move from the current centralized, discrete publishing model, to a distributed, continuous, and self-publishing model, while still preserving the best aspects of the current model such as peer review.

[Search](#)

[Seminar](#)

[Calendar](#)

[What's New](#)

### Technologies

♦ [Image Retrieval by Image Content](#)

♦ [New Document Models](#)

including [Web-based GIS](#)

♦ [Document Image Analysis](#)

♦ [Distributed Search](#)

♦ [NLP for Information Access](#)

### Collections

♦ [Quick Access to the Collections](#)

♦ [Overview of the Collections](#)

♦ [Usage and Copyright Information](#)

### About the Project

♦ [Publications, People, Systems, etc.](#)

♦ [Info for Project Members](#)

♦ [Related Projects](#)

The UC Berkeley Digital Library Project is part of the [Digital Libraries Initiative](#), sponsored by the National Science Foundation and many others. Additional funding at Berkeley comes from the [CNRI](#)-sponsored [D-Lib Test Suite](#), and the NSF-sponsored [National Partnership for Advanced Computational Infrastructure \(NPACI\)](#).

This page is dedicated to the memory of [Gary Kopec](#).

This server is powered by a [SUN Microsystems](#) Enterprise 450 Server, backed by an [IBM](#) 7013 RS 6000 and 3494 Tape Library Dataserver running AMASS software by [EMASS](#). See [About Our System](#) for details.

comments and questions: [www@elib.cs.berkeley.edu](mailto:www@elib.cs.berkeley.edu)



# Digital Library Tours

*Berkeley Digital Library Project*

---

## Guided Tours:

 [Documents](#)

 [Images](#)

 [GIS Viewer](#)

*(tours require frames support)*

---



[Berkeley DL](#)



[AccessMatrix](#)



[Information](#)



[Comments](#)

---

# Quick Access to the Collections

See also: [Disclaimer](#) | [Usage](#) | [Botanical Data](#) | [Geographical Data](#) | [Zoological Data](#)

	Description	More Information
Photographs	<ul style="list-style-type: none"> <li>● CalPhotos: <a href="#">All</a>, <a href="#">Plants</a>, <a href="#">Animals</a>, <a href="#">Landscapes</a>, <a href="#">Africa</a></li> </ul>	<ul style="list-style-type: none"> <li>● <a href="#">About the Image Collection</a> <ul style="list-style-type: none"> <li>● <a href="#">Blobworld</a></li> </ul> </li> <li>● <a href="#">computer vision research</a> <ul style="list-style-type: none"> <li>● <a href="#">FAQ</a></li> </ul> </li> </ul>
	<ul style="list-style-type: none"> <li>● <a href="#">Cal. Water Resources</a> (DWR)</li> </ul>	
	<ul style="list-style-type: none"> <li>● <a href="#">Corel Stock Photos</a>, <a href="#">BlobWorld query</a></li> </ul>	
	<ul style="list-style-type: none"> <li>● <a href="#">Aerial Photos</a> Sacramento River Delta region</li> </ul>	
	<ul style="list-style-type: none"> <li>● <a href="#">Photographers</a> who contributed photos</li> </ul>	
	<ul style="list-style-type: none"> <li>● <a href="#">All the Photos</a> in the Berkeley DLP collection</li> </ul>	
Databases	<ul style="list-style-type: none"> <li>● <a href="#">Bay Area Streets with an index*</a>, or <a href="#">without an index*</a> (both use an active map)</li> </ul>	
	<ul style="list-style-type: none"> <li>● <a href="#">California dams</a>, <a href="#">static map</a>, <a href="#">active map*</a></li> </ul>	<ul style="list-style-type: none"> <li>● <a href="#">about the dams</a></li> </ul>
	<ul style="list-style-type: none"> <li>● <a href="#">CalFlora</a>: <a href="#">species</a>, <a href="#">observations</a>, <a href="#">synonymy</a></li> </ul>	<ul style="list-style-type: none"> <li>● <a href="#">about Calflora</a>, <a href="#">FAQ</a></li> </ul>
	<ul style="list-style-type: none"> <li>● <a href="#">Museum of Vertebrate Zoology</a> specimen records</li> </ul>	
	<ul style="list-style-type: none"> <li>● <a href="#">AmphibiaWeb</a></li> </ul>	<a href="#">About AmphibiaWeb</a>
	<ul style="list-style-type: none"> <li>● <a href="#">California Gazetteer: active map*</a></li> </ul>	
	<ul style="list-style-type: none"> <li>● Standard Names: <a href="#">continents</a>, <a href="#">countries</a>, <a href="#">US states</a>, <a href="#">Cal. counties</a></li> </ul>	

<b>Documents</b>	<ul style="list-style-type: none"> <li>● <a href="#">California Environmental</a> reports, plans, ordinances, EIRs, etc., <a href="#">browse lists</a></li> <li>● <a href="#">World Conservation Union (IUCN)</a> Action Plans</li> </ul>	<ul style="list-style-type: none"> <li>● <a href="#">about the collection</a></li> <li>● <a href="#">document image analysis</a></li> <li>● <a href="#">new document models</a> <ul style="list-style-type: none"> <li>● <a href="#">about TileBars</a></li> </ul> </li> </ul>
<b>Geographical Layers</b>	<ul style="list-style-type: none"> <li>● <a href="#">GIS Viewer Example List *</a></li> <li>● Street finder for the S.F. Bay Area: <a href="#">with</a> and <a href="#">without</a> street index (both active map*)</li> <li>● <a href="#">California Gazetteer active map*</a></li> <li>● <a href="#">Delta Fish Flow</a> active map *</li> </ul>	<ul style="list-style-type: none"> <li>● <a href="#">user manual</a></li> <li>● <a href="#">downloading</a> <ul style="list-style-type: none"> <li>● <a href="#">tour</a></li> </ul> </li> </ul>

\* java is required for active maps

[Overview of the Collections](#)   [About the Database](#)   [About the Digital Library Project](#)  
[Data Statistics](#)   [Disclaimer & Usage](#)



[Digital Library Project](#)

University of California, Berkeley

questions & comments: [www@elib.cs.berkeley.edu](mailto:www@elib.cs.berkeley.edu)



# Advanced Structured Document Examples

## *Berkeley Digital Library Project*

Below are links to examples of advanced structured documents created using document-specific image decoders. Each of the examples consists of a collection of interlinked pages that provide three representations of the scanned document. The first representation is a sequence of simple scanned page images, with the usual "Previous" and "Next" type links to adjacent pages. This representation is similar to the "page image" form offered by our document server. The second representation is the corresponding sequence of ascii text pages generated by a commercial omni-font OCR program, XIS ScanWorX. This representation is similar to the "hyperocr" form offered by our document server. The third representation is an advanced structured document created using document-specific image decoders, following the document image decoding (DID) approach described in Kopec and Chou, "Document Image Decoding Using Markov Source Models", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 6, June, 1994.



Document-specific decoding is an active research area in the Berkeley Digital Library Project. Comments and suggestions for additional documents to process are welcome at [www@elib.cs.berkeley.edu](mailto:www@elib.cs.berkeley.edu)

- [DWR Bulletin 17, Dams Within Jurisdiction of the State of California](#)
- [DWR Bulletin 155, General Comparison of Water District Acts](#)
- [IESP Technical Report 9, Fishes of the Sacramento-San Joaquin Estuary...](#)

## More Information about Document-Specific Decoding

The text content for each advanced structured document was obtained using a DID recognizer whose bitmap templates were generated from sample pages of the scanned document. A recently-developed template training system was used that generates templates from a set of page images plus errorful, whole-page transcriptions that are not aligned with images. The significance of this training system is that it allows document-specific character models to be developed with relatively little user effort. It is widely known that document-specific models can provide an order of magnitude improvement in OCR error rate, compared with typical omni-font OCR devices. However, training an OCR system for a particular font typically involves considerable manual effort. As a result, specialized recognition systems have only been cost-effective for relatively large homogeneous document collections.

The operating scenario supported by the training system is that a user prepares a transcription of a small number of pages from a document, containing samples of characters in the fonts present in the document. These transcriptions may contain errors and can be created using an omni-font recognizer, for example. The system uses the transcriptions and page images to generate a set of document-specific character templates. These templates are then used to recognize the remaining pages of the document.

A quantitative performance evaluation of the system has been completed using DWR Bulletin 155

(B155). This document was selected because of the availability of the WordPerfect source file, which was used as ground truth in assessing OCR performance.

The template estimation procedure was applied to a set of 20 page images from B155, using the corresponding WordPerfect source as the training transcription. The training data contained about 40,000 glyphs and 212 different characters (where characters in different fonts are considered distinct). The resulting templates were used to decode 375 additional pages of material from B155, which contained 543,779 glyphs.

The table below summarizes OCR character error rates (substitution, deletions, insertions) using these templates in DID decoders with 3 different language models. The recognition performance of ScanWorX is also given for comparison. The "DID unigram" decoder allows any of the 212 possible characters to follow any other. This is the weakest possible language model and puts the entire recognition burden on the character templates. The "DID bigram" model is the minimal bigram model that includes all of the bigrams that occur in the document, as determined from the WordPerfect source. This model was included to provide an upper bound on the performance of any bigram model for this data. Finally, the "DID uni+bigram" decoder is a modification of the unigram decoder in which a bigram model, trained on the training data, is used for selected fields of the tables.

Decoder	Substitutions	Deletions	Insertions	Errors (%)
ScanWorX	2149	1069	1061	4279 (0.79%)
DID unigram	430	73	80	583 (0.11%)
DID uni+bigram	289	72	68	429 (0.079%)
DID bigram	87	57	53	197 (0.036%)

The character error rate of the DID unigram model is factor of 7 less than that of ScanWorX while the error rate using the unigram/bigram hybrid is a factor of 10 less. The "ideal" bigram model provides more than a factor of 20 improvement.

---

[Berkeley Digital Library Project](#) / [www@elib.cs.berkeley.edu](http://www@elib.cs.berkeley.edu) / Last Modified August 24, 1995



# Demos: Content-based Queries

## Berkeley Digital Library Project

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

### Finding Objects in Pictures

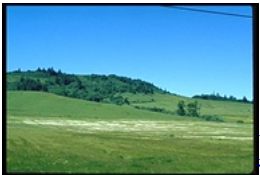


see [Finding horses using body plans](#)

### Colored Blobs and Color Percentages



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



*green % > 25 and lt. blue % > 25*

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



*sm. purple dots > 3*



*very sm. yellow dots > 15*



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



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

[Berkeley Digital Library](#) | [www@elib.cs.berkeley.edu](mailto:www@elib.cs.berkeley.edu)

# Welcome to Blobworld!

## Why Blobworld?

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

- While users generally want to find images containing particular objects ("things"), most existing image retrieval systems represent images based only on their low-level features ("stuff"), with little regard for the spatial organization of those features.
  - Systems based on user querying are often unintuitive and offer little help in understanding why certain images were returned and how to refine the query. Often the user knows only that he has submitted a query for, say, a bear but in return has retrieved many irrelevant images and very few pictures of bears.
- 

## What is Blobworld?

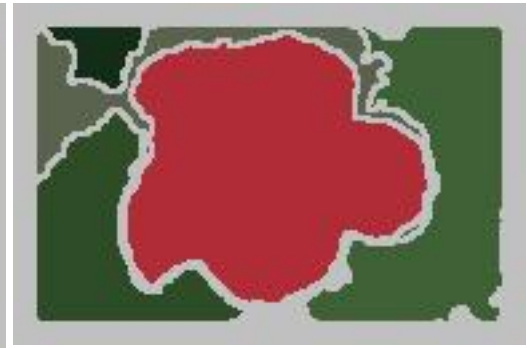
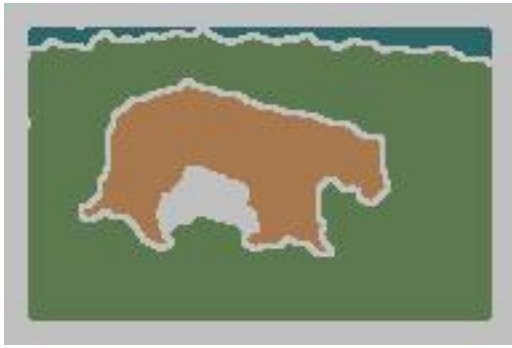
We have developed a new image representation, "Blobworld," and a [retrieval system](#) based on this representation. While Blobworld does not exist completely in the "thing" domain, it recognizes the nature of images as combinations of objects, and querying and learning in Blobworld are more meaningful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

To segment an image, we model the joint distribution of the color, texture, and position features of each pixel in the image. We use the Expectation-Maximization (EM) algorithm to fit a mixture of Gaussians model to the data; the resulting pixel-cluster memberships provide the segmentation of the image. After the image is segmented into regions, a description of each region's color, texture, and spatial characteristics is produced.

**Original  
image:**



## Blobworld:



---

## What can we use Blobworld for?

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

---

## Want to learn more?

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

Blobworld was developed by [Chad Carson](#), [Serge Belongie](#), and [Jitendra Malik](#).

---

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

---

Last updated October 29, 1999, by [Chad Carson](#)



# Image Classification

*Berkeley Digital Library Project*

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

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

<a href="#">Zebras</a>	<a href="#">Classified as zebras by Blobworld</a>	<a href="#">Classified as zebras by color histograms</a>
------------------------	---	--



[Berkeley DL](#)



[AccessMatrix](#)



[Information](#)



[Photographs](#)



[Comments](#)

---

# DLI - Santa Barbara:

---

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

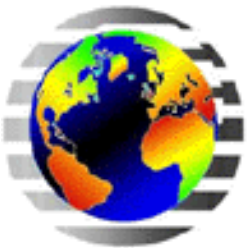
---

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

---

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

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



# Alexandria Digital Library Project

<a href="#">Home</a>	<a href="#">Services</a>	<a href="#">Documentation</a>	<a href="#">Research</a>	<a href="#">People</a>	<a href="#">Related Links</a>
----------------------	--------------------------	-------------------------------	--------------------------	------------------------	-------------------------------

## Welcome

Welcome to the Alexandria Digital Library Project. The name *Alexandria* comes from the library of Alexandria, Egypt, which was considered the center of all knowledge/learning. No one place now can claim that distinction - but all data sources together (libraries, academic institutions, private companies, government agencies, etc.) are *Alexandria*. The project began in 1995 with the development of the Alexandria Digital Library, a working digital library with collections of geographically referenced materials and services for accessing those collections. The Alexandria Digital Library Project is headquartered on the campus of the [University of California at Santa Barbara](#).

## Alexandria Digital Earth Prototype (ADEPT)

The National Science Foundation has announced funding from 1999-2004 for the next stage of the project, the Alexandria Digital Earth Prototype (ADEPT).

## Related Projects

## Digital Library Interfaces

[Alexandria Digital Library \(ADL\) \(1994-1999\)](#)[California Digital Library \(CDL\): ADL Web Client](#)  
[ADL Gazetteer Development](#)[ADL Gazetteer Server](#)

THE ALEXANDRIA DIGITAL LIBRARY  
University of California, Santa Barbara  
1205 Girvetz Hall  
Santa Barbara, CA 93106, USA  
TEL: 805.893.7665 FAX: 805.893.3045  
URL: [www.alexandria.ucsb.edu](http://www.alexandria.ucsb.edu)

Last Modified: February 13, 2000  
[Email](#) about general project inquiries  
[Email](#) about data, metadata, and access issues  
[Email](#) about web-related comments

# Universe



Alexandria Digital Library: [ADL](#)

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

## Universe

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

<a href="#">Earth</a>	<a href="#">Jupiter</a>	<a href="#">Mars</a>	<a href="#">Moon</a>
<a href="#">Saturn</a>	<a href="#">Sun</a>	<a href="#">Venus</a>	

## Universe

### Aerial photographs

- [Sources of Earth and Planetary Photography](http://www.nasm.edu/ceps/RPIF/RPIFsources.html)::<http://www.nasm.edu/ceps/RPIF/RPIFsources.html>

### Artificial satellites

- [Mission and Spacecraft Library](http://leonardo.jpl.nasa.gov/msl/home.html)::<http://leonardo.jpl.nasa.gov/msl/home.html>
- [STScI/HST Public Information](http://oposite.stsci.edu/)::<http://oposite.stsci.edu/>

### Astronomical - Observations

- [ESO and Space Telescope Science Archive Facilities](http://archive.eso.org/)::<http://archive.eso.org/>
- [European Southern Observatory Astronomical Information and Events](http://www.eso.org/outreach/info-events/)::<http://www.eso.org/outreach/info-events/>

- [Mapping the Heavens: The Next Generation of Celestial Surveys](http://spider.ipac.caltech.edu/staff/jarrett/talks/pomona/pres.html)::<http://spider.ipac.caltech.edu/staff/jarrett/talks/pomona/pres.html>

- [The Web Window to the Invisible Universe - the Radio Sky](http://www.pkts.atnf.csiro.au/databases/surveys/aitoff/aitoff.html)::<http://www.pkts.atnf.csiro.au/databases/surveys/aitoff/aitoff.html>

- [U.S. Infrared Space Observatory Science Support Center](http://www.ipac.caltech.edu/iso/)::<http://www.ipac.caltech.edu/iso/>

### Astronomical photometry

- [Latest Hubble Space Telescope Observations](http://www.stsci.edu/pubinfo/Latest.html)::http://www.stsci.edu/pubinfo/Latest.html
- [NASA JSC Digital Image Collection](http://images.jsc.nasa.gov/)::http://images.jsc.nasa.gov/
- [Space Image Libraries](http://www.okstate.edu/aesp/image.html)::http://www.okstate.edu/aesp/image.html
- [Stereoscopic Maps of Nearby Stars](http://www.clockwk.com/stars/index.html)::http://www.clockwk.com/stars/index.html
- [The Best of the Hubble Space Telescope](http://www.seds.org/hst/hst.html)::http://www.seds.org/hst/hst.html
- [Today@NASA](http://www.hq.nasa.gov/office/pao/NewsRoom/today.html)::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html

## Astronomy

- [Astronomical Applications Department: Data Services](http://aa.usno.navy.mil/AA/data/)::http://aa.usno.navy.mil/AA/data/
- [Astronomical Data Center](http://adc.gsfc.nasa.gov/)::http://adc.gsfc.nasa.gov/
- [CyberAstronomy](http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html)::http://reality.sgi.com/sambo/Oobe/CyberAstronomy/CyberAstronomy/intro.html
- [NASA/IPAC Extragalactic Database \(NED\)](http://ned.ipac.caltech.edu/)::http://ned.ipac.caltech.edu/
- [NCSA Astronomy Digital Image Library](http://imaginglib.ncsa.uiuc.edu/imaginglib/imaginglib.html)::http://imaginglib.ncsa.uiuc.edu/imaginglib/imaginglib.html
- [SEDS Internet Headquarters](http://seds.lpl.arizona.edu/)::http://seds.lpl.arizona.edu/
- [SEDS Messier Database](http://www.seds.org/messier/)::http://www.seds.org/messier/
- [SkyView Virtual Observatory](http://skyview.gsfc.nasa.gov/)::http://skyview.gsfc.nasa.gov/
- [Space.com](http://www.space.com/)::http://www.space.com/
- [Views of the Solar System](http://www.hawastsoc.org/solar/)::http://www.hawastsoc.org/solar/

## Astrophysics

- [Compton Gamma-Ray Observatory \(CGRO\) Science Support Center](http://coss.gsfc.nasa.gov/coss/)::http://coss.gsfc.nasa.gov/coss/
- [HEASARC/GSFC Home Page](http://guinan.gsfc.nasa.gov/)::http://guinan.gsfc.nasa.gov/
- [NASA Data Archive and Distribution Service \(NDADS\)](http://nssdca.gsfc.nasa.gov/)::http://nssdca.gsfc.nasa.gov/

## Atlases

- [Atlas celeste](http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg)::http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg
- [L'Atlas Catalan](http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm)::http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm
- [Out of This World: The Golden Age of the Celestial Atlas](http://www.lhl.lib.mo.us/pubserv/hos/stars/welcome.htm)::http://www.lhl.lib.mo.us/pubserv/hos/stars/welcome.htm
- [Planetary Image Atlas](http://www-pdsimage.jpl.nasa.gov/PDS/public/Atlas/Atlas.html)::http://www-pdsimage.jpl.nasa.gov/PDS/public/Atlas/Atlas.html

## Calendars

- [Astronomical Applications Department: Data Services](http://aa.usno.navy.mil/AA/data/)::http://aa.usno.navy.mil/AA/data/

## Cartography

- [Exposition Virtuelle - Le Ciel et la Terre](http://www.bnf.fr/web-bnf/expos/ciel/index.htm)::http://www.bnf.fr/web-bnf/expos/ciel/index.htm
- [L'Atlas Catalan](http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm)::http://www.bnf.fr/web-bnf/expos/ciel/catalan/index.htm

## Comets

- [Comets and Meteor Showers](http://comets.amsmeteors.org)::http://comets.amsmeteors.org

## Early maps - graphic

● [Atlas celeste](http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg)::http://www.oir.ucf.edu/wm/map/atlas-celeste.jpg

## Early maps

● [Out of This World: The Golden Age of the Celestial](#)

[Atlas](#)::http://www.lhl.lib.mo.us/pubserv/hos/stars/welcome.htm

● [The Earth & the Heavens](#)::http://portico.bl.uk/exhibitions/maps/overview.html

## Earth sciences

● [PlanetScapes](#)::http://planetscapes.com/

● [Space.com](#)::http://www.space.com/

● [Windows to the Universe](#)::http://www.windows.umich.edu/

## Eclipses

● [Astronomical Applications Department: Data Services](#)::http://aa.usno.navy.mil/AA/data/

## Galaxies - Spectra

● [Multiwavelength Milky Way](#)::http://adc.gsfc.nasa.gov/mw/milkyway.html

## Geology

● [Center for Earth and Planetary Studies](#)::http://www.nasm.edu/ceps/homepage.html

## Glossaries

● [Glossary of Cartographic Terms](#)::http://www.lib.utexas.edu/Libs/PCL/Map\_collection/glossary.html

## Historical geography - Maps

● [The Earth & the Heavens](#)::http://portico.bl.uk/exhibitions/maps/overview.html

## Maps

● [A Space Library](#)::http://samadhi.jpl.nasa.gov/

## Meteors

● [Comets and Meteor Showers](#)::http://comets.amsmeteors.org

## Moon - Phases

● [Astronomical Applications Department: Data Services](#)::http://aa.usno.navy.mil/AA/data/

## Nautical astronomy

● [Celestial Navigation](#)::http://peck.ipph.purdue.edu/al/space.html

## Planets - Geology

● [Center for Earth and Planetary Studies](#)::http://www.nasm.edu/ceps/homepage.html

## Planets - Orbits

● [Inner Planets Orbiting](#)::http://www.ac.wvu.edu/~stephan/Astronomy/planets.html

## Planets

● [Exploring the Planets](#)::http://www.nasm.edu/ceps/ETP/

## Remote-sensing images

● [EROS Selected Image Gallery](#)::http://edcwww.cr.usgs.gov/bin/html\_web\_store.cgi

- [Exploring the Planets](http://www.nasm.edu/ceps/ETP/)::http://www.nasm.edu/ceps/ETP/
- [Latest Hubble Space Telescope Observations](http://www.stsci.edu/pubinfo/Latest.html)::http://www.stsci.edu/pubinfo/Latest.html
- [NASA JSC Digital Image Collection](http://images.jsc.nasa.gov/)::http://images.jsc.nasa.gov/
- [NASA's Observatorium](http://www.rspac.ivv.nasa.gov/nasa/core.shtml)::http://www.rspac.ivv.nasa.gov/nasa/core.shtml
- [NSSDC Photo Gallery](http://nssdc.gsfc.nasa.gov/photo_gallery/photogallery.html)::http://nssdc.gsfc.nasa.gov/photo\_gallery/photogallery.html
- [Planetary Photojournal: NASA's Image Access Home Page](http://photojournal.jpl.nasa.gov/)::http://photojournal.jpl.nasa.gov/
- [Planetary image finders](http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/)::http://ic-www.arc.nasa.gov/ic/projects/bayes-group/Atlas/
- [STScI/HST Public Information](http://opposite.stsci.edu/)::http://opposite.stsci.edu/
- [Solid State Imaging \(SSI\) Education and Public Outreach Website](http://www.jpl.nasa.gov/galileo/sepo/)::http://www.jpl.nasa.gov/galileo/sepo/
- [Sources of Earth and Planetary Photography](http://www.nasm.edu/ceps/RPIF/RPIFsources.html)::http://www.nasm.edu/ceps/RPIF/RPIFsources.html
- [Space Image Libraries](http://www.okstate.edu/aesp/image.html)::http://www.okstate.edu/aesp/image.html
- [The Best of the Hubble Space Telescope](http://www.seds.org/hst/hst.html)::http://www.seds.org/hst/hst.html
- [Today@NASA](http://www.hq.nasa.gov/office/pao/NewsRoom/today.html)::http://www.hq.nasa.gov/office/pao/NewsRoom/today.html
- [U.S. Infrared Space Observatory Science Support Center](http://www.ipac.caltech.edu/iso/)::http://www.ipac.caltech.edu/iso/
- [Windows to the Universe](http://www.windows.umich.edu/)::http://www.windows.umich.edu/

## Space environment

- [Space Environment Center](http://www.sec.noaa.gov/)::http://www.sec.noaa.gov/

## Space sciences

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

## Stars - Rotation

- [Apparent Stellar Rotation](http://www.ac.wvu.edu/~stephan/Astronomy/stars.html)::http://www.ac.wvu.edu/~stephan/Astronomy/stars.html

## Views

- [PlanetScapes](http://planetescapes.com/)::http://planetescapes.com/

## Volcanoes

- [Volcano World](http://volcano.und.nodak.edu/)::http://volcano.und.nodak.edu/



Alexandria Digital Library: [ADL](#)

Last modified on 2000-05-31 at 22:47 GMT by [the Systems Engineering Team](#)

# DLI - Illinois:

---

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

---

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

---

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

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

[D-Lib Forum](#) • [D-Lib Test Suite](#) • [DLI at UIUC](#) • [About DeLiver](#) • [Quick Tips](#) • [Help](#)  
< [Search DeLiver](#) > • < [Browse Journals](#) > • [Download Software](#) • [Related Resources](#)



## ACHIEVEMENTS

[Progress Reports](#)

[Overview Papers & Talks](#)

[Publications & Reports](#)

[Workshops](#)

[Nat'l Synchronization Effort](#)

## RESEARCH GROUPS

[Repositories \(testbed\)](#)

[Social Science \(User Studies\)](#)

[Semantic Research](#)

[Interspace Prototype](#)

[System Evaluation](#)

## PARTNERSHIPS

[Publishers](#)

[Software Providers](#)

## PEOPLE

[Contact Information](#)

## TECHNOLOGY HIGHLIGHTS

[The 5 other DLI Projects](#)

[UIUC Digital Libraries](#)

[DL Related Information](#)

[Global Cultural Memory Project](#)



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

**DeLiver**  
**USAGE STATISTICS: updated daily**

The [NSF/DARPA/NASA Digital Libraries Initiative](#) (DLI) project at the [University of Illinois at Urbana-Champaign \(UIUC\)](#), 1994-1998, had the goal of developing widely usable Web technology to effectively search technical documents on the Internet. Our efforts were concentrated on building an [experimental testbed](#) with tens of thousands of [full-text journal articles](#) from physics, engineering, and computer science, and making these articles available over the World Wide Web, often before they were available in print. The DLI Testbed focused on using the [document structure](#) to provide federated search across [publisher collections](#). Our [sociology research](#) included the evaluation of its effectiveness under use by over one thousand UIUC faculty and students, a user community an order of magnitude bigger than the last generation of research projects centered on search of scientific literature. Our [technology research](#) developed indexing of the contents of text documents to enable federated search across multiple sources, testing this on millions of documents for

## **Computing the Future** **A National Research Council** **Report**

[semantic federation.](#)

Our testbed of [Engineering and Physics journals](#) is based in the [Grainger Engineering Library](#). We are placing article files into the digital library on a production basis in Standard Generalized Markup Language ([SGML](#)) from engineering and science [publishers](#).

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

**ONLINE SUMMARIES  
of the UIUC DLI PROJECT**

| [DLI Glossary](#) | [DLI National Synchronization](#) | [DL Related Information](#) | [UIUC Libraries](#) | [UIUC](#) |



[D-Lib Forum](#) • [D-Lib Test Suite](#) • [DLI at UIUC](#) • [About DeLiver](#) • [Quick Tips](#) • [Help](#)  
< [Search DeLiver](#) > • < [Browse Journals](#) > • [Download Software](#) • [Related Resources](#)

University of Illinois at Urbana-Champaign Digital Libraries Initiative  
Comments and Questions to: External Relations Coordinator: [Tom Habing](#)  
1999.11.15 may



# Glossary

## ARPA (DARPA)

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

## **Broad System of Ordering (BSO)**

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

## **Collection Interface Agent**

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

## **Collection Registry**

The database in which descriptions of collections are stored.

## **Concept Space**

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

## Corporation for National Research Initiatives (CNRI)

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

## **Derived Data**

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

## **DID**

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

## **Digital Libraries**

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

## Digital Library Federation

The Federation is comprised of leaders of fifteen of the nation's largest research libraries and archives and the Commission on Preservation and Access ([CPA](#)). A primary goal of the Federation

is the implementation of a distributed, open digital library accessible across the global Internet. The library will consist of collections expanding over time in number and scope to be created from the conversion of digital form of documents contained in founding member and other libraries and archives, and from the incorporation of holdings already in electronic form.

## **DLI**

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

## **ESRI**

Environmental Systems Research Institute

## **European Digital Library Consortium (ERCIM)**

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

## **Federated Repositories**

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

## **HTML**

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

## **IETF**

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

## **IIPA**

International Intellectual Property Alliance.

## **IITA**

Information Infrastructure Technology and Applications

## **IITF**

Information Infrastructure Task Force.

## **Information Visualization**

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

## **Intellectual Property Usage License**

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

## **Intellectual Work**

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

## **Interoperability**

The ability of software and hardware on multiple machines from multiple vendors to communicate.

## **Interspace**

The Interspace is a vision of what the Internet will become, where users cross-correlate information in multiple ways from multiple sources. It is an applications environment for interconnecting spaces to manipulate information, much as the Internet is a protocol environment for interconnecting networks to transmit data. Navigating information paths and grouping related items is a fundamental operation. So is semantic retrieval and community classification, with interactive support for vocabulary switching across domains and subject indexing for amateur classifiers.

## **IR**

Information Retrieval

## **ISO 12083**

The new international standard for electronic manuscript preparation and markup. ISO 12083 speeds computerized text from author to publisher to typesetter without retyping and transforms the document into a searchable database.

## **JAVA**

Java is a simple, object-oriented, distributed, interpreted, robust, secure, architecture-neutral, portable, high-performance, multithreaded, dynamic, buzzword-compliant, general-purpose programming language.

## **Machine Learning**

The ability of a machine to improve its performance based on previous results.

## **Magic Lenses**

This is an idea out of [Xerox PARC](#) where a region of the display (the "lens"), positioned by the mouse, is rendered in a special way. Lenses are specialized local views which might show labels where none were before, or handles on objects, or highlight certain subsets of items.

## **Metadata**

Data about data. Includes information describing aspects of actual data items, such as name, format, content, and the control of or over data.

## **Middleware**

Software that mediates between an applications program and a network. It manages the interaction between disparate applications across the heterogeneous computing platforms. The Object Request Broker (ORB), software that manages communication between objects, is an example of a middleware program.

## **Multiple View User Interface**

Multiple views means that phrases can be drag-and-drop across each individual interface for each information source.

## **Multivalent Document (MVD)**

A single document made of multiple layers of difference but intimately related material. Each layer is of homogeneous content, but is of a relatively limited scope and functionality. Layers have dynamically loaded program objects associated with them called behaviors, that manipulate the content, often communicating with other layers and other behaviors to achieve a desired effect.

## **NASA**

National Aeronautics and Space Administration. NASA's mission is to advance and communicate scientific knowledge and understanding of the Earth, the solar system, and the universe and use the environment of space for research.

## **NetBill**

The NetBill project at CMU's Information Networking Institute is designing the protocols and software to support network-based payment for goods and services delivered over the Internet. NetBill acts as a third party to provide authentication, account management, transaction processing, billing, and reporting services for network-based clients and users.

## **NII**

National Information Infrastructure.

## **NSF**

National Science Foundation. An independent agency of the U.S. government with the mission of promoting science and engineering.

## **NTIA**

National Telecommunications and Information Administration. Responsible for the Information Superhighway.

## **OCR**

Optical Character Recognition

## **Ontology**

An explicit formal specification of how to represent the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them.

## **PAD++**

Software which provides a virtual infinite extent, infinitely zoomable work surface, being developed under an ARPA grant at the University of New Mexico. Its multiscale interface, allowing interaction at many scales, is expected to allow the visualization of large scale information structures, and the organization of large and complex work activities. It is integrated with the Tcl/Tk prototyping environment and is being used as the development platform for the University of Michigan's Advanced User Interface ([AUI](#)).

## **PAT**

Indexing software developed by the OpenText Corp. which serves as the basis for its products

used for searching the WWW, intranets, etc.

## **Portals**

Windows on a zooming work surface which can be used to bring distant regions close, to give simultaneous views at multiple scales, or, when given special active functionality, to create Magic lenses.

## **Query Planning Agent**

A kind of Task Planning Agent. In many contexts, this means task planners who specialize in query tasks. Some select only from a library of existing plans for executing queries, others construct new plans.

## **Registration**

The process of adding new descriptions to the registry database.

## **Registry Database**

The database in which descriptions of agents (including collections) are stored. Also called the Conspectus database or the registry.

## **Remora Agents**

An agent which, given a URL, will check the links of a homepage at a specified interval of time, check a specified homepage for any changes in the homepage at a specified interval and notify the user of any changes, and/or search a specified homepage for key phrases, results of which are emailed to the user.

## **Scaffolding**

This concept is based on the idea that at the beginning of learning, students need a great deal of support, gradually, this support is taken away to allow students to try their independence. Providing support takes place in a number of ways - the way in which the selections are organized in a theme, the amount of prior knowledge activation that is provided, the way in which the literature is read by students, and the types of responses students are encouraged to make.

## **Semantic Retrieval**

Searching for words within a concept space (graph of terms occurring within objects linked to each other by the frequency with which they occur together).

## **Semantic Zooming**

In a multiscale interface like PAD++, normal, geometric zooming simply changes the size of objects in the view. In semantic zooming, objects change appearance or shape as they change size. For example, a growing dot will become a simple box, then a box with a one-word label, then a box with a longer label, then a rectangle filled with text and pictures. The goal is to give the most meaningful presentation at each size.

## **SGML**

Standard Generalized Markup Language. SGML is a platform-neutral standard for creating documents and information archives--it's a series of rules that everyone can follow in order to make their documents publishable in different media (print, CD-ROM, the Web) and to make their documents readable with different kinds of computers. SGML is also a structure for storing information which eases information-management and manipulation. It supports very powerful searching and allows large information repositories to be repurposed, broken down, and rearranged

intelligently into individual documents. For more information, see [SGML info](#).

## **Testbed**

A platform on which an assortment of experimental tools and products may be deployed and allowed to interact in real-time. Successful tools and products may be identified and developed in an interactive, evolutionary, interdependent process.

## **TestTiles**

TextTiling is a method for partitioning full-length text documents into coherent multi-paragraph units.

## **Thesaurus**

A controlled vocabulary with a syndetic structure within a circumscribed subject field used to organize material or information.

## **TileBars**

An interface for document that allows the user to make informed decisions about which documents to view based on the distribution of search terms in the document.

## **URC**

Uniform Resource Characteristic

## **Uniform Resource Citation**

A collection of attribute/values about an object. Some of the values may be URIs. URCs are not formally defined, yet.

## **URI**

Universal Resource Identifier - an address of some sort. See [IETF URI-WG](#) and the [W3.org](#).

## **URL**

Uniform Resource Locator. URLs are a particular kind of URI.

## **URN**

Uniform Resource Name. URNs are another kind of URI. Names are more persistent than Locations. A location may change, but a name rarely will.

## **Vocabulary Switching**

The mapping of vocabulary from one discipline onto the vocabulary of another discipline.

## **Z39.50**

The American National Standard Information Retrieval Application Service Definition and Protocol Specification for Open Systems Interconnection. The National Information Standards Institute (NISO), an American National Standards Institute (ANSI) accredited standards developer that serves the library, information, and publishing communities, approved the original standard in 1988 (referred to as Z39.50-1988 or Version 1). NISO published a revised version of the standard in 1992 (Z39.50-1992 or Version 2). ANSI/NISO Z39.50 defines a standard way for two computers to communicate for the purpose of information retrieval. Z39.50 makes it easier to use large information databases by standardizing the procedures and features for searching and retrieving information. Specifically, Z39.50 supports information retrieval in a distributed, client and server environment where a computer operating as a client submits a search request (query) to another computer acting as an information server. Software on the server performs a search on one

or more databases and creates a set of records that meet the criteria of the search request as a result. The server returns records from the resulting set to the client for processing. The power of Z39.50 is that it separates the user interface on the client side from the information servers, search engines, and databases. Z39.50 provides a consistent view of information from a wide variety of sources and offers client implementers the capability to integrate information from a range of databases and servers.

[The Acronym Expander](#) | [Free On-Line Dictionary of Computing](#)

[DLI Home](#) | [DLI National Synchronization](#) | [DL Related Information](#)  
[Information Science](#) | [Interspace](#) | [Testbed](#) | [User Evaluation](#)



University of Illinois at Urbana-Champaign Digital Libraries Initiative  
Comments to: External Relations Coordinator, [Tom Habing](#)

11/23/98



[Academic Press, Inc.](#)

[American Association for the Advancement of Science \(AAAS\)](#)

[American Astronomical Society \(AAS\)](#)

[American Chemical Society \(ACS\)](#)

[American Institute of Aeronautics and Astronautics \(AIAA\)](#)

[American Institute of Physics \(AIP\)](#)

[American Physical Society \(APS\)](#)

[American Society of Agricultural Engineers \(ASAE\)](#)

[American Society of Civil Engineers \(ASCE\)](#)

[American Society of Mechanical Engineers \(ASME\)](#)

[Institution of Electrical Engineers \(IEE\)](#)

[Institute of Electrical and Electronics Engineers \(IEEE\)](#)

[IEEE Computer Society](#)

[John Wiley & Sons](#)

---

[DLI Home](#) | [DLI National Synchronization](#) | [DL Related Information](#)

[Glossary](#) | [Information Science](#) | [Interspace](#) | [Testbed](#) | [User Evaluation](#)

University of Illinois at Urbana-Champaign Digital Libraries Initiative

Comments to: External Relations Coordinator, [Tom Habing](#)

01/18/98

## INTERSPACE

### Summary

KEVIN R. POWELL, PROJECT DIRECTOR

NEW

Darpa PI Meeting Project Summary

*The Interspace Prototype:*

*An Analysis Environment for  
Semantic Interoperability*

for more information

INTERSPACE

architectures

proposal

research & demonstration

**The Net of the Twenty-First Century** must permit users to directly solve their information problems. Hypermedia browsing has now become widespread and search facilities are beginning to appear. Users are now building information repositories on a grand scale. This will soon lead to a global information space consisting of a billion repositories.

(See [Evolution of the Net.](#)) What will this future world be like? How will we locate and correlate information in such a vast space?

The **Interspace Research Project** is developing a prototype environment for semantic indexing of multimedia information in a testbed of real collections. The semantic indexing relies on statistical clustering for concepts and categories. Interactive navigation based on semantic indexing enables information retrieval at a deeper level than previously possible for large, diverse collections. We are in the process of developing algorithms for automatically [extracting concepts](#) and computing [Concept Spaces](#), [Category Maps](#), and performing [Concept Assignment](#). Our collections include engineering literature, map images, and medical literature. The Interspace Prototype will thus enable scalable, interactive semantic interoperability across subject domain, media type, and collection size.





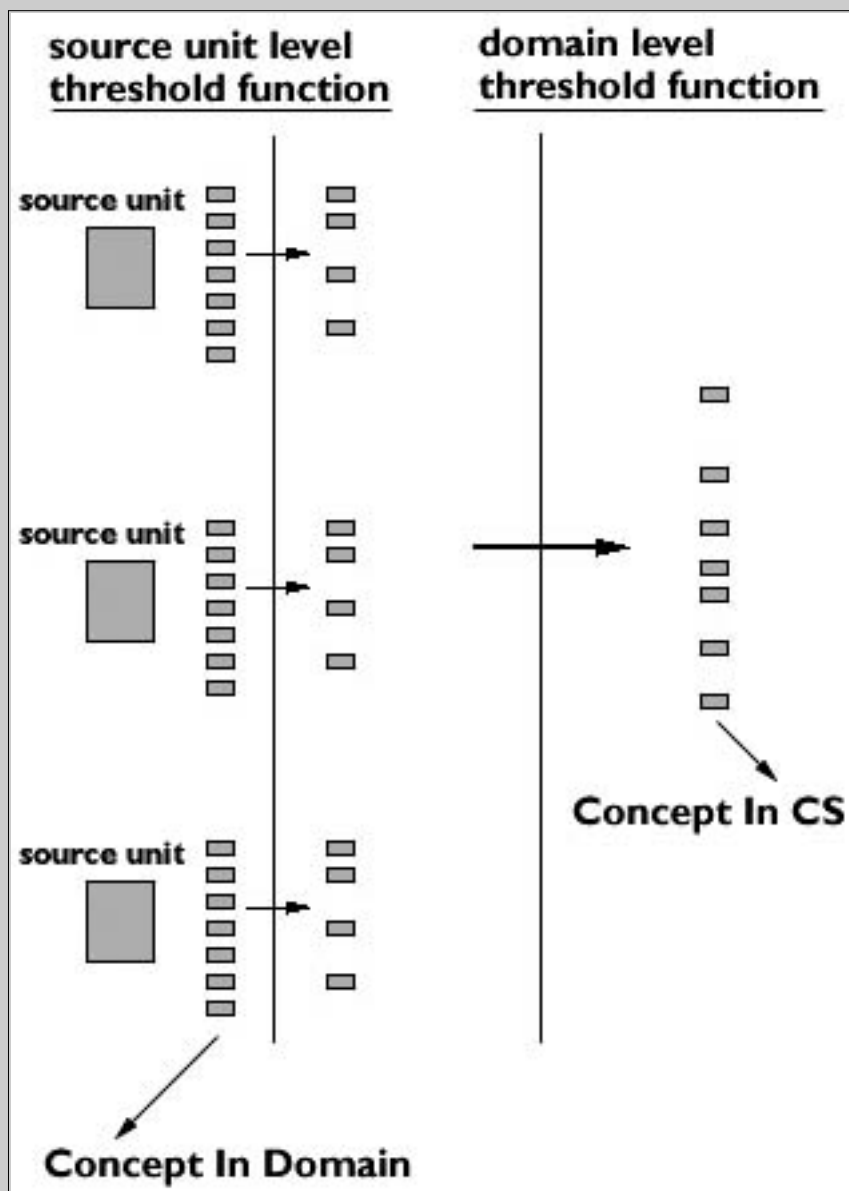
*Summary* *News* *Publications & Talks* *Reports* *Highlight*

INTERSPACE

*Interspace Architecture***CONCEPT SPACE GENERATOR**

The automatic generation of thesauri represents an area of growing importance in the field of computational science. Developed under the auspices of the federal NII [Digital Library Initiative \(DLI\)](#) at the Universities of Illinois and Arizona.

Concept Space thesauri are based on a hybrid symbolic/numeric computation that determines relationships between concepts in a collection of source units. The resulting map between concepts is designated a Concept Space and is useful in the refinement of queries presented to the collection. Concept Spaces are used, for example, in interactive query sessions as part of the DLI testbed at the University of Illinois, Urbana-Champaign. Algorithms to perform iterative search refinement which incorporate the computation of Concept Spaces.



In the conceptspace generation process called by the domain manager, a

ConceptInCS object is created for each ConceptInDomain objects that pass the domain threshold functions. Two threshold functions are used in this step: a *sourceunit* level threshold function and a *collection* level threshold function. The generation process uses a similarity function and only updates the new ConceptInCS objects. A detail discussion of the algorithm can be found in our technical report. The threshold function is needed to eliminate those ConceptInCS objects that are relatively unimportant in the conceptspace computation. We do this because it would be computationally infeasible to compute every ConceptInCS objects at this time. However, this may change in the future when hardware computation power is higher. The occurrence list of each of these ConceptInCS objects will be updated if the old one was saved or recreated if none exists. Since the cooccurrence list objects take up a large amount of space, there is an option to save it or to discard when the computation is done. After the cooccurrence is list computed, the similarity list is generated from it. This list represents the similarity matrix for each ConceptInCS object. These matrixes together form the conceptspace.

(1) Chen, H. and Lynch, K. J., "**Automatic Construction of Networks of Concepts Characterizing Document Databases**", *IEEE Transactions on Systems, Man and Cybernetics*, 22(5) 885-902, Sept./Oct., 1992.

(2) Schatz, B.R., and Chen, H., "**Building Large-Scale Digital Libraries**", *IEEE Computer*, Special Issue on Building Large-scale Digital Libraries, 29(5) 22-27, May 1996.

(3) Schatz, B.R., [Information Retrieval in Digital Libraries: Bringing Search to the Net](#), *Science*, 275(5298), 327-334, cover story and lead article, January 17, 1997.

# Welcome to the DLI Social Science Team Home Page

[Index](#)[Diary](#)[Internal](#)[Reports](#)[Completed](#)[Papers](#)[Papers in](#)[Progress](#)[Conference](#)[Presentations](#)[Site Visit](#)[and](#)[Quarterly](#)[Reports](#)[Main DLI](#)[Page](#)[Web Client-](#)[DeLiver](#)

This page consists of links to working papers and [a brief overview of the social science team](#) projects that we have been working on as the social science team for the NSF/ ARPA/ NASA [Digital Library Initiative project](#) being conducted at the University of Illinois.

Our subgroup of the Illinois Digital Library Initiative (DLI), the Social Science Team, has a mandate to study potential and actual use of prototype systems that other subgroups of the DLI build. In addition, we study the web more generally, and how the work of engineers and other scientist will be impacted by and will impact the growth of the information infrastructure.

Our Social Science Team has articulated, from the beginning, a commitment to a three way relationship between users, designers and social scientists, following in a general way the principals of participatory design. We are especially concerned with trying to fit our formative evaluation work to the ideal of this method: close contact and communication between designers and users via a series of mutually generated, iterative prototypes. To this end, we have conducted usability studies with the emergent testbed; observations of current users of electronic systems in the traditional library and beyond; focus groups, interviews and observations with faculty and staff who are potential users; and as use of the testbed continues to grow, transaction log analyses. One of our major concerns is finding a means to fit these all together.

Members of the team include: Ann Bishop, primary investigator; [Leigh Star](#), investigator; Emily Ignacio, graduate assistant; Laura Neumann, graduate assistant; [Cecelia Merkel](#), a graduate assistant; [Bob Sandusky](#), graduate assistant; and Eric Larson, graduate assistant.

---

send comments or  
questions to:  
[l-neuma1@uiuc.edu](mailto:l-neuma1@uiuc.edu)

# DLI Social Science Team Home Page

[Index](#)[Diary](#)[Internal](#)[Reports](#)[Completed](#)[Papers](#)[Papers in](#)[Progress](#)[Conference](#)[Presentations](#)[Site Visit](#)[and](#)[Quarterly](#)[Reports](#)[Main DLI](#)[Page](#)[Web Client-](#)[DeLiver](#)

send comments or  
questions to:  
[l-neuma1@uiuc.edu](mailto:l-neuma1@uiuc.edu)

## The University of Illinois DLI Evaluation Team: Who We Are and What We're About

---

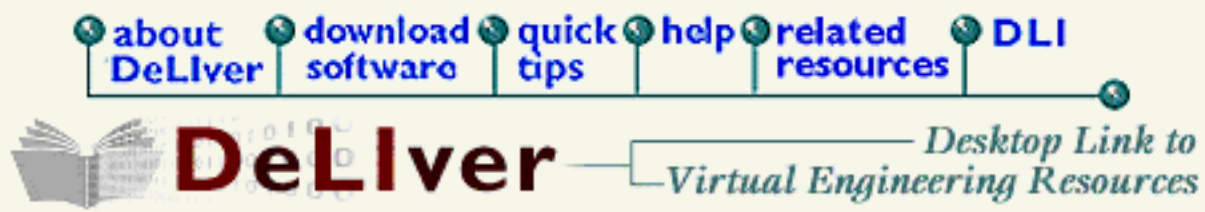
A fundamental component of our DLI project is sociological research and evaluation related to digital library use. The mandate of our testbed evaluation team is to:

- Provide ongoing user feedback for developing retrieval mechanisms, charging schemes, and other DL system features and functions
- Document and analyze extent and nature of testbed use, satisfaction, and impacts
- Identify reasons for use and non-use of the testbed, to gain a more complete understanding of testbed successes and failures
- Contribute to theoretical understanding of the changing information infrastructure and how it is transforming engineering and library work, communication, and learning practices
- Develop and assess new methods for conducting user-based digital library research, i.e., for capturing user data in the distributed repository environment
- Contribute to understanding of how large-scale information system design work is conceived and carried out.

---

To pursue these goals, we have developed an integrated research program that combines broad study of use with deep study of social phenomena. Over the course of the DLI project, we will conduct ongoing observations of engineering work and learning activities and how they intersect with the use of distributed, digital information. Individual and group interviews will be conducted with a range of potential and actual testbed users from the engineering community. We will conduct usability tests of various components and versions of the DL prototype and experiment with economic models and charging mechanisms. Extensive data on use will be gathered through large-scale user surveys and system instrumentation (i.e., the creation of testbed transaction logs).

---



## Connecting from Off-Campus IP Address

Welcome to [DeLiver](#), a **FREE**, grant supported system, providing access to the full-text of articles from over 50 journals in civil engineering, computer science, electrical engineering, and physics. Off-campus access to the DeLiver testbed is currently limited to University of Illinois at Urbana-Champaign faculty, staff, and students and to selected other users directly affiliated with the DeLiver project. Faculty and students at other institutions participating in the trial of DeLiver will only be able to connect to the testbed from computers located on their home campus. Select the type of user you are from the following choices:

[[about deliver](#)] - [[download software](#)] - [[quick tips](#)] - [[help](#)] - [[related resources](#)] - [[DLIhome](#)]

University of Illinois at Urbana-Champaign Digital Libraries Initiative  
Comments and Questions to: [DeLiver Web Master](#)

# University of Michigan Digital Library Activities

## DLI General Information

- [Home Page](#)
- [IEEE Computer article](#)
- [Introduction](#)
- [Current Status](#)
- [Technologies](#)
- [Agents, Ontologies](#)

## Campus Strategy

- Partnership of
  - [University Library](#)
  - [Information Technology Division](#)
  - [School of Information](#)
- combine: R&D; technology infrastructure; content access & user services; outreach
- shift to 21st century library model
  - user-centric, collaborative teams, global reach
  - distributed collections, heterogeneous access protocols, just-in-time information delivery
  - mixed funding models, value = access + services
- [Gateway Registry](#)
- [Electronic Reserve Shelf](#)
- [Knowledge Navigation Center](#): develop and support teaching and learning projects
- Questions:
  - How does the infrastructure at U. Michigan compare to that at your university?
  - How does this strategy relate to previous services of libraries?

## Projects

- [JSTOR](#): Journal Storage: over 1.2M pages
- [Making of America](#): with Cornell - 5K volumes, [D-Lib article](#): scanning, OCR, SGML encoding, tif2gif, interface
- [DLPS Image Services](#): see also V. 5 N. 8 Oct. 1996 [Information Technology Digest](#)
- [Humanities Text Initiative](#) and [Collaboratory for the Humanities](#)
- [Papryology](#)

- [Middle English Compendium Demo](#)
- [American Verse](#)
- [DLF](#)
- Questions:
  - Which of these projects do you find most interesting? Why?
  - Which of these projects should your university become involved in?

## Technical Approaches

- [see especially 1996 Ann Arbor Conf. on Electronic Records R & D](#)
  - Problem scenarios (see bullet list under **The Importance of Digital Preservation**)
  - Research questions (see **The 10 Research Questions**)
  - Research results: possible, requires changes and new types of efforts (see bullet list under **Research Projects and Results**)
  - [International Council on Archives](#): see **Guide for Managing Electronic Records from an Archival Perspective**, survey, literature review
- [Advanced Interfaces](#)
- [Ontology - Concept Descriptions](#) and [May 1997 slides](#)
- [Learning Agents](#)
- [Teaching and Learning Project](#)
- [SGML creation and delivery](#)
  - enormous collection: 2M pages
  - [flowchart](#)
  - [SGML Server Program](#): middleware, training
  - cross collection searching
  - multiple representations
  -
- [Leveraging rich document formats](#)
  - patterns of use
  - ease of changing delivery: new standards (HTML), new rendering/packaging
  - collection management
  - Panorama, XML support by W3C
- Questions:
  - Will the agent and ontology approach work? Soon? For production DLs?
  - What is the support needed for establishing a digital library following the UMDL approach? Training?
  - What interfaces for DLs will be usable?

THE NSF/DARPA/NASA SPONSORED  
UNIVERSITY OF MICHIGAN  
**DIGITAL LIBRARY**  
PROJECT

If you can see this list, you are using a Java-incompatible browser. This site is best viewed with a Java-compatible browser.

- [Mission](#)
  - [Introduction and Overview](#)
- [Accomplishments](#)
  - [Recent Events](#)
  - [Current Status](#)
  - [Coming Soon](#)
  - [Publications](#)
  - [Presentations](#)
- [UMDL In Action](#)
  - [Test Drive Artemis](#)
- [UMDL Technologies](#)
  - [Architecture: Agents and Ontologies](#)
  - [Access: Artemis Interface](#)
  - [Content: Collections](#)
  - [Economy: Computational Markets](#)
  - [Advanced User Interface](#)
  - [Conspectus & IR](#)
  - [Production System](#)
- [Impact](#)
  - [Education](#)
  - [Technology Transfer](#)
- [Team](#)
  - [Funders](#)
  - [Partners](#)
  - [Researchers](#)
- [Other](#)
  - [Other DLI Sites](#)



WELCOME TO THE  
UNIVERSITY OF  
MICHIGAN'S DIGITAL  
LIBRARY. HERE YOU  
WILL FIND THE LATEST  
NEWS IN WHO WE  
ARE, WHAT WE ARE  
DOING, AND WHERE  
WE ARE GOING.

- [Other Digital Library Sites](#)

[Text Only Version](#)

[About the KNC](#)

[Facilities](#)

[Services](#)

[Guides and Tutorials](#)

[Workshops](#)

[Site Map](#)



- Stop by and check our Lost and Found!
- Our spring and summer hours will remain 11-5 M-F

Recent additions to the site:

[Search feature available](#)

---

Last updated on April 26, 2000

[knc-info@umich.edu](mailto:knc-info@umich.edu)



University of Michigan Library.

Copyright © 2000. The Regents of the University of Michigan. All rights reserved.



# Guides and Tutorials

**about the knc facilities services guides workshops contact us search site map**

- [Database and Info Retrieval](#)
- [Graphics and Layout](#)
- [Internet Communications](#)
- [Miscellaneous](#)
- [Text and Presentation](#)
- [Web Site Creation](#)



You may need to download [Adobe Acrobat Reader](#) to view documents marked "PDF."

## Database and Information Retrieval

- [Bibliographic Management Information Pages: Using EndNote and ProCite](#)
- [Getting Started with EndNote: An Introduction to EndNote Features and Commands](#) (PDF 87K)
- [Getting Started with ProCite 4: An Introduction to ProCite Features and Commands](#) (PDF)
- [Getting Started with ProCite 5: An Introduction to ProCite Features and Commands](#) (PDF)

## Graphics and Layout

- [Adobe Photoshop Workshop and Tutorial](#)
- [Images Workshops](#)

## Internet Communications

- [Accessing IFS Space via Chooser from a Macintosh](#) (PDF 27K)
- [Using FETCH on Macintosh to transfer files](#) (PDF 27K)
- [Using WS\\_FTP to transfer files on Windows 95/NT](#)
- [Sending and Receiving E-Mail Attachments](#)

## Miscellaneous

- [Frequently Asked Questions](#)
- [Technical Glossary](#)
- [Copyright on the Internet](#)
- [Ethical and Legal Use of Digital Media](#)
- [Hoaxes, Chain Letters and Urban Legends](#)

## Text and Presentation

- [OCR Text Scanning at the KNC](#)
- [Microsoft PowerPoint Tutorial](#)

## Web Site Creation

- [Developing a Course Web Page](#)
- [Developing an Interactive Web Site](#)



[knc-info@umich.edu](mailto:knc-info@umich.edu)

Last updated on May 11, 2000

Copyright © 2000

The Regents of the University of Michigan.

All rights reserved.



University of Michigan Library

# Making of America



**M**aking of America (MOA) is a digital library of primary sources in American social history from the antebellum period through reconstruction. The collection is particularly strong in the subject areas of education, psychology, American history, sociology, religion, and science and technology. The collection currently contains approximately 1,600 books and 50,000 journal articles with 19th century imprints. The project represents a major collaborative endeavor in preservation and electronic access to historical texts.

**MOA**  
*about*



The Making of America collection is made up of images of the pages in the books and journals. When you find something you want to look at, you will see a scanned image of the actual pages of the 19th century volume. Optical Character Recognition (OCR) has been performed on the images to enhance searching and accessing the texts -- for more on the OCR process see [About MOA](#). A [small, but growing, group of texts](#) has also been fully processed and can be viewed either as page images or electronic text.



In the next two years, we will be adding about 7,500 more volumes to the Making of America. The first additions should appear in January, 2000.

Making of America is made possible by a grant from the Andrew W. Mellon Foundation.

Making of America is best viewed with a frames-capable browser.

---

Current online holdings:

**Pages:** 634,068

**Volumes:** 4,058

---

[Search](#) || [Advanced Search](#) || [Browse](#) || [About](#) || [Help](#)

---

© 1996 MOA. Comments and questions to [moa-feedback@umich.edu](mailto:moa-feedback@umich.edu).

## D-Lib Magazine October 1999

Volume 5 Number 10

ISSN 1082-9873

# Reference Linking in a Hybrid Library Environment

## Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment

Herbert Van de Sompel  
Los Alamos National Laboratory - Research Library  
[herbert.vandesompel@rug.ac.be](mailto:herbert.vandesompel@rug.ac.be)

Patrick Hochstenbach  
Automation Department of the Central Library  
University of Ghent, Belgium  
[patrick.hochstenbach@rug.ac.be](mailto:patrick.hochstenbach@rug.ac.be)

### Abstract

This is the third part of our papers about reference linking in a hybrid library environment. The [first part](#) described the state-of-the-art of reference linking and contrasted various approaches to the problem. It identified static and dynamic linking solutions, open and closed linking frameworks as well as just-in-case and just-in-time linking. The [second part](#) introduced SFX, a dynamic, just-in-time linking solution we built for our own purposes. However, we suggested that the underlying concepts were sufficiently generic to be applied in a wide range of digital libraries.

In this third part we show how this has been demonstrated conclusively in the "SFX@Ghent & SFX@LANL" experiment. In this experiment, local as well as remote distributed information resources of the digital library collections of the [Research Library of the Los Alamos National Laboratory](#) and the [University of Ghent Library](#) have been used as starting points for SFX-links into other parts of the collections. The SFX-framework has further been generalized in order to achieve a technology that can easily be transferred from one digital library environment to another and that minimizes the overhead in making the distributed information services that make up those libraries interoperable with SFX.

This third part starts with a presentation of the SFX problem statement in light of the recent discussions on reference linking. Next, it introduces the notion of global and local relevance of extended services as well as an architectural categorization of open linking frameworks, also referred to as frameworks that are supportive of selective resolution. Then, an in-depth description of the generalized SFX solution is given.

## Rephrasing the SFX problem statement

### The problem statement

It is relevant to rephrase the SFX problem statement in the context of the meetings and the subsequent reports and publications on reference linking organized by the [Digital Library Federation](#) (DLF), the [National Information Standards Organization](#) (NISO), the [National Federation of Abstracting and Indexing Services](#) (NFAIS), and the [Society for Scholarly Publishing](#) (SSP) (Caplan 1999a; Caplan 1999b; Caplan & Arms 1999; Needleman 1999).

The generic statement of the reference linking problem, as defined by the working group on reference linking was (Caplan 1999a; Caplan & Arms 1999):

*Given the information in a standard citation, how does one get to the thing to which it refers?*

However, the working group concentrated on a specific variation on this:

*Given the information in a citation to a journal article, how does a user get from the citation to an appropriate copy of the article?*

The SFX research also addresses these problems, but only as an instance of a more general problem that can be formulated as:

*Given bibliographic metadata, how does one present relevant extended services for it?*

## Bibliographic metadata as a starting point

Clearly, the SFX research is not only concerned about information in a standard citation. Its starting point is bibliographic metadata in general. As such, information entities originating from typical scholarly resources such as records from abstracting & indexing databases, OPAC systems and preprint archives can be used as a starting point in the SFX problem statement. This is also the case for citations to both journal articles and books found in journal articles or books. But even fractional bibliographic metadata such as an author name taken from an e-mail message is a valid starting point in the SFX problem statement.

## Extended services as a goal

A similar generalization holds for the target of the problem statement since the SFX research is not only concerned about linking to the full-text that corresponds to a citation in a journal article. It aims at the presentation of a variety of extended services for whichever metadata is used as a starting point. Extended services are services that present an information entity in a digital library -- defined as the link-source -- in the context of the entire information environment ([Van de Sompel & Hochstenbach 1999a](#)). For instance, for a given link-source record from an abstracting & indexing database, extended services can -- amongst others -- be the presentation of:

- the full-text of the paper that is abstracted in the link-source;
- a record abstracting the same publication taken from another abstracting & indexing database;
- citation information corresponding with the link-source;
- library holdings for the journal in which the article described by the link-source appeared.

## Global and local relevance of extended services

The adjective *relevant* is of particular importance in the notion *relevant extended services* as used in the SFX problem statement. It actually has two meanings: relevance as a global notion and relevance as a local notion. In order to explain this, the following types of extended services are considered:

- *full\_text*: a service providing the full-text that is referred to by a link-source;
- *review*: a service showing a book review for the item referred to by a link-source;
- *abstract*: a service that provides the abstract from an abstracting & indexing database for a link-source.

Relevant as a global notion must be interpreted as being opposed to irrelevant in every context. Certain aspects of extended services are independent of the context of an individual collection; they actually apply on a global level:

- *full\_text*: If the publication year of an article is equal to or higher than that of the first electronic issue of the journal in which the article was published, a *full\_text* service has global relevance. On the other hand, it never makes sense to present a *full\_text* service for a link-source referring to a paper in a journal if the publication year of the paper is lower than the publication year of the first issue of the journal for which full-text is globally available. As such, the publication year of the first electronic issue is a constraint of global significance to the *full\_text* service.
- *review*: It is always irrelevant to present a book *review* service if the link-source refers to a journal article. But, if the link-source describes a book, such a *review* service is globally relevant. In this case, the material type is a constraint of global significance for the *review* service.
- *abstract*: A constraint of global significance rules the relevance of an *abstract* service that looks up the abstract of a citation to a journal article in a particular abstracting & indexing database. Such a service is globally relevant if the journal in which the article is published is actually indexed in that abstracting & indexing database and is globally irrelevant otherwise.

Relevant as a local notion, refers to the fact that other aspects of extended services are dependent on the boundaries of a certain digital library collection. Local relevance has two manifestations:

- Relevance related to the content of a local collection:

While certain services are relevant in a global sense, they can become irrelevant if the digital library collection does not contain the information resource(s) required to implement them. Even if a *full-text* service is globally relevant for a certain link-source, it might be considered to be irrelevant in the context of a certain digital library collection if the journal referred to by the link-source is not part of that collection. In the same way, an *abstract* service pointing to a particular abstracting & indexing database for a given link-source can be globally relevant, as described above. Still, such a service is of no local relevance if the user's digital library does not provide access to an implementation of that particular database, while it can be of local relevance if the digital library does.

- Relevance related to the implementation of a local collection:

The relevance of extended services will also depend on the technical implementation of the information resource(s) required to create the services. When a *full\_text* service is globally relevant -- an electronic edition of an article exists -- as well as relevant in relation to the content of a certain collection -- the users of the digital library are authorized to access the electronic edition -- it can be regarded inappropriate to let the *full\_text* service link to a full-text instance at a publisher's site, when the digital library holds an instance in its local storage. In the DLF reference linking discussion,

this issue was given the name of "the Harvard problem" ([Caplan 1999a](#)). Similar problems occur in the broader scope of extended services. For instance, as shown before, an *abstract* service can be globally relevant -- the journal in which an article was published is abstracted in a particular abstracting & indexing database -- as well as relevant in relation to the content of the collection -- the local digital library does provide access to the particular database. Still, the service might be irrelevant in relation to the implementation, if the actual implementation of the database does not support a mechanism to link into it using the parameters required to do an *abstract* look-up.

Systems supportive of selective resolution

Both issues regarding the local relevance of extended services indicate the need for open linking solutions that take the context of the local collection into account when links are presented to a user ([Van de Sompel & Hochstenbach 1999a](#)). When addressing the Harvard problem the DLF reference linking discussions have referred to open linking solutions as being supportive of selective resolution ([Caplan & Arms 1999](#)). From the above, it can be seen that the problem of local relevance of extended services is actually a generalization of certain aspects of the Harvard problem. As such, when a framework is able to present an approach to deal with the broader problem, the approach will also contain valuable elements to address the narrower Harvard problem.

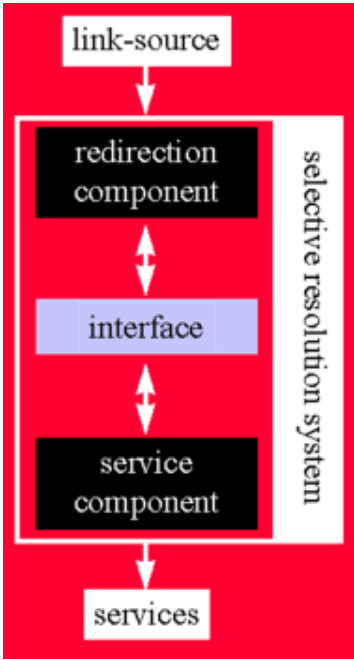


Figure 1: Systems supportive of selective resolution

In relation to the Harvard problem, Caplan and Arms divide systems that support selective resolution into two categories:

- Systems with a non-local location database, to which institutions provide a profile describing their full-text collection. The profile controls the selection of links returned to users of that profile.
- Systems with an institutional location database describing the local full-text collection and a global location database as a fall back. In addition to that, there is a mechanism to pass resolution requests to the local resolver first and in the event of a local full-text instance not being found there, to the global resolver.

This categorization can further be generalized by:

- Broadening the scope of the services to be provided beyond the restriction to the full-text, taking into account all kinds of extended services.
- Identifying the crucial components of systems supportive of selective resolution (see [Figure 1](#)):
  - The redirection mechanism that brings metadata of the link-source for which extended services are requested from the information resource to which the link-source belongs to the service component. The redirection mechanism addresses the problem that has been referred to as grabbing the link-source ([Van de Sompel & Hochstenbach 1999a](#)).
  - The service component that takes metadata from whichever information resource in the digital library collection as an input, delivering extended services as an output. The service component is an extension of the location database referred to by Caplan and Arms.
- Recognizing that the order of the redirection is subject to variation:
  - Redirection of the link-source metadata to the local service component first, using a central service component as a means to complete the set of services that can be presented.
  - Redirection of the link-source metadata to the central service component, whose default services can be overwritten and/or completed after communication with the local service component.

CATEGORY			
Category 1		central	central
Category 2	a	central & local	local => central

	b	central & local	central => local
Category 3		local	local
		<b>SERVICE COMPONENT</b>	<b>REDIRECTION ORDER</b>

**Table 1: categorization of systems supportive of selective resolution**

The resulting categorization is represented in [Table 1](#), where 3 main categories of systems supporting selective resolution are shown, based on the nature of the service component and the redirection order:

- Category 1 only has a central service component and hence a central redirection mechanism. To some extent, this is the category under which the NCBI [LinkOut](#) solution resides. Still, since that solution is tied in with the PubMed database and cannot be used in connection with other resources, it can hardly be seen as a real service component in the sense described earlier.
- Category 2 has both a central and a local service component that contribute to the presentation of the services. Also, there is some form of communication between both. For this Category, it is possible to imagine both approaches regarding the redirection order mentioned above.
- Category 3 builds purely on a local service component and hence also needs a local redirection mechanism. The SFX implementations of both the Elektron and "SFX@Ghent & SFX@LANL" experiments fall within this Category.

## The "SFX@Ghent & SFX@LANL" experiment

In the "SFX@Ghent & SFX@LANL" experiment (April 1999 - June 1999; henceforth referred to as Ghent&LANL), the [Library Without Walls](#) team of the [Research Library at the Los Alamos National Laboratory \(LANL\)](#) and the Automation Department of the [Central Library at the University of Ghent](#) have cooperated to illustrate the feasibility of the SFX approach as a means to provide extended services in a realistic and complex information environment.

The information environment in which Ghent&LANL has been conducted is dramatically different from the one of the first Elektron SFX experiment. To illustrate, [Table 2](#) presents an overview of the information resources used in Ghent&LANL. The rows show the names of the information resources used in the experiment, the columns refer to the digital library collection. For each resource/collection combination the table indicates:

- The Type of resource: OPAC system, abstracting & indexing database (A&I), full-text collection (FTXT) or web-service (WWW);
- The Authority running the resource;
- Whether within the digital library collection, the resource is used as a Source. If so, information entities from the resource can be link-sources for which extended services can be requested. If a resource is a Source, the authority running it has made it SFX-aware;
- Whether within the digital library collection, the resource is used as a Target. If so, the resource is used to be linked into in order to provide extended services. If a resource is a Target, a link-to syntax has been developed by the authority running the resource, in order to allow for it to be the Target of dynamic SFX-links.

RESOURCE		GHENT			LANL		
	Type	Authority	Source	Target	Authority	Source	Target
Advance	OPAC	-	-	-	LANL	yes	yes
Aleph 500	OPAC	Ghent	yes	yes	-	-	-
Amazon.com	WWW	Amazon	no	yes	Amazon	no	yes
Antilope	OPAC	UA	no	yes	-	-	-
APS PROLA	FTXT	APS	yes	yes	APS	yes	yes
the arXiv	FTXT	LANL	yes	yes	LANL	yes	yes
BIOSIS	A&I	Ghent	yes	no	LANL	yes	no
Books in Print	A&I	Ghent	yes	yes	Ghent	yes	yes
Compendex	A&I	Ghent	yes	no	LANL	yes	no

Current Contents	A&I	Ghent	yes	yes	Ghent	yes	yes
EconLit	A&I	Ghent	yes	no	-	-	-
Genome base	A&I	NCBI	no	yes	NCBI	no	yes
Inspec	A&I	-	-	-	LANL	yes	no
		SP	no	yes	SP	no	yes
Ulrich's	A&I	Ghent	yes	yes	-	-	-
LiSa	A&I	Ghent	yes	yes	-	-	-
MathSci	A&I	Ghent	yes	no	-	-	-
Medline	A&I	Ghent	yes	no	-	-	-
		NCBI	no	yes	NCBI	no	yes
SciSearch	A&I	LANL	yes	yes	LANL	yes	yes
ScienceServer	FTXT	LANL	no	yes	LANL	no	yes
Various	FTXT	various	no	yes	various	no	yes
Wiley InterScience	FTXT	Wiley	yes	yes	Wiley	yes	yes

Table 2: information resources in Ghent&amp;LANL

Some considerations regarding [Table 2](#):

- As can be seen, some resources are available in both digital library collections, but run on different technical implementations. This is the case for BIOSIS and Compendex, which in Ghent run on a [SilverPlatter](#) ERL platform, while LANL -- at the time of the experiment -- used a [Geac](#) Advance implementation.
- For the purpose of this experiment, Ghent and LANL share some of their resources. Ghent makes its SilverPlatter ERL version of Books in Print and Current Contents available for LANL, whereas LANL opens access to its Topic implementation of the [ISI](#) Science Citation Index (SciSearch) and its [ScienceServer](#) storing the full-text of all Elsevier journals.
- Ghent uses two Medline versions: a locally stored ERL version as Source and the NCBI PubMed version as Target. Similarly, LANL uses two Inspec versions: the local [Geac](#) Advance implementation as a Source and an ERL implementation run by [SilverPlatter](#) in Boston as a Target. Time constraints that prevented the development of appropriate link-to syntaxes for the local versions are the reason for this peculiarity.
- Of special importance is the fact that some journals from the [Wiley](#) InterScience collection as well as the complete [PROLA](#) archive of the [American Physical Society](#) are made SFX-aware ([Halstead 1999](#); [Spilka 1999](#)). Both Ghent and LANL can use citations in the full-text of these repositories as link-sources for SFX requests. Also, in the course of this experiment Wiley has implemented a link-to syntax that will be brought into production later in 1999. For the PROLA archive such a link-to syntax was already available.
- Some resource names require a little more explanation. [Aleph 500](#) is the Ghent Integrated Library System, [Advance](#) is the one for LANL, while [Antilope](#) is the Belgian Union Catalogue of Serials run by the University of Antwerp. The header "various" refers to a variety of full-text repositories to which dynamic links are available in this experiment. This is -- amongst others -- the case for [Academic Press](#), [Company of Biologists](#), [HighWire](#), [Springer](#), [American Chemical Society](#), etc. The [arXiv](#) is the Topic implementation of the Ginsparg arXiv e-print repository, developed by the [Library Without Walls](#) team of the LANL Library. It has also been made SFX-aware.
- As can be seen from careful exploration of [Table 2](#), from the point of view of each digital library collection, the SFX-aware information resources are highly distributed. Some resources are run by the institutional library automation team while others are run remotely, actually by three external authorities. From the point of view of Ghent these authorities are LANL, Wiley and the American Physical Society; from the point of view of LANL, they are Ghent, Wiley and the American Physical Society.

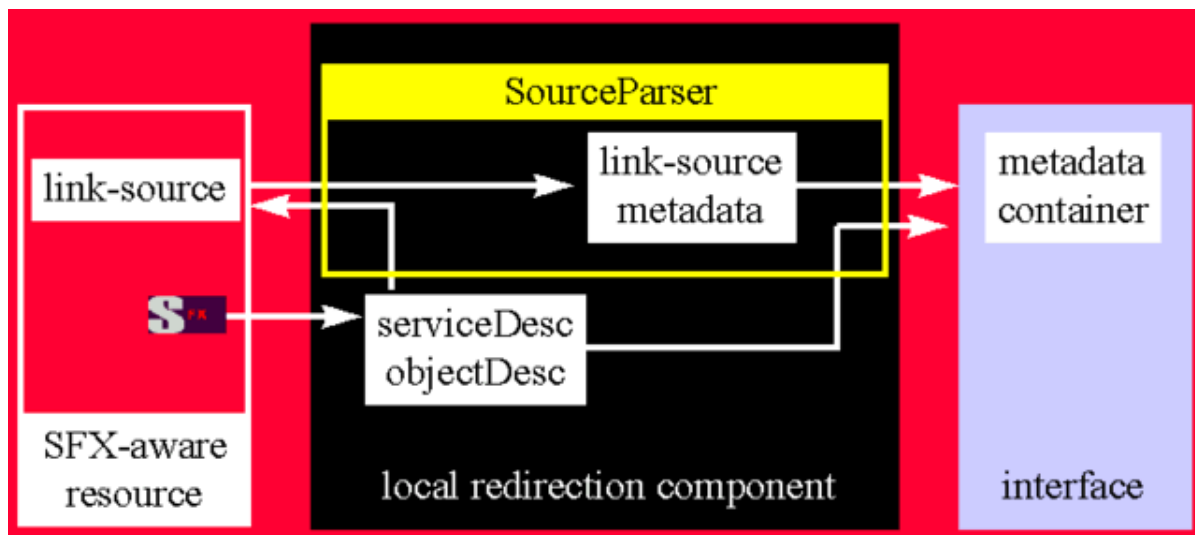
From the above, it can be concluded that from the point of view of the amount of resources that are involved, and given their distributed nature and the availability of multiple SFX service components, Ghent&LANL is a very realistic experiment.

## The need for a generalization of the SFX components

Although the fundamental concepts of SFX -- dynamic linking, just-in time linking and conceptual services (see ([Van de Sompel & Hochstenbach 1999b](#))) -- have been left untouched for the Ghent&LANL experiment, the nature of its working environment and its goals have led to a strong generalization of the SFX components. The main impulses that inspired such a generalization and that distinguish the Ghent&LANL project from the Elektron experiment are:

- The extension of the digital library collection in which SFX was being tested beyond a well-controlled sub-collection of one institution. In Ghent&LANL, SFX is introduced in the realistic, complex and dissimilar digital libraries of two autonomous institutions each running their local SFX components;
- The extension of the scope of data for which extended services can be requested beyond the internally stored collections. Link-sources in Ghent&LANL also originate from resources held by external authorities;
- The extension of the datatypes for which extended services can be requested beyond abstracting & indexing databases and OPAC systems. Link-sources in Ghent&LANL can also be citations in journal articles;
- The accommodation of extended services linking into target resources, based on metadata in general, not only SICI-related metadata;
- The need for high transportability of the SFX solution between the digital library environments that are involved.

The redesign of the SFX solution for Ghent&LANL leads to an architecture with a clear separation between the redirection component and the service component. Both components obviously interoperate in order to achieve a functional system. But the redirection component can potentially operate in an environment with non-SFX service components, while the SFX service component can equally function with another redirection mechanism, as long as that supports delivery of link-source metadata to the SFX service component. Several functional building blocks in both components have also been generalized in order to address the problems that arise from the complexity of the Ghent&LANL environment. The overall approach of the generalized solution is shown in [Figure 2](#) and will be explained in more detail in the remainder of this paper. Information resources that can interoperate with SFX -- from now on referred to as SFX-aware systems -- insert an SFX-button for each link-source in the result set of a query. The just-in time approach of SFX requires the user to click such an SFX-button when requesting extended services for a specific link-source record. In response to this click, the local SFX redirection component will fetch link-source metadata -- usually -- from the origin resource using whichever protocol it takes to do so. Next, link-source metadata as well as information on its origin will be converted into an interfacing format. At this point, the local redirection mechanism has fulfilled its task and is able to deliver this information in a consistent representation to the local SFX service component.



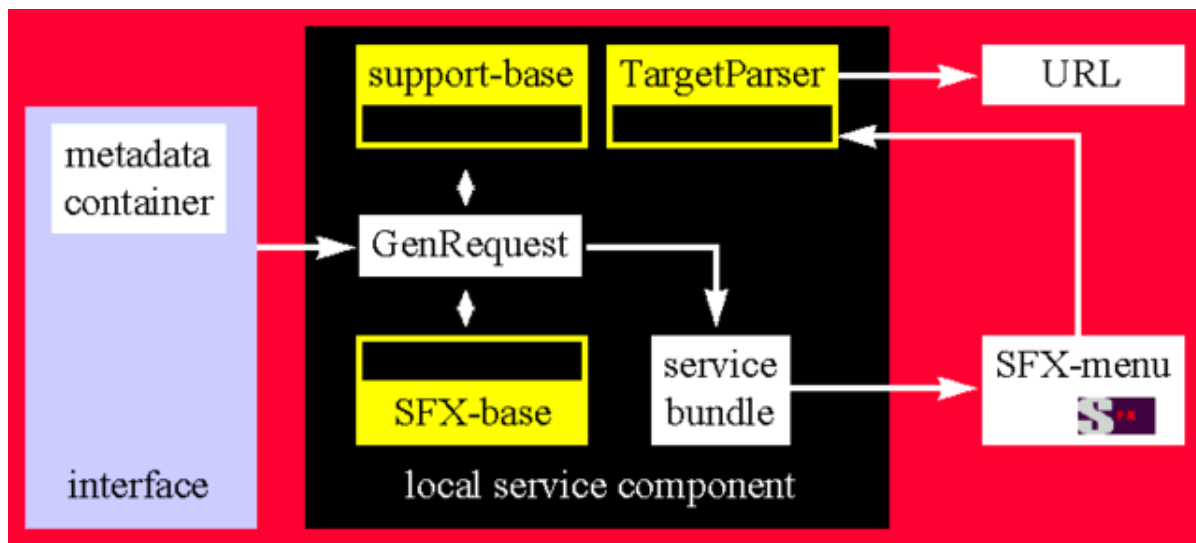


Figure 2: the local redirection and service components of the generalized SFX solution

The first task of the local service component is to parse the information, handed over by the local redirection component, into a normalized internal representation object. During this process, the original content can be enhanced and/or augmented. The resulting information object is then fed into the SFX evaluation process in which it will be compared to the SFX-database. The SFX-database is a special kind of linking database. Unlike traditional linking services, it does not contain any static links between "documents" (records/citations/full-text/etc.) of a collection. Rather, it contains a collection of conceptual services that express potential inter-relationships between documents at the level of the resource from which they originate. The SFX evaluation process determines the relevance of each of these conceptual services using the -- lack of -- content in the information object. Next, the resulting bundle of relevant services is sent back to the user in the SFX-menu-screen. Consistent with the just-in-time approach of SFX, only when the user decides to use a service from the bundle, will the service be resolved into a URL to which the user is being redirected.

## The SFX mechanism for local redirection

The task of the local redirection mechanism is to transport link-source metadata to the local redirection component, that interfaces with the local service component. In order to be able to interoperate with the SFX redirection mechanism, information resources need to be enhanced by the authorities running them in order to make them SFX-aware. The aim of this is to create the ability for information resources to insert an SFX-button targeted at the local redirection component for each link-source in the result set of a query into the resource. In the context of Ghent&LANL, the following are important considerations with this regard:

- a. Many information resources that are involved in the experiment are also used in normal production at the very same time. This means that they are also approached by users that do not have access to an SFX service component. In order to prevent such a user from seeing an irrelevant SFX-button, an SFX-aware resource must be able to recognize whether the user has access to an SFX service component or not. Based on that information, the resource can insert an SFX-button or not.
- b. Some information resources are approached by users from both digital library environments, hence with access to different SFX service components. An SFX-aware resource must be able to target the SFX-button at the appropriate local redirection component, in order for it to be able to deliver the link-source metadata from the origin information resource to the doorstep of the appropriate service component. This means that an SFX-aware resource must be able to parameterize the target of an SFX-button.
- c. Upon receipt of a request for extended services from a user, the local redirection component must be able to fetch the link-source metadata from its origin resource. This means that the local redirection component has to be informed about the origin and the identity of the link-source in order to be able to take the appropriate steps. Given the amount, distribution and diversity of the SFX-aware resources in Ghent&LANL, a consistent manner to communicate such information to the local service components is required.
- d. Link-source metadata must be fetched from a wide variety of distributed information resources that support different access protocols. In addition to that, those resources will respond by sending link-source metadata formatted according to different metadata schemes. In order for the local redirection component to be able to interface in a generic manner with the local service component, a unique metadata interchange format is desirable.

As will be shown, in the detailed description below, these issues are approached by:

- For (a) and (b): the CookiePusher mechanism;
- For (c): the consistent SFX-URL structure;
- For (d): the SourceParser solution.

## Making information resources SFX-aware

The authorities running information resources need to enhance their systems in order to make them SFX-aware. The complexity of the Ghent&LANL environment has urged for a thoughtful exploration of ways to make resources SFX-aware, since only approaches that minimize the overhead in doing so for the authorities running the resources can be acceptable and workable. In the current implementation of the SFX redirection mechanism, they have to do this by:

- Installing the CookiePusher script delivered by the project managers of Ghent&LANL;
- Hyperlinking the SFX-buttons for link-sources using a URL that complies to a predefined format.

### The CookiePusher

The CookiePusher script is a pragmatic solution introduced to dynamically notify an information resource about the existence and location of a local SFX redirection component in the environment of the user consulting the resource. The underlying idea is that an information resource could at any time access the location of a local redirection component, if its URL were written as a cookie in the browser of the user consulting the resource. The availability of this URL is essential, since the resource must be able to dynamically target the SFX-button at the appropriate local component. However, for reasons of security and privacy, such browser cookies can maximally be read within the Internet domain of the server that has set the cookie (see [Shishir 1996](#) pages 203-204). As such, it is impossible to set such a cookie so that it can be read by all information systems in a digital library collection when it consists of resources distributed over several domains, typically resources that are local and remote to the user's institution.

In order to solve this problem, the first step in connecting to a resource is to request a server in the domain of the information resource to create an HTTP cookie. This detour is called the CookiePusher. The very simple CookiePusher script is installed in the domain of the information resource that has to be made SFX-aware. Rather than connecting immediately to the desired URL in the information resource, a connection is made to the resource's CookiePusher first, sending values for the two parameters of the CookiePusher script:

- SFX\_location: the URL of the local redirection component of the SFX solution;
- Redirect: the desired URL in the resource.

Upon receipt of these parameters, the CookiePusher will first read the URL of the local redirection component and will use it to set a cookie in the user's browser. Since the CookiePusher is in the domain of the resource, that cookie will be readable by the resource. Next, the CookiePusher will redirect the user to the desired URL in the resource.

As such, once the CookiePusher has been installed for a resource, the URL to connect to that resource will be changed to:

**CookiePusher\_URL?SFX\_location= local\_SFX& Redirect= service\_URL**

Where

- **CookiePusher\_URL** is the URL of the CookiePusher script;
- **local\_SFX** is the URL of the local SFX redirection component;
- **service\_URL** is the desired URL in the information resource as used under normal -- non-SFX -- conditions. Such a URL can point at the initial search screen for an abstracting & indexing database, it can be a URL linking to an article at a publishers site, etc.;
- **local\_SFX** and **service\_URL** are URL-encoded.

For instance:

```
http://publish.aps.org/edaccess/prolatest/cookiepusher?
SFX_location=http%3A%2F%2Fiserv.rug.ac.be%2Fcgi-bin%2Fsfx%2Fbin%2Fmenu.cgi
&Redirect=http%3A%2F%2Fpublish.aps.org%2Fedaccess%2Fprolatest%2Ftext%2FPRD%2Fv52%2Fi1%2Fp15_1
```

is the URL used to connect to an item in the [APS/PROLA](#) domain. The APS/PROLA CookiePusher will read the location of the local redirection component from the SFX\_location parameter and will use this to set a cookie named local\_SFX with value:

```
http%3A%2F%2Fiserv.rug.ac.be%2Fcgi-bin%2Fsfx%2Fbin%2Fmenu.cgi
```

which is the encoded location of the Ghent local SFX redirection component. Next, it will redirect the user to the desired location in the APS/PROLA:

```
http://publish.aps.org/edaccess/prolatest/text/PRD/v52/i1/p15_1
```

From now on, at any point in the consultation, APS/PROLA will be able to read this cookie and use it to target -- in this case -- the Ghent redirection component.

### The consistent SFX-URL structure

The essence of the detour made via the CookiePusher is the ability it creates for an information resource to know at any point whether the consulting user has access to a selective resolution system and, if so, what the location of its redirection

component is. Based on that information, the resource can dynamically decide whether or not to insert an SFX-button for search results and if it does, which redirection component to target with the SFX-button. In order to make the many systems involved in the Ghent&LANL experiment interoperable with SFX, authorities running the systems have been asked to make the URL targeted by the SFX-button -- the SFX-URL -- compliant to the following format:

GENERAL	target?serviceDesc&objectDesc
DETAILED	local_SFX?vendorId=<theVendor>&databaseId=<theBase>&objectDesc=<theIdentifier>

Table 3: the syntax of the SFX-URL

In [Table 3](#)

- target is the URL of the local redirection component of the SFX solution;
- serviceDesc uniquely defines the origin resource. It contains information on the vendor of the resource and on the resource itself. It is of the form:

vendorId=<theVendor>&databaseId=<theBase>.

serviceDesc information will play a crucial role at later stages of the SFX local redirection mechanism, as well as in the SFX-base which is central to the SFX service component.

- objectDesc contains information that relates to the identity of the link source. Its syntax and content is extremely flexible and it will be defined by the authority running the resource, making it dependent on the vendor and his database implementation. objectDesc typically contains the unique record identifier for a link-source in its origin resource. Alternatively or in addition to that, it can contain SICI-like metadata. In some cases, it can even contain all metadata of the link-source.
- The parameter values <theVendor>, <theBase> and <theIdentifier> are URL-encoded.

[Figure 3](#) to [Figure 6](#) show examples of link-sources taken from Sources in the Ghent and/or LANL collections, mentioning their SFX-URL. For reasons of readability, the parameter values are not shown as being URL-encoded. Rather, it is mentioned that parts should be URL-encoded by enclosing them in a URLEncode function.

Record 2 of 33 in Biological Abstracts 1999/01-1999/03

TI

Long term outcome of patients with hairy cell leukemia treated with pentostatin.

AU

[Ribeiro-Patricia](#); [Bouaffia-Fadhela](#); [Peaud-Pierre-Yves](#); [Blanc-Michel](#); [Salles-Bruno](#)

AD

{a} Serv. Hematol., Cent. Hosp. Lyon-Sud, 165 Chemin du Grand Revoyet, 69495

SO

[Cancer](#). Jan. 1, 1999; 85 (1) 65-71..

PY

1999

DT

Article-

IS

0008-543X

LA

English

AI

Y

ST

Hominidae-; Primates-, Mammalia-, Vertebrata-, Chordata-, Animalia-

RN

53910-25-1: PENTOSTATIN


AN

199900063465

UD

19990223 .

AV



SFX-URL for this link-source, pointing at the Ghent local redirection component:

http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?vendorId=ERL&databaseId=BX


&objectDesc=URLEncode(BX02 A:199900063465 I:0008-543X V:00085 S:000001 P:000065 Y:1999)

In the serviceDesc part of the URL, ERL refers to the SilverPlatter ERL implementation of BIOSIS, while BX is the family name of BIOSIS databases in the ERL environment. The objectDesc component contains several information elements in a tagged and fixed length representation. BX02 is the volume of the BIOSIS database where the link-source originates, while 199900063465 is the accession number, a unique record number of the link-source in BIOSIS. Other elements in the objectDesc are ISSN number, volume, issue, starting page and publication year.

[http://www.dlib.org/dlib/october99/van\\_de\\_sompel/10van\\_de\\_sompel.html](http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html) (9 of 30) [6/1/2000 3:53:29 AM]

Figure 3: a link-source from the Ghent ERL implementation of BIOSIS and its SFX-URL


Record **3** of **91** [Mark](#) [Full Record](#)

**Article:**  [Article](#)

**Title:** [The identification of cDNAs that affect the \*\*mitosis\*\*-to-\*\*interphase\*\* transition in Schizosaccharomyces pombe, including sbp1, which encodes a spilp-GTP-binding protein.](#)

**Author:** [HE, XIANGWEI](#) ; [HAYASHI, NAOYUKI](#) ; [WALCOTT, NATHAN G.](#) ; [AZUMA, YOSHIO](#) ; [PATTERSON, THOMAS E.](#) ; [BISCHOFF, F. RALF](#) ; [NISHIMOTO, TAKEHARU](#) ; [SHELLEY, SHELLEY](#)

**Source:** [Genetics; Feb. 1998; v.148, no.2, p.645-656.](#)

**Other Links:** 

---

Location	Holdings
MAIN	Holdings: v.109- (1985- ) ; Missing: v. 134 no. 4 (1993) ; Last rec'd: VOL.152 NO.2 / J issue on display ; Shelved as: GENETICS.
WWW	<a href="http://www.genetics.org/">http://www.genetics.org/</a> ; Holdings: v.148, no.1- (Jan.1998- ) ; Abstracts and table of contents: v.148, no.2-v.147, no.4 (Feb.1980-Dec.1997)

---

SFX-URL for this link-source, pointing at the LANL local redirection component:

[http://vole.lanl.gov/cgi-bin/sfx/bin/menu.cgi?vendorId=ADVANCE&databaseId=Biosis&objectDesc=URLencode\(fetchId=21179970&objectId=PREV199800135979&SICI=0016-6731\(1998\)148:2<645:TIOCTA>2.0.TX;2-P\)](http://vole.lanl.gov/cgi-bin/sfx/bin/menu.cgi?vendorId=ADVANCE&databaseId=Biosis&objectDesc=URLencode(fetchId=21179970&objectId=PREV199800135979&SICI=0016-6731(1998)148:2<645:TIOCTA>2.0.TX;2-P))

The serviceDesc part of this URL is self-explanatory. The objectDesc component is tagged and fields can have variable lengths. The fetchId is the unique number of the link-source in the LANL implementation of BIOSIS, while the part of objectId after "PREV" is the BIOSIS accession number which is comparable to the A field in the SilverPlatter objectDesc of [Figure 3](#). The SICI part contains a SICI for the link-source, from which ISSN, volume, issue, pagination and publication year can be derived.

**Figure 4: a link-source from the LANL Advance implementation of BIOSIS and its SFX-URL**

**WebSPIRS: Search - Netscape**

**Special Effects: University of Ghent - Netscape**

**Article Abstract - Netscape**

**WILEY InterScience®**

PERSONAL HOMEPAGE JOURNAL FINDER SEARCH HELP CONTACT US LOGOUT

ALL JOURNALS PREVIOUS ARTICLE NEXT ARTICLE

**Article Abstract**

**CANCER**

Online ISSN: 1097-0142 Print ISSN: 0008-543X

**Cancer**

**Volume 85, Issue 1, 1999. Pages: 65-71**

**Original Article**

**Long term outcome of patients with hairy cell leukemia treated with pentostatin**

Patricia Ribeiro, M.D. <sup>1</sup>, Fadhela Bouaffia, M.D. <sup>1</sup>, Pierre-Yves Peaud, M.D. <sup>2</sup>, Michel Blanc

**References**

- 1 Saven A, Piro L. Treatment of hairy cell leukemia. *Blood* 1992; **79**: 1111-20. [Medline](#)
- 2 Jaiyesimi I, Kantarjian H, Estey E. Advances in therapy for hairy cell leukemia. A review. *Cancer* 1993; **72**: 5-16. [Medline](#)
- 3 Saven A, Piro L. The newer purine analogues for the treatment of hairy-cell leukemia. *N Engl J Med* 1994; **330**: 691-7. [Medline](#)

SFX-URL for the third reference as a link-source, pointing at the Ghent local redirection component:

<http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?vendorId=Wiley&databaseId=WIS>

&objectDesc= URLEncode(TYPE=JCIT& SNM=Saven&FNM=A&SNM=Piro&FNM=L&ATL= The newer purine analogues for the treatment of hairy-cell leukemia.&JTL=N Engl J Med &PYR=1994&VID=330&PPF=691&PPL=7)

The serviceDesc component now refers to the Wiley InterScience collection. The objectDesc is tagged and starts with an indication on the material type of the reference -- journal citation in this case -- followed by a tagged repetition of the full citation.

**Figure 5: a link-source from Wiley InterScience and its SFX-URL**

**Preprints Retrieval Results - Netscape**

File Edit View Go Communicator Help

**Los Alamos Research Library**  
NATIONAL LABORATORY

New Search Up Comments  
Mark All List Marks Help

**Preprints**

## Preprints Retrieval Results

23 out of 100245 records matched the query below. 23 records displayed:  
*(fractal <in> tissuab <and> year >= 1991)<and>year <= 1999<and>physics <in> archcode*

Marks	Score	Title, Author, Eprint
<input type="checkbox"/>	0.97	<a href="#">Integers and Fractions</a> Diptiman Sen (Indian Institute of Science, Bangalore, India) physics/9811004 (03 Nov 1998)
<input type="checkbox"/>	0.96	<a href="#">Irrational Numbers of Constant Type --- A New Characterization</a> Manash Mukherjee and Gunther Kerner physics/9706009 (04 Jun 1997)
<input type="checkbox"/>	0.95	<a href="#">Adaptive Ising Model and Bacterial Chemotaxis</a> Yu Shi physics/9901013 (28 Jan 1999)

SFX-URL for the first link-source in the above result screen, pointing at the LANL local redirection component:

`http://vole.lanl.gov/cgi-bin/sfx/bin/menu.cgi?vendorId=LANLTopic&databaseId=arXiv`  
`&objectDesc= URLEncode(fetchId=phys-9811004&objectId=physics/9811004)`

The serviceDesc refers to the LANL Topic implementation of the Ginsparg e-print archive. The fetchId is the unique key for the record in that implementation, while the -- very similar -- objectId is the unique record number in Ginsparg's implementation of the archive. No further metadata is available in the objectDesc.

**Figure 6: a link-source from the arXiv and its SFX-URL**

### Fetching link-source metadata from an SFX-aware information resource with SourceParsers

The CookiePusher mechanism enables a resource to insert an SFX-button for each of the link-sources that are transferred to a user consulting the resource. The structure of the SFX-URL targeted by these SFX-buttons has been made consistent across resources to be of the form `target?serviceDesc&objectDesc`. When a user requests extended services by clicking such an SFX-button, a request is sent to his local SFX redirection component, which will receive serviceDesc and objectDesc values as parameters for the target script. The local component holds a collection of SourceParser scripts with names corresponding to valid serviceDesc's (see Table 4). Having analyzed the serviceDesc information, the target script will launch the appropriate SourceParser. This serviceDesc-specific SourceParser uniquely implements:

- The interpretation of the information contained in the objectDesc parameter based upon the syntax defined by the vendor (see examples in Figure 3 to Figure 6);
- The mechanism to fetch the link-source from its origin resource based on its origin and on the content of its objectDesc. Table 4 shows those fetch mechanisms for the examples of Figure 3 to Figure 6. As can be seen, no real fetching is required for the Wiley citations, since these are completely transferred in the objectDesc part of the SFX-URL. The same technique is used for citations in the PROLA archive. Both the Ghent and LANL BIOSIS implementations deliver some -- SICI related -- metadata in the objectDesc. But since several extended services that SFX aims to deliver require more metadata, a fetch is required in order to obtain more complete information. Since the objectDesc for the arXiv only contains an identifier, a fetch is definitely required;
- The conversion of the fetched link-source metadata, that is expressed in the metadata scheme supported by the authority running the origin resource, into a metadata container compliant with the scheme of the unique metadata interchange format. This metadata container is the interface between the local redirection component and the local service

RESOURCE	serviceDesc		SourceParser	Fetch protocol	Fetch key
the arXiv	LANLTopic	arXiv	S::LANLTopic:arXiv	HTTP	fetchId
BIOSIS	ERL	BX	S::ERL::BX	Z39.50	A
BIOSIS	ADVANCE	Biosis	S::ADVANCE::Biosis	Z39.50	fetchId
Wiley	Wiley	WIS	S::Wiley::WIS	none	none

**Table 4: Some SFX-aware resources with their serviceDesc, Fetch protocol and Fetch key**

## The SFX service component

The task of the local SFX service component starts at the point where the local redirection mechanism hands over the metadata container that contains, in a consistent representation:

- link-source metadata that became available through the local redirection mechanism;
- information on the origin of the link-source, basically serviceDesc information.

It is the task of the SFX service component to deliver extended services based on this information. The following are important considerations regarding the SFX service component in Ghent&LANL:

- The amount and quality of link-source metadata that becomes available in the metadata container is dependent on the type of resource from which its link-source originated and on the amount of information that the authority running the origin resource allows and/or supports to be fetched. In some cases such metadata can be corrupt or lack information that is essential for the SFX evaluation process to adequately perform its task;
- The SFX service component must be easily transportable between different digital library environments and remain easily manageable;
- The SFX service component must ultimately deliver service links in a just-in-time manner.

As can be seen from a detailed description of the SFX service component, these problems have been approached by:

- For (a): the GenericRequest object;
- For (b): a generalization of the implementation of the SFX-database, that explicitly reflects the notion of global and local relevance of conceptual services as well as the notion of global and local Thresholds;
- For (c): the TargetParser solution.

### The GenericRequest object

The service component will take the metadata container delivered by the local redirection mechanism as input and turn it into a normalized internal representation, called the GenericRequest object. [Table 5](#) shows a representation of the GenericRequest object for the third citation in [Figure 5](#). The GenericRequest object is an intelligent object, that is able to self-check the validity of its information elements based on pre-configured rules. It can also augment/enhance its content using information from a supporting database. For instance, the citation of [Figure 5](#) does not contain an ISSN number nor a journal title, but rather an abbreviated journal title. In this case, the GenericRequest object augments its content, by adding the missing information via communication with a supporting database. Obviously, the GenericRequest object also contains a normalized version of the link-source metadata, as well as information about its origin.

At the time of the experiment, interoperability between the SFX local service component and non-SFX local redirection mechanisms was not an issue, since none were existing. As such, for reasons of simplicity, the metadata scheme of the GenericRequest object has fulfilled the role of interfacing metadata scheme between the local redirection and the local service component in Ghent&LANL.

```
<perldata>
<hash>
<item key="rec$vendorId">Wiley</item>
<item key="rec$databaseId">WIS</item>
<item key="rec$dbId">Wiley:WIS</item>
<item key="objectType">JOURNAL</item>
<item key="@abbrevTitle">
<array>
<item key="0">N ENGL J MED</item>
</array>
</item>
<item key="journalTitle">NEW ENGLAND JOURNAL OF MEDICINE</item>
<item key="ISSN">0028-4793</item>
<item key="year">1994</item>
<item key="volume">330</item>
<item key="startPage">691</item>
<item key="endPage">7</item>
<item key="@authLast">
<array>
<item key="0">Saven</item>
<item key="1">Piro</item>
</array>
</item>
<item key="@authInit">
<array>
<item key="0">A</item>
<item key="1">L</item>
</array>
</item>
<item key="articleTitle">The newer purine analogues for the treatment of hairy-cell
leukemia.</item>
</hash></perldata>
```

Table 5: Representation of an augmented GenericRequest object for the link-source of [Figure 5](#)

The SFX linking service and the SFX-base

As a result of the above, an instance of the GenericRequest object for the link-source for which extended services have been requested has become available to the SFX service component. It will be the task of this component to deliver the extended services to the user that has requested them. In this sense, the SFX service component is a linking service that, given a certain input "document", outputs "documents" related to the input. The SFX linking service is special, however, since it does not store static relationships between individual documents. Rather, it stores relationships between the resources from which the documents originate. In SFX, these relationships are called conceptual services and they are stored in the SFX-base. The SFX evaluation process will determine the relevance of each of these conceptual services based upon the information and origin of a link-source.

The requirement imposed on the Ghent&LANL implementation of the SFX service component to be easily transportable between different digital library environments has led to an important generalization of the design of the SFX-base. This has been achieved by explicitly reflecting the notion of global and local relevance of services in the implementation. A synthesized representation of the lay-out of the Ghent&LANL SFX-base is given in [Figure 7](#).



Figure 7: Simplified lay-out of the SFX-base

### Splitting the Colli table

As in the Elektron version of the SFX-base, the Source table contains the information resources that can be origins for link-sources. They are SFX-aware resources. In the Elektron version, the Colli contained conceptual services, directly coupled with the Target resources. (see [Table 2](#) in ([Van de Sompel & Hochstenbach 1999b](#))). Such a set-up was not adequately generic and, in the current design, this Colli has been split. One table has kept the name Colli, the other has been named the Target table. The Target table contains those resources into which linking is possible. The Colli table that connects the Source and Target tables now expresses the type of service that relates Source with Target resources. [Table 6](#) shows the type of services implemented in Ghent&LANL.

COLLI SERVICES	FUNCTION
<i>abstract</i>	look-up of abstract information in an abstracting & indexing database for the item represented by the GenericRequest object
<i>author</i>	look-up of references by an author of the item represented by the GenericRequest object in an abstracting & indexing database
<i>cited_author</i>	look-up of citations to work by an author mentioned in the GenericRequest object
<i>cited_reference</i>	look-up of works citing the item represented by the GenericRequest object
<i>full_text</i>	link to the full-text of the item represented by the GenericRequest object
<i>genome</i>	look-up of sequence information found in the GenericRequest object
<i>holding</i>	holdings look-up in an OPAC system for the item represented by the GenericRequest object
<i>review</i>	look-up of a book review for then item represented by the GenericRequest object

Table 6: Services in the Colli and their function

### Taking advantage of the global relevance of conceptual services

It is not a coincidence that the resources shown as Source and/or Target carry their globally common names rather than those of their local implementations in Ghent or LANL. This is actually a reflection of the conclusion that services relating Source and Target resources have global relevance. It is globally relevant to deliver an *abstract* service that, given a link-source from BIOSIS shows the corresponding abstract from Medline. Such a conceptual service can be imagined regardless of the implementations of each of these resources in a specific digital library. Therefore, the Ghent&LANL SFX-base expresses the relationships between Sources and Targets at the level of global relevance: there is an *abstract* service connecting BIOSIS and Medline, regardless of their local implementations. A very limited number of examples of how such services of global

relevance connect Source and Target is shown in [Table 7](#).

COLLI		
SOURCE	SERVICE	TARGET
APS/PROLA	<i>abstract</i>	Inspec
the arXiv	<i>author</i>	Inspec
BIOSIS	<i>abstract</i>	Medline
BIOSIS	<i>genome</i>	Genome Base
Current Contents	<i>abstract</i>	LiSa
EconLit	<i>review</i>	Books in Print
Inspec	<i>full_text</i>	Springer
Wiley	<i>abstract</i>	Medline
Wiley	<i>cited_reference</i>	Science Cit. Base

**Table 7: Examples of service relationships between Sources and Targets**

### **Localization of services of global relevance**

While the services shown in [Table 7](#) are of global relevance, they do not take into account issues of relevance in relation to the local digital library collection. This localization of services of global relevance is achieved by:

- The introduction of fields referring to the local implementations, next to the globally common names.

As shown in [Table 8](#) and [Table 9](#), a key reflecting the serviceDesc values of the local implementations of resources -- found in the rec\$dbId field of the GenericRequest object -- is added next to the global common name of the Sources. In the same way, at the Target side, the name of a local TargetParser is added next to the global name of which the local Target is an implementation. The TargetParser procedure implements the link-to syntax into the local implementation of the Target resource. It can be seen from [Table 8](#) and [Table 9](#) that Ghent and LANL use a different SourceParser for BIOSIS, which reflects that they have a different implementation. However, they share a TargetParser to provide the *abstract* service into Medline, since both have chosen the PubMed implementation as a Target to achieve this.

- Deactivating services of global relevance when they are not of local relevance.

When the Source or Target resource required to implement a certain service is not available in the digital library collection, when the local implementation of the Target resource does not support the link mechanism required to implement the service, or when local librarians decide the service to be of no use to their end-users, its flag will be set to inactive. The service will no longer be taken into account in the SFX evaluation process deciding on the local relevance of conceptual services. In [Table 8](#) this is the case for services with Inspec as a Source since Ghent does not have an Inspec implementation in its collection. In [Table 9](#), this is the case for services with LiSa as a Target, since LANL does not have access to a LiSa implementation.

SOURCE		COLLI	TARGET	
local	global		global	local
S::APS::PROLA	APS/PROLA	<i>abstract</i>	Inspec	T::ERL::IN
S::LANLTopic:arXiv	the arXiv	<i>author</i>	Inspec	T::ERL::IN
<b>S::ERL::BX</b>	<b>BIOSIS</b>	<i>abstract</i>	<b>Medline</b>	<b>T::NCBI::PubMed</b>
<b>S::ERL::BX</b>	<b>BIOSIS</b>	<i>genome</i>	<b>Genome Base</b>	<b>T::NCBI::Genome</b>
S::ERL::CCO	Current Contents	<i>abstract</i>	LiSa	T::ERL::LI
S::ERL::EC	EconLit	<i>review</i>	Books in Print	T::ERL::BOIP

<b>inactive</b>	Inspec	<i>full_text</i>	Springer	T::Springer::LINK
S::Wiley::WIS	Wiley	<i>abstract</i>	Medline	T::NCBI::PubMed
S::Wiley::WIS	Wiley	<i>cited_reference</i>	Science Cit. Base	T::CIC15:SciSearch

**Table 8: Localization of services from [Table 7](#) for Ghent**

Source		Colli	Target	
local	global		global	local
S::APS::PROLA	APS/PROLA	<i>abstract</i>	Inspec	T::ERL::IN
S::LANLTopic:arXiv	the arXiv	<i>author</i>	Inspec	T::ERL::IN
<b>S::Advance::Biosis</b>	<b>BIOSIS</b>	<i>abstract</i>	<b>Medline</b>	<b>T::NCBI::PubMed</b>
<b>S::Advance::Biosis</b>	<b>BIOSIS</b>	<i>genome</i>	<b>Genome Base</b>	<b>T::NCBI::Genome</b>
S::ERL::CCO	Current Contents	<i>abstract</i>	LiSa	<b>inactive</b>
<b>inactive</b>	EconLit	<i>review</i>	Books in Print	T::ERL::BOIP
S::Advance::Inspec	Inspec	<i>full_text</i>	Springer LINK	T::Springer::LINK
S::Wiley::WIS	Wiley	<i>abstract</i>	Medline	T::NCBI::PubMed
S::Wiley::WIS	Wiley	<i>cited_reference</i>	Science Cit. Base	T::CIC15:SciSearch

**Table 9: Localization of services from [Table 7](#) for LANL**

### Global and local Thresholds

The relationships between Source and Target resources expressed by a service connection in the Colli is made subject to restrictions called Thresholds. These Thresholds are the way to fine-tune conceptual services in order to minimize the presentation of services that are considered not to be appropriate to be presented. In order to illustrate this concept, two types of Thresholds are described:

- Thresholds expressed in terms of boundaries for the metadata elements that make up the GenericRequest object structure. Technically, these Thresholds are expressed as conditional statements using field names of the GenericRequest object. Such Thresholds are in many cases very simple, but they can as well be scripts of whichever degree of complexity. For instance:
  - *cited\_author*: In order for a *cited\_author* service to be relevant, the lowest Threshold that has to be passed is the existence of an author name in the GenericRequest object. Such a Threshold could be expressed as \$GenericRequestObject->need('authLast').
  - *book\_review*: A *book\_review* service is only relevant for link-sources that describe books: \$GenericRequestObject->need('objectType', 'eq', 'BOOK').
  - *genome*: A *genome* service is only relevant if the link-source contains genome sequence identifiers: \$GenericRequestObject->need('genID')
  - *abstract*: The Threshold for an *abstract* service might express that the link-source should describe a journal article and should at least have year, volume and issue information: \$GenericRequestObject->need('objectType', 'eq', 'JOURNAL') && \$GenericRequestObject->need('year') && \$GenericRequestObject->need('volume') && \$GenericRequestObject->need('issue').
- objectLookup Thresholds: The *abstract* service is clearly also subject to another type of boundary requiring that the Target resource into which the abstract service intends to link, actually abstracts the journal in which the item referred to by the GenericRequest object was published. This requirement explains the existence of the Objects table in [Figure 7](#) and of a special objectLookup threshold. This type of Threshold will also be required to determine whether a journal in which the item referred to by link-source was published is part of a specific full-text repository in order to decide on the relevance of a *full\_text* service into the repository.

Just as with the conceptual services, there is a global and a local component to these Thresholds. The global objectLookup

Threshold for a *full\_text* service linking into the [Springer](#) full-text collection, will learn whether a certain journal is a Springer e-journal or not. The local part of this Threshold will learn whether the journal is part of the actual digital library collection. In the same context, a global Threshold can express the fact that a journal is available in electronic form since 1996, while the local component might show that the local subscription only starts in 1998. In the same way, the BIOSIS-*abstract*-Medline service is subject to a global objectLookup Threshold. But there is also a global Threshold expressing that the publication year of the GenericRequest object has to be greater than 1965, reflecting the full coverage of Medline. Still, the local Threshold component for this service might be set to a more recent year if the local Medline implementation stores less data.

## The SFX evaluation process

In order to present extended services for a given GenericRequest object the SFX evaluation process will determine the relevance of each of the conceptual services stored in the SFX-base using the content, or lack thereof, in the GenericRequest object. There are two phases to this evaluation process.

### **Phase 1: Selection of active services with the origin resource of the GenericRequest object as a Source**

The interface between the redirection component and the service component delivers both link-source metadata and information on the origin of the link-source. The latter is stored in the `rec$dbId` field of the GenericRequest object that is created by the service component. During the evaluation phase, the value of this field becomes the key for a lookup in the local component of the Source table of the SFX-base. The global common name of the resource is detected there, next to this key which refers to the local implementation of a resource. This global name is now connected via services of the Colli to various global names of Target resources, as already shown in [Table 7](#). Hence, the result of this lookup is a bundle of services that might be relevant for the current GenericRequest object, as judged upon by its origin. Inactivation of certain services during the localization of the SFX-base guarantees that the resulting bundle already reflects the local digital library situation.

In [Table 8](#) and [Table 9](#) the mechanism is shown in bold for a GenericRequest object representing an item originating from the Ghent implementation of BIOSIS. Its `rec$dbId` value is `ERL::BX`. In this phase of the evaluation process, `S::ERL::BX` -- a prefix `S` is added as a means to refer to Source -- becomes the key for a look-up in the local component of the Source table. There, BIOSIS is detected as the global common name of the resource. Several services are leading out from BIOSIS into Target resources. For instance, *abstract* connects BIOSIS with Medline & *genome* connects BIOSIS with Genome Base. These are the services that remain as potentially relevant.

### **Phase 2: Filtering out selected active services by comparing the content of the GenericRequest object with the Thresholds**

Phase 1 of the SFX evaluation process filters out services of the SFX-base that do not have BIOSIS as an origin. For each of the resulting services, the information in the GenericRequest object will be matched against the Thresholds -- global and local -- that are connected to these services. The *genome* service connecting BIOSIS and Genome Base will be filtered out if the GenericRequest object does not contain an entry for the `genID` parameter. The *abstract* service connecting BIOSIS and Medline will be filtered out if an objectLookup for the `ISSN` value in the GenericRequest object learns that the journal is not abstracted in Medline. Or it could be filtered out if the GenericRequest object does not contain a value for year, volume or issue.

Again, since some Thresholds express the local situation, and since these Thresholds can overrule the global ones, the result of this filtering process will reflect the situation of the local digital library collection. Those services remaining from Phase 1, for which at least one of the Threshold evaluations fails will be rejected as not being relevant. The ones that make it through the complete evaluation process will be presented to the user in the SFX-menu-screen as locally relevant extended services for the current GenericRequest object, hence for the link-source for which the whole process has been initiated by clicking the SFX-button

## Resolving locally relevant extended services into URLs with TargetParsers

Consistent with the just-in-time linking philosophy of SFX, the bundle of relevant services that is obtained as a result of the SFX evaluation process described above, is not resolved into URLs at the moment of their presentation to the user that launched a request for extended services. Rather, for each menu-item in the SFX-menu-screen, the following elements are sent as parameters for a script that will be initiated when a user selects a menu-item:

- The identifier of the GenericRequest object;
- A name referring to the service and its Target, that is represented by the menu-item;
- For some services, overwritable metadata elements from the GenericRequest object (see SFX-menu-screens in [Figure 8](#), [Figure 10](#), [Figure 12](#) and [Figure 14](#)).

When the user clicks a menu-item, the appropriate TargetParser script corresponding to the chosen service and Target is launched. These TargetParsers implement resource-specific link-to syntaxes. They take data from the GenericRequest object as input and compute the URL to which the user will be redirected.

## Comments

### The impact of the redesign of the SFX service component

The new design of the SFX service component, reflecting aspects of global and local relevance, has a considerable impact on the transportability and manageability of the SFX service component. Once an SFX-base containing conceptual services of global relevance with appropriate global Thresholds has been compiled, the localization of the set-up requires minimal efforts. As an illustration of this, it is interesting to look once again at the BIOSIS-*abstract*-Medline example. This service was initially localized in Ghent by filling out the names of the local SourceParser for BIOSIS and the local TargetParser for Medline. Both parsers implement the desired connection with the [SilverPlatter](#) ERL platform that locally hosts the databases. LANL has a different implementation of BIOSIS, currently running on a [Geac](#) Advance system, while no local Medline is available. Therefore, LANL chose to use the PubMed implementation as a Target. When transporting the SFX-base -- that had been localized for Ghent first -- over to LANL, very limited editing of the SFX-base had to be done to activate the BIOSIS-*abstract*-Medline service in the new environment: the global service and its global Thresholds remained valid. The parser values for the Los Alamos version of BIOSIS and Medline were used to overwrite the Ghent values, as shown in [Table 9](#). The Threshold indicating that Medline is only available from 1985 onwards in the Ghent collection, was overwritten by a 1965 Threshold for LANL. In this case, the local and global Threshold are equal. The elegance of the PubMed [Entrez](#) link-to mechanism and the availability of the complete Medline collection caused Ghent to reconsider the Target -- hence TargetParser -- to be used in favor of the PubMed implementation. Upon this decision, again, very limited editing of the Ghent SFX-base had to be done.

In Ghent&LANL, most of the TargetParsers are implemented as Perl scripts. Towards the end of the experiment, advantage has been taken of the launch of a preliminary version of the S-Link-S Calculator ([Openly Inc. 1999](#)). This Calculator is designed to compute URLs based on input metadata and XML templates that describe link-to syntaxes in an S-Link-S compliant manner ([Hellman 1998](#)). As such, SFX TargetParser scripts that perform the computation of the URLs can eventually be replaced by templates describing the link-to syntax and that are used as input for the S-Link-S Calculator. The experiment ended with a hybrid solution, in which the SFX service component was adapted to dynamically choose between the two mechanism that became available to compute URLs: the TargetParsers and S-Link-S templates. TargetParsers can be shared between different digital library implementations, since many will link into the same resources or family of resources. Again, this reduces the overhead in running the solution. But the tie-in between the SFX and the S-Link-S work, will eventually further diminish the administration of the SFX solution as publishers start and contribute link-to templates and corresponding metadata to the S-Link-S framework. Sharing of TargetParsers will then be replaced by using S-Link-S templates from the framework that has already been set up by Eric Hellman to collect them.

Also, SourceParsers are easily transferable between different digital library implementations, requiring little or no enhancements to be made. For instance, OPAC systems worldwide support the [Z39.50](#) protocol and respond to requests with [MARC](#) formatted records. Making abstraction of the regrettable idiosyncrasies of Z39.50 and MARC implementations, SourceParsers for such OPAC systems can be reused with little editing, apart from the adaptation of Z39.50 parameters such as host, target, port etc... to the local situation. Also, Z39.50 can be used to fetch link-sources from all implementations of databases on a [SilverPlatter](#) ERL platform, and -- again -- only the Z39.50 parameters will be different. All implementations of MathSci on such ERL platforms can use the same parsing procedures, making the parsing part of the SourceParser even universal. This is also the case for the SourceParser used for the APS [PROLA](#) archive and the [Wiley](#) journals, since there is only one implementation of those, worldwide. This approach opens attractive possibilities of sharing SourceParser on a large scale, reducing the overhead in running the solution. Furthermore, it allows information vendors to provide SourceParsers for their resources, keeping full control of the amount of information that they allow to be fetched.

### The impact of a service component building on conceptual services

The consequences of the introduction -- in the Elektron experiment -- of a linking service built upon a database of conceptual services, have only now become fully evident, due to the complex and distributed nature of the environment in which Ghent&LANL was conducted. Actually, the more resources are added to the environment, the more the elegance and feasibility of the SFX service component become apparent. The introduction of a new SFX-aware resource in an environment requires very limited editing of the SFX-base, in order for all the existing conceptual services that had already been registered in the SFX-base to become immediately available for the new resource too. The dynamic manner in which the SFX system brings up a list of extended services for link-sources of a newly added resource can only be called remarkable. Even the designers of the system found it increasingly difficult to manually predict the outcome of a request for extended services, even when knowing the system and its underlying database in and out, and when studying the content of the link-source record or its GenericRequest object in detail. As an illustration of this, it is noteworthy to mention how SFX delivers a PubMed link for a citation in the Information Science journal JASIS, where [Wiley](#) themselves do not have such a link (see Screencam 3 of the examples). This is odd, since Wiley does insert static PubMed links for citations in their journals. Probably they only do so for those with a biomedical subject, since the cost/benefit balance of sending all their citations into the NCBI [PubRef](#) process would be negative. The dynamic and conceptual SFX approach does not require such precomputational processes and is able to recognize the validity of presenting a PubMed link for a citation in JASIS on the fly. Another remarkable and appealing example is where full-text links presented by SFX lead from a citation in a Wiley journal to an article in another Wiley journal. At the time of the experiment, Wiley did not offer such a linking service within its own collection, even if they fully

control the data to implement it. These examples do not only illustrate the strength of the SFX solution, but more importantly, they are a strong indication of the problems of scale of static linking solutions.

## The SFX redirection mechanism and namespace specific identifiers

Important efforts are under way to enable reference linking using [DOIs](#) ([Paskin 1999a](#)). Publishers will contribute metadata of their publications along with the corresponding DOIs to the [DOI](#) metadatabase. Other publishers can then match references in their own publications to the DOI metadatabase enabling them to insert the DOI of the work represented by the reference next to the reference. Currently, due to lack of support of the [handle](#) protocol in mainstream browsers, such a DOI -- say 10.1000/123456789 -- is being hyperlinked as <http://dx.doi.org/10.1000/123456789> and the DOI handle proxy will resolve this link into a single URL, being the one of the publication at the publisher's site. ([Paskin 1999b](#)). This hyperlink is a perfect example of a closed link that does not take into account the local context in which the link will be used ([Van de Sompel & Hochstenbach 1999a](#)). This closed link causes the Harvard problem to arise, since it does not take into account the possibility of storage of the same work at another, preferred location. Also, this mechanism does not allow other, locally relevant extended services to be provided for the reference at hand, since their provision requires the full metadata -- not only the identifier -- of the reference.

The local redirection approach presented by the SFX work does present a pragmatic way to open such a closed linking framework. DOIs can be carried in an SFX-URL pointing at the preferred SFX-server. For instance, for a citation in a [Wiley](#) InterScience article that has a static link to DOI 10.1000/123456789 the link can be dynamically rewritten if the existence of a service component is detected. In the case of SFX as a service component, it can become:

<http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?>

[vendorID=Wiley&databaseId=WIS&nameSpace=DOI&objectDesc=URLencode\(DOI=10.1000/123456789\)](http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?vendorID=Wiley&databaseId=WIS&nameSpace=DOI&objectDesc=URLencode(DOI=10.1000/123456789))

In essence, such a pragmatic mechanism can redirect the identifier to the redirection component of a selective resolution system, that can decide what to do with it, based on the local context. In the case of SFX, receipt of the above URL causes the launch of a SourceParser. As shown before, typically, this will be the SourceParser corresponding to the serviceDesc part of the URL. Still, upon detection of the "nameSpace" parameter, this default could be overwritten to become the namespace-specific SourceParser, in this example the one for the DOI namespace. This DOI SourceParser would do a so-called reverse look-up in the DOI metadatabase, using the DOI value as a key to fetch the corresponding metadata. Both the fetched metadata and the DOI can then be used in a process to determine the relevant extended services for the citation, including a link to the most appropriate full-text instance. As will be discussed in the next section, for other types of service components, redirection of the DOI without metadata-fetch could be sufficient.

The same mechanism can be used to open the links that are connected to citations carrying identifiers originating from other namespaces such as [PubMed](#) and [Astrophysics Data System](#). This can easily be seen by taking the following citation from a Wiley journal that has a static link to PubMed:

Rainer RO, Geisinger KR. Beyond sensitivity and specificity. Am J Clin Pathol 1995; 103: 541-2. [Medline](#)

The Medline hyperlink points at:

<http://www4.ncbi.nlm.nih.gov:80/htbin-post/Entrez/query?uid=95259660&form=6&db=m&Dopt=r>

and uses the PubMed ID as a look-up key into the NCBI PubMed. This hyperlink can be pointed at a local resolution solution if it is dynamically rewritten as:

<http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?>

[vendorID=Wiley&databaseId=WIS&nameSpace=Medline&objectDesc=URLencode\(Medline=95259660\)](http://isiserv.rug.ac.be/cgi-bin/sfx/bin/menu.cgi?vendorID=Wiley&databaseId=WIS&nameSpace=Medline&objectDesc=URLencode(Medline=95259660))

In a similar manner as in the [DOI](#) example, this URL can deliver the PubMed ID to a local redirection component. In the case of SFX, this would cause the Medline-namespace SourceParser to be launched, which would fetch the corresponding record from the Medline database. This SourceParser can actually fetch the metadata either from the PubMed implementation of Medline (using the HTTP protocol and the [Entrez](#) link-to-syntax) or from a local implementation of Medline if one exists. Once that metadata has been turned into a GenericRequest object, the SFX evaluation process can deliver locally relevant extended services. In this way, the initial service of [Wiley](#) is augmented with locally relevant services. Also, the quality of the metadata resulting after such a fetch will be higher than that in the original citation, as can be seen from the fact that no issue number is available in the citation of this example, as is common in the medical literature. The fetched record, however, will contain the missing issue number and much more valuable information.

This consideration illustrates that link-source metadata does not necessarily have to be fetched from its origin resource. In both examples, for a link-source originating in a Wiley journal, the link-source metadata is fetched from the authoritative resource for the namespace of which the link-source carries an identifier.

## Identifiers, metadata and service components

It is interesting to further reflect on the nature of the service component and the requirements it imposes on the redirection mechanism of a selective resolution solution. To commence this consideration, service components that only aim at the delivery of the appropriate full-text instance for a given link-source are considered. Such a service component could operate solely on an underlying database of identifiers, meaning it could be a traditional linking service building on static links between documents. For such service components, it would be sufficient if the redirection mechanism would transfer identifiers only, without bothering about the associated metadata. Such a service component might be sufficient to address the Harvard problem. It might contain a repository of identifiers and locations of full-text for which a local or preferred warehouse other than the default one exists. However, such a repository of identifiers could quickly become very large and difficult to maintain. It can even be considered awkward to maintain such an institutional repository if no full-text is stored locally, but is only being accessed from preferred external aggregators. This consideration points at the desirability of a service component of a different, more abstract nature. Such a service component can build on the logic underlying the distribution of the collection rather than on individual identifiers of material in the collection. Under normal conditions, such logic might tell that all journals of a certain publisher are accessed at a certain warehouse, that certain ISSN numbers have to be accessed from another one, and that an ISSN number has to be accessed in one repository before a certain date and at another one after that date. This level of abstraction drastically reduces the amount of information to be maintained in the service component and hence makes it more scaleable. But it requires metadata of link-sources to act as operators, not only identifiers. Adding to this the fact that the identifiers required to make such a scenario work for link-sources originating from full-text repositories, abstracting & indexing databases, OPAC systems and e-print archives are not available and will most probably not become universally available any time soon, it must be concluded that service components will have to be able to operate on the basis of metadata in general with identifiers being a special instance of metadata. This imposes a requirement on the redirection mechanism to be able to deliver link-source metadata, not only identifiers.

Adding to the task of the service component the delivery of other extended services, it becomes hard to imagine that a scaleable solution in a highly distributed environment could build on an architecture with a static linking database. Several illustrations of this consideration have resulted from the Ghent&LANL experiment. A more abstract and dynamic service component is required, which will perform some rule-based decision making, that tends towards the evaluation of the relevance of conceptual services as introduced in the SFX work. As will be clear from the SFX experiments, this type of service component requires the availability of link-source metadata in order to be able to function. Again, this imposes a requirement on the redirection mechanism to be able to deliver link-source metadata. This does not mean that identifiers are irrelevant to this type of solution. Quite to the contrary, since it has been shown that identifiers -- from whichever namespace -- are a welcome tool to enable the local redirection component to adequately deliver high-quality metadata.

The general conclusion of the above is that, realistically, identifiers will not be sufficient to address the problem of delivering extended services in a distributed digital library collection. Metadata is required for scaleable service components to be able to perform their tasks. Moving from left to right on the scale of service components ranging from traditional static linking systems to dynamic linking systems building on conceptual services, the data required for the service components to adequately do their jobs ranges from identifiers to full metadata.

## Illustrations of project results

The concrete results of the project are illustrated by Lotus Screencam movies that show how a Ghent and a LANL user navigates in his institutional SFX-aware digital library collection. The Screencams are provided as stand-alone executables that can only be run on WinTel computers. Since the Screencams are large files, their size is mentioned. The Screencams do not contain audio. In addition to these Screencams, some examples are also given by means of screendumps.

**Screencam 1: Ghent implementation of BIOSIS** (Lotus Screencam executable for WinTel computer; no audio; size 52 Mb)

The user starts from the Ghent implementation of BIOSIS, and requests the services of the Ghent SFX solution. Services are shown linking from BIOSIS into the Ghent OPAC, into full-text collections, into the LANL implementation of the Science Citation Database, into PubMed and into Ulrich's Serials Directory. At a certain point the user links out to a paper in Cancer, a Wiley InterScience journal that is SFX-aware. Upon request, the user receives similar services from the Ghent SFX solution for citations in that paper. They link him to the LANL OPAC, to full-text at other publishers sites, to PubMed and to the LANL implementation of the Science Citation Database.

**Screencam 2: LANL implementation of BIOSIS** (Lotus Screencam executable for WinTel computer; no audio; size 46 Mb)

The user starts from the LANL implementation of BIOSIS and requests services from the LANL SFX solution. He uses links into the Ghent implementation of Current Contents, into several full-text collections and into the LANL implementation of the Science Citation Database. From the Citation Database, again, he requests SFX-services that lead him into more full-text collections as well as into PubMed. Next, the user returns to his result set in BIOSIS, where he requests services for other records. These lead him into the Journal Citation Reports and to more full-text. For one of the BIOSIS records, the *genome* service appears leading the user to the Genome database.

**Screencam 3: Ghent implementation of Current Contents** (Lotus Screencam executable for WinTel computer; no audio; size 36 Mb)

The user starts from the Ghent implementation of Current Contents, and requests the services of the Ghent SFX solution. The user links out to a paper in JASIS, an SFX-aware Wiley InterScience journal, and -- upon request -- is receiving SFX-services from the Ghent SFX server for citations found in the JASIS paper. Shown are links to PubMed, LISA and to the LANL implementation of the Science Citation Database. For another record from Current Contents, the user links into the Science Citation Database. There, he requests extended services for several records in the result set, which link him to full-text collections etc.

**Screencam 4: Ghent Aleph 500 OPAC** (Lotus Screencam executable for WinTel computer; no audio; size 13 Mb)

The user starts from the Ghent OPAC. He requests SFX-services for a record describing a book and links into Amazon.com. For a journal, he links to the full text collection and to the Journal Citation Reports.

**Screencam 5: LANL Advance OPAC** (Lotus Screencam executable for WinTel computer; no audio; size = 21 Mb)

The user starts from the LANL OPAC and requests services from the LANL SFX solution. For a book record, he links into Amazon.com. For the OPAC record of the journal Cancer he links to the full-text repository and browses towards a paper. In that paper, citations have SFX-buttons since the journal is SFX-aware. The user requests SFX-services for some citations, that bring him to PubMed and to the LANL implementation of the Science Citation Database. For records resulting from the latter service, the user again requests SFX-services that take him to the Journal Citation Reports, the LANL OPAC and to various full-text collections. One of those is -- again -- the Wiley InterScience collection.

**Screencam 6: LANL Inspec** (Lotus Screencam executable for WinTel computer; no audio; size = 29 Mb)

The user starts from the LANL Inspec and requests services from the LANL SFX solution. Services bring him to full-text collections, the LANL OPAC, the Journal Citation Reports. At a certain point the user goes out to the PROLA archive and finds SFX-aware citations for which he requests extended services. Due to the current implementation of the SFX-URL for PROLA little metadata makes it into a GenericRequest object and as such little services result. Back in the Inspec, the user links to the LANL implementation of the Science Citation Database, from which he further links to the LANL OPAC and to full-text collections.

**Screencam 7: the arXiv, consulted by a LANL user** (Lotus Screencam executable for WinTel computer; no audio; size = 19 Mb)

The user searches the Topic interface for the arXiv e-print repository and requests SFX-services for search results. These link him into the Inspec database, searching for references by the authors of the e-prints.

**Screendump example 1:**

[Figure 8](#) shows how the link-source record from [Figure 3](#), originating from the Ghent implementation of BIOSIS yields a variety of extended services, amongst other *full\_text*, *holding*, *abstract*, *author*, *cited\_author* and *cited\_reference* that are presented in the SFX-menu-screen. From this menu-screen, the user chooses the full-text link to Wiley, which leads him into an article of the journal Cancer that is SFX-aware ([Figure 9](#)). [Figure 10](#) shows how the same user has requested extended services for the third citations in that Wiley article and has received a Ghent SFX-menu-screen. For this citation similar services are available, all linking dynamically from the external Wiley resource back into the Ghent digital library collection. As can be seen, a link to the on-line version of the New England Journal of Medicine has become available. The user chooses to use the *cited\_reference* service that is also available for this citation. This leads him into the SFX-aware LANL implementation of the Science Citation Database ([Figure 11](#)) from where he can request extended services, again.

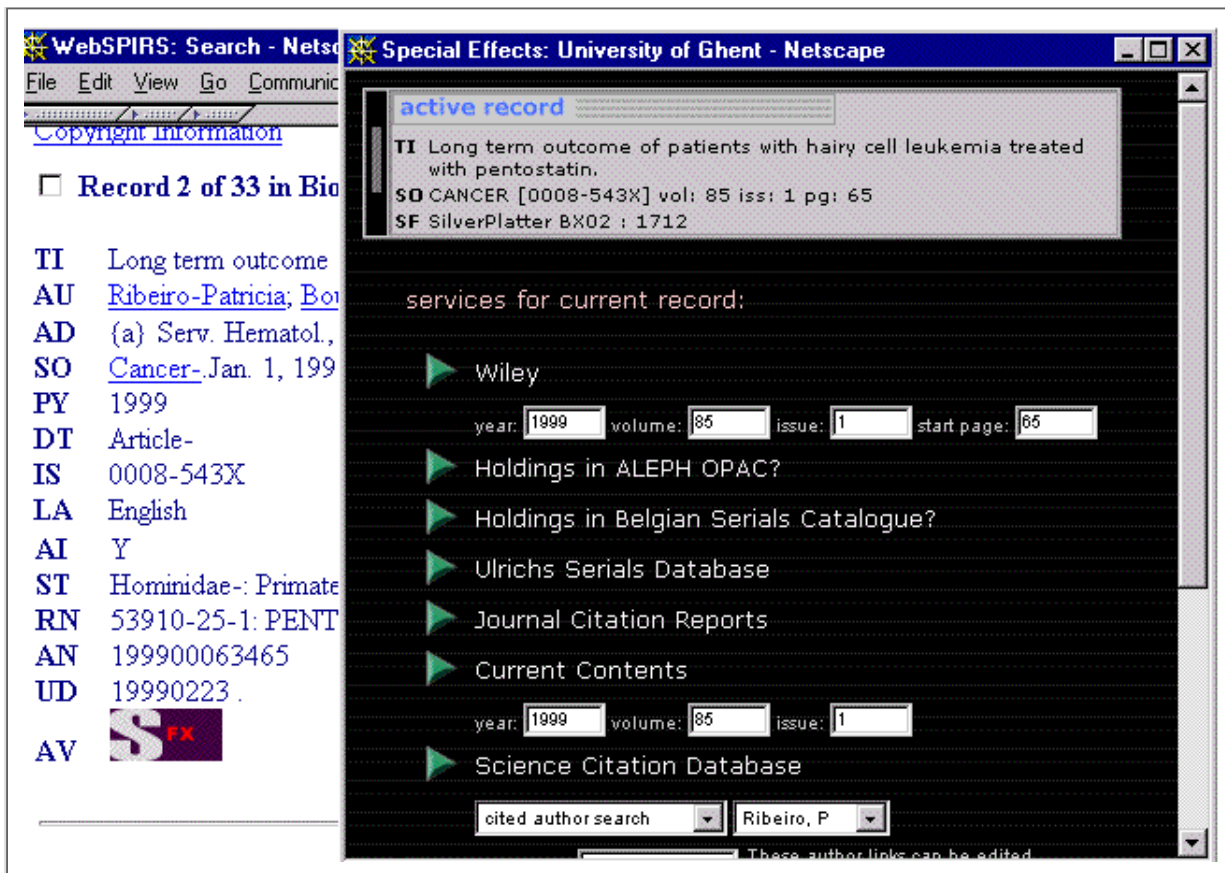


Figure 8: Ghent SFX-menu-screen for link-source from Ghent BIOSIS (record from Figure 3)

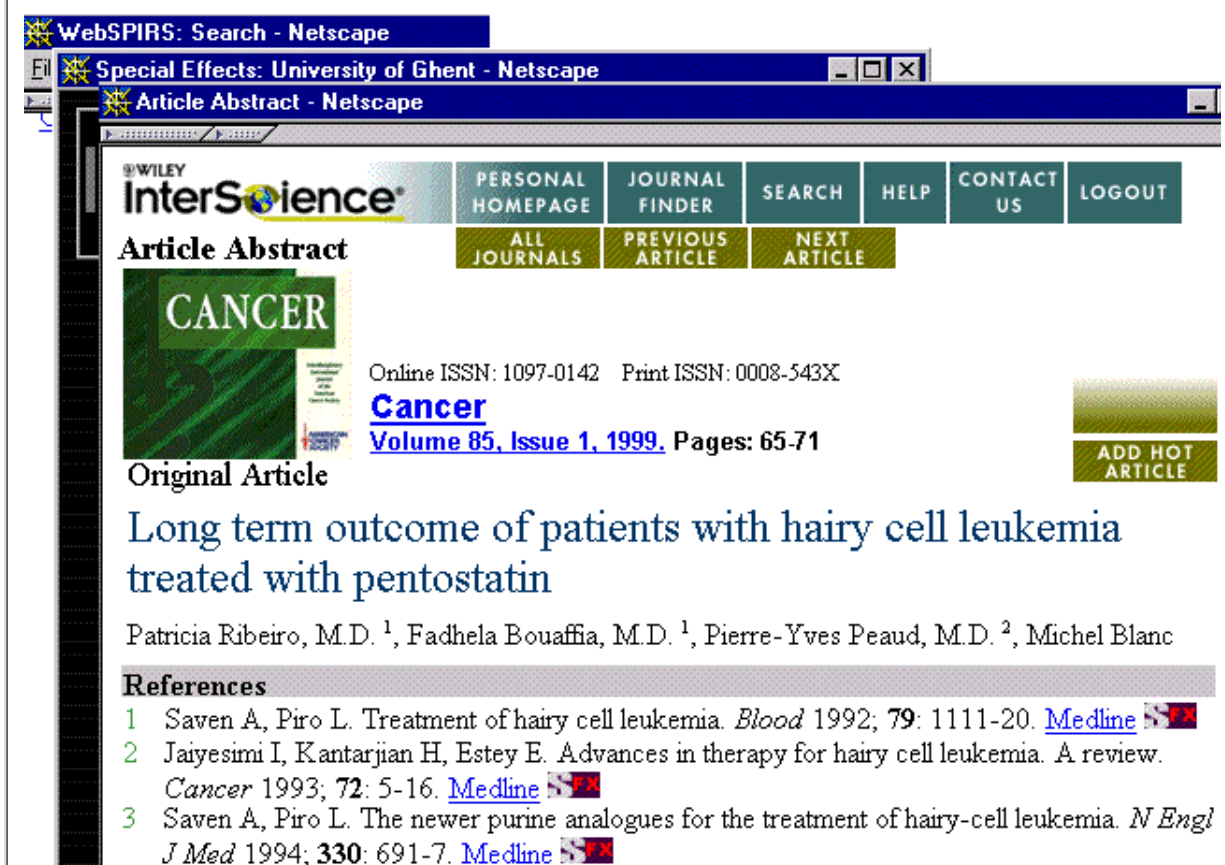


Figure 9: Ghent user follows Wiley full\_text service from the SFX-menu of Figure 8

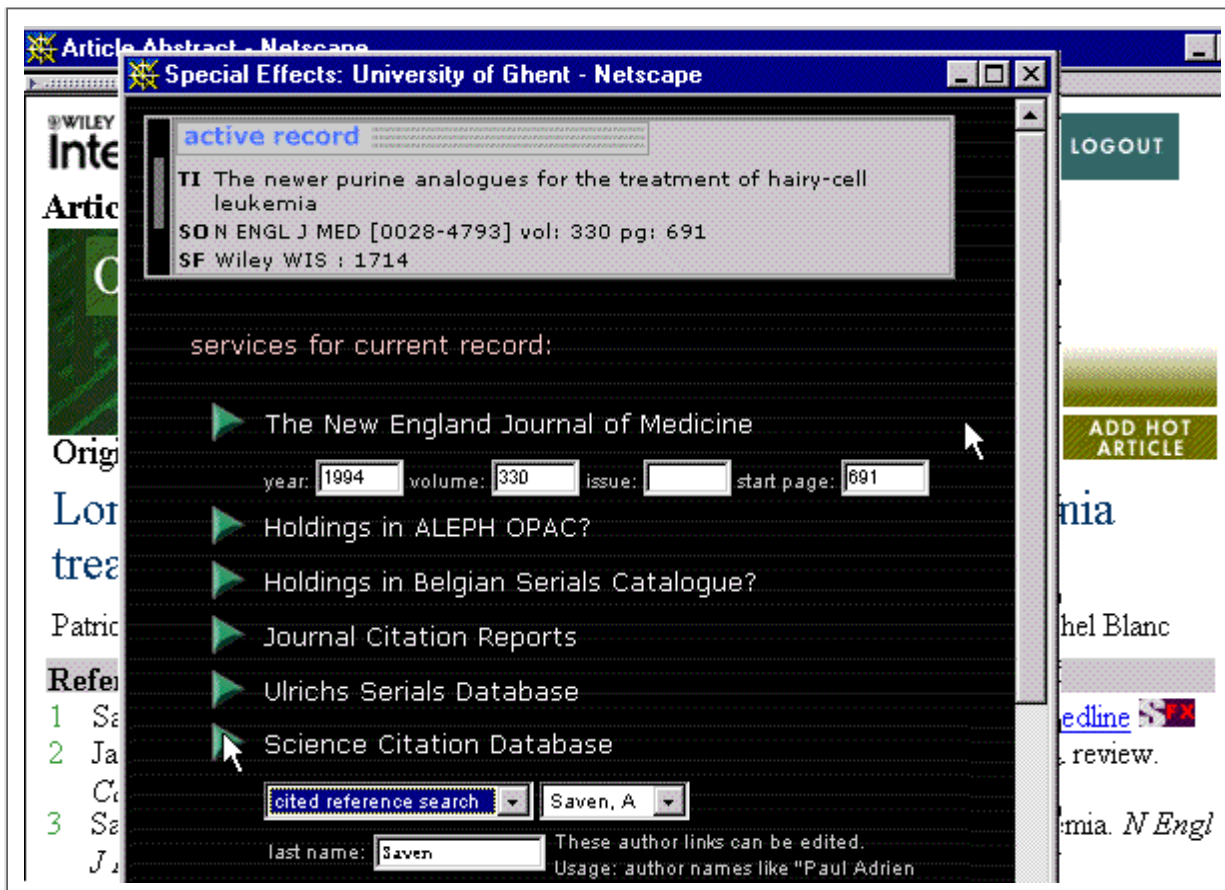


Figure 10: Ghent SFX-menu-screen for link-source from Wiley InterScience (third citation from Figure 9)

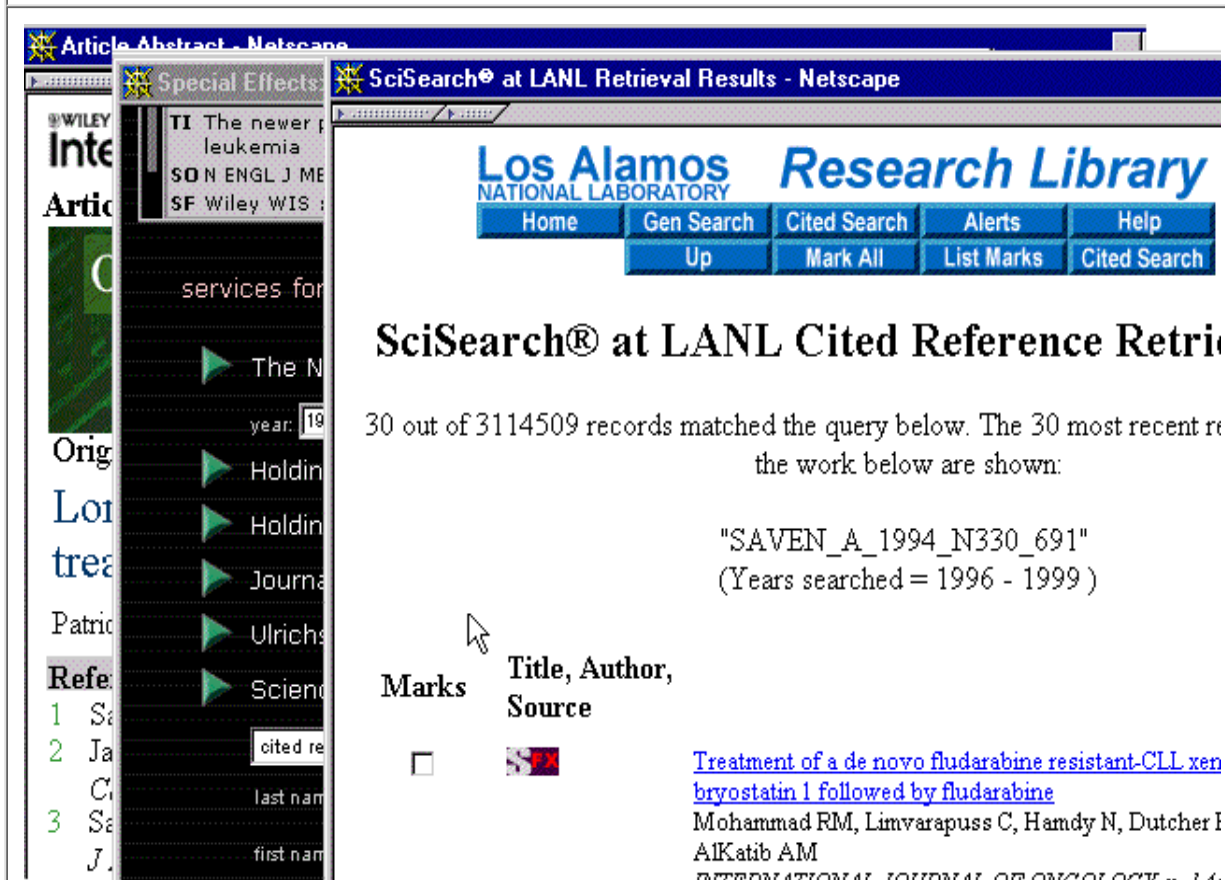
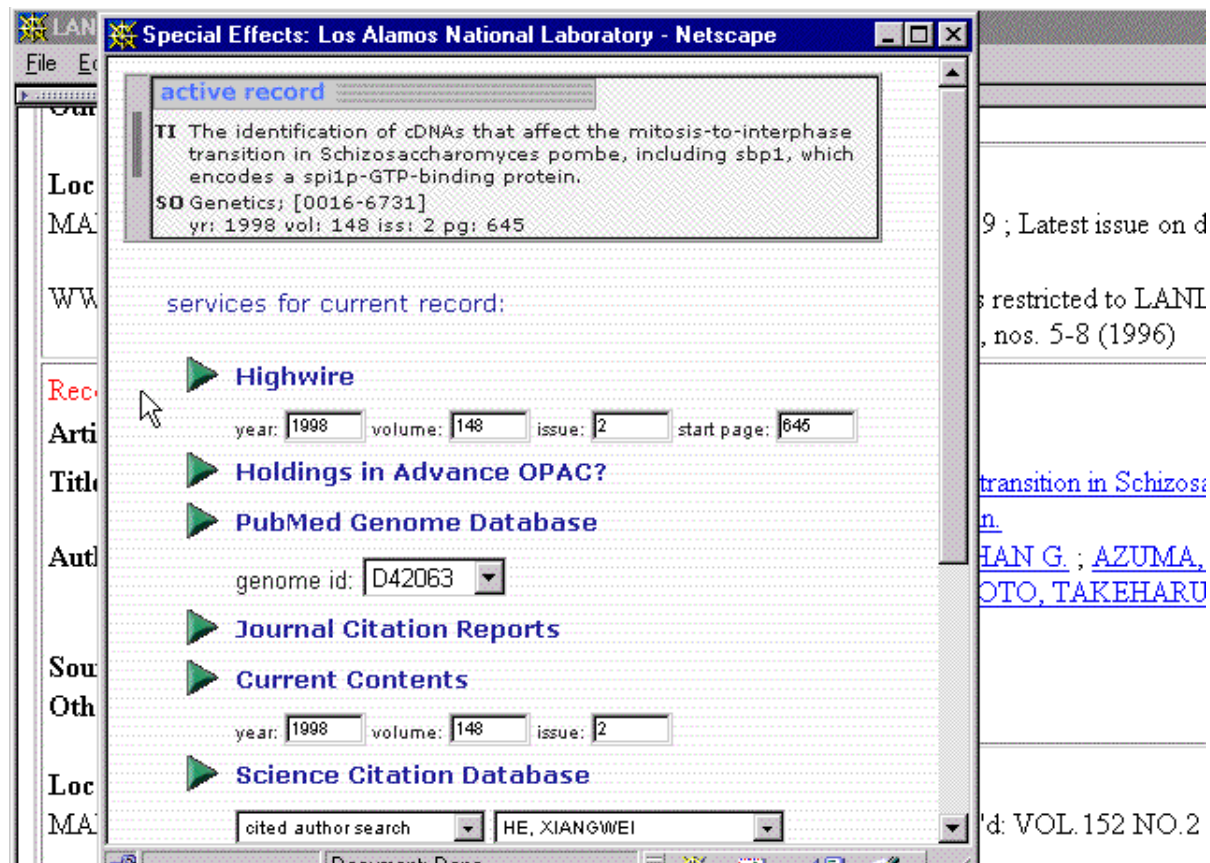


Figure 11: Ghent user follows the cited\_reference service from the SFX-menu of Figure 10

**Screendump example 2:**

In [Figure 12](#), the record from [Figure 4](#) originating from the LANL implementation of BIOSIS is used as a link-source. The resulting SFX-menu-screen looks a little different, as an illustration of the fact that another SFX system is being consulted to provide extended services. The LANL localization can also be derived from the OPAC link, that now leads into the Los Alamos Advance catalogue. Another service appears in this screen too: it provides a look-up of sequence information for genome identifiers that were found in the link-source metadata. This service leads the user to the NCBI Genome database using the Entrez link-to syntax ([Figure 13](#)).



**Figure 12:** LANL SFX-menu-screen for link-source from LANL BIOSIS (record from [Figure 4](#))

LANL Special Effects: PubMed Sequence query - Netscape

active record

TI The identific  
transition in  
encodes a s  
SO Genetics; [0  
yr: 1998 vol

services for

Highw  
year: 199

Holdir  
PubM  
genome

Journ  
Curre  
year: 199

Scienc  
cited au

NCBI Entrez Sequence QUERY BLAST Ent

Other Formats: FASTA Graphic

Links: MEDLINE

LOCUS M79174 314 bp mRNA EST

DEFINITION EST01322 Subtracted Hippocampus, Stratagene (c  
sapiens cDNA clone HHCPO45, mRNA sequence.

ACCESSION M79174

NID g273487

VERSION M79174.1 GI:273487

KEYWORDS EST.

SOURCE human.

ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;  
Eutheria; Primates; Catarrhini; Hominidae; Homo

REFERENCE 1 (bases 1 to 314)

AUTHORS Adams, M.D., Dubnick, M., Kerlavage, A.R., Morenc  
Utterback, T.R., Nagle, J.W., Fields, C. and Vent

TITLE Sequence identification of 2,375 human brain c

JOURNAL Nature 355, 632-634 (1992)

MEDLINE [92168112](#)

COMMENT Contact: Kerlavage, AR

Figure 13: LANL user follows the *genome* service from the SFX-menu of [Figure 12](#)

(after selection of identifier M79174 from the pop-down)

### Screendump example 3:

[Figure 14](#) shows the LANL SFX-screen for the first record from the Topic implementation of the [arXiv](#) e-print repository shown in [Figure 6](#). Here, an *author* service is available, that provides the option to look-up records in the Inspec database, that abstract publications authored by the e-print authors as a means to support the evaluation of the reliability of the non-peer-reviewed e-print. The author that is being looked-up has 160 references in the Inspec database ([Figure 15](#)).

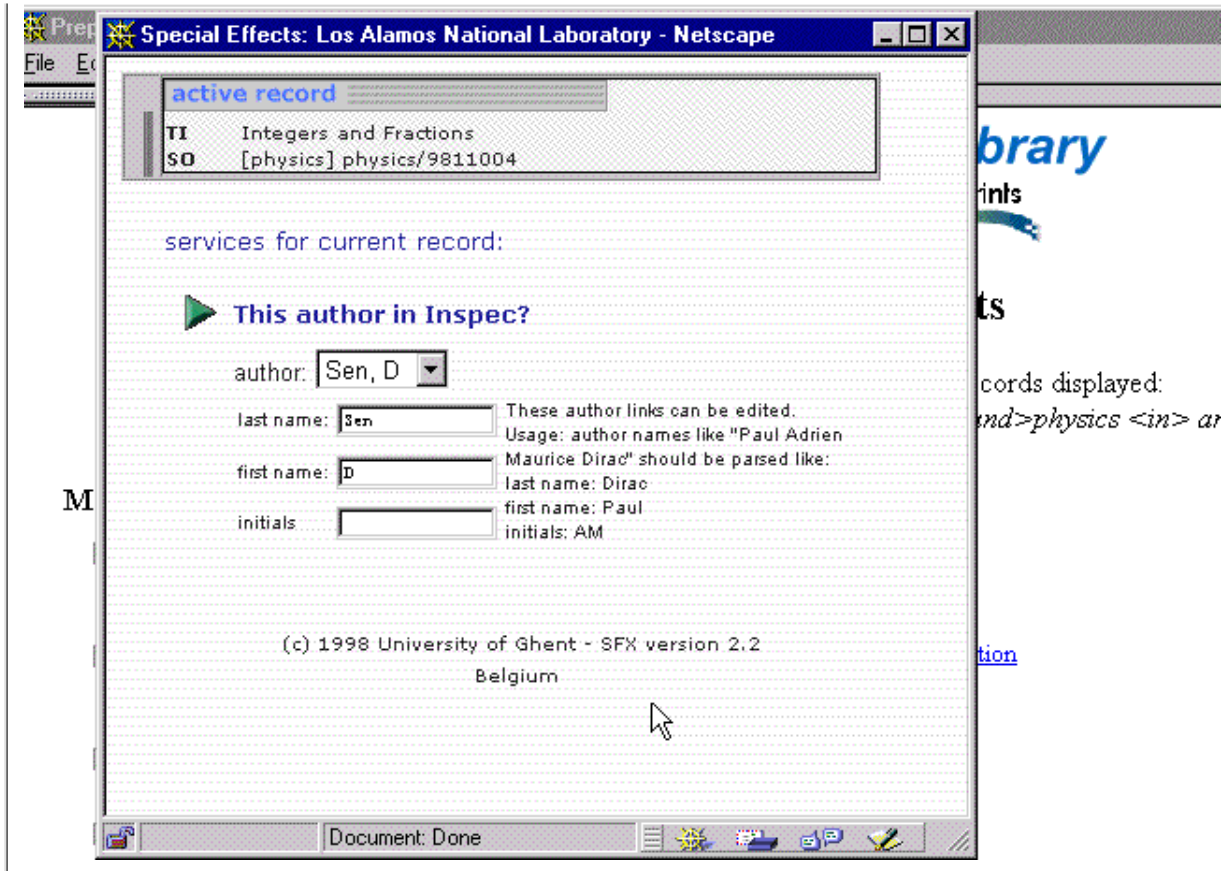


Figure 14: LANL SFX-menu-screen for link-source from the arXiv (first record from Figure 6)

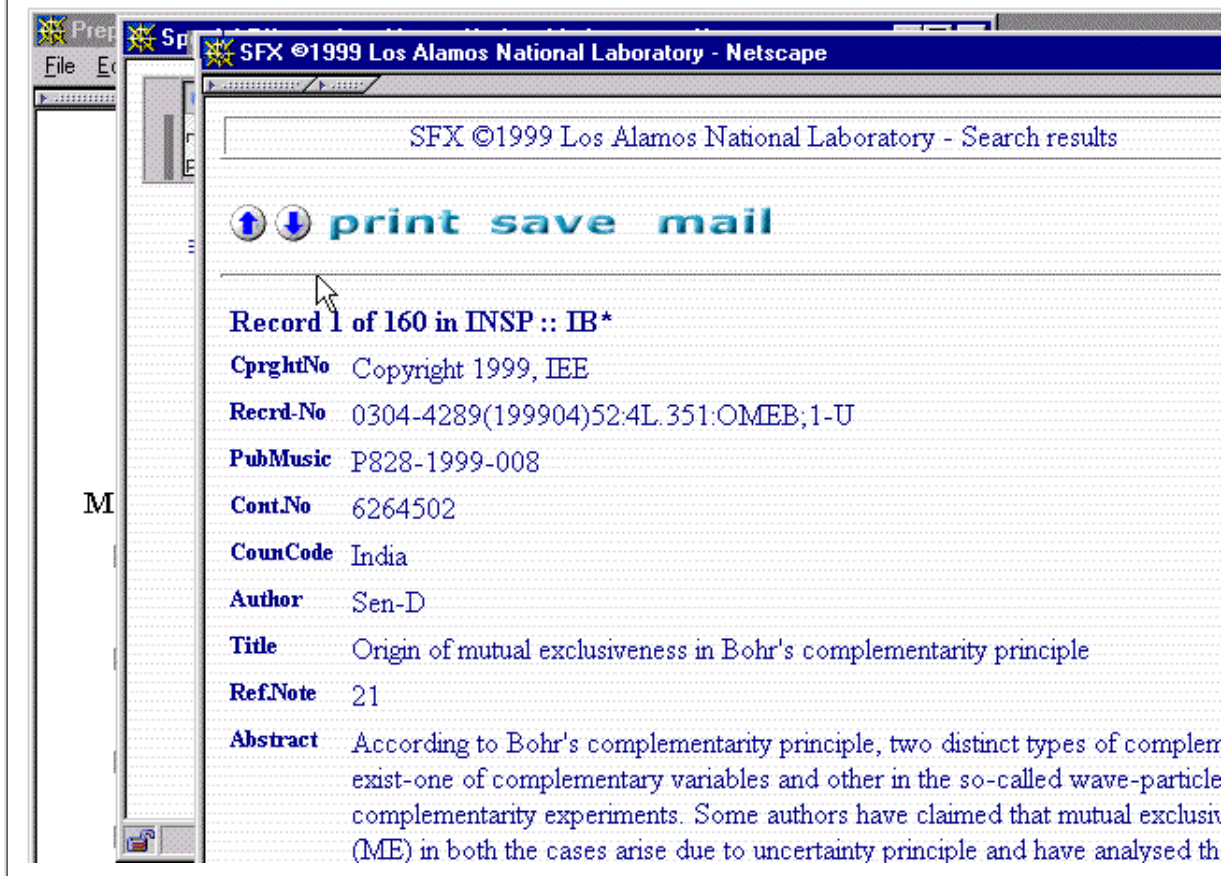


Figure 15: LANL user follows the Inspec author service from the SFX-menu of Figure 14

## Conclusion

The "SFX@Ghent & SFX@LANL" experiment has led to important generalizations of two components that had been introduced in the Elektron experiment and that are essential for systems supportive of selective resolution: the redirection mechanism and the service component. Although both have been discussed in relation to one another, it has also been shown that they can be separate components that exchange information in a unique metadata interchange format. For the experiment, this format was internally defined and inspired on the structure of the GenericRequest object, since -- by lack of non-SFX local redirection components -- interoperability at this level was not an issue. If more local redirection solutions and more local service solutions become available, standardization of this interchange format will become important. Also, a standardization of a local redirection URL -- like the SFX-URL -- and of vendorId, databaseId and nameSpace values is required. Such a standardization will also be valuable for other ongoing work in the digital library environment.

In essence, the SFX redirection mechanism can be combined with a service component of a very different architecture, even one that builds on a static linking database of identifiers. The current implementation of the SFX local redirection mechanism builds on the CookiePusher mechanism, a consistent SFX-URL and SourceParsers. Each of these buildings blocks can be replaced by more robust alternatives, as long as they perform the same function. The investment required to make systems SFX-aware using the CookiePusher and the SFX-URL has been minimized. Still, it will be easier to implement SFX-awareness in resources that deliver information in a dynamic rather than in a static manner. It has been shown that the SFX local redirection mechanism can be used to redirect namespace-specific identifiers to a local service component. This leads to the capability of the SFX-URL to open closed linking frameworks, which is seen as a powerful illustration of the feasibility of the approach.

The SFX service component can also operate with a different redirection method, as long as that supports delivery of link-source metadata and its origin to the service component. The Ghent&LANL information environment, with its many resources and different technologies running those resources, has led to a design in which the SFX linking service has become a totally neutral module in the digital library that can potentially interoperate with every other system in the environment. Its redesign, reflecting the notions of global and local relevance of services, has led to an important reduction in the overhead of running the solution. In addition to that, the possibility to share SourceParsers, TargetParsers and S-Link-S templates further diminishes the administrative overhead.

As far as can be verified, Ghent&LANL has been the first experiment in which bi-directional context-sensitive linking between distributed resources under control of different authorities has been demonstrated in the scholarly communication environment. As can be seen from the examples, the net result of making systems SFX-aware and delivering extended services for link-sources originating from those systems via the SFX-menu-screen, is a fully hypertextual scholarly information environment in which jumping between related distributed resources becomes possible, hopefully in the way Gardner had imagined it ([Gardner 1990](#)). As with the most renowned hypertext system -- the World Wide Web -- the ease with which this navigation occurs, can lead to getting lost in the information space. At this point, this feature is seen as a compliment to the solution, since no comparable navigational capability has been demonstrated before.

## Future directions

The SFX solution will be brought into production in both Ghent and Los Alamos around January 2000. This will hopefully lead to user feedback that has so far been limited and has definitely not been investigated in a systematic manner. It will also lead to more involvement of librarians, which can only result in improvement of user aspects in the system. The current SFX solution is also ready to be beta-tested in digital libraries that are run by staff with reasonable technical skills. However, in order to be able to successfully coordinate such beta-testing, some central resources would be required. SFX is also being tested as a tool to integrate the subversive scholarly communication mechanisms ([Okerson & O'Donnell 1995](#)) with the established ones. This is, amongst others, done in the [UPS](#) protoproto work, that aims at building a multidisciplinary digital library service for major e-print initiatives ([Van de Sompel 1999](#)). In that project, the SFX work is combined with the Smart Objects Dumb Archives research ([Maly, Nelson & Zubair 1999](#)). Work is also underway to investigate the feasibility of making the [SLAC/SPIRES HEP system](#) SFX-aware. The JISC/NSF project ([Harnad 1999](#)) that looks into inter- and intralinking citations in the [arXiv](#) e-prints also intends to experiment with SFX as a building block. Discussions with scholarly publishers about interoperating with SFX are also under way. It is believed that a wider distribution of the solution is feasible, but will probably require commercial support or alternative financial investments.

The current design of the SFX-base in Ghent&LANL, implementing the notions of global and local relevance, suggests the possibility of an architectural redesign that builds on a division of the set-up in a central component, that describes the global level, and a local component for the localization. Such a redesign would bring the SFX solution into category 2 of the categorization of systems supportive of selective resolution introduced in [Table 1](#). This could considerably lower the redundancy of information in various implementations of SFX-bases and hence could make their local administration easier.

Research is on its way to build a recommendation system based on user activities in an SFX-aware environment. The fact that relevant user traversals in both external and internal information resources can be tracked leads to a wealth of log data that can be exploited by the recommendation system. Further research is also required in the area of the insertion of SFX-buttons for citations in closed information containers like PDF files and Word documents.

## References

- Caplan, Priscilla. 1999a. A model for reference linking. Report of the working group of the reference linking workshop; May 1999. [<http://www.lib.uchicago.edu/Annex/pcaplan/reflink.html>].
- Caplan, Priscilla. 1999b. Report of the second workshop on linkage from citations to journal literature; June 9th 1999, Boston. [<http://www.niso.org/linkrept.html>].
- Caplan, Priscilla and William Y. Arms. 1999. Reference linking for journal articles. *D-Lib Magazine* 5, no. 7/8. [<http://www.dlib.org/dlib/july99/caplan/07caplan.html>].
- Gardner, William. 1990. The electronic archive: scientific publishing for the 1990s. *Psychological Science* 1, no. 6.
- Halstead, Amy. 1999. PROLA: More Than Just a Pretty Acronym. *APS News* 8, no. 8. [<http://www.aps.org/apsnews/0899/089914.html>].
- Harnad, Stevan. 1999. Integrating and navigating eprint archives through citation-linking (NSF / JISC-eLib Collaborative Project). [<http://www.princeton.edu/~harnad/citation.html>].
- Hellman, Eric. 1998. Scholarly Link Specification Framework (SLinkS). [<http://www.openly.com/SLinkS/>].
- Maly, Kurt, Michael Nelson, and Mohammad Zubair. 1999. Smart Objects, Dumb Archives: A User-Centric, Layered Digital Library Framework. *D-Lib Magazine* 5, no. 3. [<http://www.dlib.org/dlib/march99/maly/03maly.html>].
- Needleman, Mark. 1999. Meeting report of the NISO linking workshop; February 11th 1999, Washington DC. [<http://www.niso.org/linkrpt.html>].
- Okerson, Ann and James O'Donnell. 1995. Scholarly Journals at the Crossroads; A Subversive Proposal for Electronic Publishing. Washington, DC: Association of Research Libraries. [<http://www.arl.org/scomm/subversive/toc.html>].
- Openly Inc. 1999. S-Link-S Calculator. June 1999. [<http://www.openly.com/SLinkS/Calculator/>].
- Paskin, Norman. 1999a. DOIs used for reference linking. Washington & Geneva. [<http://dx.doi.org/10.1000/143>].
- Paskin, Norman. 1999b. DOI: Current Status and Outlook. *D-Lib Magazine* 5, no. 5. [<http://www.dlib.org/dlib/may99/05paskin.html>].
- Shishir, Gunavaram. 1996. CGI Programming on the World Wide Web. Sebastopol, CA.: O'Reilly and Associates, Inc.
- Spilka, Susan. 1999. Wiley InterScience Update. June 1999. [<http://www.wiley.com/about/corpnews/wisupdate.html>].
- Van de Sompel, Herbert. 1999. the Universal Preprint Service initiative. July 1999. [<http://vole.lanl.gov/ups/>].
- Van de Sompel, Herbert and Patrick Hochstenbach. 1999a. Reference linking in a hybrid library environment. Part 1: Frameworks for linking. *D-Lib Magazine* 5, no. 4. [[http://www.dlib.org/dlib/april99/van\\_de\\_sompel/04van\\_de\\_sompel-pt1.html](http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt1.html)].
- Van de Sompel, Herbert and Patrick Hochstenbach. 1999b. Reference linking in a hybrid library environment. Part 2: SFX, a generic linking solution. *D-Lib Magazine* 5, no. 4. [[http://www.dlib.org/dlib/april99/van\\_de\\_sompel/04van\\_de\\_sompel-pt2.html](http://www.dlib.org/dlib/april99/van_de_sompel/04van_de_sompel-pt2.html)].

## Acknowledgments

The authors wish to thank the following parties for their active cooperation in the experiment:

- Lieve Rottiers at the [Ghent Library Automation team](#) for art work
- Miriam Blake, Johan Bollen, Doug Chafe, Mariella Di Giacomo, Frances Knudson, Dan Mahoney and Mark Martinez at the [LANL Library Without Walls](#) team
- Abe Lederman at the LANL CIC-15 group
- Mark Doyle at the American Physical Society / [PROLA](#) archive
- Andy Stevens, Andy Townsend and Craig Van Dyck at [Wiley Interscience](#)
- Denis Lynch, Jenny Walker and Andrew Wilkins at [SilverPlatter](#)
- Oren Beit-Arie and Yohanan Spruch at [ExLibris](#)
- Eric Hellman at [Openly](#)

Herbert Van de Sompel wishes to thank:

- the [Belgian Science Foundation](#) for a special PhD grant
- Rebecca Graham, Deanna Marcum and Don Waters at the [Council on Library & Information Resources](#) and the [Digital](#)

[Library Federation](#) for a travel grant and active support

- Donna Berg and Rick Luce at the [LANL Research Library](#)
- Paul Ginsparg at the LANL e-print [arXiv](#)
- William Y. Arms at [D-Lib Magazine](#) & Cornell University
- Jennifer De Beer for proofreading

Copyright © 1999 Herbert Van de Sompel and Patrick Hochstenbach

*(At the request of the authors, reference to "Orion ScienceServer" has been changed to "ScienceServer" and a link to Science Server has been added; in the acknowledgement to Mark Doyle, "American Institute of Physics" has been corrected to read "American Physical Society"; and in the abstract, "digital library collections of the Los Alamos National Library Research Library" has been corrected to read "digital library collections of the Research Library of the Los Alamos National Laboratory". Corrections made 10/19/99, 1:34 pm.)*

---

[Top](#) | [Contents](#)

[Search](#) | [Author Index](#) | [Title Index](#) | [Monthly Issues](#)

[Previous Story](#) | [Next story](#)

[Home](#) | [E-mail the Editor](#)

---

[D-Lib Magazine Access Terms and Conditions](#)

[DOI](#): 10.1045/october99-van\_de\_sompel

## National Archives and Records Administration

NARA

... to ensure ready access to essential evidence . . . that documents the rights of American citizens,  
the actions of federal officials, and the national experience . . .

[Search](#)[Research Room](#)[Records Management](#)[Federal Register](#)[Exhibit Hall](#)[NHPRC & Grants](#)[Digital Classroom](#)[Archives & Preservation](#)[About NARA](#)[Home](#)

## Treasures of Congress



[Online Exhibit](#)

### Quick Links To:

[Nationwide Facilities:  
Locations, Hours, &  
Accessibility](#)

[News & Events](#)

[What's New at NARA's Web  
Site](#)

[Opportunities for Public  
Comment](#)

[Presidential Libraries](#)

[Employment, Internships, and  
Volunteering.](#)

[NARA's Magazine: \*Prologue\*](#)

[Freedom of Information Act  
\(FOIA\)](#)

[The NARA Gift Shop](#)

[NARA Publications](#)

**Welcome to NARA:** Find speeches from the Archivist and Hot Topics. Learn about NARA's mission, history, values, Strategic Plan and performance measurements, program goals, partnerships, and more. . .

**The Research Room:** Discover NARA's nationwide holdings, learn about family history/genealogy research and veterans' service records, learn how to order reproductions, search the NARA Archival Information Locator (NAIL) database, locate Government documents and library materials, and more. . .

**Records Management, Storage, and Centers:** Find Federal records schedules, records management guidance, drafts for public comment, Federal records officers, **NEW!** Records Center Program **NEW!**, and more. . .

**The Federal Register:** Read the official text of Federal laws, regulations, notices and Presidential documents, get a list of documents appearing in upcoming Federal Register issues, learn about the Electoral College, and more. . .

**The Online Exhibit Hall:** See American Originals, the *Declaration of Independence*, the *Constitution of the United States of America*, and the *Bill of Rights*, World War II Posters, "When Nixon Met Elvis," and more. . .

**Digital Classroom:** Find teaching curriculum, students activities, and prepare for National History Day in The Digital Classroom, and more. . .

**NHPRC & Grants:** Discover available grants from the NHPRC (National Historical Publications and Records Commission) and Presidential Libraries, learn about the NHPRC, and more. . .

**Archives and Preservation Resources:** Find technical guidance concerning archival preservation and management, training for archivists and preservation professionals, and resources for at-home

record-keepers, genealogists, and more. . .

**Privacy Statement:** We do not provide personal data about our customers to any other parties without permission. For site improvement purposes, we log temporary information about the Internet capabilities of our online customers. [Read more.](#) .

**Terms and Conditions** for Using NARA's Web Site

**Contact NARA electronically:**  
[questions and comments.](#)

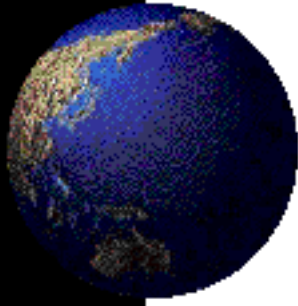
National Archives and Records Administration  
700 Pennsylvania Avenue, N.W.  
Washington, D.C. 20408

---

[National Archives and Records Administration home page](#)

URL: <http://www.nara.gov/index.html>  
[webmaster@nara.gov](mailto:webmaster@nara.gov)

Last Modified on May 31, 2000



# Digital Library Technology



DLT  
Projects

Project  
Sites

Program  
Reports

Digital  
Studio

IITA  
Activities

Affiliated  
Links

## Digital Library Technology Projects

- Projects Funded through the IITA Cooperative Agreement:  
["Public Use of Earth and Space Science Data Over the Internet"](#)
- Projects Funded through the NSF-ARPA-NASA Joint Initiative:  
["Research on Digital Libraries"](#)

---

[\[ Back to Main Menu \]](#)

Curator: Margaret Williams, [Margaret.E.Williams.1@gsfc.nasa.gov](mailto:Margaret.E.Williams.1@gsfc.nasa.gov) Responsible  
Official: Dr. Nand Lal, Project Manager, [Nand.Lal@gsfc.nasa.gov](mailto:Nand.Lal@gsfc.nasa.gov) Updated  
September 22, 1997

# The LIBRARY of CONGRESS

[SEARCH THE CATALOG](#) | [SEARCH OUR WEB SITE](#) | [ABOUT OUR SITE](#)

[America's Library: New Site for Kids & Families!](#) "Log On ... Play Around ... Learn Something"

## USING the LIBRARY

*Catalogs, Collections  
& Research  
Services*



## THOMAS

*Congress  
At Work*

## COPYRIGHT OFFICE



*Forms &  
Information*

## BICENTENNIAL 1800-2000

*Libraries • Creativity • Liberty*



## HELP & FAQs

*General Information*

## AMERICAN MEMORY

*America's Story in  
Words, Sounds  
& Pictures*



## EXHIBITIONS

*An On-Line  
Gallery*



## THE LIBRARY TODAY

*News, Events  
& More*



Above, the interior of the dome of the Main Reading Room of the Library of Congress

101 INDEPENDENCE AVE. S.E.  
WASHINGTON, D.C. 20540  
(202) 707-5000

Comments: [lcweb@loc.gov](mailto:lcweb@loc.gov)

[Please Read Our Legal Notices](#)

[USING the LIBRARY](#) | [THOMAS](#) | [COPYRIGHT OFFICE](#) | [AMERICAN MEMORY](#) | [EXHIBITIONS](#) | [The LIBRARY TODAY](#) | [BICENTENNIAL](#) | [HELP & FAQs](#) | [AMERICA'S STORY from AMERICA'S LIBRARY](#) | [TOP of PAGE](#)

# People:

---

[Rob Akscyn](#) of [Knowledge Systems Incorporated](#) with its [PetaPlex Project](#)

[William Arms](#), at [Cornell CS](#), formerly at [CNRI](#)

[Dan Atkins](#) [University of Michigan, DLI-1 Digital Library Project](#) Director.

[Howard Besser](#) of [School of Information Management and Systems at Berkeley](#)

[Bill Birmingham](#): [University of Michigan, DLI-1 Digital Library Project](#) Researcher.

[Chris Borgman](#) of [Information Studies at UCLA](#)

[Hsinchun Chen](#) Head of the [AI Lab of U. Arizona](#) and director of new [DLI-2 project](#)

[Stephan Fischer](#) - working on multimedia and metadata

[Edward A. Fox](#) Director of the [Digital Libraries Research Group](#) at Virginia Tech.

[Rick Furuta](#) of [CS at Texas A&M Univ.](#)

[Hector Garcia-Molina](#) In the [Stanford DB Group](#)

[Henry Gladney](#) at [IBM Almaden Research Laboratory](#)

[Robert Kahn](#) of [CNRI](#)

[Judith Klavans](#) of [Digital Libraries Projects at Columbia](#)

[Carl Lagoze](#) of [DL Research Group](#) of [CS at Cornell Univ.](#)

[John Leggett](#) of [CS at Texas A&M Univ.](#)

[Michael Lesk](#) Director of [NSF' IIS program](#) that runs the [Digital Libraries Initiative](#)

- [Images: Quantity is not always Quality - U. KY talk](#)
- [digital libraries](#)
- [library preservation](#)
- [information retrieval](#)
- [networking, etc.](#)
- [Projections for Making Money on the Web](#)

[Richard Lucier](#), University Librarian and Executive Director, [California Digital Library](#). See his related D-Lib [article](#)

[Clifford Lynch](#) Director of [CNI](#)

## [Gary Marchionini](#)

- Previously at [U. Md.](#) with its [DL Home Page](#)
- Now at [U. NC Chapel Hill School of Information and Library Science](#)
- [Encyclopedia article draft](#)
- [CACM April 1995 article](#)

[Michael Mauldin](#) ([home page](#), [Lycos](#), [CMU School of Computer Science](#))

[Bruce Schatz](#) Principal Investigator of [University of Illinois at Urbana-Champaign, DLI Project](#)

[Robin Sewell](#), co-PI with Hsinchun Chen (see above) on U. of Arizona DLI-2 project

[Marvin Sirbu](#) of [CMU Engineering and Public Policy](#)

- [publications available online](#)

[Terry Smith](#) from [Geography](#), Director of [Alexandria project](#) at [U. CA Santa Barbara](#)

[Robert Wilensky](#) Principal Investigator of [Berkeley DLI Project](#)

---

Note: for an extensive list of people involved in digital libraries, see the [Author Index](#) of D-Lib Magazine.

Note: for a list of some of the key people in the digital libraries field, see the report on this from a Delphi Study at [http://www.coe.missouri.edu/~is334/projects/Delphi\\_DL/StatementAnalysis.htm](http://www.coe.missouri.edu/~is334/projects/Delphi_DL/StatementAnalysis.htm): "By consensus, those identified in the rounds of the Delphi as the top ten (10) include: William Arms, Christine Borgman, Hector Garcia-Molina, Edward A. Fox, Carl Lagoze, Michael Lesk, Richard Lucier, Clifford Lynch, Gary Marchionini, Bruce Schatz, and Terence R. Smith."

---

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

---

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

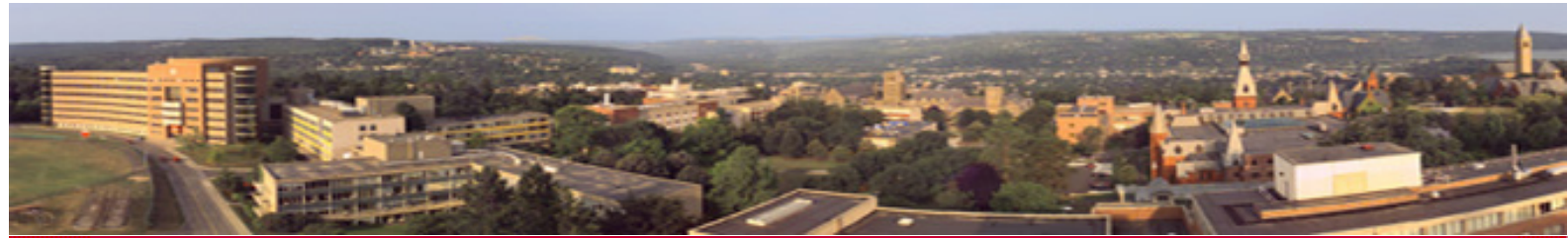
**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**

# President

Robert Akscyn is founder and President of Knowledge Systems

---

- [Address](#)
- [Research Interests](#)
- [Professional Activities](#)
- [Experience](#)
- [Education and Training](#)
- [Awards and Achievements](#)
- [Family](#)
- [Publications](#)



[Message](#)
[Interdisciplinary People](#)
[Facilities](#)
[Interactions Speakers](#)
[Highlights Publications](#)
[Data Table of Contents](#)
[Research](#)
[Faculty](#)

## William Y. Arms

Professor

[wya@cs.cornell.edu](mailto:wya@cs.cornell.edu)

**D.Phil University of Sussex,  
U.K., 1973**

My research interests concentrate on digital libraries and electronic publishing. These fields integrate methods from many disciplines, so that the work ranges from technical topics, such as distributed computing and information representation, to the economic and social aspects change.

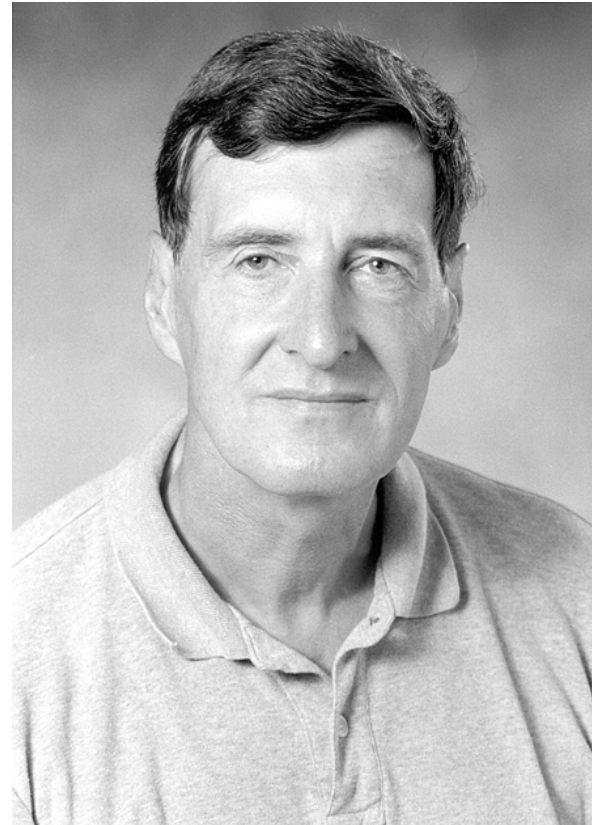
This year has seen great progress in reference linking, the

generalization of citations and hyperlinks to references among digital works. A general framework has been developed and is being applied to several large sources of information. We plan to use these results to analyze changes in use patterns of scientific literature. This follows a long-standing interest in quantitative research into information systems. With support from DARPA, we have created a test suite for digital libraries research, to develop systematic metrics and replicable research results.

Several people in the department are involved in the changes in computer science publication as online material replace printed journals as the primary means of creating, storing, and distributing research information. I chair the ACM publications board, Joe Halpern has led the creation of the CoRReprint repository for computer science, and Carl Lagoze heads the Networked Computer Science Research Library (NCSTRL).

### Professional Activities

- Chair: Publications Board, Association for Computing Machinery.
- Chair: National Science Foundation workshop on a National Digital Library for Undergraduate Science, Mathematics, Engineering and Technology Education.
- Program committee: ACM Digital Libraries '99 conference
- Series editor: MIT Press series on Digital Libraries and Electronic Publishing.
- Editor-in-chief: D-Lib Magazine



## Publications

"The D-Lib Test Suite: testbeds for digital libraries research." *D-Lib Magazine*, February 1999 5(2).



---

# **Daniel E. Atkins**

## **Professor**

Also Dean, School of Information

## **Electrical Engineering and Computer Science**

300 West Hall 1092

(313) 647-3576

[atkins@umich.edu](mailto:atkins@umich.edu)

Homepage: <http://www.sils.umich.edu/People/atkins.html>

## **Degrees:**

BS '65, Bucknell; MS (EE) '67, PhD (CompSci) '70, U-Illinois

## **Research Interests:**

Computer architecture, computer-support cooperative work, digital libraries

---

Please send questions or feedback to [engin-web-info@umich.edu](mailto:engin-web-info@umich.edu)

# Howard Besser

Spring 2000 office hours: Mon 4-5, 241 GSE&IS Building, (310)825-8975 *and by appointment*

[-9/99 Vietnam Trip](#)

[-address and travel schedule](#)

[Vita](#) | [partial list of publications](#) | [online papers and articles](#)

-Spring 00 class on [Visual Materials: Metadata, Standards, and Best Practices for Digital Libraries](#),

Winter 00 course on [Digital Collections of Still & Moving Images](#)

and Fall 99 course on [Impact of New Information Environments](#)

-Spring 99 courses on: [Digital Visual Materials in Cultural Heritage](#); [Web Design and Development](#)

-Fall 98 courses on: [Web Design and Development](#); [Impact of New Information Technologies](#)

[-Howard's new LA image](#) | [Howard's first week in LA](#)

## Sections on this page

- [Writings on Social Effects of Information Technologies](#)
- [Teaching with technology and Distance-Independent Learning activities](#)
- [Digital Libraries, Standards, & Longevity Activities](#)
- [Work on Multimedia DBs](#)

Howard Besser (howard@sims.berkeley.edu) is Associate Professor at [UCLA's School of Education and Information Studies](#) where he teaches, does research, and supervises projects. His four main interest areas are Multimedia Databases (particularly in cultural institutions), the social and cultural effects of information technology, digital library issues (particularly around standards, longevity, and intellectual property), and the development of new ways to teach with technology (including web-based instruction and distance learning). He is particularly interested in design issues and the use of critical theory perspectives.

Dr Besser has been on the faculty of UC Berkeley's [School of Information Management & Systems](#), and is affiliated with the [Berkeley Multimedia Research Center](#). From 1994-96 he was on the faculty of the University of Michigan's [School of Information](#) where he headed a committee developing a curriculum in multimedia and digital publishing. He has also been an Assistant Professor at the University of Pittsburgh.

Howard is also actively involved with museums and the art community. He was one of the founders and served on the Management Committee of the [Museum Educational Site Licensing Project](#), and directed a Mellon-sponsored [study of image distribution from museums to universities](#). For several years he was in charge of long-range information planning for the [Canadian Centre for Architecture](#) in Montréal, and for many years he headed information technology for Berkeley's University Art Museum. His most recent work involves examining issues of organization, access, and longevity for new media art in collaboration with the [Electronic Café International](#) and a [group of museums with electronic art](#)



[collections.](#)

He [travels](#) a lot, speaks frequently at professional conferences, gives workshops on Image Databases or on Metadata about half a dozen times a year, and consults for libraries, museums, and other institutions. For several years he served as co-chair of the American Library Association's [Technology & the Arts](#) Interest Group (co-sponsored by the Association of College & Research Libraries and the Library Information Technology Association).



[See what's planned at upcoming conferences and what he's learned at selected conferences](#)

## Social Effects of New Information Technology

Versions of some of his papers and talks on the social effects of new information resources:

- [The Changing Role of Photographic Collections With the Advent of Digitization](#), draft of chapter to appear in Katherine Jones-Garmil (ed.), *Museums and Emerging Technologies*, Washington: American Association of Museums, 1996
- [The Transformation of the Museum and the Way it's Perceived](#), draft of chapter to appear in Katherine Jones-Garmil (ed.), *Museums and Emerging Technologies*, Washington: American Association of Museums, 1996
- [The Information SuperHighway: Social and Cultural Impact](#) Chapter from [Resisting the Virtual Life: The Culture and Politics of Information](#), edited by Jim Brook and Iain Boal, City Lights Books, 1995
- [Movies-on-demand May Significantly Change the Internet](#) appeared in the October 1994 *ASIS Bulletin* theme issue on Entertainment Technology and Information Services
- [A Clash of Cultures on the Internet](#) Op Ed piece appeared in *San Francisco Chronicle* August 25, 1994
- [The Changing Role of Photographic Collections With the Advent of Digitization](#) Discussion Paper for Working Group for Digital Image in Curatorial Practice, George Eastman House, June 4, 1994; (get [complete conference proceedings](#))
- [The Information Highway must be a Two-Way Street: The Arts and Humanities Communities Cannot be merely Consumers](#) Presentation to the Convergence Conference: Arts and Humanities and the NII, Oct, 1994
- [Use of Non-Broadcast Channels to Communicate Information In Social Change Situations: Berkeley Anti-Apartheid and Solidarity Poland](#)
- [Elements of Modern Consiousness](#) (excerpts from his doctoral dissertation which he forces his

students to read)

- **Older Activities**

- Documentation on his Winter 1995 course on the [Impact of Multimedia and Networks](#)
- 1995 conference on [Ethics and the Internet](#) and [images from Howard's presentation](#)
- [Other interesting sites](#)

## Teaching with Technology

For many years Howard has been actively involved in developing and testing new methods for using technology to teach. In recent years those efforts have focused on Distance-Independent Learning. Howard edited a [Perspectives issue on distance education](#) for the **Journal of the American Society for Information Science** which appeared in Nov of 1996. In Winter 1995 he taught a course on the [Impact of Multimedia and Networks](#) in which students in both Ann Arbor and Berkeley used a wide variety of technologies to interact with one another both in class as well as to collaborate on projects. In 1999 he is co-teaching a class with half the students in Berkeley and the other half at UCLA.

Since Spring 1994 Howard has been using the Internet as the key delivery system for instructional support, placing curricular materials on the WorldWide Web, having students engage in online discussion groups, and making students read the online work of previous students and incorporate this work into their own Web pages and online discussions. And since 1997 he has been teaching a course in good Web Design and directing a grant project that hires students from his department to develop well-designed online web-based delivery systems for course materials for large undergraduate classes on the UC Berkeley campus.

Howard helped found the [Museum Educational Site Licensing Project](#) (MESL) as a way to provide digital images for instruction. Recently he has been examining the new instructional strategies being developed to teach with these images. At Michigan Dr. Besser also worked both to develop new curriculum that relied extensively on technology, and he worked on design of instructional technology labs to support extensive teaching with technology. Under a grant from the Kellogg Foundation to the University of Michigan [to revamp curriculum](#), Howard chaired a [subcommittee](#) that [examined the creation and design of digital documents](#), and another committee that was designing a new Information Studies Media Lab.

### Howard's Distance Learning papers, journal articles, etc.

- [Special Issue on Distance-Independent Education](#) , **Journal of the American Society of Information Science** 47(11), Nov 1996 ([official table of contents](#))
- [Issues and Challenges for the Distance-Independent Environment](#), **Journal of the American Society of Information Science** 47(11), Nov 1996 *online access restricted to users within the UC Berkeley domain*
- [The Impact of Distance-Independent Education](#), **Journal of the American Society of Information Science** 47(11), Nov 1996 *online access restricted to users within the UC Berkeley domain* (co-authored with Maria Bonn)

- *Interactive Distance-Independent Education: Challenges to Traditional Academic Roles*, **Journal of Education for Library and Information Science** 38 (1), Winter 1997, pages 35-42 (co-authored with Maria Bonn)
- *Multimedia and Networks Teach about Museums: Issues in Maintaining a WWW Site to Facilitate Distance Learning*, in David Bearman (ed.), **Multimedia Computing and Museums** (Selected Papers from the Third International Conference on Hypermedia and Interactivity in Museums), Pittsburgh: Archives & Museum Informatics, 1995, pages 124-140
- [Distance Learning in the Humanities & Social Sciences: Doing it, Supporting it, and Looking at its Impact](#), Howard Besser's lecture to Advanced Information Technologies Group and Digital Library Research Program, University of Illinois, February 10, 1997 ([Real Audio](#))
- [Howard's Distance Education links](#)

## Curricular Support Material on the Web

### Howard's Projects, Courses

- [Experimental Michigan Berkeley Distance class](#), Winter 1995 ([course evaluation](#) | [classroom images](#))
- Externally funded [project to provide online web-based delivery of course materials for UC Berkeley undergraduate courses](#) (also taught as a course from 1997 on); [Powerpoint summary of project](#)
- Ongoing course on the [Impact of Multimedia and Networks](#) (content on the Web since 1994)
- The [Museum Educational Site Licensing Project](#) (MESL), an experimental project to provide digital images and metadata for campus instruction and research.

### Howard's Papers

- [Difficulties of Implementing and Maintaining a WorldWide Web Site to Support Instruction](#), *Revue Informatique et Statistique dans les sciences humaines*, 1996, 32(1-4), pages 11-28 *online access restricted to users within the UC Berkeley domain*
- The [UC Berkeley Mellon Study of the Cost and Use of Digital Images](#) as part of the MESL project

## Misc

- [Education as Marketplace](#), book chapter from Muffoletto, R., Knupfer, N. (1993) **Computers in education: Social, historical, and political perspectives**, New Jersey: Hampton Press. (*online access restricted to users within the UC Berkeley domain*)
- Howard's [Web Teaching, Instructional Technology links](#)
- [Links to other distance-independent education sites](#) (from 1996)

# Digital Libraries, Standards, Metadata, & Longevity Activities

Howard is involved in a variety of activities around Digital Libraries, Standards, and Longevity, including:

- A member of the National Research Council/National Academy of Sciences' committee on [Intellectual Property Rights and the Emerging Information Infrastructure](#)
- Member of the [Technological Standards and Architectures Working Group](#) of the California Digital Library
- Author of [Best Practices for Image Capture](#) (focused on scanning and administrative metadata) for the California Digital Library (7/99)
- Co-organizer of the [National Information Standards Organization Invitational Meeting on Technical Metadata Elements for Image Files](#) ([Howard's opening presentation](#) | [Howard's meeting summary](#))
- Participant in [Time & Bits](#), a small meeting on longevity of digital information organized by the Getty Information and Conservation Institutes in association with Stuart Brand
- Co-author (with Peter Lyman) of a paper on [issues of longevity of digital information](#) as part of the Time & Bits Meeting
- Was a member of the [task force examining the archiving of digital information](#) (sponsored by the Commission on Preservation & Access and the Research Libraries Group)
- Maintains a set of [links to resources on Digital Longevity](#)
- Co-author of [Making of America II White Paper](#), and participant in the MOA2 project to define structural and administrative metadata standards for digital representations of photographs, photo albums, diaries, letterbooks, and other archival materials (Sponsored by the [Digital Library Federation](#))
- Member of the [Metadata Working Group](#) of the joint NSF/European Community [Digital Library Collaboratory](#)
- Maintains a list of [current important standards activities](#)
- Participated in the development of the [Dublin Core](#), the metadata standard to describe Network Objects (first meeting sponsored by NCSA and OCLC)
- Produced a document on [Standards for Images](#) (an effort that is being coordinated with both the Coalition for Networked Information and the Computerized Interchange of Museum Information)

## Multimedia Databases

Here are some of his current and recent multimedia database activities:

- Maintains a list of [links to image database-related information on the WWW](#).
- Has written an [Introduction to Image Databases](#) published both electronically and in print form by the [Imaging Initiative](#) of the [Getty Information Institute](#).
- Has co-authored [The Museum Educational Site Licensing Project: Technical Issues in the](#)

[Distribution of Museum Images and Textual Data to Universities](#) with Christie Stephenson

- Is recipient of a \$250,000 grant from the Andrew W. Mellon Foundation [to study the costs and benefits of networked distribution of digital museum information for educational use](#) as part of the Museum Educational Site Licensing Project
- On the Governing Board of the [Museum Education Site Licensing Project](#).
- [Guidelines for Special Collections](#) contemplating putting images up on the WWW
- Since Spring 1994 has taught a course on the [Impact of New Information Technologies: Multimedia and Networks](#) ([Winter 95 Distance Learning class](#), [Fall 96 class primarily oriented towards MBAs](#))
- Keeps a [short list of interesting jobs](#)
- **Older Activities**
  - Created a narrative description of [interesting image database resources and projects](#) (around 1994-95)
  - Helped develop [criteria for evaluation of multimedia programs](#) for the National Engineering Education Delivery System's Premier Awards
  - [Interesting readings and resources](#) compiled for his Image Database class
  - Arranged for his students to produce an [index to Internet accessible museum information from around the world](#). [newest museum entries](#)
  - Compiled an index to [Moving Image Resources on the Net](#).
  - Taught a [Fall 1995 Image Database class](#)
  - Taught a Fall 96 class on the [Protection of Digital Information](#)
  - [interesting Web sites](#)
  - [additional interesting Web sites](#) prepared for 7/18/95 talk to Korean visitors
  - [Impact on Haas Server](#)



[Irv Besser dies 5/7/97](#)

[Howard's T-Shirt image database](#)

[1996 Ann Arbor T-Shirt Exhibition](#)

[UCB Lecture Series on Art, Technology, & Culture](#), organized by Ken Goldberg

[Rants on current events](#) (including the [Unabom](#))

[Anarchist links](#)



[Course material from previous multimedia classes](#)

[Loki's HomePage](#)

# Bill Birmingham

---

Associate Professor

- [Artificial Intelligence Laboratory](#)
- [Department of Electrical Engineering and Computer Science](#)
- [University of Michigan](#)

Joint Appointment with [School of Information \(SI\)](#)

Office Address:

128 ATL Building  
1101 Beal Avenue  
Ann Arbor, MI 48109-2110

Email: [wpb@eecs.umich.edu](mailto:wpb@eecs.umich.edu)

Phone: (734) 936-1590

FAX: (734) 763-1260

A [brief bio](#) is available.

The call for participation for [AAAI 2000 Workshop on Artificial Intelligence and Music: Towards formal models for composition, performance, and analysis](#) is now available. (Check out the official [website](#)).

---

## Research Interests

It is not a question of *whether* or *when* the arts and technology will become integrated - anthropologists attest to the entwining of the two throughout human history. Rather, it is a question of *where* in the next century this integration will lead us. We have started **MusEn** [research program](#) that integrates music with engineering and a graduate program in the Media Arts. We believe the future lies in the development and use of media technologies that support the creation and experience of art.

## Research Objectives

To advance a comprehensive interdisciplinary research program that develops technologies in support of the arts and explores artistic expression as a model for new technologies.

- Advance the fields of music theory, musicology, composition and performance by developing and integrating technologies and research methodologies that provide user-friendly and transparent systems for the analysis, creation and realization of music.
- Advance the fields of signal processing and computer science by developing computable formalisms for musical style and expanding the domain of knowledge engineering.

· Develop technologies for the advancement of performance pedagogy, analysis of twentieth-century electronic music, and new performance and compositional interfaces.

[Hierarchical Concurrent Engineering](#) (HCE) is a model of concurrent engineering that attempts to do two things: maximize concurrency in a concurrent-engineering process through decentralized, distributed decision making, and optimize through shared preference structures and constraint networks. In HCE, designers are represented as rational decision makers that are part of a network composed of constraints and (partially) shared, hierarchical preference structures. A key aspect of HCE is that it stresses decentralized decision making by designers: decentralization provides increased concurrency during the design process, makes modeling the design process easier, and has the potential to scale well. A fledgling activity involving AI and Fine Arts is just beginning. The overall activity is exploring computational methods for analyzing, synthesizing, and evaluating (the aesthetics) artistic expression.

---

### Graduates students in my research group:

- [Eric Glover](#)
  - Bryan Pardo
  - Stuart McAlpin
- 

### Teaching Activities:

- [EECS 592](#): Advanced Artificial Intelligence ( [Winter 2000](#) Winter '97, Winter '96)
  - [EECS 492](#): Introduction to Artificial Intelligence ([Fall 99](#) Winter 99, Fall 97)
  - EECS 370 ( [Fall 98 Winter '98](#) )
  - EECS 598-7: Computation, Science and the Arts ( [Fall '97](#) )
  - [School of Information](#) : SI 612 Agent Systems Design Lab ([Winter '97](#))
  - [School of Information](#) : SI 609: "[Foundations](#)" (Fall '96)
  - [EECS 373: Design of Microprocessor-based systems](#) (Fall '96)
  - [EECS 543: Knowledge Systems](#) (Fall '95)
- 

### Associated activities

I am the editor of the journal **AIEDAM** ( [Artificial Intelligence for Engineering Design, Analysis and Manufacture](#) ).

If you are interested in submitting a paper to the journal, **please send it to me at the address listed in this page** (do not send the paper to the Emeritus Editor, Prof. Dym).

- List of upcoming [special issues](#)
  - Interested in becoming a special issue editor? Here are the [guidelines](#).
-

Last Updated: January 18, 2000



# Researcher



About Us

Teaching

Research

*People*

Jobs

Location

Events

News

Student Info

Search



## Dr. Stephan Fischer

Merckstr. 25  
64283 Darmstadt, Germany

Phone: +49-6151-166161  
Fax: +49-6151-166152

Email: [Stephan.Fischer@KOM.tu-darmstadt.de](mailto:Stephan.Fischer@KOM.tu-darmstadt.de)

[Link to Private Homepage](#)



---

Please contact our Webmaster with questions or comments.  
A quick overview of KOM is available here.

# Dr. Richard Furuta



## Associate Professor

Department of Computer Science  
Texas A&M University  
College Station, TX 77843-3112

**Office:** 402C H. R. Bright Building

**Phone:** (409) 845-3839

**Fax:** (409) 847-8578

**Email:** [furuta@cs.tamu.edu](mailto:furuta@cs.tamu.edu)

- [Education and Experience](#)
- [Honors and Professional Activities](#)
- [Research](#)
- [Publications](#)
- [Personal Home Page](#)

# Hector Garcia-Molina

**Professor, Departments of Computer Science and  
Electrical Engineering**

---



---

## Contact Information

Office: Gates Hall 4A, Room 434  
Phone: (650) 723-0685  
Fax: (650) 725-2588  
Email: [hector@cs.stanford.edu](mailto:hector@cs.stanford.edu)

Address: Department of Computer Science  
Stanford University  
Gates Hall 4A, Room 434  
Stanford, CA 94305-9040 USA

Assistant: Marianne Siroker  
(650) 723-0872  
[siroker@cs.stanford.edu](mailto:siroker@cs.stanford.edu)

---

# Biographical Sketch

Hector Garcia-Molina is the Leonard Bosack and Sandra Lerner Professor in the Departments of Computer Science and Electrical Engineering at Stanford University, Stanford, California. From August 1994 to December 1997 he was the Director of the Computer Systems Laboratory at Stanford. From 1979 to 1991 he was on the faculty of the Computer Science Department at Princeton University, Princeton, New Jersey. His research interests include distributed computing systems and database systems. He received a BS in electrical engineering from the Instituto Tecnologico de Monterrey, Mexico, in 1974. From Stanford University, Stanford, California, he received in 1975 a MS in electrical engineering and a PhD in computer science in 1979. Garcia-Molina is a Fellow of the ACM, received the 1999 ACM SIGMOD Innovations Award, and is a member of the President's Information Technology Advisory Committee (PITAC).

---

## Publications

You may search for my recent publications at  
[DB Group Publications](#).

---

[My son David's](#) home page

[Database Group](#) home page

[Computer Science Department](#) home page

---

# Henry Gladney:

---

- Access Control Articles in D-Lib Magazine:  
Gladney et al., Safeguarding Digital Library Contents and Users:
  - [Assuring Convenient Security and Data Quality](#),
  - [Document Access Control](#)
  - [Digital Images of Treasured Antiquities](#)
  - [A Note on Universal Unique Identifiers](#)
  - [Storing, Sending, Showing, and Honoring Usage Terms and Conditions](#)
- [Gladney et al. report on DL requirements and architecture \(PostScript\)](#)

---

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

---

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

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



**Robert E. Kahn**

President, CNRI  
1895 Preston White Drive  
Suite 100  
Reston, Virginia 20191-5434  
rkahn@cnri.reston.va.us  
(703)620-8990 (tel.)  
(703)620-0913 (fax)

---

Robert E. Kahn is Chairman, CEO and President of the Corporation for National Research Initiatives (CNRI), which he founded in 1986 after a thirteen year term at the U.S. Defense Advanced Research Projects Agency (DARPA). CNRI was created as a not-for-profit organization to provide leadership and funding for research and development of the National Information Infrastructure.

After receiving a B.E.E. from the City College of New York in 1960, Dr. Kahn earned M.A. and Ph.D. degrees from Princeton University in 1962 and 1964 respectively. He worked on the Technical Staff at Bell Laboratories and then became an Assistant Professor of Electrical Engineering at MIT. He took a leave of absence from MIT to join Bolt Beranek and Newman, where he was responsible for the system design of the Arpanet, the first packet-switched network. In 1972 he moved to DARPA and subsequently became Director of DARPA's Information Processing Techniques Office (IPTO). While Director of IPTO he initiated the United States government's billion dollar Strategic Computing Program, the largest computer research and development program ever undertaken by the federal government. Dr. Kahn conceived the idea of open-architecture networking. He is a co-inventor of the TCP/IP protocols and was responsible for originating DARPA's Internet Program which he led for the first three years. Dr. Kahn also coined the term National Information Infrastructure (NII) in the mid 1980s which later became more widely known as the Information Super Highway.

In his recent work, Dr. Kahn has been developing the concept of a digital object infrastructure as a key middleware component of the NII. This notion is providing a framework for interoperability of heterogeneous information systems and is being used in several applications such as the electronic copyright registration system at the Library of Congress and its National Digital Library Program. He is a co-inventor of Knowbot programs, mobile software agents in the network environment.

Dr. Kahn is a member of the National Academy of Engineering and a former member of its Computer

Science and Technology Board, a Fellow of the IEEE, a Fellow of AAAI, a recipient of the AFIPS Harry Goode Memorial Award, the Marconi Award, the ACM SIGCOMM Award, the President's Award from ACM, the IEEE Koji Kobayashi Computer and Communications Award, the ACM Software Systems Award, the Computerworld/Smithsonian Award, the ASIS Special Award and the Public Service Award from the Computing Research Board; he was twice the recipient of the Secretary of Defense Meritorious Civilian Service Award. Dr. Kahn is a former member of the Board of Regents of the National Library of Medicine and the Presidents Advisory Council on the National Information Infrastructure. He has been designated as recipient of the 1997 IEEE Alexander Graham Bell Medal.

---

*Last updated 1/6/97*

[ [home](#) | [officers & directors](#) | [programs & activities](#) | [publications](#) ]





***Judith L. Klavans, Ph.D.***

Director, [Center for Research on Information Access](#)

---

## Research Interests

My research lies in computational linguistics and natural language processing. I am currently working on ways to represent meaning from texts and to link meaningful segments via semantic nets. I have been involved in the Text Encoding Initiative, an SGML based set of guidelines for indepth text representation, which has led me to be interested in mark-up languages in general. I am also working on ways to automatically determine lexical information via statistical means over parsed texts. I have worked on linguistic and statistical methods for extracting and linking bilingual phrasal verbs from corpora (i.e. the many to one problem in bilingual corpora), and am also currently involved in a joint project on the expansion and conflation of technical terms in French for monolingual text indexing and for use in bilingual information access.

Prior to arriving at Columbia, I spent nearly ten years at the TJ Watson IBM Research Division, where I worked on extracting information from machine-readable dictionaries. I have also worked on speech synthesis and text-to-speech systems.

The kinds of systems that have used the results of my research have been: natural language parsers, machine translation systems, text critiquing systems, and topic identification systems.

---



[Resume](#)



[Recent Publications](#)



**Selected Ongoing Research Projects.** See also [Activities of the Center for Research on Information Access](#)

- NSF STIMULATE - [Generating Coherent Summaries of On-Line Documents: Combining Statistical and Symbolic Techniques](#)
  - Project report from 1997 PI Meeting
  - Slides from March 6, 1998 PI Meeting

- [Document Segmentation](#)
- [Role of Verbs in Document Analysis](#)
- [Digital Libraries Integration](#)
- [Linguistic Techniques for Bilingual Terminology Conflation \(joint with Christian Jacquemin and Evelyne Tzoukermann\)](#)
- [Spring 1998 Student Research Projects](#) (for participants only)



### **Selected Recent Professional Activities**

- 
- [Presentations from the European Conference on Digital Libraries at Crete, 1998.](#)
- [NSF Information and Data Management Workshop: Research Agenda for the 21st Century.](#) - March 1998
- National Science Foundation and Digital Library Federation Workshop on Rights Management - 4/6/98 [Rights Management](#)
- [Multilingual Information Access Working Group](#) - 1997-98
- [PODS Tutorial - June 1998](#)
- Grace Hopper Conference - September 1997
  - [Abstract of Keynote](#)
- Keeping Up with Information Growth on the Web  
SCIP - March 28, 1996
  - [Course Outline](#)
- Workshop on the Text Encoding Initiative and Digital Libraries - March 23, 1996  
In conjunction with DL-96, sponsored by the Association for Computing Machinery ACM
  - [Call for Papers](#)
  - [Final Program](#)
- Tutorial on the Reusability, Interchangeability, and Compatibility: Answering the Questions of Text Encoding Standards (ACM-SIGIR) - July 1995
  - [Abstract](#)
- Symposium on Computational Lexical Semantics - April 1995  
American Association for Artificial Intelligence AAAI
  - [Call for Papers](#)
  - [Final Program](#)
- Workshop on Computational Linguistics -  
Linguistics Society of America LSA
  - [Final Program](#)

### **Funding**

- Workshop on Technology Issues for Terms and Conditions (NSF) PI: Judith L. Klavans & James R. Davis
- STIMULATE: Generating Coherent Summaries (NSF) PI: Judith L. Klavans & Kathleen McKeown
- Significant Topics (NSF) PI: Judith L. Klavans
- Multilingual Information Access Working Group (NSF) PI: Judith L. Klavans
- Workshop - Information and Data Management (NSF) PI: Judith L. Klavans



### **[Department of Computer Science](#)**



### **[Columbia University Libraries](#)**



**Contact Information:**

Director, CRIA - Center for Research on Information Access  
Department of Information Services  
535 West 114th Street, New York, NY 10027  
212-854-7443 (phone)  
212-854-9099 (fax)  
Research Scientist, Department of Computer Science  
500 West 120th Street, New York, NY 10027

*Send any comments to [klavans@cs.columbia.edu](mailto:klavans@cs.columbia.edu)*

006338 accesses since 3/1/96

*This page is located at <http://www.cs.columbia.edu/~klavans/home.html>*

**This page was last updated on 2/5/98**



## Carl Lagoze Digital Library Scientist Cornell University

*Department of Computer Science  
4112 Upson Hall  
Cornell University  
Ithaca, NY 14850-7501  
Phone: +1-607-255-6046  
Fax: +1-607-255-4428  
Internet: [lagoze@cs.cornell.edu](mailto:lagoze@cs.cornell.edu)*

I lead Digital Library research efforts in the [Computer Science Department](#) at [Cornell University](#). I am also affiliated with the [University Library](#) and [Cornell Information Technologies](#), with whom I collaborate on a number of Digital Library and Electronic Publishing activities.

[Research](#)

[Publications](#)

[Professional Activities  
and Talks](#)

[Personal](#)



*Page last updated: June 1, 1999*

# Michael Lesk's Grade Crossing on the Information Superhighway

Please change any address/link to this page to **<http://www.purl.net/NET/lesk>**. The address `purl.net' refers to `permanent URL' and this address should survive local administrative changes. Thank you.



Professional



Amateur



Coin-operated

---

Now out: my new book *Practical Digital Libraries: Books, Bytes and Bucks*, [Morgan Kaufmann](#), July 1997.

---

**Position:** Division Director, Information and Intelligent Systems, National Science Foundation, <http://www.cise.nsf.gov/iis>.

**Also:** Visiting Professor, University College London, Department of Computer Science.

## Biography

In the 1960's I worked for the SMART project, wrote much of their retrieval code and did many of the retrieval experiments, as well as obtaining a PhD in Chemical Physics. In the 1970's I worked in the group that built Unix and I wrote Unix tools for word processing (*tbl*, *refer*), compiling (*lex*), and networking (*uucp*). In the 1980's I worked on specific information systems applications, mostly with geography (a system for driving directions) and dictionaries (a system for disambiguating words in context), as well as running a research group at Bellcore. And in the 1990s I have worked on a large chemical information system, the CORE project, with Cornell, OCLC, ACS and CAS.

I am also Visiting Professor in computer science at University College London; I'm on the Visiting Committee for the Harvard University Library; and I've worked with the Commission on Preservation and Access addressing digital preservation issues. I received the ``Flame" award for lifetime achievement from Usenix in 1994, and I am a Fellow of the ACM. You can read my [publication list](#) if you wish. The

previous paragraph is available in [Japanese](#).

## Where?

Michael Lesk

[National Science Foundation](#)

4201 Wilson Boulevard, Room 1115

Arlington, Virginia 22230

703 306-1930 [Voice]

703 306-0599 [Fax]

lesk@acm.org

---

## Interests

[Digital Libraries](#)



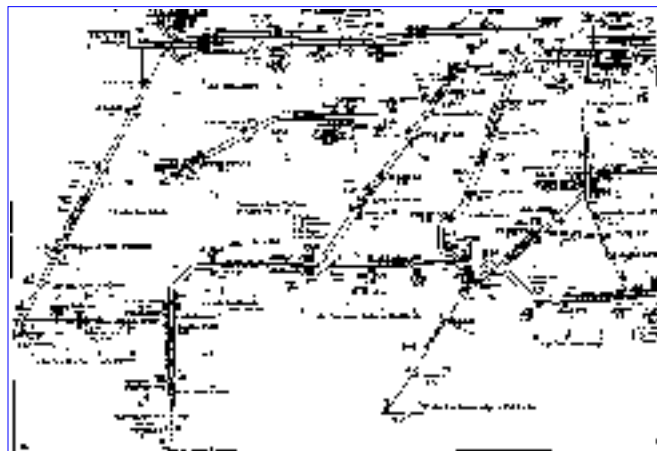
[Library preservation](#)



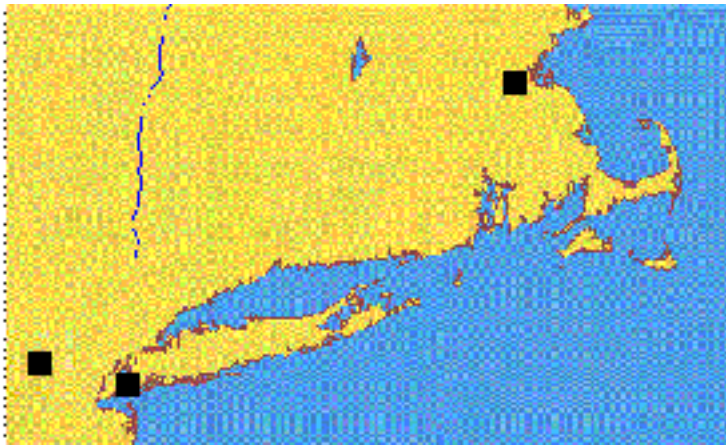
[Information Retrieval](#)



[Networks & Misc.](#)



## Places I have lived



[New Jersey](#) . . [Brooklyn](#) . . . . . [Cambridge, Mass.](#) . . . . . [London](#)  
..... [transport](#)

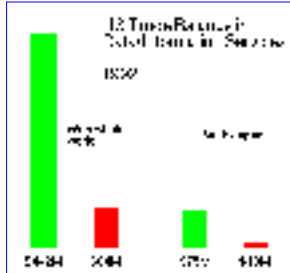
---

[Serving Human Needs Through Human Centered Systems](#). Draft of NSF subgroup report from workshop held February 1997. Contributors to text include Gio Wiederhold, Ben Shneiderman, and Jim Hollan.

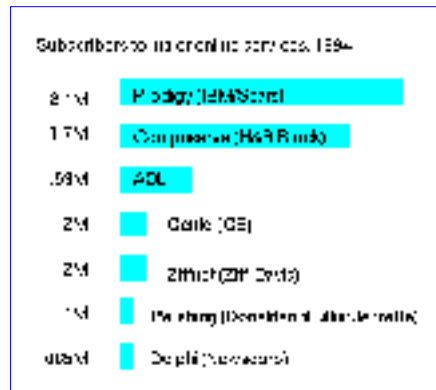
lesk@bellcore.com Michael Lesk  
Last changed: 3 June 1998

# Factoids

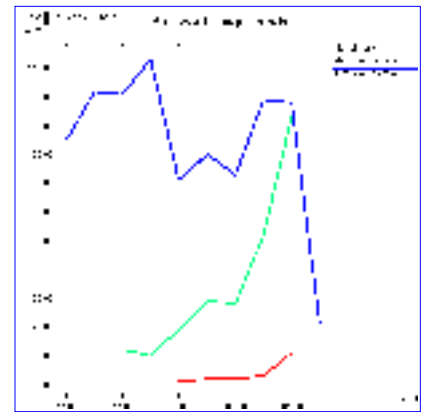
United States balance of trade in information services



Number of customers of online services



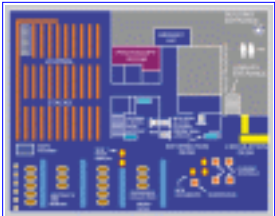
Trends in buzzwords



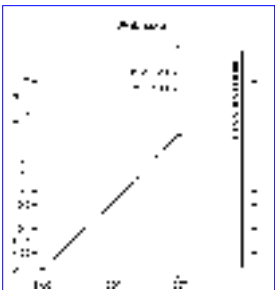
## Material on digital libraries



[\*US Digital Library Programs: What Goals?\*](#)



[\*The Organization of Digital Libraries\*](#)



[\*How Much Information Is There in the World?\*](#)



[\*Digital Libraries: A Unifying or Distributing Force?\*](#), to be presented at *Scholarly Communication and Technology*, a conference sponsored by the Andrew W. Mellon Foundation, Atlanta, Georgia (April 24, 1997).



[\*Mad Library Disease: Holes in the Stacks\*](#), Lazerow Lecture, given at University of California Los Angeles, 18 April 1996, to appear in print later in 1996



[\*Libraries and the Web\*](#), to appear *Libraries and Information World Wide*, 1996



[\*Economics of Digital Libraries\*](#), course outline for lectures given Jan-Apr 1996, Columbia University, New York.



[\*Why Digital Libraries\*](#), Follett lecture on electronic libraries, given 19 June 1995, BBC Conference Center, London, England.



[\*The Future Value of Digital Information and Digital Libraries\*](#); lecture given 9 November 1995 at the Kanazawa Institute of Technology Roundtable on Libraries and Information Systems, Kanazawa, Japan.



[\*Making a Digital Library: The Contents of the CORE Project\*](#); draft paper, October 1994; to appear, ACM TOIS

# US Digital Library Programs: What Goals?

## Introduction.

Humanities scholars in the United States can look forward to three kinds of benefits from the very diverse digital library research in the country. The NSF/DARPA/NASA Digital Library Initiative is producing new ways to access material; the Digital Library Federation is digitizing considerable amounts of special collections; and the JSTOR effort started by the Andrew W. Mellon Foundation looks at economic access to traditional materials. Of these, JSTOR may well have the largest immediate effect, as it both provides a new kind of access to the materials humanities scholars have often used, and a way of extending access to places that have never had it.

The three programs described above have widely differing goals. The Digital Library Initiative is six large projects, mostly consisting of computer science research. Each project receives \$1M per year from the government and supplements it with private contributions from partnering organizations. Primary emphasis is on improving our ability to search, organize, and display either new kinds of media or old material in new ways. The focus is on the technology rather than the material handled. The Digital Library Federation is more loosely connected, with the member libraries carrying on projects mostly based on their own funding. The largest single project is that of the Library of Congress, whose effort on the American Memory project is digitizing 5 million items over five years. The typical DLF project is converting special collections material. JSTOR, by contrast, is an economics-focussed effort. The material in the project is key journals, widely held in libraries, and the main question to be answered is whether a subscription model for access can become self-supporting. JSTOR tries to avoid a future in which NEH is asked for more and more money to digitize everything, producing a Congressional image of welfare queens in tweed jackets.

From the standpoint of humanists, these projects want to answer quite different questions. DLI asks 'how can we find new things?' while NDLF asks 'what old things can we digitize?' and JSTOR asks 'how can we make conversion into a self-supporting activity?' Both JSTOR and NDLF are collection-based; it makes sense to start off asking *what* will be available. DLI, in contrast is about things will be available, and less about what they are. Humanists can look more to JSTOR to make conventional humanistic materials more easily studied, and NDLF to make a wider range of primary subject materials more readily available. DLI is potentially an expansion of what is now considered subject material, but it's harder to know just how it will develop.

## The Digital Library Federation

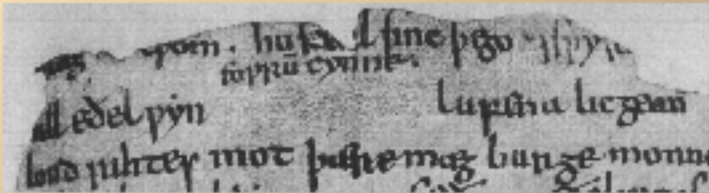
Until recently this was known as the National Digital Library Federation, but it shortened its name (although so far it has not yet taken in any non-US members). It has fifteen members: Columbia, Cornell, Emory, Harvard, the Library of Congress, the National Archives and Records Administration, the New York Public Library, Penn State, Princeton, Stanford, UC-Berkeley, Michigan, USC (University of Southern California), Tennessee, and Yale. In many cases, the material being digitized is from special collections. There are several reasons for choosing to digitize special collections instead of conventional

books.

- Typically, special collections items are unique, so it is more of a service to scholars to put them on the Web than to digitize items which existed in many copies. Printed works often exist in a library near a scholar; manuscripts or photographs are usually only in one place.
- Special collections items are often fragile, oversize, or otherwise in need of particular care, and as unique items they are irreplaceable, so that replacing their use with the use of some digital surrogate helps with preservation. It can also become much faster to look at them digitally than to browse materials which must be handled slowly and carefully to avoid deterioration.
- Sometimes, special collections materials present fewer copyright problems than conventional materials. A library may have obtained a large amount of material all of which is controlled by one copyright holder, and been able to obtain permission for the use of the complete collection. As an example of digitization of non-book materials, the first project in the Emory University virtual library listing is a conversion effort for African art images. Other libraries are focussing on music scores or photographs. Here, for example, is a photograph from the American Memory Project at the Library of Congress (Ulysses S. Grant).

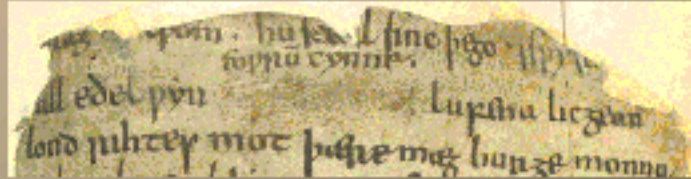


Sometimes a digital conversion can not only provide access at a distance, or access to fragile materials, but actually better access than would be provided by physical inspection. Here, for example, are three images of the *Beowulf* manuscript at the British Library, photographed in three different ways, as arranged by Kevin Kiernan of the University of Kentucky:

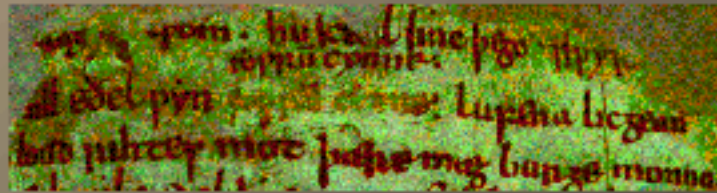


B & W Photo, Sean

Daylight Scan



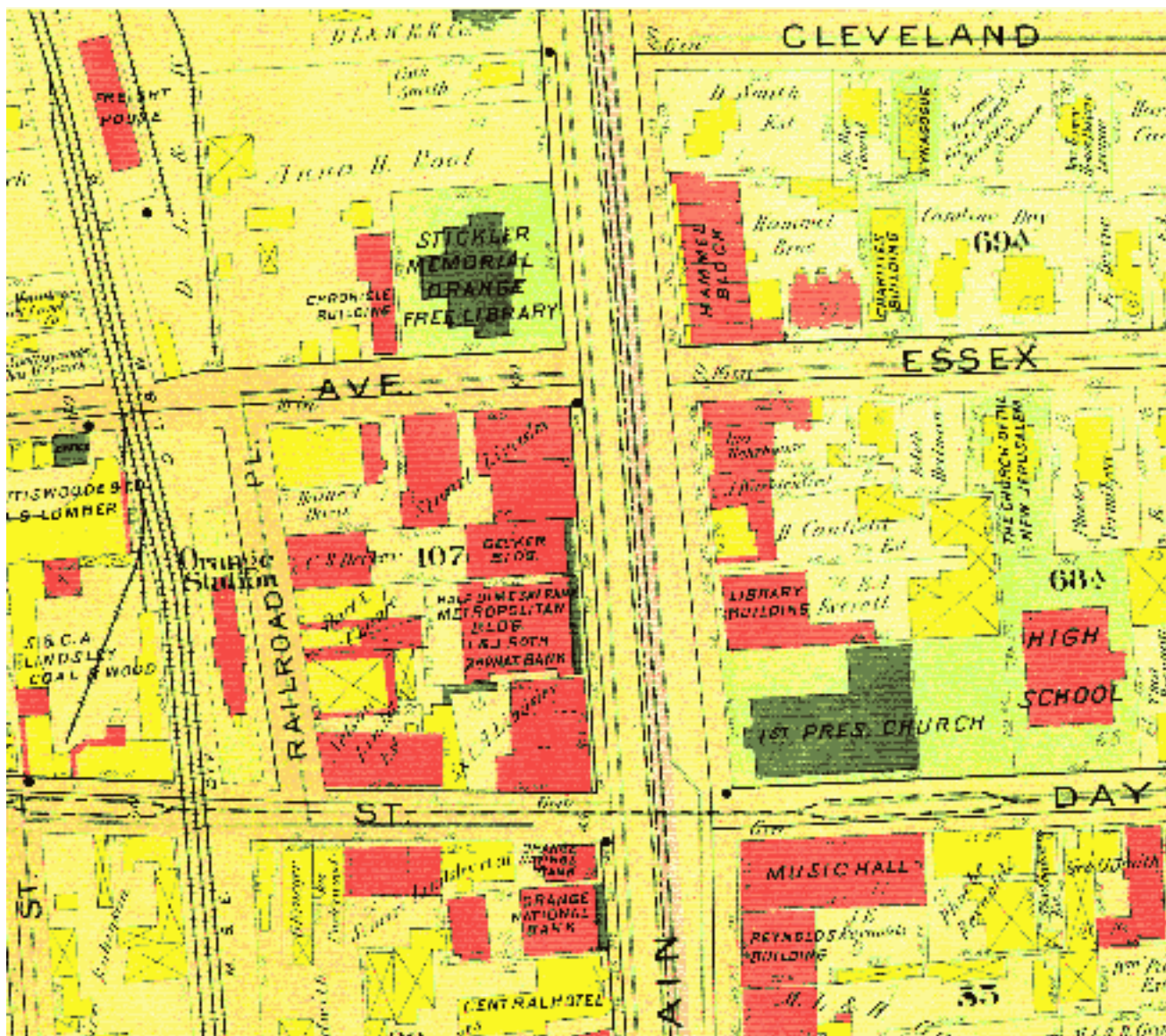
Ultraviolet Scan



This manuscript was damaged by an 18th-century fire and then by an inadequate restoration in the 19th century (predating its preservation in the British Library). As a result there are parts of the manuscript which are not readable in normal light today; in some cases they can be read in one of the digital versions. Since there is no other source for *Beowulf* (aside from one copy of the manuscript made before the fire) these images offers great advantages to scholars. Of course, this project is very expensive, and it is not possible to digitize all materials with the care taken for *Beowulf*; nevertheless it shows what can be done when the need justifies the cost.

Another example of a project done by a DLF library which shows economy as well as scholarly advantage is the digitization of Judaica poster material at Harvard by Charles Berlin. In this case some 130,000 posters were digitized via the Photo-CD process and converted to CD-ROM. This allows the originals to be moved to better storage and makes it much easier for scholars to look at these inconveniently bulky items. Conversion was relatively inexpensive, and yet the ability to study these posters has completely revolutionized the attitude towards them by some historians of Israel.

A project which will interest many specialists in local history, genealogy, and similar subjects is the conversion of the fire insurance map collection at the Library of Congress to digital form. Fire insurance maps show every building and its construction; the Library of Congress has some 700,000 of them. Here is a sample showing Orange, New Jersey.



Many other institutions are engaged in other digitizations. Here for example are pictures of a plant record converted by the National Institute of Biodiversity in Costa Rica and a modern vase made by Sidney Hutter and digitized by the National Museum of American Art.



One large effort combining Cornell and Michigan is a project called the *Making of America*, which is

digitizing material relating to American history 1850-1900. In this case the source publications are conventional magazines such as *Harpers*, *Scientific American*, and *Scribners*. The effect on humanities research will be one of accessibility to conventional publications rather than introducing unusual material.

The largest project is that of the Library of Congress, which is engaged in a wide variety of collection digitization. Photographs, architectural drawings, maps, sounds, and movies are all included in the *American Memory* project. Much of this material has been of restricted use because of preservation concerns, and can become much more widely available via their digital library. If one thinks of the Ken Burns television documentary *The Civil War* as a kind of highlights film of the Library of Congress photograph collection, it will in the future be possible for those who are attracted by it to view much of the remainder of the collection. A particular value of the LC work is the completeness of much of their conversion, in which whole collections are being digitized, as opposed to excerpts requiring users to consult the majority of material on paper anyway.

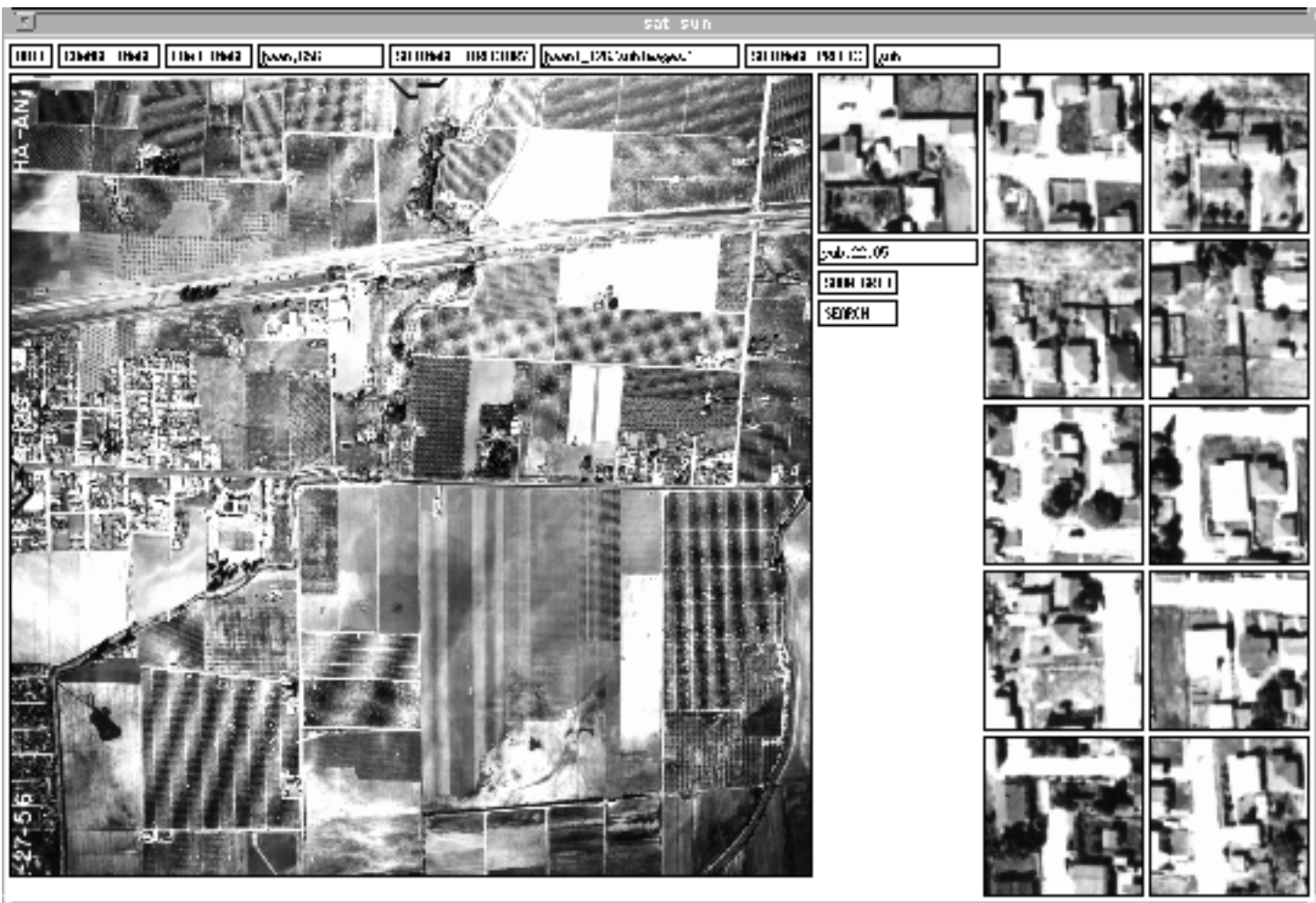
In addition to the library oriented work, humanists must notice the corresponding activities in the worlds of archives and museums. The Getty Foundation has supported the digitization and use in classes of some 8,000 artwork images under their Museum Educational Site Licensing program (MESL). This is being followed up by two different groups, one centered around art museums and one including historical and specialty museums (the Museum Licensing Cooperative). Archivists, although somewhat short of money, are at least studying the conversion of their catalogs and in some cases the primary materials.

The general impact of the Digital Library Federation in terms of converting material for use by humanities scholars is to extend the kind of material that is readily available for study. For a long time we have found it easier to get printed works than anything else, and true integration of paintings, sculpture, music or other aspects of culture into literary studies has been impeded by inconvenient access. If special collections become more widely digitized, we can expect less sharp boundaries between subjects such as music, theatre, art and literature.

## **The NSF/DARPA/NASA Digital Library Initiative**

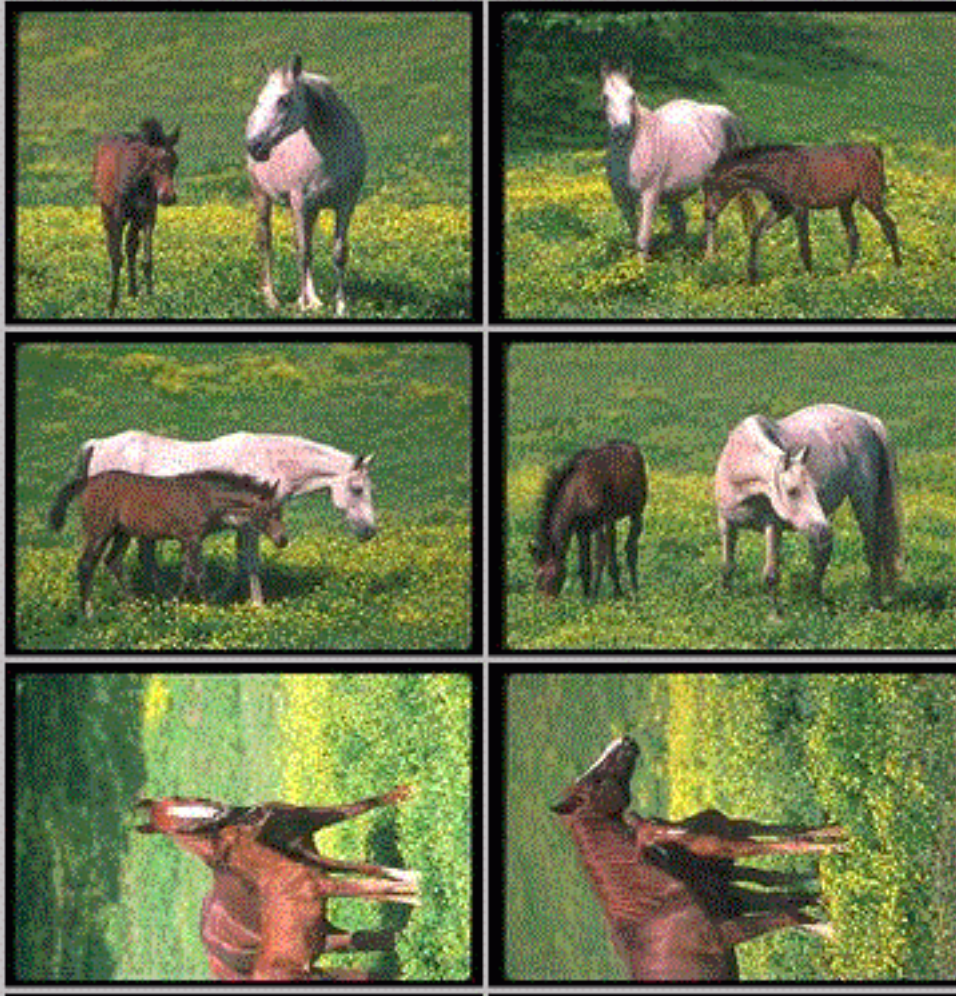
There are six DLI projects, at Berkeley, CMU, Illinois, Michigan, Santa Barbara and Stanford. In general, each is looking at new ways of retrieving material. Much of the work is really computer science oriented rather than library oriented, and so few complete collections are being converted. Furthermore, the subject matter covered in the collections used for research are not usually humanities related. Instead, the research in the DLI is valuable for its production of new ways to index and search.

The most library-like project is the one at University of California Santa Barbara, which is building a collection of geographically indexed data (including in particular maps and aerial photographs) relating to Ventura County, California. This project really is trying to accumulate all information about geography in their area, and so it does aspire to the kind of comprehensiveness found in some libraries. UCSB has search technology indexing by location, and also research on classifying imagery by content. A sample illustration is shown below, in which an aerial photograph has been automatically divided into different regions, with a dictionary of textures shown on the right.



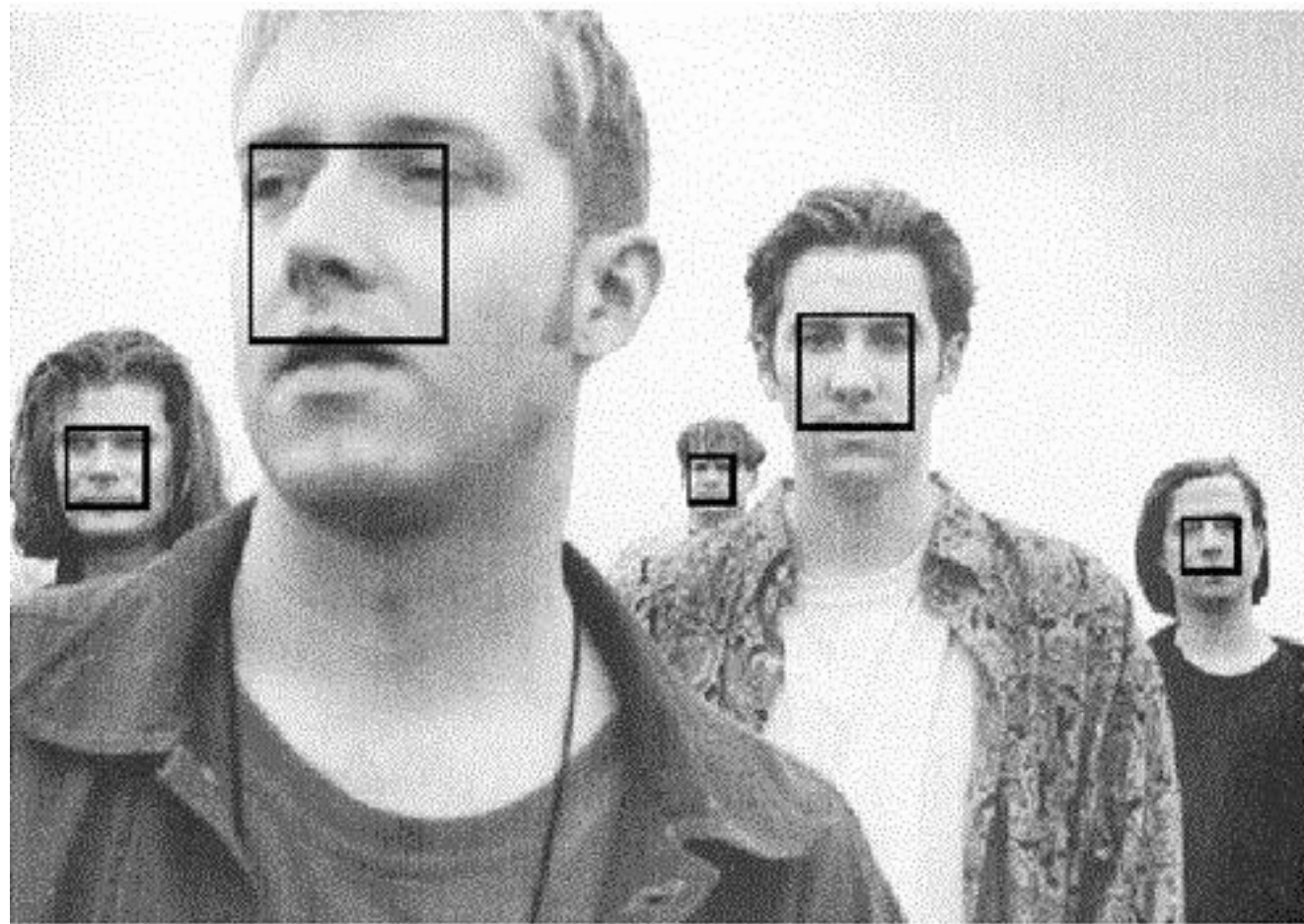
Another project with a geographical area focus and much work on images is the project at the University of California Berkeley, whose subject area is environmental reports about California. The Berkeley researchers are ranging widely over many technologies, however, including ways of displaying multiple views of the same document or image, and in particular content-based image search. They have implemented shape and color searching allowing them to look through images for sunsets, flags, lakes, and the like. Here, for example, is part of their result from writing a search routine for objects that look like horses (it does not always work this well, of course):

## Finding horses using body plans



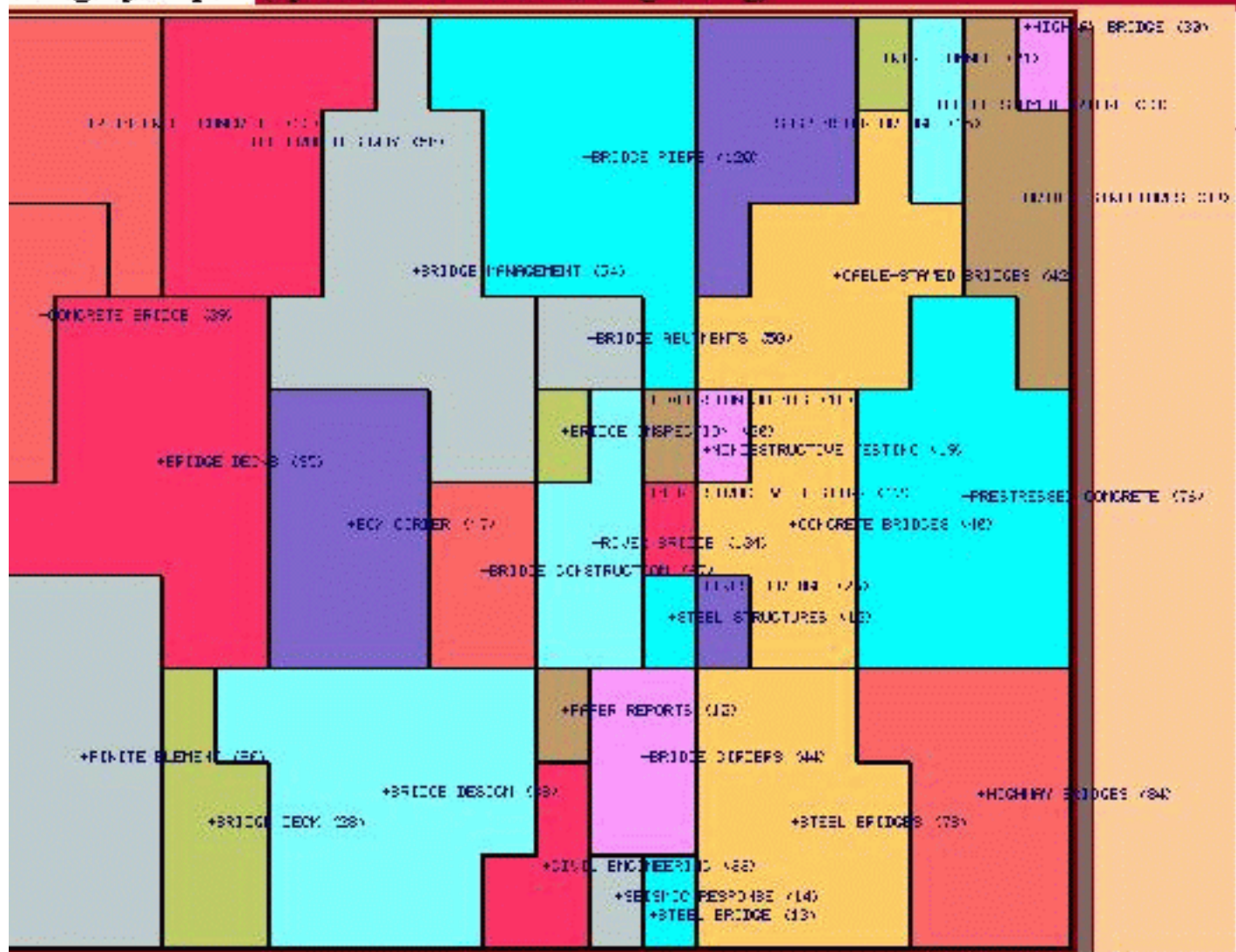
The Berkeley collection, although not a traditional library collection and somewhat fuzzier of definition, is actually large enough to be useful for many practical applications. After winter flooding last year in California, for example, people came to the Berkeley project to find aerial photographs of the areas affected before the inundation.

Carnegie-Mellon University has focussed its efforts on video indexing. They have a collection of some hundreds of hours of broadcast television news, which they search using closed-captioned text, speech recognition, and image searching. For example, they have built image analysis software which looks for text superimposed on an image and tries then to do OCR of that text. They also have an algorithm to identify faces and then to search for matches. The image below shows boxes where the system has identified a face in the television picture.



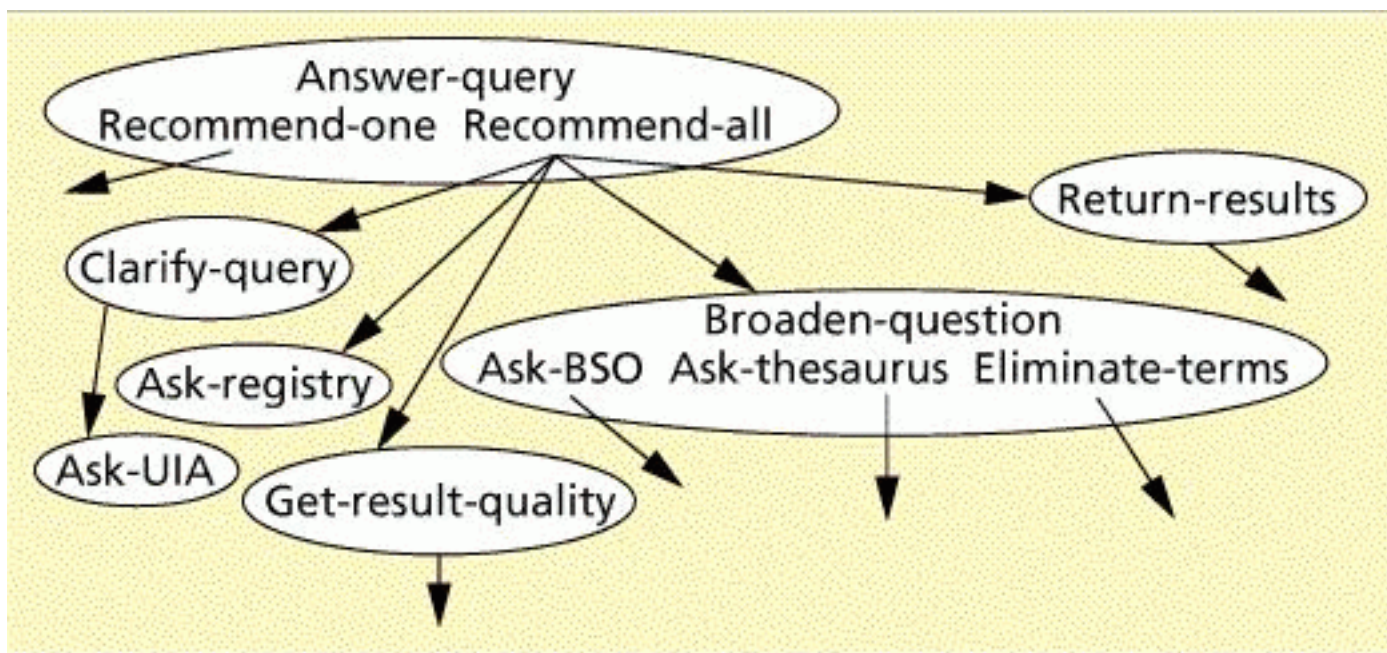
The University of Illinois is looking at scientific journals in digital form. They are working closely with publishers to build effective systems for access to primary journal articles in electronic form. The focus on scientific journals makes this project relatively more distant from the humanities. However, there are some very interesting automatic classification algorithms being studied here; the image below shows an automatic partitioning of a document collection into subject areas. The use of a two-dimensional representation instead of a linear hierarchy changes the view of information classification, with consequences as yet unstudied.

## Category Maps (dynamic automatic self-organizing)



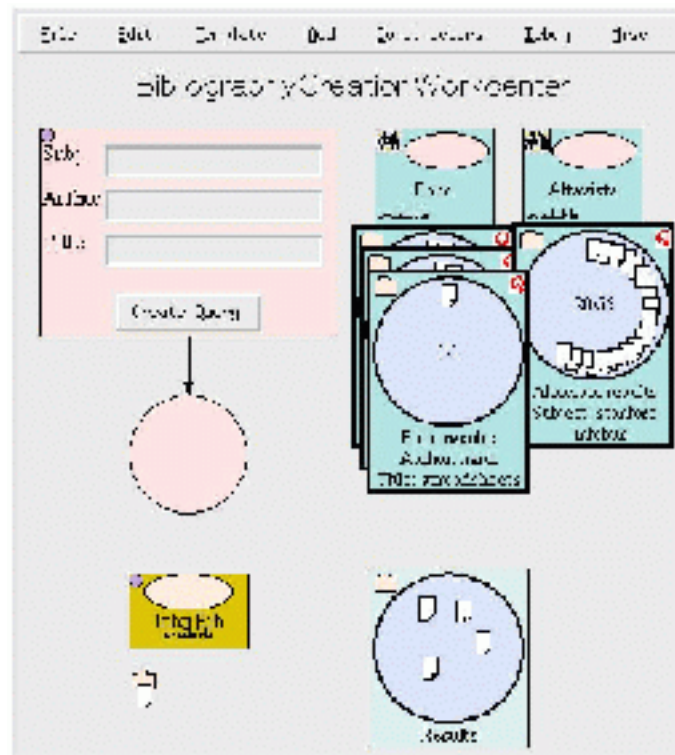
© 1996 Bruce Schatz. All rights reserved.

The University of Michigan project subject area is also scientific, with an emphasis on earth and space sciences. The most relevant part of this digital library effort to the humanities is the work attempting to define a set of agents which can represent user needs. The diagram below shows some of the agent roles and functions they imagine.

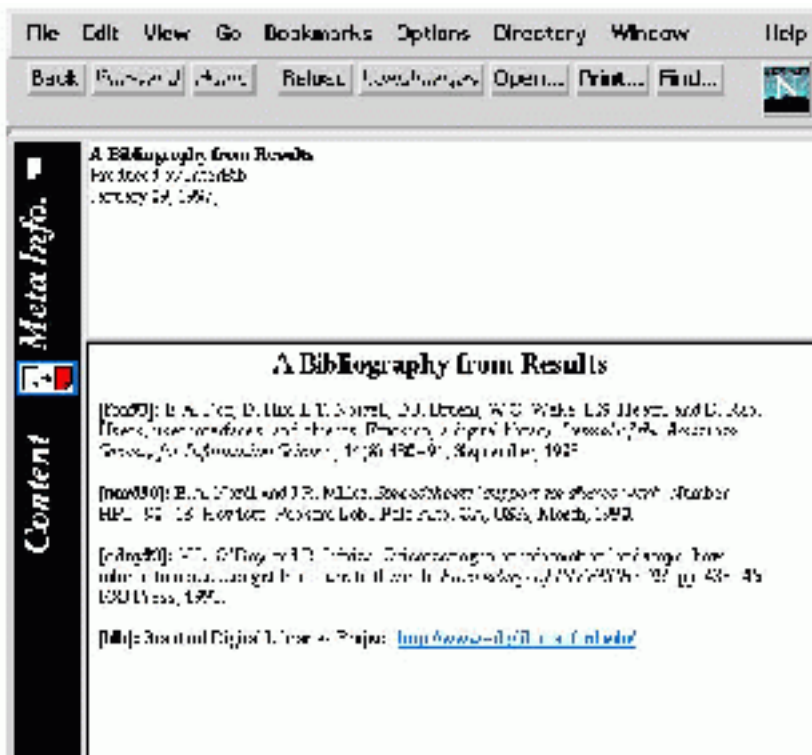


Success in this project might simplify the problems of accessing very diverse materials, or materials in places which have different rules for users. Instead of having to learn different procedures for each collection, agents could handle the economic or technical translations needed.

Finally, the Stanford project has no actual collection at all. It does have some very interesting techniques for database merger and rating. They have studied ways of sending a search to many search engines, which do not necessarily support the same search syntax or even the same searching operations. They are also looking at ways of ranking documents, for example, considering documents which have a great many links pointing to them as probably more valuable than those with fewer links.



(D)



(E)

In summary, most of these projects are of less immediate interest to humanists. They are developing new search techniques that may be very valuable in the future, in particular the ability to search images in

artistic or photographic collections. But the actual material studied in these projects is not focussed on the current uses of humanities scholars.

## **JSTOR**

The JSTOR project, now an independent non-profit organization, is not so much a technology project as an attempt to make a self-sustaining digitization organization. With the aid of startup funding from the Andrew W. Mellon Foundation, JSTOR was able to digitize an initial ten journals in both image and text format. They plan to continue with digitizing 100 journals, and to sell these on a subscription basis to university and public libraries. If about 500 libraries sign up, they expect to be able to continue going indefinitely, continuing to scan additional back issues.

The journals were chosen to be of very wide interest, and the scanning is high quality both in terms of the image appearance, and also in terms of making sure that each journal is complete. Libraries often find that their set of some journal is incomplete, with either occasional articles or volumes stolen or missing. The JSTOR set is checked and known to be complete, and it is hoped that at some point libraries will save shelf space by not keeping the original paper versions.

The illustration below, from the American Economic Review, is reduced-size and thus less sharp than the original JSTOR screen.

possible to use the rolling stock as the basis of its own purchase money loans.

In addition to these two main reasons there are, as has been said, others which sometimes influence the issue of equipment obligations. The tax laws of a state may subject bonds to a personal property tax, while the equipment certificates, being certificates of part ownership in physical property, escape. Again the bankers of the road may believe there is a better demand for the road's obligations with banks than with private investors, and equipment obligations are especially favored by banks. Still again the car and locomotive manufacturers are often willing to accept an equipment obligation in part or for nearly the whole payment of railway purchases on cheaper terms than the railroad can obtain by selling its own bonds and using the proceeds to reimburse the manufacturers. This is especially true at a time of slack business activity combined with high interest rates.

In substance, all equipment obligations are direct liens on rolling stock, but as now issued they may be divided into two great classes—those issued under the Philadelphia plan of a lease and those issued under a direct mortgage, sometimes, without reason, called the New York plan. As the Philadelphia plan is at once the most individual and the most complex, giving rise to the strongest kind of railroad obligation, it will be described first in considerable detail.

In addition to viewing the original page image, JSTOR provides a complete searchable text of each document, obtained via OCR and correction.

JSTOR provides desktop access to a wide variety of the backfile of important journals. A particularly interesting aspect of the project has been that many sales have been to small libraries. Originally, the sponsors thought that the most eager customers would be large libraries who owned all the journals on paper and were anxious to save on the costs of shelf space. Instead, many smaller colleges which never had been able to afford these journals found the JSTOR prices so attractive that they have subscribed.

The most important question asked by the JSTOR project is whether their pricing mechanism will suffice to keep the project going. At the moment they focus on sales to libraries, not to departments within universities, and most universities do not additionally charge users within the university. Thus the typical patron in the university sees a great improvement in the service, with desktop access and full free-text search, at no additional cost. Libraries bear the burden of the subscriptions, admittedly at a much lower price than to buy and shelve the paper equivalents, but nevertheless not a trivial amount.

The JSTOR interface is designed to help read the journal articles. Although the database contains the information to answer questions such as "how many times does the letter *q* appear in the *American Economic Review*, year by year?" the interface does not support that question. Thus JSTOR may avoid

some of the complaints that users of machine-readable full text tend towards low level statistical analysis of the texts, rather than a higher level understanding.

JSTOR supplements a existing collections of important primary texts. There are something over 6,000 literary works available online via the ``Online Books Page" <http://www.cs.cmu.edu/books.html> including full text of many important authors. Similarly the commercial LiON service of Chadwyck-Healey <http://lion.chadwyck.com/> is advertised to contain more than 250,000 texts. However, until now most critical and review material has been missing. The extension of digital libraries into journals and related material offers a new kind of support to scholarship.

JSTOR is a very important economic experiment. It is delivering a large, useful collection of important material, and attempting to do so in a way that will be self-supporting. Many of the other digital library projects are basically supported by research funding and must in the end make a transition to some kind of operational support or they will not be able to convert enough material to satisfy library users. The JSTOR project is the prototype of such a transformation.

## Conclusions

The likely impact on humanities research of digital library work is to extend the range of material that is regularly used. The various image processing and image digitization projects will promote the use of visual material. The extension of available material to more journals and more campuses via projects like JSTOR will mean that many more humanists have easy access to a wide variety of critical material. The major need in the humanities is to be sure that adequate amounts of material are converted to machine-readable form under terms scholars or their universities can afford. At the moment too many projects, from the viewpoint of a humanities scholar, are still investigating research in sample collections, rather than comprehensive conversion of the material needed by a practical scholar. Librarians and scholars need to be more active in identifying the material which should be provided in digital libraries and in seeing that it can be made available in an affordable way. Otherwise scholars may feel that digital libraries, although in principle revolutionizing the way work is done, offer great promise but not enough performance.

## References

DLI: <http://www.cise.nsf.gov/iris/DLHome.html>.

DLF: <http://lcweb.loc.gov/loc/ndlf/>.

JSTOR: <http://www.jstor.org>.

# How Much Information Is There In the World?

Michael Lesk

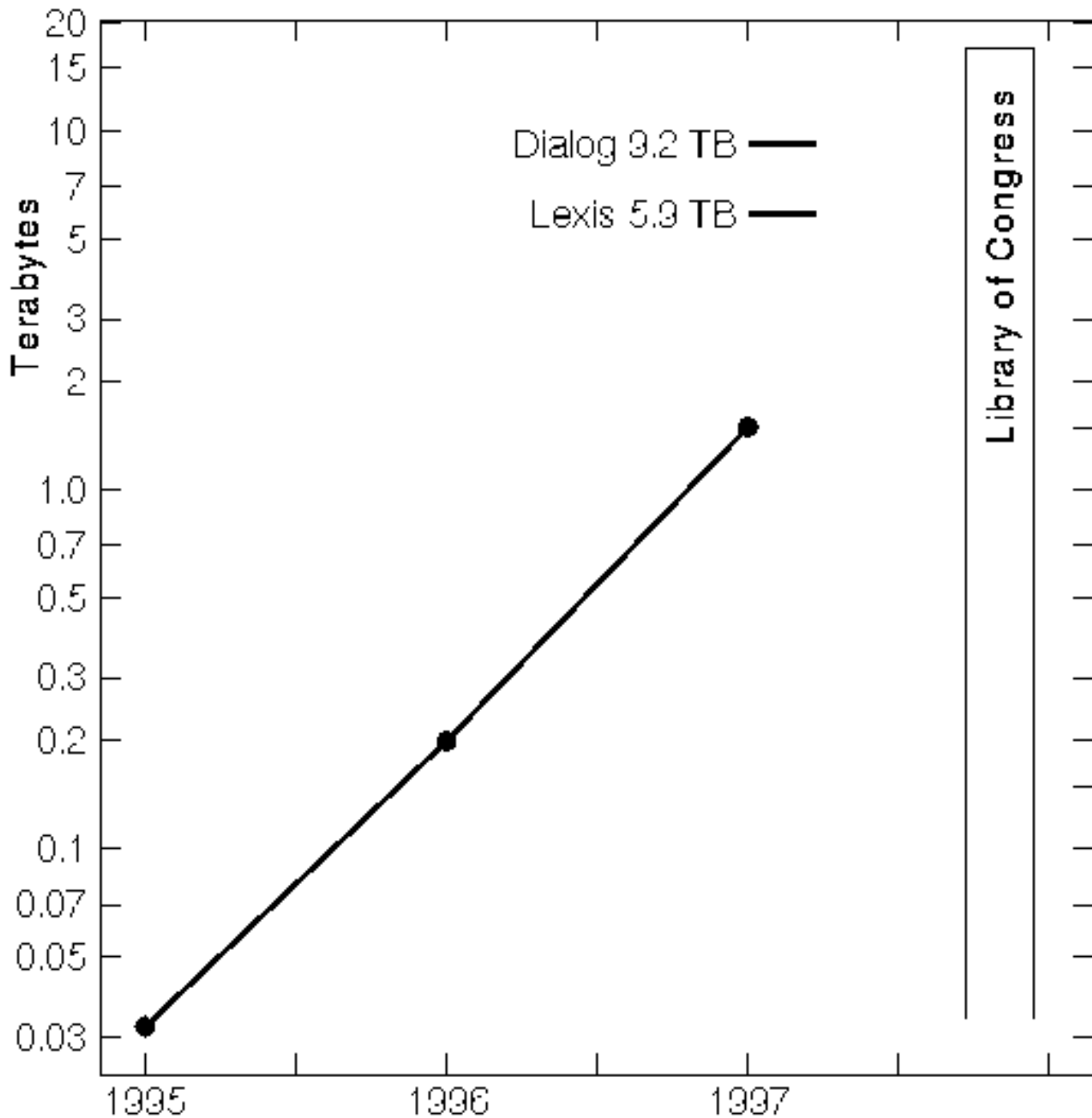
## *Abstract*

How much information is there in the world? This paper makes various estimates and compares the answers with the estimates of disk and tape sales, and size of all human memory. There may be a few thousand petabytes [\*] of information all told; and the production of tape and disk will reach that level by the year 2000. So in only a few years, (a) we will be able save *everything* \- no information will have to be thrown out, and (b) the typical piece of information will *never* be looked at by a human being.

---

Here is a chart of the current amount of online storage, comparing both commercial servers [Tenopir 1997]. and the Web [Markoff 1997]. [Mauldin 1995]. with the Library of Congress. These numbers involve Ascii text files only. This chart suggests that next year the Web will be as large as LC.

# Web size



The Web has been growing 10-fold each year. Can it continue to do so and for how long? Current estimates of the number of Internet users run in the tens of millions, perhaps 50 M, and this might grow to one billion; thus a factor of twenty is available by increasing the number of people on the Web, but not more. Can people put more and more of their life online? Perhaps, but I suspect not more than another factor of 20. This suggests that the amount of Ascii on the Web might increase to 800 terabytes. Is there that much text around? What about images, movies, and sounds?

**How much traditional information is there?**

The 20-terabyte size of the Library of Congress is widely quoted and as far as I know is derived by assuming that LC has 20 million books and each requires 1 MB. Of course, LC has much other stuff besides printed text, and this other stuff would take much more space.

1. Thirteen million photographs, even if compressed to a 1 MB JPG each, would be 13 terabytes.
2. The 4 million maps in the Geography Division might scan to 200 TB.
3. LC has over five hundred thousand movies; at 1 GB each they would be 500 terabytes (most are not full-length color features).
4. Bulkiest might be the 3.5 million sound recordings, which at one audio CD each, would be almost 2,000 TB.

This makes the total size of the Library perhaps about 3 petabytes (3,000 terabytes).

Of course the most important discrepancy in comparing the Web and the Library of Congress is that the Library of Congress predominantly contains published materials. The Web has more text than LC already, if you only ask for English-language material written in the last 18 months. I tried to guess what fraction of Web material represents something that has been published, however, by sampling fifty random English-language URLs. I found fourteen which looked to me as if they were probably in a large conventional library, or 28%. By contrast most of the contents of Lexis-Nexis and Dialog are versions of published material, albeit much more easily searched.

What other kinds of traditional information might be around? The United States manufactures 38 million tons a year of the kind of paper used for writing and printing. If a typical pound of paper is 220 A4 pages and each sheet held 5000 bytes, that would be about 8,000 terabytes of text each year. Of course many of the sheets are copies of other sheets, and many of them do not contain words. How much could reasonably be written fresh? Suppose that half the pages have text and that we assume 100 copies of the average sheet; that would be 40 terabytes of fresh information. If 40 million U. S. 'knowledge workers' each wrote 1 megabyte a year, that would also be 40 terabytes a year. Since the US gross domestic product of \$7T is about one-quarter of the world GDP (\$30.8B) I will in general multiply the US by 4 to extrapolate to the earth, and suggest that the entire world's writing amounts to 160 terabytes each year. Of this the published books are about 863,000 (in 1991), plus 9,315 newspapers, [UNESCO 1995]. making perhaps a terabyte of professionally written or refereed material, not even 1% of the total.

Other kinds of information, compared with Ascii text, are bulkier.

1. *Cinema*. There were 4,615 films made world-wide in 1989; at 5MB/sec and 7200 seconds average, that would be 166 terabytes.
2. *Images*. There are about 52 billion (thousand million) photographs taken each year in the world. [Mills 1996]. If each of those is a 10 KB JPG, that is 520,000 terabytes, or 520 petabytes, and these are actually all different. Again, less than 1% represent professionally taken or reviewed pictures, probably less than 0.1%. By comparison even the NASA earth observing project, expected to accumulate 11,000 terabytes, [Fargion 1996]. doesn't affect the numbers.
3. *Broadcasting*. In the US, we have 1593 television stations. If each sends out 5 MB/sec for 30 million seconds per year, that is over 200 petabytes. However, one might expect that only about 1/10 of the programming is actually different for different stations; that is 20 petabytes of distinct programming, and extrapolated to the world would be 80 petabytes. Radio, by contrast, is insignificant; the US has 6,956 radio stations and if each sends out 30 million seconds per year at 8 KB/sec we would have only 1.7 TB in the United States.

4. *Sound.* Sales of recorded music in the US in 1992 were 407 million CDs and 336 million cassettes (and 20 million vinyl disks, still). Assuming 550 MB for each CD and cassette that would be 400 petabytes, much duplicated of course. If the number of different recordings for sale is about 30,000 this would be 15 terabytes in the US and 60 terabytes world-wide.
5. *Telephony* The largest storage requirement would come from converting all telephone conversations to digital form. In the US in 1994 there were 500 billion call-minutes of 'interlata toll' and there is about 20 times as much local calling, so at 56 kbits/sec this would be 4,000 petabytes of digitized voice. The only thing I am not considering is consumer videotape, on the grounds that much of it is used to record off-the-air TV and duplicates the TV stations.

The conclusion is that in terms of text there are terabytes of information and perhaps one terabyte of professional information. Including sounds and images there are thousands of petabytes of information. The letter from Sincerbox which started all of this suggested that there would be 12,000 petabytes of information in the world, perhaps not an unreasonable guess. Only a small part of this, dominated by the TV stations, is commercially produced or validated in some way; perhaps that amounts to 100 petabytes.

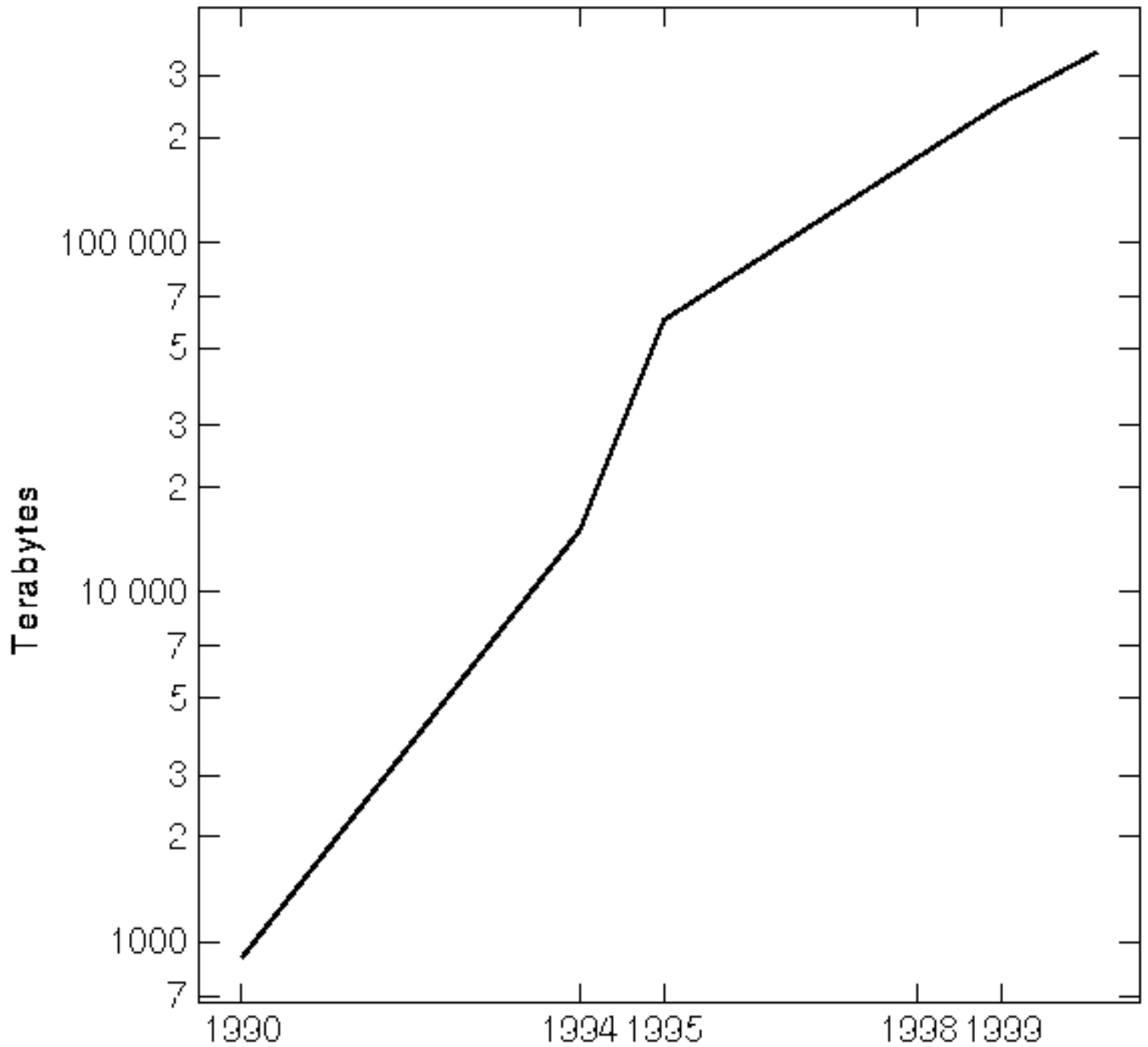
### **How much computer storage space is there?**

The single largest data storage system I have seen described is a year-old description of the Accelerating Strategic Computing Infrastructure project at Livermore, Los Alamos and Sandia Laboratories, which has 75 terabytes of disk, and a plan for hundreds of petabytes of tape archive. [Louis 1996 ]. The Los Alamos HD-ROM project using scanning electron microscopes to etch bits into stainless steel in a vacuum, which has been transferred to the startup company Norsam Technologies, has achieved 200 GB/square inch. They hope to put 12 terabytes on a single CD-size disk.

One way of guessing the total size of the world's computer storage is simply to view the single largest establishment as one point on a log-normal curve. To oversimplify, the largest city in the world has about 1/300 the population of the world. and the largest company in the world has about 1/300 the world's GDP. So this suggests that if the largest disk farm in the world in 1996 was 75 terabytes, the total disk space in the world was 22,500 terabytes.

Of course, there are statistics on the disk drive industry. The chart below makes a guess at how many terabytes of disk space are sold per year, using data from Computerworld, [Radding 1990]. IBM, [Bell 1994]. and Optitek. [Optitek]. The different uncoordinated sources for this table make it fairly irregular; I've been unable to find good numbers from a single source. But it is clear the answer today is tens of thousands of terabytes of disk sold each year.

# Disk space sold



Optitek predicts 1998 sales and capacities of different storage media:

Device	Price	Total market	Total size
Magnetic disk	\$100/GB	\$25B	250 petabytes
RAID disk	\$200/GB	\$13B	65 petabytes
Optical disk	\$20/GB	\$0.5B	25 petabytes
Optical jukeboxes	\$20/GB	\$5B	250 petabytes
Magnetic tape	\$1/GB	\$10B	10,000 petabytes
Tape stackers	\$1/GB	\$2B	2,000 petabytes

Both Alan Bell of IBM and Jim Gray of Microsoft estimate that 200 petabytes of tape storage were sold in 1995.

Note that these numbers added up are all comparable to the size of the numbers for the total amount of information in the world. So the implication is that in the year 2000 we will be able to save in digital form everything we want to - including digitizing all the phone calls in the world, all the sound recordings, and all the movies. We'll probably even be able to do all the home movies in digital form. We can save on disk everything that has any contact with professional production or approval. Soon after the year 2000 the production of disks and tapes will outrun *human* production of information to put on them. Most computer storage units will have to contain information generated by computer; there won't be enough of anything else.

Of course, this has already true despite the lower size of computer memory today. The typical computer disk byte is probably part of some Microsoft object module. After that, it's probably some kind of database. But we still see complaints that relatively little of the data in many large archives (the NASA files or the Palomar sky survey) has ever been looked at by anyone. That will be normal in the future: computer memory will be mostly for other computers. Today this memory is highly duplicative, with tens of millions of copies of popular programs. Tomorrow, with everyone on-line with high speed connections, and extended use of site license agreements, it may be common for PCs to fetch on demand object modules of software needed once in a while, as we already do at Bellcore. The disks on our machines will then be available for our own personal information. A fast author might write a megabyte a year; not even Trollope wrote 100 MB in his life; but we'll all have at least a gigabyte of personal storage by 2000, when we have about as many petabytes of disk sold as there are millions of computers in the world (300 each, roughly).

### **How much human memory is there?**

And to look at a third measure, how much does human memory hold? Tom Landauer tried to estimate this some years ago and concluded that the brain held about 200 megabytes of information. [Landauer 1986]. He got this number partly by looking at the rate at which people could take in information, both by reading and by looking at pictures. He also studied estimates of the rate at which people forget things, and the amount of information adults need in order to do the tasks they normally do. His numbers (expressed in gigabits, not gigabytes), were 1.8, 3.4, 2.0, 1.4 and .5 gigabits. Averaging these and dividing by 8 yields 227 MB. Since there are between  $10^{12}$  and  $10^{14}$  neurons, this suggests that the brain contains 1,000 to 100,000 neurons for each bit of memory. Of course, much of the brain is used for perception, motor control, and the like; but even if only 1% of the brain is devoted to memory Landauer pointed out that it looks like your head accepts considerable storage inefficiency in order to be able to make effective use of the information.

With something like 6 billion people on earth, that makes the total memory of all the people now alive about 1,200 petabytes. To the accuracy with which these calculations are being done, the results are comparable. We can store digitally everything that everyone remembers. For any single person, this isn't even hard. Landauer estimated that people only take in and remember about a byte a second; a typical lifetime is 25,000 days or 2 billion seconds (counting time asleep). The result is 2 gigabytes, or something that fits on a laptop drive.

Would it be hard to remember every word you heard in your lifetime, including the ones you forgot? The average American spends 3,304 hours per year with one or another kind of media. [Census 1995]. 1,578

hours are with TV; adding in 12 hours a year of movies, at 120 words per minute that's 11 million words, perhaps 50 megabytes of Ascii. And 354 hours a year of reading newspapers, magazines and books at 300 words per minute reading speed would be another 32 megabytes of text. In seventy years of life you would be exposed to around six gigabytes of Ascii; today you can buy 23 gigabyte disk drives.

Could we simply make a wearable device that would record everything? Yes, if either (a) we had decent speech recognition and OCR, or (b) books move to electronic form and TV sets provide access to the closed-captioned Ascii form of the scripts. Perhaps both of these choices are likely in the near future. School children no longer need to do arithmetic without calculators; perhaps they will soon no longer need to memorize anything either. If you think this is horrible remember that Plato (in the *Phaedrus*) suggested that writing would 'create forgetfulness in the minds of those who learn to use it' and would create 'the show of wisdom without the reality.' If writing something down isn't cheating, why is recording it? It is now common for speakers to use transparencies, for a conference to hand out printed proceedings, and for people to sit at talks with cassette recorders. Would it be that terrible if each attendee had a laptop doing speech recognition, and the laptop kept the transcript and provided a small vibration to wake up the attendee when a promising topic was mentioned?

Two years ago I heard Ted Nelson at a conference suggest that we should keep the entire record of everyone's life \- all the home snapshots, videos and the like. Some six-year-old, he said, is going to grow up to be President; and then the historians will wish we knew absolutely everything about his or her life. The only way to do this is to save everything about everyone's life. I laughed, but it's indeed possible. Whether it is worthwhile is another question: are we better off having all possible information and giving it the most sketchy consideration, or having less information but trying to analyze it better? Computers do not use log tables, and chess computers have dictionaries of opening and endgame positions but not whole games. We need to understand our ability to model more complex situations to know how to make best use of stored information.

## Conclusion

There will be enough disk space and tape storage in the world to store everything people write, say, perform or photograph. For writing this is true already; for the others it is only a year or two away. Only a tiny fraction of this information has been professionally approved, and only a tiny fraction of it will be remembered by anyone. As noted before the storage media will outrun our ability to create things to put on them; and so after the year 2000 the average disk drive or communications link will contain machine-to-machine communication, not human-to-human. When we reach a world in which the average piece of information is *never* looked at by a human, we will need to know how to evaluate everything automatically to decide what should get the precious resource of human attention.

Today the digital library community spends some effort on scanning, compression, and OCR; tomorrow it will have to focus almost exclusively on selection, searching, and quality assessment. Input will not matter as much as relevant choice. Missing information won't be on the tip of your tongue; it will be somewhere in your files. Or, perhaps, it will be in somebody else's files. With all of everyone's work online, we will have the opportunity first glimpsed by H. G. Wells (and a bit later and more concretely by Vannevar Bush) to let everyone use everyone else's intellectual effort. We could build a real 'World Encyclopedia' with a true 'planetary memory for all mankind' as Wells wrote in 1938. [Wells 1938]. He talked of 'knitting all the intellectual workers of the world through a common interest;' we could do it. The challenge for librarians and computer scientists is to let us find the information we want in other people's work; and the challenge for the lawyers and economists is to arrange the payment structures so

that we are encouraged to use the work of others rather than re-create it.

## Acknowledgment.

This paper was suggested by a query from Glenn Sincerbox of the University of Arizona.

---

\* Here are the names of the units of very large storage sizes:

gigabyte 1,000 megabytes

terabyte 1,000 gigabytes

petabyte 1,000 terabytes

exabyte 1,000 petabytes

[Bell 1994]. Alan Bell; *IBM Academy Digital Library Workshop* (Sept 12-13, 1994).

[Census 1995]. United States Census Bureau *Statistical Abstract of the United States* Government Printing Office (1995).

[Fargion 1996]. G. S. Fargion, R. Harberts, and J. G. Masek An Emerging Technology Becomes an Opportunity for EOS From the online file; see the URL:  
<http://ecsinfo.hitc.com/cdwg/datamining/overview.html>.

[Landauer 1986]. T. K. Landauer; "How much do people remember? Some estimates of the quantity of learned information in long-term memory," *Cognitive Science*, **10** (4) pp. 477-493 (Oct-Dec 1986).

[Louis 1996 ]. Steve Louis *Cooperative High-Performance Storage in the Accelerated Strategic Computing Initiative* 5th NASA Goddard Conference on Mass Storage Systems and Technologies (Sept. 17-19, 1996 ). As reported by Ron Van Meter, <http://www.isi.edu/~rdv/conferences/goddard96.html> .

[Markoff 1997]. John Markoff; "When Big Brother is a Librarian," *The New York Times* pp. 3, sec. 4 (March 9, 1997).

[Mauldin 1995]. Matt Mauldin, "Measuring the Web with Lycos," *Third International World-Wide Web Conference*, April 1995.

[Mills 1996]. Mike Mills; "Photo Opportunity," *Washington Post* pp. H01 (January 28, 1996).

[Optitek]. The Need for Holographic Storage [http://www.optitek.com/hdss\\_competition.htm](http://www.optitek.com/hdss_competition.htm).

[Radding 1990]. Alan Radding; "Putting data in its proper place," *Computerworld* pp. 61 (August 13, 1990).

[Tenopir 1997]. Carol Tenopir, and Jeff Barry; "The Data Dealers," *Library Journal* pp. 28-36 (May 15, 1997).

[UNESCO 1995]. *UNESCO Statistical Yearbook* Bernan Press (1995).

[Wells 1938]. H. G. Wells *World Brain* Methuen (1938).

## Organization of the California Digital Library

CDL QuickLinks

### **Richard Lucier**

#### **University Librarian and Executive Director**

Richard Lucier has an extensive background in information studies, computers and digital technology. He was University Librarian at the San Francisco campus from 1991 through 1997. During that same period he served as assistant vice chancellor for academic information management, director of UCSF's Center for Knowledge Management and assistant clinical professor in the School of Pharmacy.

Prior to joining UC, Lucier served as founding director of the Laboratory for Applied Research in Academic Information at Johns Hopkins University in Baltimore, where he spearheaded the development of the Genome Data Base and the Online Mendelian Inheritance in Man, a book in support of the international Human Genome Initiative. He was also an associate director for research and computing at Johns Hopkins and director of academic information resources management at the University of Cincinnati.

He holds a master's degree of library science from Rutgers University and a bachelor's degree in music and philosophy from the Catholic University of America.

### **Shared Content**

#### **Beverlee French, Associate Director**

Coordinates the selection, integration, and management of shared digital knowledge resources for the CDL. Shared digital content includes a wide range of forms from metadata to text, still and moving images, numeric data, and audio, among others. Content may be published or unpublished, purchased or licensed, selected from publicly available materials (e.g., governmental), or created within the University of California such as by building digital formats of published and unpublished materials owned by UC (e.g., rare books, unique publications, manuscripts, archival materials, maps, slides, audio and video collections, etc).

Beverlee has a wide range of expertise in collection development, library management and technology, and public services. Most recently she has served as the Assistant/Associate University Librarian for Sciences and Systems at UC Davis (1992-present), where she administered the science libraries (Biological and Agricultural Sciences Department, Physical Sciences, Health Sciences, and Medical Center Libraries), Government Information and Maps, and library computing services.

She has also been Acting Assistant University Librarian for Collections (1988-89) and Assistant University Librarian for the Sciences (1987-92) at UC Davis, as well as Chair of two systemwide committees, Heads of Public Services and the Computer Files Committee.

Prior to her appointment at UC Davis, she served as Head of the Science and Engineering

Library, and as a reference librarian and cataloger at UC San Diego. She holds an A.B. in social sciences and an M.L.S. from UC Berkeley.

The Online Archive of California (OAC) is a unique resource which is an integral part of the CDL's shared content. The OAC is a California-wide digital archive that integrates finding aids to and digital facsimiles of selected content from archival collections dispersed throughout the state in a single searchable database. Critical issues range from ensuring that policies, procedures, and mechanisms are in place for contributing digital finding aids to the OAC database to coordinating the deployment of select digital content and the development of documentation, tools, and training needed to advance the growth and management of the OAC. Robin Chandler is the Manager of the OAC.

### **Business Development**

#### **Cate Hutton, Assistant Director**

Establishes, extends, and oversees the CDL's relationships with parties outside of the University. This includes ensuring the complete and correct execution of licenses and other agreements with content providers, as well as analyzing options for mutually beneficial service arrangements with a variety of educational and private sector partners.

Cate's background, in addition to her external and governmental relations work at UCOP, includes an MLIS from Berkeley and positions as a research associate with Andersen Consulting and as an American Library Association Book Fellow on assignment in Tibet.

### **Education and Applied Research**

#### **John Ober, Assistant Director**

Ccoordinates education and evaluation programs that increase understanding of and innovation in digital resources. This includes coordinating partnerships with the digital library research community and (working with the Digital Library Technologies unit) identifying opportunities for the transfer and use of new technologies. Working with Digital Library Services staff and public service staff at each campus, educational programs for the CDL foster independent and successful use of the CDL and encourage an environment of continuous learning for CDL staff and partners. Education also includes outreach and communication of CDL goals and activities to a variety of audiences. Evaluation activities, such as needs assessments, surveys and focus groups, develop an understanding of the needs of CDL patrons and their use of CDL resources and help focus enhancements to collections and services.

John has broad experience and knowledge in librarianship, teaching, and computer technology. He has served as a faculty member and the Development Librarian for Electronic Resources at California State University (CSU), Monterey Bay, as the Acting Director, Library Systems, at UC Berkeley, the Network Resources Librarian at UC Berkeley, an ALA Library Bookfellow in Benin, west Africa, and Assistant Professor at the UC Berkeley School of Library and Information Systems.

He holds a B.A. in English/Philosophy, an M.S. in Sociology from the University of Houston, an M.L.I.S. from UC Berkeley, and a Ph.D. in Library and Information Systems Management, also from UC Berkeley.

### **Scholarly Communication Initiatives**

#### **Catherine Candee, Assistant Director**

Coordinates the application of digital technologies to influence and support innovations in scholarly communication throughout its life cycle, including production and dissemination. The eScholarship program is the focal point. Its goals are to influence and support innovations in scholarly communication. It includes the establishment of open archives for scholarly communities based upon an e-print server infrastructure, as well as supporting services and new scholarly products and publications drawn from the archives. More information is available at [ <http://www.cdlib.org/eschol> ].

Catherine has a strong commitment to innovation in scholarly communication and a history of working with faculty and scholarly societies toward those innovations. Prior to working for the CDL, she was the head of Stanford's Physics Library and Program Officer for their "Access to Information" Committee. She brings UC and scholarly publications experience to the position, having also served as UC Berkeley's head of the Astronomy/Math/Statistics Library at UC Berkeley and at one time having developed a publications program for the Institute for Social and Economic Studies in Berkeley.

Catherine holds a B.A. from California State University and an M.L.I.S. from UC Berkeley.

### **Digital Library Services**

#### **Laine Farley, Assistant Director**

Designs and coordinates the implementation of the CDL web site in support of the CDL's goals, and as the delivery mechanism for the CDL's collections, tools, and services. In conjunction with CDL Technologies, this unit develops the CDL's online systems, including the Melvyl Catalog and other CDL-hosted databases. DLS works closely with CDL Technologies to create an application development framework and co-investment model to leverage campus initiatives for developing tools and services and integrating them into the CDL. The unit ensures that the process for managing releases to the CDL web site are well integrated with collection planning, testing and feedback activities, and CDL Technologies resources and planning.

Laine was formerly the User Services Coordinator and most recently the Coordinator of Bibliographic Policy and Services in the UC Division of Library Automation. Previously, she was a reference librarian and coordinator of bibliographic instruction at UC Riverside, and head of the Humanities department at the Steen Library, Stephen F. Austin State University.

She holds a B.A. in liberal arts (Plan II) and an M.L.S. from the University of Texas at Austin.

### **Digital Library Technologies**

#### **David Walker, Director**

Coordinates the technical design, management, and operation of the California Digital Library online systems, including the Melvyl catalog, in support of the plans and policies formulated by the University Librarian of the CDL. Along with other technology centers at the University of California, CDLT evaluates and tests emerging technologies and standards for potential use in meeting the needs of the users of the California Digital Library, works with the University community to define architectural guidelines and standards and implements new systems within this framework as required.

David brings to his position broad technical and management experience, successes in planning and implementing large systems, deep understanding of the technical requirements for a digital library, and a long commitment to the University. He comes to the University of California Office of the President (UCOP) from Cox Communications, Orange County, where he has been Engineering Manager for Commercial Broadband Services since 1997. Prior to entering the private sector, David spent 20 years in increasingly responsible technical positions at UC. He rose from systems programmer to Assistant Director for Strategic Activities and Advanced Technology at Irvine, where he planned and implemented a number of major strategic initiatives to provide computing and communications services to the campus. He then moved to the San Diego campus to become Director of Telecommunications.

David participated in the planning phase of the UC Digital Library, which eventually led to the CDL, and is the author of numerous network plans, technical architecture plans, and technical policies. He frequently speaks and writes on these and related issues. In addition to the CDL planning process, David participated in UC's Task Force on Telecommunications Needs for Distance Learning, and was a member of two review teams for campus network projects. He is active in the California Internet Federation, CERFnet, and other technical groups.

David is a graduate of UC Irvine with a B.A. in Mathematics and a B.S. in Information and Computer Science. He also completed five years of graduate work in Information and Computer Science at Irvine.

Last updated: Friday, 28-Apr-00 12:37:04

[COLLECTIONS & SERVICES](#)[GUIDES](#)[NEWS](#)[ABOUT THE CDL](#)[home](#) • [search](#) • [feedback](#)

© 1999 The Regents of the University of California



[ARL](#) / [EDUCAUSE](#)

Coalition for Networked Information

To Advance Scholarship  
and Intellectual Productivity

## Clifford A. Lynch

### Clifford A.

**Lynch** has been the Executive Director of the Coalition for Networked Information (CNI) since July 1997.

CNI, jointly sponsored by the Association of Research Libraries and Educause, includes about 200 member organizations concerned with the use of information technology and networked information to enhance scholarship and



intellectual productivity. Prior to joining CNI, Lynch spent 18 years at the University of California Office of the President, the last 10 as Director of Library Automation, where he managed the MELVYL information system and the intercampus internet for the University. Lynch, who holds a Ph.D. in Computer Science from the University of California, Berkeley, is an adjunct professor at Berkeley's School of Information Management and Systems. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science. Lynch currently serves on the Internet 2 Applications Council and the National Research Council Committee on Intellectual Property in the Emerging Information Infrastructure.

- [Some Publications and Recent Talks.](#)

- [Longer Bio and Recent Activities.](#)

## Clifford A. Lynch

Executive Director, Coalition for Networked Information

21 Dupont Circle

Washington, DC, 20036

202.296.5098

202.872.0884 (fax)



[clifford@cni.org](mailto:clifford@cni.org)



[What is CNI?](#)

[Projects](#)

[Meetings](#)

[Conferences](#)

[What's New?](#)

[Net Services](#)

[Search](#)

[Archives](#)

# CNI

21 Dupont Circle Suite #800

Washington, DC 20036-1109

202.296.5098

<http://www.cni.org/>



Developed & Maintained  
by:

[webmgr@cni.org](mailto:webmgr@cni.org)

[Any comments, or feedback?](#)

© 2000 Coalition for Networked Information  
ALL RIGHTS RESERVED.

Last Update: Monday, 12 April, 1999 - 01:20 PM - EDT

# Gary Marchionini

Email: [march@ils.unc.edu](mailto:march@ils.unc.edu)

Position: Cary C. Boshamer Professor, [School of Information and Library Science](#).

Research interests: Information seeking, human-computer interaction, digital libraries, information design, information policy

Gary Marchionini is Cary C. Boshamer Professor in the School of Information and Library Science at the [University of North Carolina](#) where he teaches courses in communications, interaction design, and digital libraries. His Ph.D. is from Wayne State University in mathematics education with an emphasis on educational computing. He was previously professor in the College of Library and Information Services at the University of Maryland, a member of the [Human-Computer Interaction Laboratory](#). He heads the [Interaction Design Laboratory](#) at SILS.

He was PI for a U.S. Department of Education Challenge Grant project, the [Baltimore Learning Community](#). He has served for ten years as the Director of Evaluation for the [Perseus Project](#) (a digital library devoted to classical culture) and served for two years as the General Editor of Hypertext Publications for the Association of Computing Machinery. He was the Conference Chair for [ACM Digital Library '96 Conference](#).

Professor Marchionini has had grants or contracts from the National Science Foundation, Council on Library Resources, the National Library of Medicine, the Library of Congress, Bureau of Labor Statistics, Kellogg Foundation, and NASA, among others. He has published over seventy articles, chapters and reports in a variety of books and journals. He is author of a book titled *Information Seeking in Electronic Environments* published by Cambridge University Press.

He serves on the editorial boards of the *Journal of the American Society for Information Science*, *Information Processing and Management*, *Library Quarterly*, *Library and Information Science Research*, *Journal of Network and Computer Applications*, *Journal of Digital Information*, *Educational Technology*, and the *Journal of Educational Multimedia and Hypermedia*.

Current intests and projects are related to: interfaces that support information seeking and information retrieval; understanding statistical tables; alternative representations for electronic documents; multimedia browsing strategies; digital libraries; information design; and evaluation of interactive media, especially for learning and teaching.

---

## Selected Talks

- [The Sharium: A Distributed Learning Space](#) American Association for the Advancement of Science Annual Conference, Feb. 18, 2000
- [Supporting Citizen Access to Statistical Data: WWW Interfaces for Tables](#) National Health Statistics Conference, Aug., 2, 1999

- [Expanding Library Services in the Digital Age: The Search for \[Almost\] Equilibrium](#) Digital Library Federation Forum, July 17, 1999
  - [Overview of Digital Libraries](#) SILS Info-to-Go Workshop March 1999
  - [The Baltimore Learning Community: An Evolving Sharium](#) ASIS 98 Panel slides
  - [Information Visualization Interfaces: The Alchemist's Workbench](#) ASIS 98 Panel slides
  - [Designing for End Users](#) Online World 98 Conference slides
  - [Focusing on the User](#) Workshop on Knowledge Management Opportunities at the U.S. Department of Labor slides
  - [Dynamic Key Frame Presentation Techniques for Augmenting Video Browsing](#) AVI '98 slides
  - [Can we build a sharium?](#) Digital Library '98 panel slides
  - [Digital Libraries and Digital Government: Challenges and Opportunities \(NAL March 19 1998\)](#)
  - [Teaching With Technology Symposium Keynote: Don't Let Technology Get in the Way of Your Teaching](#)
  - [Association of American Publishers Panel: What do Users Want?](#)
  - [NLM Evaluation Talk](#)
  - [Allerton Digital Library Workshop](#)
  - [Digital Library Interfaces](#) (Spring 95 digital library research seminar)
- 

## Selected Recent Papers and Reports

- [Extending Understanding of Federal Statistics in Tables](#) (submitted to Conference on Universal Usability with Carol Hert, Liz Liddy, and Ben Shneiderman)
- [Agileviews: A Human-Centered Framework for Interfaces to Information Spaces \(with Gary Geisler & Ben Brunk\)](#) SILS Technical Report 2000-01
- [Augmenting Library Services: Toward the Sharium](#) Invited paper presented at the International Symposium on Digital Libraries 1999
- [An Alternative Site Map Tool for the Fedstats Website](#)
- [The people in digital libraries: Multifaceted approaches to assessing needs and impact](#) with C. Plaisant & A. Komlodi. Chapter to appear in A. Bishop, B. Battenfield, & N. VanHouse (Eds.) Digital library use: Social practice in design and evaluation. MIT Press.
- [Evaluation Report on the Perseus Project Publication Model 1998-1999](#)
- [Consider a Sharium DRAFT!](#) (an evolving concept paper)
- [Educating Responsible Citizens in the Information Society](#) appeared in Educational Technology Spring 1999
- [Advanced Interface Designs for the BLS Website: Final Report to the Bureau of Labor Statistics.](#)
- [Digital Library Research and Development](#) (article in Encyclopedia of Library and Information

Science)

- [Overviews and Previews for Multimedia Instructional Resources](#) (with Wei Ding, in ASIS 98 Proceedings, describes the BLC interface, including video browsing features)
- [Seeking Statistical Information in Federal Websites: Users, Tasks, Strategies, and Design Recommendations \(with Carol Hert\) Report to BLS](#) (summer 1997)
- [Public Access and Use of Government Statistical Information \(with Stephan Greene\)](#) (White Paper presented to the Federal Information Services NSF Workshop, Spring 1997)
- [Content+Connectivity=Community: Digital Resources for a Learning Community](#) (ACM DL '97 paper describing the Baltimore Learning Community Project)
- [Bringing Treasures to the Surface: Iterative Design for the Library of Congress National Digital Library Program](#) (ACM CHI '97 Design Briefing ftp'ed from the HCIL)
- [Costs of Educational Technology: A Framework for Assessing Change](#) (ED-MEDIA 95 Invited Talk)
- [The roles of digital libraries in teaching and learning \(with Hermann Mauer\)](#) (Communications of the ACM, April, 1995 (HTML) OR [full paper with color images at ACM Digital Library](#)),
- [Resource Search and Discovery](#) (Getty AHIP paper 1995)
- [Overviews and Previews for the Library of Congress National Digital Library program \(with Catherine Plaisant & Anita Komlodi\)](#) (1997 Technical Report summarizing LC NDL Project, final version appeared in Information Processing & Management 1998)



---

march@ils.unc.edu  
School of Information and Library Science  
University of North Carolina at Chapel Hill  
Chapel Hill, NC 27599-3360  
march@ils.unc.edu 919 966-3611

**updated 1/15/00**

# Research and Development in Digital Libraries

Gary Marchionini

## I. Digital Library Perspective

Digital library is a concept that has different meanings in different communities. To the engineering and computer science community, digital library is a metaphor for the new kinds of distributed data base services that manage unstructured multimedia data. To the political and business communities, the term represents a new marketplace for the world's information resources and services. To futurist communities, digital libraries represent the manifestation of Wells' World Brain. The perspective taken here is rooted in an information science tradition.

Digital libraries are the logical extensions and augmentations of physical libraries in the electronic information society. Extensions amplify existing resources and services and augmentations enable new kinds of human problem solving and expression. As such, digital libraries offer new levels of access to broader audiences of users and new opportunities for the library and information science field to advance both theory and practice.

High levels of attention and funding were first given to digital libraries in the early and mid 1990s leading to a plethora of visions and projects invariably driven first by finding ways to apply the many technologies developed in the 1980s and second by desires to create new technologies for managing distributed information resources. This perspective is best illustrated by the mission statement of the Digital Library Initiative Interagency Coordinating Committee charged with monitoring the progress of six large-scale projects funded by the US government. "The broad goal of the Digital Libraries Initiative is to dramatically advance the means to collect, store, organize and use widely distributed knowledge resources containing diverse types of information and content stored in a variety of electronic forms." This technical emphasis stands in contrast to the mission statement of a typical large public library. "The mission of the Carnegie Library of Pittsburgh is to be a force for education, information, recreation, and inspiration in the communities it serves." Thus, much of the early attention related to digital libraries was technology-centered and content-centered rather than people and community centered. In some cases, notably the efforts of national libraries or large academic libraries, efforts focused on extending access to existing collections through digitization and network access.

It is surely the case that all libraries will have some digital collections or finding aids and there will be some libraries that offer digital collections exclusively. At present the term digital library has focused on digital collections and limited access services. Depending on the source, digital libraries include anything from simple repositories of huge volumes of homogeneous electronic data with primitive access services

to the electronic extensions of the world's most prominent libraries (see CACM, April 1995 for briefings on plans by the Library of Congress and the British Library, see Representations, Spring 1993 for several commentaries on the Bibliotheque de France [Jamet & Waysbord, 1993], see CACM, April 1998 for briefings on various digital library projects internationally). To be called a library, an entity must be rooted in one or more communities of practice and be guided by a service mission that is manifested in policies of acquisition (collection development), organization, and access. Libraries offer both content and services guided by such policies and exist in a social-political context that influences policies and operations. To be modified by the term digital, a library must have some electronic content and services. In practice, a digital library makes its digital objects and services accessible remotely through networks such as the Internet or limited-access intranets. In some cases, the digital library objects and services may be distributed transparently to users from a variety of machines and locations. Thus, digital libraries are defined by mutually dependent attributes, which include content, services, technology, and socio-political culture.

## II. Content

Much digital library research, especially in the private sector has been driven by the aphorism "Content is king." Theoretically, any object from a text fragment to an animal in a zoo may be rendered digitally. Thus, there is no limit to the types of content that may be held by a digital library, however, there is a wide range of levels of practical difficulty in rendering different objects and in the efficacy such renderings offer to people. All content share intellectual, technical, and cultural challenges as well as offering specific challenges. Authority, surrogate creation, formats, intellectual property rights and costs of acquisition and maintenance are issues for all digital objects, but different types of content present special challenges.

### A. Types of Content

Most digital libraries provide renderings for textual objects. Whenever possible, text is scanned and optical character recognition (OCR) technology used to create digitally coded (e.g., ASCII, Unicode) renderings. Having text in digital form allows easy character string search as well as more sophisticated search and linguistic pattern matching analysis. Manuscripts that are not easily character recognized or have inherent value in the actual script, and the figures and images from typeset documents are scanned and provided as bit mapped image renderings. OCR accuracy and scanning resolution are bound by costs for textual documents and manuscripts respectively, i.e., perfect OCR requires costly human validation; very high resolution bitmaps are more expensive to store and transmit. An additional systems management effort is needed to coordinate the retrieval and display of ASCII/Unicode and bit mapped files for large collections. In some cases, texts are manually marked up using the Standard Generalized Markup Language (SGML) so that structural content as well as formats are provided to users. The edition of the text used is often an issue, especially for translations, as is the inclusion of critical commentaries or apparatus criticuses. How surrogates such as bibliographic records, abstracts, keywords/phrases, indexes, or concordances are created (e.g., automatically or manually) and displayed to users also vary across digital libraries. Consortia such as the Text Encoding Initiative address many of

these issues but different digital libraries use a variety of techniques and formats for their textual components.

Specialized digital libraries provide renderings for single medium objects such as images, statistical data, sound recordings, or silent films. Determining which formats to use is one challenge for such collections. The Museum Educational Site Licensing Project worked with images from seven prominent museums that provided images to the project in one of four distinct formats at seven distinct resolutions. Image collections provided by stock photo companies or projects such as the Library of Congress National Digital Library program currently provide multiple formats (e.g., GIF, TIFF, JPEG) to accommodate the wide range of platforms and software people may bring to the collection. For each image, multiple files must be stored, maintained, and linked to indexes and catalogs. Similar redundancies are currently necessary for digitized sound (e.g., AU, WAV, AIFF) and digitized film or video (e.g., Quicktime, AVI, MPEG, Shockwave). Although techniques to store high-quality data and create the required formats on the fly will surely be developed, digital librarians today must often juggle multiple files for the same object. Another decision librarians must make is what resolution to use for digitized objects. Is 300 or 600 dots per inch sufficient for photographs or must higher resolutions be acquired and stored? For example, an 8 bit digital rendering for a color slide of a vase in the Perseus digital library at 640 by 480 pixel resolution may be sufficient for students studying vase shapes and styles but may not be adequate for the art historian examining fine details.

Indexing for non-textual objects is particularly troublesome. Most digital libraries depend on textual captions or titles for retrieval and these distinct textual objects are themselves stored in different files. Creating and using new surrogates for non-textual objects is an active area of digital library research (see the services section below). Additionally, authority concerns are exacerbated with non-textual content (a mustache on the Mona Lisa is not an issue as it is an obvious alteration) and techniques for digital watermarking or information hiding are becoming commercially available.

Much of the interest in digital libraries stems from the possibilities of providing interactive multimedia content to users. Although video programming is the most obvious type of multimedia content, animated texts, hypermedia corpuses, on-demand video, and collaborative scenarios (MUDs and MOOs) for work or play are possible in digital environments and these dynamic digital events and objects will become part of digital libraries. Increasingly, the fruits of creativity and expression are inherently digital in nature. Computer simulations, games, and virtual worlds are objects collected in digital libraries. Interactive multimedia go beyond combining more than one medium to provide people with control mechanisms for making choices over multiple iterations, i.e., they are interactive. Digital libraries may provide access to a standardized entry point and leave it up to the user to deal with various components (much like libraries index a book rather than chapters or paragraphs). Interactive media allow the possibility of indexing at much finer granularities (e.g., words or video frames). Whether such fine-grained access is actually useful for information seekers remains to be determined. Surely, new kinds of surrogates such as document vectors or color histograms will be useful to the system for searching and text summaries and keyframe video extracts will be useful for user browsing.

All digital libraries must cope with making metadata available to users. Metadata are another level of content to librarians but a means to the content for users. Not only do digital librarians face challenges in standardizing metadata to insure interoperability across digital libraries, but the range and distinctiveness of metadata are problematic. In some cases, it is only the metadata that is made available digitally. In such cases, users search through pointers and must acquire the primary information physically or through a different (e.g., fee based) system. Such libraries are more properly considered as referral services rather than digital libraries. In more typical cases, metadata for objects of different granularity (e.g., titles for collections and titles for single objects) are mixed together on computer displays with full texts or objects. In physical libraries, the card catalog or OPAC is physically distinct from the items on shelves. These distinctions are difficult to make in electronic environments because everything is displayed on the same physical screen; thus the boundaries between metadata and primary data are often blurred.

Metadata are used primarily as intermediate steps to retrieving content but creators and digital librarians are creating new types of surrogates for objects to allow users to quickly preview and browse content. Huge challenges remain in creating surrogates for digital content. Today, most retrieval is facilitated through words--titles, captions, manually created descriptions, automatically extracted keywords, etc. There is enormous attention focused on creating non-textual surrogates such as color and shape characterizations for images and speaker identification schemes for audio recordings, but there are more difficult metadata issues looming as more content is not stored at all but created on the fly according to the specifications of the user. For example, today's web sites create specialized graphs from enormous varieties of statistical data in government repositories such as the Bureau of Labor Statistics. These graphs are generated on the fly according to the variables users specify. These new objects are impossible to uniquely title or index in advance as the permutations of hundreds of variables allow huge variants. As more digital libraries support sophisticated user profiles or agents, customized, original information objects will be provided to users from the library's "collection of possibilities." Physical libraries do not generally save and index results of reference activities except to create tickler files to help reference librarians the next time a question is asked or to use in creating pathfinders for popular topics. Characterizing what content is possible rather than what exists is a much larger challenge in digital libraries.

## **B. Managing Content**

Many digital library efforts devote the bulk of their resources to managing content. The key content management functions in any library are selection and acquisition; indexing, storage, and access; and collection maintenance. Most of the research and development activity in early digital library efforts were devoted to these functions, although it is likely that more attention will be given to user services as digital libraries mature.

### **Selection and Acquisition.**

Libraries select content according to a collection development policy. Such policies manifest the missions of the library and determine how materials budgets are expended. Many digital libraries, especially those in governmental agencies, have arisen out of the need to take an existing body of electronic materials and make them available to users. Some digital libraries are strictly opportunistic, selecting objects to digitize from the existing collection and those for which intellectual property rights are held. For example, the Library of Congress National Digital Library Program selected objects from a variety of reading rooms that were out of copyright (historical collections) or that were produced by US government agencies where copyright is not claimed. Some digital libraries mainly acquire or develop materials according to specific missions and policies and augment the collection opportunistically. For example, the Perseus project texts were mainly acquired opportunistically from the Loeb collection at Harvard University which holds rights. In some cases, new translations were commissioned. On the other hand, the bulk of the images in the Perseus Library are from original photography of museum objects that were selected to meet scholarly and pedagogical goals. Another example of a specialized digital library is the Alexandria Project at the University of California Santa Barbara that focuses on spatial information. This project aims to make existing maps and other spatial material more broadly available by combining digital representations for visual objects such as maps with the text-based attributes of names and geographic features (e.g., Smith, 1996). It is likely that as digital libraries continue to evolve, new, specialized collections will be built according to institutional missions and well-defined collection development policies.

There are two key challenges for content selection: cost and quality. First, librarians consider the costs of acquisition. Intellectual property rights are an important first consideration, but the costs of digitization and maintenance must also be taken into account. Libraries that receive collection gifts often require that donors supply funds for cataloging, shelving, and preservation and digital gifts bring their own one-time and ongoing costs. Second, librarians consider the quality of the content before acquiring it-- If content is king, quality is its lineage. This is a more problematic consideration because issues of authenticity as well as veracity arise. Which of Monet's lily studies best represent his style at the end of his life? Is a transcribed Latin text from one 15th century Italian monastery superior to a second transcription from a neighboring city? Which of the many best seller lists are most authoritative for adult fiction? Which web site collaborative rating service is most useful for selecting eighth grade science simulations? The issues of cost and quality are addressed in numerous, labor-intensive ways in traditional libraries and it is unlikely that this will change fundamentally in digital libraries, although collaborative ratings and better communication facilities will certainly augment librarians' as well as patrons' abilities to make informed judgments about what they select.

Once decisions about selection are made, content must be acquired. Payments in physical libraries are already mainly computerized but delivery may be easier in digital libraries. For objects already in digital form, then file transfer through networks or mass storage is straightforward as long as file formats are well-specified. In the case of physical objects, digitization must be conducted. Scanners for text and images range in quality on several dimensions: output resolution, value and condition of the physical objects (e.g., brittle, rare manuscripts must be handled differently than technical reports), speed (digitizing 100 images is quite a different task than digitizing a million images). In addition to the engineering challenges digitization provide, policy decisions must be made. For example, which

resolutions and formats to adopt, which text to OCR and error correct, how to link different digital representations for multiple media from single collections (e.g., a manuscript collection that includes field notes, photographs, and audio tapes). The complexities and tradeoffs involved in digitization and user access are well-illustrated by the CORE Project that systematically applied different digitization schemes to published chemistry materials and then conducted multiple user studies (Entlich, et. al., 1996).

### **Indexing, Storage, and Access.**

Once content has been selected and acquired, it must be added to the collection in such a way that users will be able retrieve it effectively. Indexing, storage and access are perhaps the most active areas of research and development in digital libraries. Digital libraries have given new life to work on automatic indexing as manual indexing of huge volumes of data are beyond the resources of most libraries. In many cases, texts are "indexed" using vector-space or probabilistic information retrieval models that provide access through weighted values for all but a few common words. Thus, the classification system itself is empirically determined from the data as a by-product of the indexing. Perhaps the most successful example of this approach to date is the Inquiry system (Croft, Cook, & Wilder, 1995) used as the retrieval engine in many digital library projects (e.g., the Library of Congress). These approaches stand in stark contrast to the traditional approach of manually assigning objects to a limited number of manually constructed concept classes (classification system) represented in a controlled vocabulary (e.g., Library of Congress Subject Headings or Medical Subject Headings). Other automatic techniques index objects to mathematically abstract concept classes, for example, Latent Semantic Indexing assigns documents to "concepts" composed of term-vector document singular vectors (e.g., Deerweister, et. al, 1990) Several WWW-based services use a hybrid approach by manually creating a classification system and then using automatic techniques to assign objects. Perhaps the most ambitious effort to automatically index large volumes of documents to date is the work of Schatz and Chen (Schatz et. al, 1996; Chen et. al, 1997) who have applied supercomputer resources to indexing scientific and engineering documents. Additionally, as digital libraries become more global, multiple language documents, documents in different languages, and multiple language versions of documents are concurrently available to users who bring queries expressed in different languages. Researchers are actively applying existing text retrieval techniques to the cross-language retrieval problem (e.g., Sheridan & Ballerini, 1996).

Most retrieval systems for images, video, audio recordings and other non-textual objects have depended on text items such as title, creator name, or manually assigned subject headings for retrieval. Digital libraries have generated enormous research interest in inventing indexing techniques that do not depend on text representations. One line of research is to adapt the statistical techniques used in text retrieval to characterize objects by feature vectors for characteristics such as color (color histograms are commonly used) and brightness. Researchers also have begun to tap the research in robotics (vision systems) and signal processing to automatically extract unique attributes such as shapes, optical flow, and pitch that may be used for retrieval.

For example, image segmentation is a fundamental problem in image processing in general and also in

creating surrogates for retrieval and use. Various feature analysis techniques have been exploited to identify images and provide a basis for queries. These include edge and corner detection, foreground/background separation (e.g., Rosenfeld & Smith, 1981), texture analysis (texture energy determined by filters (e.g., Jain, Ratha, & Lakshmanan, 1997), and color (e.g., Jain & Vailaya, 1996). These feature analyses are augmented by measures of optical flow in moving images (e.g., Sim & Park, 1997). The Informedia Digital Library Project at Carnegie Mellon University has applied several of these techniques to support video search and browsing (e.g., Wactlar et. al, 1996). Some of the techniques have been integrated into commercial products such as IBM's incorporation of Query by Image Content (Flicker et. al. 1995) techniques into its Digital Library Solution. It seems certain that the digital library research and development activity of the 1990s will insure that considerable progress is made in automatically indexing non-textual objects with non-textual attributes. New indexing challenges will emerge as more dynamic objects (e.g., virtual conference proceedings, active networks) are added to digital libraries. The temporal nature of such objects will require ongoing indexing--consider, for example, how you would index the events of your life as it progresses.

There is a two-fold advantage to electronic content. First, a multiplicity of pointers are economically feasible since many separate cards or other physical devices need not be created. Thus, rather than the four or so catalog cards (author, title, and a few subject headings) in a physical system, dozens or hundreds of index terms may be assigned or many different levels of representation may be created in an electronic system. It is essential to novel and flexible access interfaces that multiple and varied indexes be available. Second, unlike physical objects which must reside in a single space, electronic objects may exist in many locations. Thus, the logical many-to-many relationships among concepts and information objects may be leveraged for both searching and browsing in electronic environments.

Storage is mainly a technical requirement although new media may complicate storage decisions and costing. When data is to be delivered continuously (e.g., streaming video or audio) rather than as discrete files, then alternative technologies are required (drives and database management software that operate continuously have different engineering requirements than drives optimized for bursts of data). Today's large digital repositories use multiple levels of mass storage media (e.g., disk, tape) and mechanical robots to locate and mount the media. Various supercomputer centers today use tape robots that provide rapid access to many terabytes of data (e.g., the Oak Ridge National Laboratory in 1997 had capacity for 100 terabytes of uncompressed data). Digital libraries will surely apply such technology just as libraries today apply movable shelving and complex conveyer systems to move physical materials.

Ultimately, users must be able to access the content digital librarians have selected, indexed, and stored. During the 1970s, large libraries invested heavily in computerizing cataloging and circulation functions to give users faster and better access and service. Online Public Access Catalogs have evolved to give library patrons remote access to the bibliographic records. Digital libraries offer access to primary content using a variety of access tools. An active area of research is user interfaces for digital collections. Access interfaces depend on the content organization and storage discussed above and serve as the bridge between internal (technical services) and patron services. Access interfaces are considered under search services in the Services section.

## **Maintenance.**

Maintaining buildings and systems, and preserving content are important and costly activities in physical libraries. Digital libraries may avoid some of the costs of wear and tear on buildings and books but still have significant maintenance costs, including some unique to electronic environments. System hardware and software upgrades have become accepted expedencies of today's workplace and there is no reason to expect that this will change. New equipment, improved or alternative network solutions (e.g., ISDN, ATM, wireless), and software upgrades will require excellent technical personnel. Archivists have long worried about the persistence of digital media. Magnetic tape life expectancies are typically less than ten years under ideal temperature and humidity conditions. Optical storage offers longer life spans, but digital librarians must plan for copying digital holdings periodically and especially plan for the inevitable obsolescences of different media types and playback devices. These maintenance issues correspond to traditional maintenance requirements but their because they apply across many industries and require rapidly changing technical skills, they tend to be much more expensive.

Just as the computational systems change, digital content may also change. A digital document may have numerous versions, especially given the ease with which electronic documents may be changed. Maintaining the most essential (not necessarily the most recent) document requires that versions be well managed, including updating and deleting the links to those objects. In addition to this version control problem, digital librarians must manage the multiplicity of indexes and file formats. More problematic are link management requirements as hypertext links are created among distinct documents. A policy such as requiring all links to point to the top of a document (e.g., main home page of a web site) aid the librarian in managing links in a database but may not serve the user who expects to go directly to the location of the relevant information.

Although most of the research and development effort in digital libraries has been devoted to building the collections and making them available to users, there is enough experience for the creation of digital librarian's tool kits. Such a tool kit might include tools for selecting, acquiring, indexing, and maintaining digital content. For example, tools would include: library building tools for viewing directory structures, converting formats, checking screen layout consistencies, quickly viewing objects, and encrypting data; interface simulators for testing interfaces on multiple platforms; database tools for property rights, file naming histories, links, and metadata definitions; and maintenance tools for automatically checking links, automating transaction log analyses, maintaining security, updating versions, and backing up the system. Although many of these tools exist, developers will surely undertake systematic efforts to augment the list and bring them all together with a common interface amenable to the widest possible set of digital libraries.

For existing libraries, many of the decisions related to managing content are questions of how many resources to divert from existing operations and what levels of redundancy to assume for physical and digital collections and services. For new, exclusively digital libraries, the decisions are driven mainly by

resource acquisition.

### **III. Services**

The range and depth of services that a library provides to patrons are driven by its service mission and policies. Policies determine who may use the library, when content and services are available, what types of services are available, and how resources are allocated to patrons and services. Digital technology offers the potential to radically change who may use a library, when they may do so, and what types of services are offered. Digital library services both amplify existing services and augment library service with new possibilities for users. Thus, libraries that offer digital content and services must reconsider their policies in light of new capabilities and patron demands.

#### **A. Who are the patrons?**

Given network capabilities, libraries must decide whether to expand their user populations beyond the usual physical limitations of time and space. Patrons can access digital content at any time of day but human library services may still be restricted to local working hours. Public libraries must decide whether to seriously consider serving the world community rather than the local population that supports the library. Like corporations that provide access through restricted intranets, public libraries now may opt to maintain local community user policies through password access. Until public intranets or some other solution emerges, it is much easier for public libraries to provide universal access to local bibliographic holding data and password access to other databases and services. The recent creation of the Gates Library Foundation aims specifically at public libraries and may have an enormous impact on electronic services in public libraries. National libraries may leverage digital technology to more realistically serve the population in its service mission, or expand its policy. The Library of Congress service mission, for example, does not explicitly serve children, however, the Library of Congress National Digital Library Program does have an explicit outreach to K-12 schools. Thus, the digital library effort has effectively broadened the scope of service for this national library. Additionally, libraries often make arrangements to serve users with special needs (e.g., access ramps, Braille books) or users from varied cultures (e.g., languages and customs) and must find ways to extend these services in digital environments.

#### **B. What types and qualities of service to offer?**

Even more difficult than who can use digital library resources are decisions about what types of services to provide digitally. Ultimately, this challenge may define the legacy of digital libraries. Libraries offer different types of reference and referral services (e.g., ready reference, exhaustive search, selective dissemination of information), instructional services (e.g., bibliographic instruction, database searching), added value services (e.g., bibliography preparation, language translation) and promotional services (e.g., literacy, freedom of expression). Although much of the impetus for digital library research and development was content, it is clear that the most used and engaging aspect of the Internet is electronic

mail and chat rooms. People want to communicate and collaborate. Libraries that develop service strategies for connecting people together in information-rich environments are most likely to prosper.

Services can be provided at different quality levels according to how resources are allocated. Libraries set policies about how much time a reference librarian may spend on reference questions, how requests are received (verbal in person, written, phone, fax, email, etc.), and what types of special services are offered. Digital technology offers new capabilities as well as different, often greater, expectations on the part of patrons. The changing expectations of service populations demands that digital libraries continue to revise service policies.

### **C. Search services**

The most basic access service is search of the library's collection. Online catalogs have long provided author, title, and limited subject access to local holdings and more recently to union holdings across multiple libraries. The expectation for digital collections is that catalogs should seamlessly link to the digital collection itself so that remotely located users can not only find and display bibliographic information but also the primary information objects. This expectation yields several challenges to librarians. Distinguishing metadata and primary data is not a trivial problem in rich collections. In homogeneous collections it is possible to define a unit of primary information (e.g., a book rather than a chapter or series) but this is more problematic in heterogeneous collections containing finding aids, manuscript bit maps, videos, and hypertexts. The challenges are first, to extract and provide multiple levels of representation and second, to provide users with control mechanisms to move from high-level surrogates to detailed objects (Marchionini, 1995). This is a basic human-system interface problem. The mechanisms digital librarians provide to users depend on the levels of representation that are available in the collection.

The most common search mechanism is a query line or form that allows users to enter a term or terms as a query. Depending on the type of indexing the library uses, ranked lists or exact-matched sets of results are returned to the user. There is a rich history of query-based searching from the information retrieval research community and online service industry that digital libraries may build upon. Limitations in the architecture of the WWW (statelessness) strongly limited many of the early WWW-based digital library search mechanisms, but server-side caches, client side caches (e.g., "cookies") and the development of Java allow the incorporation of mechanisms known to improve search capabilities such as relevance feedback and user profiles. These advances and growing experience in web-based designs have also led to support for more sophisticated search options (e.g., proximity, scope limits). One of the most pressing needs is for search mechanisms that give users more control over results--most give users simple lists with perhaps some sorting options. Interface prototypes for the Library of Congress (Plaisant et al., 1997) give users information about the level of representation of results (e.g., collection or item, media type) as well as flexible options for sorting and display.

Interactive environments have begun to force designers to accept the user's perspective that browsing is a

legitimate information seeking strategy (Marchionini, 1995). Nowhere is this more dramatic than in hypertext environments such as the WWW. Thus, many digital library access interfaces provide users with navigational mechanisms. These are often based on some high-level hierarchical classification that allows users to select categories at increasingly detailed levels of granularity to eventually reach specific information objects. Clearly, useful digital libraries will provide hybrid solutions that allow users to apply both selection and query strategies according to their specific experience and needs.

Although search forms and selection-based navigation are the default access mechanisms in most digital libraries, there are an array of novel interfaces that allow users to manipulate visualizations of collections. Shneiderman's (1994; Ahlberg & Shneiderman, 1993) dynamic query interfaces allow users to query collections through direct manipulation tools such as sliders and immediately see the results of these actions. Fox et. al. (1993) have developed a visual interface for a computer science literature digital library that allows users to manipulate search results represented as an array of icons. Hearst (1997) has merged clustering techniques (scatter gather) for search with visualization of results (tilebars) to help users search and explore digital collections. Lin (1997) creates semantic maps that depict two-dimensional maps for high-dimensional concept spaces. The map region sizes are proportional to the importance of the concept and the juxtaposition of the regions represents the similarity of concepts. Korfhage (1997) has developed several different interfaces that graphically represent users' points of interest within a concept space. Some systems provide zooming mechanisms that allow users to easily shrink or expand information spaces. Bederson's Pad++ system (Bederson & Hollan, 1994) allows continuous zooming that is highly effective for graphical objects such as timelines, hierarchies, or images. . Marchionini et. al., (1997) have combined dynamic query interface style with video preview techniques in a digital library of instructional resources.

Digital libraries will take advantage of these developments to provide users with usable yet powerful interfaces to control sophisticated computational tools behind the scenes. Informedia, for example, provides users with a variety of interface tools such as spoken queries, video walls and video skims to search and browse with advanced pattern recognition systems in the background (Wactlar et. al., 1996).

## **D. Reference and question answering services**

Although digital libraries may provide communication channels (e.g., chat rooms, Internet "News" groups) where people may interact to answer each others' questions, many patrons come to librarians for answers to questions. Librarians may provide answers, references to literature that may contain the answers, or referrals to other people or services. These reference services are an essential part of most libraries' mission and an important question is how such services will evolve as a result of technology. There are five ways that reference services are provided in digital libraries.

The most basic service is to anticipate questions and provide canned answers. Frequently asked question services (FAQ) anticipate common questions and provide answers so that users can go to the FAQ service before requesting human assistance. These services are particularly popular for system-related

questions that new users typically might have. In a more elaborate version of this solution, digital librarians may also create electronic pathfinders for specific topics that they anticipate may be useful to many patrons. These pathfinders or special collections are then featured prominently at the library's virtual entry point.

A second type of solution leverages asynchronous exchange between patrons and librarians or content experts. Certainly, electronic mail requests allow users to reach reference services more conveniently. These online reference services are logical extensions of traditional reference services that respond to written requests and facilitate multiple iterations over times convenient to users and librarians. Although technology allows digital librarians to serve patrons more conveniently, these solutions still demand substantial human attention. Moreover, the availability of digital assistance tends to increase the volume of requests and the expectations of requesters.

A third approach is to combine automated and human services. If FAQ solutions fail the user, the request is forwarded to an appropriate automated service or human expert. Services such as the Answer Garden (Ackerman, 1993, Ackerman & McDonald, 1996) not only route questions through the FAQ list as they come in, but also capture new requests and the human responses and add them automatically to the FAQ list. Such a system has the added benefit of sharing new questions and responses to the corporate memory of the particular community of practice. Moreover, it can lead to longer queries which can yield better results with today's search engines. We can expect to find many new hybrid solutions to the reference problem that take advantage of both human and machine capabilities.

A fourth solution is real time dialog with a librarian or content expert augmented by technology. Software customer service hotlines and catalog order centers use databases and telephone management software to speed their work and digital libraries will also leverage such tools to provide human reference service. In the case of system help questions, help desk tools that allow respondents to replicate what users see on their screens remotely (or in intranet environments actually allow information specialists to take over a remote machine for trouble shooting) offer new possibilities for librarians to provide remote reference service. Internet chat or video links may be very effective for specific reference advice but is very expensive since it demands concurrent human attention. As such, it will find applications first in corporate digital libraries and public fee-based services.

The most ambitious solution to the question answering problem is to create software agents that take into account the user's context and act as human surrogates. The Knowledge Navigator video created by Apple Inc is the quintessential example of such an agent. Natural language understanding is necessary but not sufficient for these automatic reference services since reference librarians often help people to clarify and articulate their information needs. As the NLU problem is itself incredibly complex, it is likely that we can expect progress to be made by teaming humans and machines and finding the best allocations of machine and human resources for answering reference requests.

## **E. Filtering and Selective Dissemination of Information**

A service that is particularly important in special libraries is selective dissemination of information--sometimes know as routing, alerting, or filtering. Users develop interest profiles and as new materials are added to the collection or become known to the library staff, they are compared to the profiles and relevant items are passed on to the users. Filtering services are particularly applicable to newswires, Internet "News", and broadcast media abstracting services. Electronic user profiles in conjunction with online database services have long been available and will surely proliferate as more library content becomes available digitally. Automatic filtering services differ from retrieval services in that in filtering the corpus changes dramatically from period to period (e.g., day to day) and the query remains relatively stable (Oard, 1997). This leads to queries (profiles) that are more carefully and fully developed, and to the need to extract the salient regularities and relationships in the corpus anew each period. In some cases, filtering services provide added value by abstracting primary information (e.g., answers to a standing question, hyperlinked threads across documents) to users whereas search services typically bring documents that may contain the primary information.

One interesting extension of this concept is to use the connectivity inherent in digital libraries to support collaborative filtering where patrons rate or add value to information objects and these ratings are shared with a large community so that popular items can be easily located or people can search for objects found useful by others with similar profiles (Maes, 1994; Resnick, 1997). Such an approach is analogous to peer review for research papers, but involves many more reviewers. Although there are privacy issues related to personal profiles, the benefits of collaborative filtering may make such services increasingly important for libraries. Eventually, specialized library services will emerge to manage large numbers of profiles. Such profile management systems will only be able to optimize performance (e.g., by leveraging redundancies in profiles), but also serve as population parameters for social scientists and historians studying group behavior. In addition, digital libraries may provide services that assist users in developing and maintaining profiles.

## **F. Instruction**

Libraries have always been an essential element of the educational infrastructure. In formal learning settings (e.g., K-12, university) libraries are the center of the school. This is evident from the often cathedral-like architecture to the certification requirements imposed by accreditation bodies. More importantly, libraries are essential in supporting informal and professional learning beyond the formal school system. We have argued (Marchionini & Mauer, 1995) that digital libraries will lead to more close integration among formal, informal, and professional learning. Digital libraries offer new opportunities to break down classroom walls and allow people to learn wherever they are and whenever they want. Many digital library projects seek to bring multimedia resources to teachers and students on demand. For example, the Earth System Science Community (<http://www.circles.org/>) and the University of Michigan Digital Library Teaching and Learning Project (<http://www.umich.edu/~aaps/>) aim to provide students with rich, interactive science materials. The Baltimore Learning Community ([www.learn.umd.edu](http://www.learn.umd.edu)) collects and indexes multimedia materials for middle school social studies and science, the Perseus Project ([www.perseus.tufts.edu](http://www.perseus.tufts.edu)) provides materials and tools for students and

teachers of classics, the Museum Site Licensing Project (<http://www.ahip.getty.edu/mesl/home.html>) brings together seven museums and seven universities to share art resources, the Library of Congress National Digital Library Project includes a Learning Page devoted to supporting K-12 schools (<http://lcweb2.loc.gov/ammem/ndlpedu/>) and the Informedia Digital Library has also been applied in high school settings (Christel & Pendyala, 1996). Such resources will continue to drive both teacher-led and self-directed learning as more high quality materials are digitized and thoughtful links and pathfinders are created by students, teachers, and librarians. In such digital libraries, all participants are learners and teachers.

In addition to providing the content to enrich learning, librarians help patrons acquire information-seeking skills (traditionally known as bibliographic instruction) which have become more essential in the informed society (many school library media specialists and public librarians collaborate on information literacy courses). Digital libraries have the potential to support collaborative distance learning and to provide intermediation services to aid participants in shaping questions, finding relevant materials, and interpreting and using information. These intermediations will surely require new types of human support services augmented by computational tools. The new learning facilitators who work in such environments will themselves be learners who are part librarian, part teacher, and part debate moderator. Their roles will range from facilitating collaborative learning to assisting individuals configure the local area networks carried on their bodies.

#### **IV. Technology Requisites**

Digital libraries are dependent on and driven by several general purpose technologies such as computer hardware, high-speed networking, security, and interoperability.

Better computers are needed on both the library and user end. Today's workstations serve thousands of users per hour but as more information is streamed (e.g., video, real-time collaborative experiences) rather than transferred as discrete files, more powerful machines and storage devices and new intermediary machines will be required. In addition to storing ever-increasing volumes of digital objects, libraries will also need additional computational resources to store billing and transaction log data. Thus, continued progress in digital libraries will benefit from faster, more powerful CPUs and cheaper, higher-density storage devices.

The trend toward new and multiple input and output devices (e.g., Jacob et. al., 1993) will also influence how digital libraries evolve and are used. Speech, gestural, and tactile input devices should allow users to more easily control access tools and library resources. Likewise, a richer array of output devices ranging from large, flat-panel displays to digital paper will open new possibilities for digital librarians to share collections more broadly and easily. Software tools for rapid prototyping and testing of interfaces will also help designers improve the quality of digital library interfaces.

High-speed, reliable, ubiquitous networking will allow libraries to become increasingly digital. Research and development in network architectures, low-cost access in homes, and performance metrics will continue to determine how digital library access moves from privileged locations in campuses, businesses, and government offices to homes. Developments that allow library access through a mix of wired and wireless paths will enable this spread of access. Engineering research directed at seamless interoperation of networks ranging from personal body LANs to the WWW is a high priority for the technical community.

Most importantly, networked computational resources are becoming more mobile and special purpose. Users will no longer be strictly tethered to workstations to use digital information. This trend will allow libraries to provide new genre of information services. As networked computational resources are commonly built into buildings, automobiles, appliances, clothing, jewelry, and various prosthetics, libraries will be able to provide users with continuous information streams rather than only discrete information objects. Obviously, SDI services will take on entirely new meanings when information can be streamed continually to users wherever they are, whatever they are doing, and without interrupting whatever activity is underway. Users will be able to choose to receive the latest information from their information service unobtrusively via earpiece or eyepiece during a meeting. Weather, traffic conditions, or other public information may be continually delivered on special-purpose displays or speakers built into homes, offices, or vehicles. Proactive libraries will invent new ways to augment discrete collections and services with accretional collections and ongoing services.

Software developments that support rapid, reliable, and secure transfers support this developing infrastructure. However, new software that supports user search and library services with easy to use interfaces demand enormous software engineering efforts. As much as half the code in today's programs is devoted to the user interface and the challenges of multiple input/output devices for a wide variety of distributed computational devices will require new paradigms for human-computer interaction and improvements in algorithms for information organization and search.

An important condition for continued development of digital libraries is seamless exchange across different digital libraries. This interoperability problem is addressed on two fronts. First, groups work to create standards for data storage and transmission, for query representation, and for vocabulary control. In this solution, digital libraries adopt standards and change content and services at the local level. The standards solution proceeds based upon shared interests but depends on agreement among vested interests and most often must follow long-term implementations adopted in the marketplace. The second approach is to allow individual digital libraries to be as innovative as necessary but to create public services that map local content and services to other digital libraries (as word processing programs read files created by other systems). The Z39.50 protocol exemplifies such an approach for mapping queries to different databases (Lynch, 1991). An extension of this approach is to publish an abstract view (an ontology) accepted by a federated community that may then be used to facilitate interoperability (Wiederhold, 1992). The Stanford Digital Library Project is addressing the interoperability problem with

an architecture called the InfoBus (<http://www-db.stanford.edu/~testbed/>).

## V. Culture

Libraries are keepers of culture. As such, they are subject to and reflect the social, political, and economic forces that shape their constituency. Likewise, libraries as institutions help to shape culture. Public and academic libraries promote values of scholarship and appreciation for culture but are subject to localized beliefs and motives--turf is often defined ideologically. Corporate libraries support the mission of the organization and vie for resources with other cost units such as data processing. Digital collections provide new challenges for the socio-political context. Public library policies must specify who may access collections, which collections are digitized, and what existing resources will be cut to support technology and digitization. Corporate and academic managers must grapple with integration of computing and library functions. Libraries will evolve in different ways in different cultures and over time, successful models for specific environments will be adopted more widely, however, digital technology will not lead immediately to standardized practices but rather to more diversity based on the socio-political forces of the constituent community. Whether the potential of information technology to promote cultural standardization will overshadow its potential for empowering individual expression is a long-term socio-political issue. See the report from the social aspects of digital libraries workshop for views on these issues (<http://www.gslis.ucla.edu/DL/>).

### A. Economic Challenges.

In addition to the challenges of community-based context, two global and interdependent issues influence research and development in digital libraries: intellectual property rights, and information security and authority. Both issues are rooted in a culture that places economic value on scarcity.

Copyright exists to promote intellectual production by providing economic incentives. Security protects unauthorized access but also must deal with the more subtle problem of insuring the veracity and authority of digital information objects. The ease with which perfect and unlimited copies of digital products may be made causes many owners of intellectual property to avoid digital distribution. Avoidance is moot for the growing set of products created in and for digital technologies (e.g., software, games, virtual worlds, hypertexts), but owners of the existing base of books, photographs, films, sound recordings, and other intellectual property have begun cautious experiments in digitizing and repurposing their assets to develop new markets. For example, the CORE Project (Entlich, et. al., 1996; ) is a collaboration between the American Chemical Society and universities and the Tulip project a collaboration between Elsevier and universities (<http://www1.elsevier.nl/homepage/about/resproj/tulip.htm>), both aimed at exploring scientific journal licensing and delivery schemes. Additionally, the JSTOR project aims to create a new non-profit electronic publishing medium (<http://www.mellon.org/jstor.html>), and the Association for Computing Machinery has developed a digital library policy and site (<http://www.acm.org/dl/>).

There are two types of questions that shape digital library research and development. The first question is: What does it mean to use intellectual property? Current practice provides for specialized fair use in socially beneficial situations (e.g., education, scholarship). Do these fair uses apply to digital objects? Publishers have developed agreements about derivative works composed from individual pieces of intellectual property but the nature of derivative work may be different in the digital realm. For example, do people have to pay for the right to link to other work? To deal with such questions, there are efforts to change copyright laws to protect digital objects. Some publishers press for elimination of the right of resale that allows people who purchase a book or other object the right to resell it. The extreme interpretation of this approach is that every random access memory representation for an information object requires payment to the copyright holder--a type of payment for the potential of using a digital object. See Pamela Samuelson's column in Communications of the ACM for a series of thoughtful discussions of the issues related to intellectual property in electronic environments.

The second question deals with the technical problems of how to protect intellectual property against illegal use. There are many efforts to develop technical solutions that protect copyright either through copy protection or automatic billing mechanisms. Research on encryption algorithms, digital watermarking, and electronic commerce are leading to the development of trusted systems that protect intellectual property rights by managing the necessary financial transactions while protecting consumers by providing authoritative information securely (Stefik, 1997). Encryption has advanced beyond the point where most government agencies or individuals can monitor or decode personal communications. Techniques to include either visible or hidden watermarks on digital objects have also been developed and incorporated into commercial products. These techniques insure the veracity of an object and may help prevent copying and distribution in the open marketplace. Digital commerce is an active area of research with different approaches under testing. Systems such as CyberCash use a third party intermediary to mediate transfer of property and payment; systems such as Digicash issue digital money in the form of bit stream tokens that are exchanged for intellectual property and recirculate throughout the network; and systems such as Netbill use a prefunded account to enable intellectual property transfer. These developments are important for digital libraries that may want to offer minuscule priced (e.g., one cent) digital objects to huge volumes of customers. Traditional credit card purchase schemes (often used for high cost items or conference registration) are relatively costly to maintain (e.g., thirty cents plus 1.75% of the purchase price) and thus the digital schemes above will continue to be developed until one or a few marketplace winners are determined.

## **B. Communities of Practice.**

Libraries are as much defined by people as by information resources. Libraries come into existence because people wish to preserve and share representations of heritage and wisdom. Libraries serve one or more communities with common interests and culture. In the past, the extent of communities was highly constrained by physical distance and the size or importance of the community. Thus, many kinds of libraries were geographically bound (e.g., local public libraries) and others (e.g., international repositories, large research libraries, corporate information centers) were limited to critical communities of practice (e.g., health, science, commerce). Digital technology facilitates specialized libraries that may

serve very diverse and unique world-wide communities with very few or disadvantaged members who would not be able to support or travel to a physical library. This represents a fundamental shift in library services and human culture more generally. Decentralized, specialized, globally dispersed special interests can blossom though shared information and communication resources.

Digital libraries, like any social phenomenon are not only shaped by the social elements that motivate them, but will eventually influence those elements. It is too soon to tell how institutions will change as a result of the digital libraries they create. Will the global mission of the Library of Congress change as K-12 students (who have traditionally not been part of the service mission) are attracted by the National Digital Library Learning Page? Will the National Library of Medicine's decision to make Medline available free to the public affect its primary mission to serve the medical research and development communities? Will government agencies like the Bureau of Labor Statistics shift resources to serving the increasing numbers of requests from the public that result from making labor statistics available through a digital library? If so, where will the resources come from? Will the service missions and resource allocation formulae of government agencies evolve toward dissemination of information to the detriment of information gathering and creation? How will corporate digital libraries reshape the nature of business in different corporations? What new communities of practice will emerge to amplify and augment the information industry in general? Will increasingly interdependent information resources change the function and form of nation states?

The digital library research and development community is only beginning to address the organizational impacts of digital libraries. It is clear that although the genesis of digital libraries was digit centered, there is increasing attention given to digital libraries as people and organization centered entities that reflect communities of practice. It is sensible to expect that the long-term implications for global culture will be reflected in the evolution of libraries--keepers of culture.

A third of a century ago, Douglas Engelbart provided a vision of electronic technology for augmenting the human intellect. Just as libraries reflect and influence human culture, digital libraries will extend and augment the collective intellect. They aim to make knowledge more equitably and universally accessible and to link people together through their information needs. Research and development in digital libraries may have been initiated by technology, but it ultimately is in the service of extending and augmenting human interactions.

Acknowledgements: The author thanks Doug Oard and Tony Tse for helpful comments on earlier versions of this paper.

## References

- Ackerman, M. (1993). Answer Garden: A Tool for Growing Organizational Memory. . Doctoral Thesis, MIT.
- Ackerman, M. & McDonald, D. (1996). Answer Garden 2: merging organizational memory with collaborative help. Proceedings of ACM Conference on Computer-Supported Collaborative Work (November 1996), 97-105.
- Ahlberg, C., Williamson, C., & Shneiderman, B. (1993). Dynamic queries for information exploration: An implementation and evaluation. In B. Shneiderman (Ed.), Sparks of innovation in human-computer interaction. Norwood, NJ: Ablex. p. 281-294.
- Bederson, B. & Hollan, J. (1994). Pad++: A zooming graphical interface for exploring alternative interface physics. Proceedings of UIST '94. Marina del Rey, CA, Nov 2-4. 17-26.
- Chen, H., Ng, T., Martinez, J., & Schatz, B. A concept space approach to addressing the vocabulary problem in scientific information retrieval: An experiment on the Worm Community System, Journal of the American Society for Information Science, 48(1), 17-31, 1997.
- Christel, M.G., & Pendyala, K. Informedia Goes to School: Early Findings from the Digital Video Library Project. D-Lib Magazine, September, 1996. <http://www.dlib.org/dlib/september96/informedia/09christel.html>
- Croft, B., Cook, R. and Wilder, D., "Providing Government Information on the Internet: Experiences with THOMAS," in Proceedings of the Digital Libraries Conference DL'95, Austin, TX. June 10-12, 1995, pp. 19-24.
- Deerwester, S., Dumais, T., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391---407.
- Entlich, R., Garson, L., Lesk, M., Normore, L. Olsen, J. & Weibel, S. Testing a digital library: User response to the CORE Project, Library Hi Tech, 14(4), 99-118, 1996.

Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D. and Yanker, P. (1995). Query by image and video content: The QBIC system. *Computer*, 28(9), 23-32.

Fox, E., Hix, D., Nowell, L., Brueni, D., Wake, W., Heath, L., & Rao, D. (1993). Users, user interfaces, and objects: Envision, a digital library. *Journal of the American Society for Information Science*, 44(5), 480-491.

Hearst, M., Interfaces for searching the web. *Scientific American*, 276:68-72 (1997).

Jacob, R., Leggett, J., Myers, B., & Pausch, R. (1993). Interaction styles and input/output devices. *Behavior and Information Technology*, 12(2), 69--79.

Jain, A. K. & Vailaya, A. (1996). Image retrieval using color and shape, *Pattern Recognition*, 29(8), 1233-1244.

Jain, A., Ratha, N. & Lakshmanan, S. (1997). Object detection using Gabor filters. *Pattern Recognition*, 30(2), 295-309.

Jamet, D. & Waysbord, H. (1993). History, philosophy, and ambitions of the Biblitheque de France. *Representations*, 42 (Spring). 74-79.

Korfhage, R. (1997). *Information storage and retrieval*. NY: John Wiley.

Lin, X. Map displays for information retrieval, *Journal of the American Society for Information Science*, 48(1), 40-54, 1997.

Lynch, C. (1991). The client-server model in information retrieval. In M. Dillon (Ed.), *Interfaces for information retrieval and online systems*. NY: Greenwood Press. pp 301-322.

Maes, P. (1994). Agents that reduce work and information overload. *Communications of the ACM*,

37(7), 31-40.

Marchionini, G., Information seeking in electronic environments, Cambridge University Press, NY, 1995.

Marchionini, G. & Mauer, H. The roles of digital libraries in teaching and learning, Comm ACM, 38(4), 67-75 (1995).

Marchionini, G., Nolet, V., Williams, H., Ding, W., Beale, J., Rose, A., Gordon, A., Enomoto, E., & Harbinson, L. (1997). Content+Connectivity => Community: Digital resources for a learning community. Proceedings of ACM Digital Libraries '97 (Philadelphia, PA, July 23-26, 1997). 212-220.

Oard, D., (1997). A Conceptual Framework for Text Filtering. UMUAI '97 (to appear). <http://www.glue.umd.edu/~oard/research.html>.

Plaisant, C., Marchionini, G., Bruns, T., Komlodi, A., & Campbell, L. (1997). Bringing treasures to the surface: Iterative design for the Library of Congress National Digital Library Program. Proceedings of ACM CHI '97 (Atlanta, March 22-27, 1997). NY: ACM Press, 518-525.

Resnick, P. (1997). Filtering information on the Internet. Scientific American, 276(3), 62-64.

Rosenfeld, A. & Smith, R.C. (1981). Thresholding using relaxation. IEEE Transactions on Pattern Analysis Machine Intelligence PAMI-3, 598-606.

Schatz, B., Mischo, W., Cole, T., Hardin, J., Bishop, A., & Chen, H. (1996). Federating diverse collections of scientific literature, Computer, May, 28-35.

Sheridan, P. & Ballerini, J. (1996). Experiments in multilingual information retrieval using the SPIDER system. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval, (Zurich, Switzerland, 58-65.

Shneiderman, B. (1994). Dynamic queries for visual information seeking. IEEE Software, 70-77.

Sim, D., & park, R. (1997). A two-stage algorithm for motion discontinuity-preserving optical flow estimation. *Computer Vision and Image Understanding*, 65(1), 19-37.

Smith, T. A digital library for geographically references materials, *Computer*, May, 54-60, 1996.

Stefik, M. (1997). Trusted systems. *Scientific American*, 276(3), 78-81.

Varian. H. (1997). Versioning information goods. *Digital Information and Intellectual Property*. (Harvard University Workshop, January 23-25, 1997).

Wactlar, H., Kanade, T., Smith, M., Stevens, S. (1996). Intelligent access to digital video: Informedia Project. *Computer*, May, 46-52.

Wiederhold, G. (1992). Mediation in the architecture of future information systems. *IEEE Computer*, March, 38-49.

## **Bibliography**

*Scientific American* (1997). Special report. The Internet: Fulfilling the promise. 276(3), March., pp. 49-83.

Dailianas, A., Allen, R.B., & England, P. (1995). Comparison of automatic video segmentation algorithms. *Proceedings of SPIE--Photonics East '95*, Philadelphia, Nov., 1995.

Elliott, E. (1993). Watch, grab, arrange, see: Thinking with motion images via streams and collages. MSVS Thesis Document. MIT Media Lab: Cambridge, MA.

England, P., Allen, R.A., Dailianas, A., Sullivan, M., Bianchi, M., & Heybey, A. (1996) The video library toolkit: A system for indexing and browsing digital video libraries. *Proceedings of SPIE Photonics West '96*, San Jose, Jan. 1996.

- Fox, E. & Lunin, L. Perspectives on digital libraries: Introduction and overview, *Journal of the American Society for Information Science*, 44(8), 441-445, 1993.
- Fox, E.A., Akscyn, R., Furuta, R., & Leggett, J. (1995). Digital libraries: Introduction. *Communications of the ACM*, 38(4), 22-28.
- Hearst, M. TextTiling: Segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 23(1), 33-64, 1997.
- Lesk, M. Practical digital libraries: Books, bytes, and bucks. Morgan Kaufmann.
- Otsuji, K., Tonomura, Y., and Ohba, Y. (1991). Video browsing using brightness data. *SPIE Visual communications and Image Processing 91: Image Processing*, 1606, 980-989.
- Salton, G. (1989). Automatic text processing: The transformation, analysis, and retrieval of information by computer. Reading, MA: Addison-Wesley.
- Salton, G., & McGill, M. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.
- Samuelson, P., & Glushko, R. J. (1991). Intellectual property rights for digital library and hypertext publishing systems: An analysis of Xanadu. *Proceedings of Hypertext '91* (San Antonio, December 15-18, 1991), pp. 39---50.
- Teodosio, L. & Bender,. W. (1993). Salient video stills: Content and context preserved. *Proceedings of ACM Multimedia 93* (Anaheim, CA, Aug. 1-6, 1993), NY: ACM Press, p. 39-46.

## BRUCE R. SCHATZ

**CANIS**  
**704 S. Sixth Street**  
**Champaign, IL 61820**

[schatz@canis.uiuc.edu](mailto:schatz@canis.uiuc.edu)  
<http://www.canis.uiuc.edu>

(217) 244-0651  
fax (217) 333-6869



### **PROJECTS** **MEDSPACE**

**The Interspace Prototype**

**Digital Libraries Initiative**

**CURRICULUM VITAE**  
**Curriculum Vitae**

**NSF Biographic Sketch**

**Bruce R. Schatz** is Director of the **COMMUNITY Architecture for Network Information SYSTEMS (CANIS) LABORATORY** at the **University of Illinois at Urbana-Champaign**, serving as Principal Investigator of the **Digital Libraries Initiative** project, a \$4M flagship effort in the Federal Program in National Information Infrastructure. This project is building a large-scale testbed of SGML documents from engineering and science journals.

He is also Principal Investigator of a \$3.5M flagship effort in the DARPA Information Management Program performing research in information systems to build analysis environments to support community repositories (**Interspace**), and in information science performing large-scale experiments in semantic retrieval for vocabulary switching.

He holds faculty appointments in **Library and Information Science**, **Computer Science**, Neuroscience, and Health Information Sciences. He is also a Senior Research Scientist at the **National Center for Supercomputing Applications** (NCSA), serving as the scientific advisor for digital libraries and information systems. He has served in this role since 1989, including the period during which NCSA developed Mosaic.

Schatz previously spent ten years in industrial R&D at Bellcore and Bell Labs, where he built prototypes of networked digital libraries which served as a foundation of current Internet services (**Telesophy**), and five years at the University of Arizona, where he was PI of the NSF National Collaboratory project which built a community system in molecular biology referenced as a national model for future science information systems (**Worm Community System**).

## **Community Architectures for Network Information Systems**

last updated: 06/16/99

contact [canis@uiuc.edu](mailto:canis@uiuc.edu) for comments and questions



# Faculty Information

[HOME](#)[MAPS](#)[QUESTIONS](#)[COMMENTS](#)[NEWS!](#)

## Terry Smith

Dept. of Geography

Ellison Hall 3611

Santa Barbara, CA 93106-4060

Office: Eng1-2155

Telephone: (805) 893-2966

FAX: (805) 893-3146



Email: [smithtr@geog.ucsb.edu](mailto:smithtr@geog.ucsb.edu)

| [Back](#) |

---

## Education:

Ph.D. The Johns Hopkins University Geography & Environmental Engineering

B.A. Cambridge University -Geography

## Research Interests:

spatial data processing, spatial analysis, spatial databases, and knowledge based approaches to GIS

## Courses Taught (since 9/95):

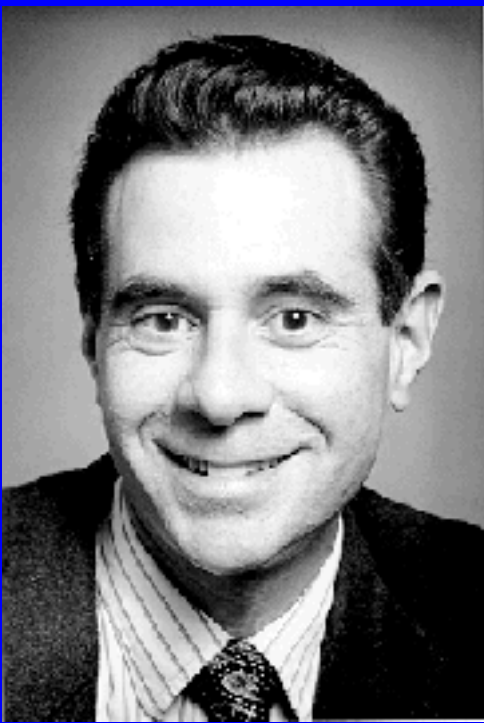
### Lower Division Courses:

- [3. Physical Geography](#)

## Publications:



**ALWAYS UNDER  
CONSTRUCTION**



# Robert Wilensky

Professor

Computer Science Division and School of Information Management and Systems

721 Soda Hall

University of California, Berkeley  
Berkeley, CA 94720

TEL: (510) 642-7034

FAX: (510) 643-1534

wilensky@CS.Berkeley.EDU

Administrative contact:

Winnie Wang

719 Soda Hall

(510) 642-9575

catwoman@cs.Berkeley.EDU

Office hours for Fall, 1999: 4-5PM Tuesdays and Thursday, 721 Soda Hall

News flash! Robust Hyperlinks Cost Just Five Words Each!

I'm working on "Digital Information Services", as exemplified by



The paper, Digital Library Resources as a Basis for Collaborative Work, describes some of our efforts. An IEEE Computer article on our previous work is available; so is an overview of our current project (in PowerPoint).

My research interests include:

- **Digital Documents**

See the [Multivalent Document](#) page. You can also check out that the [Digital Documents](#) course home page.

- **Robust Linking**

Robustness in a chaotic environment poses some interesting questions. Check out the [Robust Hyperlinks and Robust Locations](#) page.

- **[Using Natural Language Processing to Improve Information Access](#)**

Check out the home page for [CS288: An Artificial Intelligence Approach to Natural Language Processing](#).

- **Digital Information Infrastructure**

[A Framework for Distributed Digital Object Services](#) by Robert Kahn and Robert Wilensky, offers some elements.

Here are some of my other UC Berkeley Computer Science technical reports on various subjects:

- [Extending the Lexicon by Exploiting Subregularities](#)
- [Sentences, Situations and Propositions](#)
- [The Berkeley UNIX Consultant Project](#)
- [Primal Content and Actual Content: An Antidote to Literal Meaning](#)
- [Some Problems and Proposals for Knowledge Representation](#)
- [UC--A Progress Report](#)
- [Talking to Unix in English: An Overview of an On-line Consultant](#)

You can examine any UC Berkeley Computer Science Technical Report, and those of quite a few other institutions, via the [UC Berkeley Technical Report Server](#).

BAIR (Berkeley Artificial Intelligence Project) members include

[Isaac Cheng](#), [Michael X. Schiff](#), [Narciso Jaramillo](#) You can also check our some [alumni](#) pages, and some people working on the [Digital Library Project](#) .

In addition to having recently served as Chair of UC Berkeley's Computer Science Division, as well as a brief stint as [Emperor of China](#), I also [consult](#) on various topics.

UCB DLIB [overview in PowerPoint](#) and in [Acrobat](#)

[MVD talk in PowerPoint](#)

[InterLib talk in PowerPoint](#)

[Experimental Blobworld home page](#).

Some [family pictures](#)....

[RESIDU](#)

[research overview](#)

[Some Demos](#)

["See the world today..."](#)

[Directions to 560 Spruce Street](#)

[Robust locations examples](#)

# Countries & Regions:

---

(Chapter 11, page 245, "Books, Bytes and Bucks", Michael Lesk)

- **United States of America:** In the US, NSF, NASA and ARPA have funded six important Digital Library efforts, called the DLI (Digital Libraries Initiative). These programs each involve a large consortium of cooperating institutions but the six main ones are : University of California at Berkeley, University of Santa Barbara, University of Michigan, Carnegie Mellon University, Stanford University, and the University of Illinois.
  - University of California at Berkeley: Image content queries along with Xerox PARC, database extraction from documents, multivalent documents, NLP. Headed by Robert Wilensky.
  - University of Michigan: Scalability and Education. They are also investigating the use of agent architectures for Digital Libraries and trying to merge DLI with their other digital library efforts such as JSTOR and TULIP. Headed by Dan Atkins.
  - University of Illinois: Concentrating on using scientific journals as their base collection with diversity in both documents as well as publishers, making the transition process from SGML to HTML smoother, defining semantic spaces. Headed by Bruce Schatz.
  - Stanford University: concentration is on the infrastructure development such as basic networking and databases to support digital libraries. Also concerned with interoperability between different digital library projects. Headed by Hector Garcia-Molina.
  - University of California at Santa Barbara: spatial indexing and retrieval , image processing. Headed by Terry Smith.
  - Carnegie Mellon University: digital video, image analysis, speech recognition, face recognition, natural language understanding. Headed by Michael Mauldin and Marvin Sirbu.

Other than DLI, many research projects are underway at some other universities such as Virginia Tech and Texas A&M. In the near future, extensive funds are expected to be allocated for Digital Libraries.

The Library of Congress, under James Billington is digitizing 5 million of its items in a massive \$60 million effort. Other universities involved in related projects are Georgia Tech, Cornell, MIT, University of Tennessee, Washington and California and Virginia Tech (known for the Envision system of Ed Fox). Other limited efforts include University of Virginia, University of Georgia and Columbia University.

- **United Kingdom:** Though efforts are still limited to penny-pockets, 20 million pounds have been set aside for digital library projects. The program originally called FIGIT, now known as E-LIB funded 35 projects. Work includes cataloging of archives, digitization of documents and data sharing. Some of the more notable efforts are : Digitizing the Burney collection of pre-1800 newspapers and scanning of Batley News, the Canterbury Tales project that involves scanning all pre-1500 manuscripts and some other similar projects. However, the most notable is the Electronic

Beowulf project which is a US/UK collaboration between Kevin Kiernan (University of Kentucky), Paul Szarmach (Western Michigan University) and the British Library.

- **France:** Work includes some scanning of old manuscripts with the most notable being the Tresor de la Langue Francaise project at the University of Nancy. The French, along with the Japanese are also leaders in the Group 7 project which is a museum project. Other efforts are INIST and FOUDRE (1989 to 1992) followed by EDIL and ELITE.
- **The EU:** The European Union funds a large number of international efforts in digital libraries. (Please see page 255 of Michal Lesk's book for details)
- **Japan:** Japan is involved in some digitization and cataloguing efforts and has a \$50M project on. They are also working on modern document delivery and OCR.
- **Australia:** Australia has recently made a modest effort to enter into digital library research. They are planning some digitization projects with a \$10M (Australian) digitization project on the anvil. They are also interested in digitizing Aborigine scriptures and paintings.
- **Elsewhere:** Many other countries are involved in digital library research on much smaller scales. Notable amongst them are Canada, Singapore, Korea and China.

**NOTE 1:** For detailed information on any of the above please refer to Dr. Lesk's book (recommended as supplement text for this course).

**NOTE 2:** See also the table pointing to various national digital libraries from April 1998 CACM [online pages](#)

---

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

---

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

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

# Centers, sites and organisations:

---

**Some major Digital Library centers and research programs, separately described:**

- [Carnegie Mellon University](#)
  - [CNRI](#)
  - [Library of Congress](#)
  - [Stanford University](#)
  - [University of California at Berkeley](#)
  - [University of California at Santa Barbara](#)
  - [University of Illinois](#)
  - [University of Michigan](#)
  - [Texas A&M](#)
  - [Virginia Tech](#)
- 

## Selected other sites:

**[ACM DL](#)** : Tap into the ACM Digital Library, a vast resource of bibliographic information, citations, and full-text articles.

**IEEE-CS** [Digital Library](#)

**IBM**

- [IBM DL Home page](#)
- [IBM Renaissance Consortium Panel](#) and [workshop](#)
- [images - QBIC](#)

**[National Library of Medicine](#)**

**[Digital Library Research Program](#) at**

**[Lister Hill National Center for Biomedical Communications,](#)**

**[National Institutes of Health](#)**

**[OCLC](#)** (OCLC is a nonprofit, membership, library computer service and research organization dedicated to the public purposes of furthering access to the world's information and reducing information costs).

- Research <http://www.oclc.org/oclc/research/index.htm>  
SiteSearch <http://www.oclc.org/oclc/menu/site.htm>

**Xerox**

- [DL Interfaces Home Page](#)

- [Scientific American article](#)
- [Scatter/Gather examples](#)
- **Questions:**
  - **Compare**
    - **What are the various interfaces built? How do they compare? What is the best use of each?**
  - **Scatter/gather**
    - **Explain clustering, relate it to scatter/gather.**
    - **What are special problems with large category systems and how can they be solved?**

---

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

---

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

(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta

# CNRI:

---

- home page (site map) [http://www.cnri.reston.va.us/site\\_map.html](http://www.cnri.reston.va.us/site_map.html)
  - Architecture
    - Kahn-Wilensky Framework for Distributed Digital Object Services  
<http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html>
    - key architectural issues <http://WWW.CNRI.Reston.VA.US/home/cstr/arch/slides.html>
    - architecture for information in digital libraries  
<http://www.dlib.org/dlib/february97/cnri/02arms1.html>
    - Digital Object Architecture Project <http://www.cnri.reston.va.us/doa.html>
  - Handle System (<http://www.handle.net/>) and Digital Object Identifier System (<http://www.doi.org/>)
  - CS-TR Computer Science Technical Reports <http://www.cnri.reston.va.us/cstr.html>
- 

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

---

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

**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**

# Library of Congress:

---

- American Memory <http://lcweb2.loc.gov/>
  - Call/Awards about American Memory <http://lcweb2.loc.gov/ammem/award/>
  - Sponsors and Contributors to the National Digital Library Program  
<http://lcweb2.loc.gov/ammem/sponsors.html>
- 

[[Main](#)] [[Contents](#)] [[Resources](#)] [[Centers](#)]

---

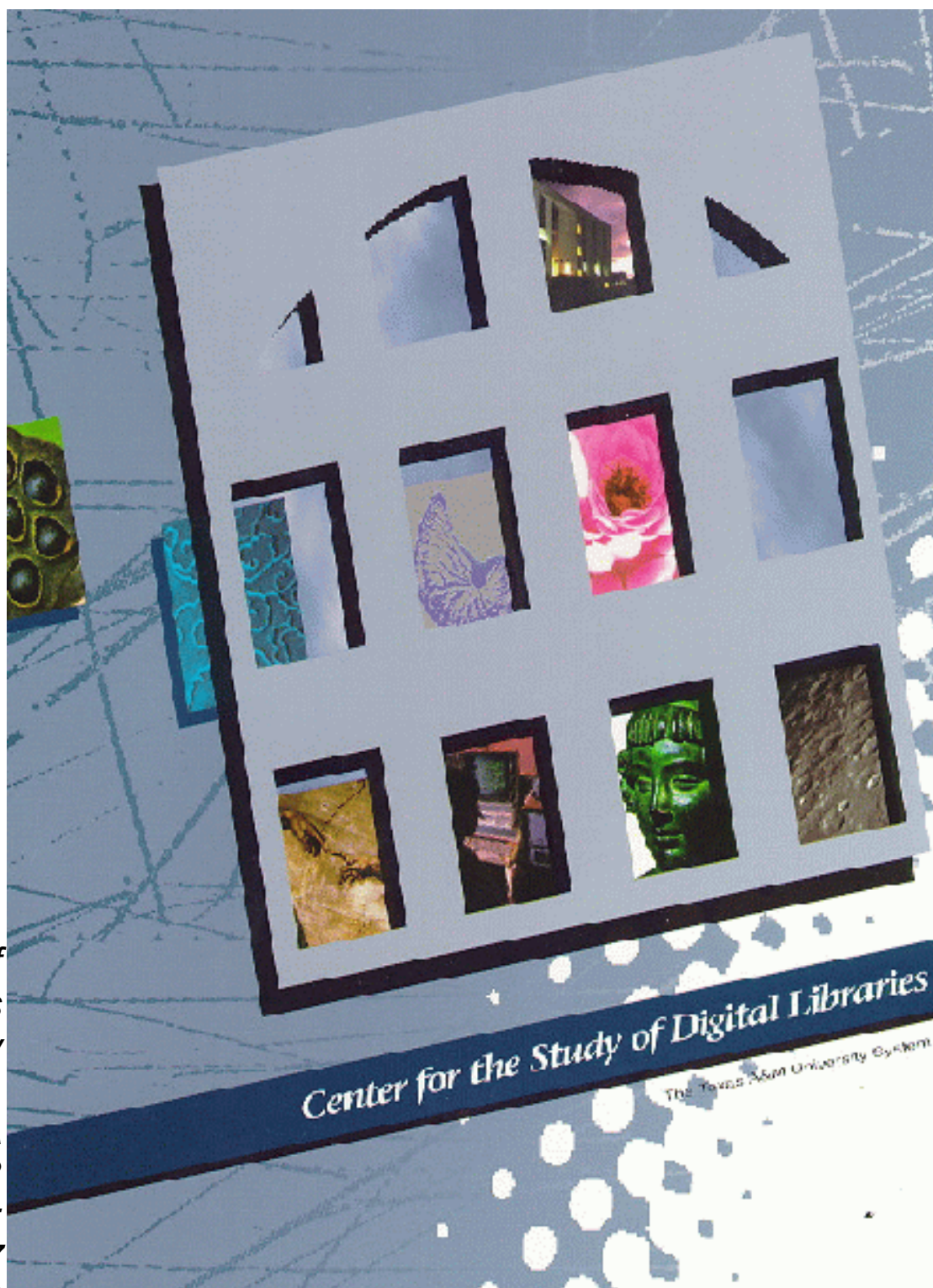
Please send comments/suggestions to [Ed Fox](#).

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



- [THE CENTER](#)
- [FACILITIES](#)
- [RESEARCH](#)
- [PEOPLE](#)
- [COURSES](#)
- [PUBLICATIONS](#)
- [CONFERENCES](#)

Center for the Study of  
Digital Libraries  
Texas A&M University  
College Station,  
Texas, USA  
77843-3112  
Telephone:  
01-409-862-3217  
Fax: 01-409-847-8578  
[csdl@csdl.tamu.edu](mailto:csdl@csdl.tamu.edu)



HEWLETT  
PACKARD



The Center for the Study of Digital Libraries gratefully acknowledges the corporate support of the

**Hewlett-Packard Company; Informix Software, Inc.; and Knowledge Systems, Inc.**



A joint venture of:

[Information Systems](#)  
[Department of Computer Science](#)  
[Internet Technology Innovation Center](#)

And don't forget our sister organizations:

[Scholarly Communications Project](#)  
[Virginia Tech Digital Libraries Project](#)  
[Multimedia and Distance Learning Lab](#)

[Members](#)

[Philosophy](#)

[Mission](#)

[Products:](#)

[MARIAN](#)   [NDLTD](#)   [VT-ETD](#)   [Envision](#)

[Research Initiatives:](#)

[5S Model](#)   [DL Logging](#)   [PetaPlex Archive](#)  
[Java MARIAN](#)   [TREC-8](#)   [Virtual Realities](#)  
[DL Taxonomy](#)   [Open Archives Initiative](#)   ...

[Resources](#)

[Publications](#)

[Reports](#)

---

**Location:** [840 Pointe West Commons, Suite 8](#) Blacksburg, VA 24061-0368 USA   Webmother: [anansi@dtheses.org](mailto:anansi@dtheses.org)

[home](#)[feedback](#)[join/renew](#)[go shopping](#)[search acm](#)

# ACM Digital Library

*ACM brings you the world of computing*

**Tap into the ACM Digital Library, a vast resource of bibliographic information, citations, and full-text articles.**

## Browse and Search the Digital Library

- ◆ Browse the library:
  - [ACM journals and magazines](#)
  - [ACM proceedings by subject](#)
  - [ACM proceedings by sponsor](#)
  - [ACM proceedings by series](#)
  - [journals and magazines by affiliated publishers](#)
  - [resources from affiliated organizations](#)

- ◆ [Search](#) the Digital Library
- ◆ [My Bookshelf](#)

## About the Digital Library

- ◆ [Content and Organization](#)
- ◆ [Terms of Usage](#)
- ◆ [How To...](#)
- ◆ [Frequently Asked Questions](#)
- ◆ [Known Problems](#)
- ◆ [System Availability](#)
- ◆ [Feedback](#)

## What's New at the Digital Library

- ◆ [Announcements](#)
- ◆ [Latest Conference Proceedings](#)

## Subscription and Access Information

If you are not yet a subscriber, you can still use the Digital Library: As a service to the computing community, the Digital Library will continue to offer its search and bibliographic database resources to all visitors, for free. All you need to do is register with us.

Access to full-text is by pay-per-view or subscription only: ACM members who are Digital Library subscribers have access to all full-text articles. Members and nonmembers who subscribe to electronic publications (but not to the entire Library) have full-text access to their subscriptions only.

- ◆ [Register](#)
- ◆ [Subscribe to the Digital Library](#)
- ◆ [Subscription Information for Institutions](#)
- ◆ [ACM Document Delivery Service](#)

---

To read full-text PDF articles, use [Adobe Acrobat Reader](#).

---

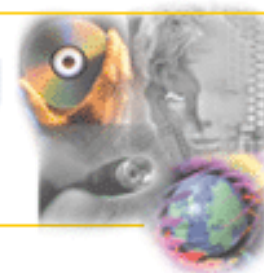
The Digital Library is published by the Association for Computing Machinery. Copyright © 1999 ACM, Inc.

[library home](#)[list alphabetically](#)[list by SIG](#)[search library](#)[register DL](#)[subscribe DL](#)[feedback](#)

[Join](#)[Publications Center](#) ▶[Channels](#) ▶[Conference Wire](#)[Standards](#)[Career Services Center](#)[Education & Certification](#)[History of Computing](#)[Awards](#)[About the Computer Society](#)[Get Involved](#)[Member Benefits & Services](#)[Volunteer Resources](#)Institute of Electrical &  
Electronics Engineers

Computer.org

## Digital Library



- ▶ [About The Computer Society Digital Library](#)
- ▶ [Subscribe to the Digital Library](#) for only \$50.

### ENTER THE DIGITAL LIBRARY

- ☒ [Search The Digital Library](#)
- ☒ [Magazines](#)
- ☒ [Transactions](#)
- ☒ [Conference Proceedings](#)

**NOTE:** You will be asked for your **Computer Society Web Account Login** when selecting an article or paper for the first time.

## Magazines

- ▶ [Computer](#)
- ▶ [Annals of the History of Computing](#)
- ▶ [Computing in Science & Engineering](#)
- ▶ [Computer Graphics and Applications](#)
- ▶ [Concurrency](#)
- ▶ [Design & Test of Computers](#)
- ▶ [Intelligent Systems](#)
- ▶ [Internet Computing](#)

▶ [IT Professional](#)

▶ [Micro](#)

▶ [MultiMedia](#)

▶ [Software](#)

---

## Transactions

▶ [Computers](#)

▶ [Knowledge & Data Engineering](#)

▶ [Parallel & Distributed Systems](#)

▶ [Pattern Analysis & Machine Intelligence](#)

▶ [Software Engineering](#)

▶ [Visualization & Computer Graphics](#)

---

## Conference Proceedings

▶ [Growing body of Conference Proceedings](#)

---

Send general comments and questions about the IEEE Computer Society's Web site to [webmaster@computer.org](mailto:webmaster@computer.org).

This site and all contents (unless otherwise noted) are [Copyright](#) © 2000, Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

computer.org Navigator

- ▶ [Information Technology](#)
- ▶ [Design & Test](#)
- ▶ [Internet](#)
- ▶ [Software Engineering](#)
- ▶ [Computer Graphics & Visualization](#)
- ▶ [Technical Councils, Committees, task forces](#)
- ▶ [Chapters](#)
- ▶ [Students](#)

- ▶ [Digital Library](#)
- ▶ [CS Store](#)
- ▶ [Computer magazine](#)
- ▶ [IT Professional](#)
- ▶ [Internet Computing](#)
- ▶ [Software magazine](#)
- ▶ [Magazines](#)
- ▶ [Transactions](#)
- ▶ [Conference proceedings](#)
- ▶ [Subscription info.](#)


[ShopIBM](#)
[+ Support](#)
[↓ Downloads](#)
[Home](#)
[Products](#)
[Consulting](#)
[Industries](#)
[News](#)
[About IBM](#)

Search

[Products](#) > [Software](#) > [Database and Data Management](#)

## DB2 Digital Library

### DB2 Digital Library

[How to buy](#)
[Support](#)
[More information](#)
[News](#)
[Case studies](#)
[Library](#)
[Services](#)
[Events](#)
[Education](#)
[IBM Business Partners](#)
[→ IBM Worldwide](#)

Tomorrow's digital asset management system is here today and, you can be part of it. Whether it's video, audio, images, or text, IBM DB2 Digital Library transforms multimedia assets into digital form which can be distributed over public or private networks.

### Features at a glance

Whether it's video, audio, images, or text, IBM DB2 Digital Library transforms multimedia assets into digital form which can be distributed over public or private networks -- like the Internet and your corporate intranets -- to users around the world.

There are [real implementations](#) of IBM DB2 Digital Library that serve the needs of archivists, film/video production groups, educators and researchers medical technologists, advertising and creative agencies, multimedia, print and Web publishers and marketing communications departments. These applications allow you to manage your analog and digital media assets centrally. Through these efforts such benefits can be brought to you...*fast*.

We invite you to take a look at [IBM DB2 Digital Library](#): the product, the architecture and industry solutions. You'll see why IBM DB2 Digital Library is revolutionizing the way you'll do business with your multimedia assets.

IBM DB2 Digital Library is available for the AIX and Windows NT operating systems. Client support includes Windows 95 or 98, Windows NT, AIX, and Macintosh.

[Buy now](#)


### Buy other versions

[→ DB2 Digital Library V2.4](#)

### Spotlight


[Announcing Content Manager](#)

### News

[Operating systems](#)

- ⇒ [IBM Content Manager announced](#)
- ⇒ [IBM DB2 Digital Library VideoCharger now supports QuickTime 4](#)
- ⇒ [IBM and The Hermitage Museum Project wins Imaging Solution of the Year award](#)
- ⇒ [Results of the State Hermitage Museum technology partnership with IBM](#)

DB2 Digital Library runs on **AIX, Mac OS, Windows 95 & Windows 98 and Windows NT.**

#### More resources

- ⇒ [IBM DB2 Digital Library Version 2.4 Brochure](#)
- ⇒ [IBM DB2 Digital Library Version 2.4 Fact Sheet](#)
- ⇒ [IBM DB2 Digital Library VideoCharger Version 2.0](#)
- ⇒ [IBM DB2 Digital Library Connection for Avid](#)
- ⇒ [ISLIP MediaKey Digital Library Video System](#)
- ⇒ [JCollaborate Distributed Education](#)
- ⇒ [IBM Cryptolope](#)
- ⇒ [Content Management Solutions](#)
- ⇒ [DB2 DL Competency Center - Gaithersburg](#)
- ⇒ [DB2 DL Competency Center - Asia Pacific](#)

[Privacy](#)

[Legal](#)

[Contact](#)



# Digital Library Research Program

[National Library of Medicine](#) / [National Institutes of Health](#)

---

The digital library research program at the [Lister Hill National Center for Biomedical Communications](#) investigates all aspects of creating and disseminating digital collections including proposed and adopted standards, emerging technologies and formats, effects on previously established processes, and protection of original materials.

Our early experiments in document management and conversion resulted in a digital library system of historical materials from the 1960's and 1970's. The [Regional Medical Programs collection](#) consists of approximately 40,000 pages comprising some 1,500 documents. Though the work on this system predated recent research in digital libraries, we addressed many of the same issues that currently face digital library projects.

Working in close collaboration with NLM's [History of Medicine Division](#), we launched the [Profiles in Science](#) site. The site focuses on the major scientific achievements of the twentieth century. The project is making the archival collections of prominent biomedical scientists available to the public through modern digital technology. The collections have been donated to the NLM and contain published and unpublished materials, including books, journal volumes, pamphlets, diaries, letters, manuscripts, photographs, audio tapes and other audiovisual materials.

## Recent publication

McCray, Alexa T., Marie E. Gallagher, Michael A. Flannick. [Extending the Role of Metadata in a Digital Library System](#). In: Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries, pp. 190-199, 1999.

## [Digital Library Resources](#)

[Digital Library Initiative - Phase 2](#) - Through its Extramural Programs Division, NLM co-sponsors the multi-agency Digital Library Initiative - Phase 2.

---

<http://www.lhncbc.nlm.nih.gov/dlb/>

*Last updated: Friday, 21-Apr-2000 17:38:24 EDT*



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

News

About OCLC

OCLC Services

Support &amp; User Doc.

Contacts &amp; Addresses

[▶ About](#)[-- What's New?](#)[▶ Programs](#)[▶ Projects](#)[▶ Publications](#)[▶ Archives](#)

**The mission of the OCLC Office of Research** is to expand knowledge that advances the goal of OCLC's commitment to improving access to the world's information resources, whatever their form, substance, subject, language, or location.

This mission is pursued through the integrated employment of the computer, library, and information sciences in research activities such as performing experiments, building prototypes, advancing standards, undertaking studies, and participating in research collaborations.

**Shaping the Future of Librarianship:** The OCLC Office of Research is one of the world's leading centers devoted exclusively to the challenges facing libraries in a rapidly changing information technology environment. Since its origin in 1978, the Office has investigated trends in technology and library practice to identify technical advances that will enhance the value of library services and improve the productivity of librarians and library users. Among the areas of study are natural language processing, information retrieval, Internet metadata standards, knowledge management, interface design, and classification theory and practice.



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

# Interfaces for Information Access

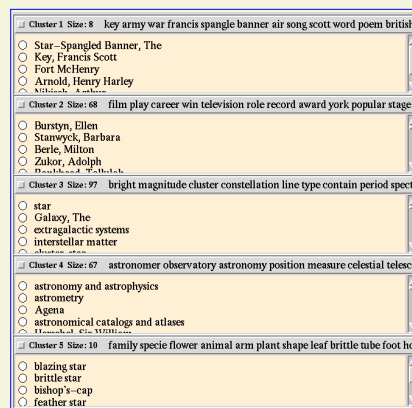
## Companion Pages for *Scientific American* Article [\*Interfaces for Searching the Web\*](#)

The field of Information Access concerns helping people find, use, understand, and create the information they need, often using computer systems as tools. Information can be found in many forms and media, although much of our research has been concerned with text in general, not focusing exclusively on the Web.

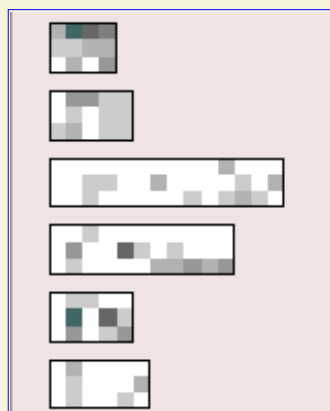
Text analysis and user interface technology must be combined with an understanding of how users work with information and computer tools when building systems to support information access.

Currently, these pages provide additional information about some of the ideas discussed in the *Scientific American* article *Interfaces for Searching the Web* by [Marti Hearst](#). There is a great deal of research in Information Access at [Xerox PARC](#), of which this pages show only a small sample.

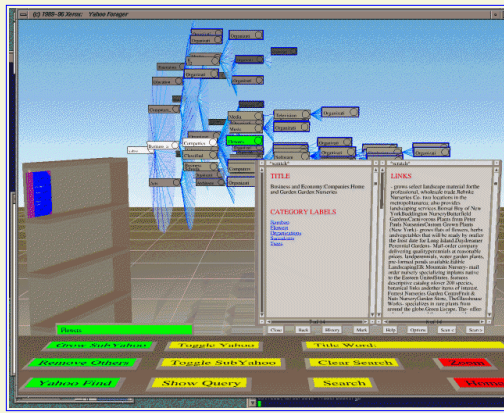
### [About Scatter/Gather](#)



### [About Tilebars](#)



### [About the Cat-a-Cone](#)



Xerox PARC  
hearst@parc.xerox.com  
6/9/97

# SCIENTIFIC AMERICAN

[Main Menu](#)[Interview](#)[Bookmarks](#)[Feedback](#)[Current Issue](#)[Explore!](#)[Ask the Experts](#)[Marketplace](#)[Search the Site](#)

## FEATURE ARTICLES



### SPECIAL REPORT

# Interfaces for Searching the Web

**The rapid growth of the World Wide Web  
is outpacing current attempts to search and organize it.  
New user interfaces may offer a better approach**

*by [Marti A. Hearst](#)*

#### SUBTOPICS:

[The \(Slow\) Speed of](#)



[Thought](#)

[Organizing Search](#)



[Results](#)

#### [FURTHER READING](#)

[BACK TO THE  
INTRODUCTION](#)

How does anyone find anything among the millions of pages linked together in unpredictable tangles on the World Wide Web? Retrieving certain kinds of popular and crisply defined information, such as telephone numbers and stock prices, is not hard; many Web sites offer these services. What makes the Internet so exciting is its potential to transcend geography to bring information on myriad topics directly to the desktop. Yet without any consistent organization, cyberspace is growing increasingly muddled. Using the tools now available for searching the Web to locate the document in Oregon, the catalogue in Britain or the image in Japan that is most relevant for your purposes can be slow and frustrating.

More sophisticated algorithms for ranking the relevance of search results may help, but the answer is more likely to arrive in the form of new user interfaces. Today software designed to analyze text and to manipulate large hierarchies of data can provide better ways to look at the contents of the Internet or other large text collections. True, the page metaphor used by most Web sites is familiar and simple. From the perspective of user interface design, however, the

page is unnecessarily restrictive. In the future, it will be superseded by more powerful alternatives that allow users to see information on the Web from several perspectives simultaneously.

Consider Aunt Alice in Arizona, who connects to the Net to find out what kind of edible bulbs, such as garlic or onions, she can plant in her garden this autumn. Somewhere in the vast panorama of the Web lie answers to her question. But how to find them?

Alice currently has several options, none of them particularly helpful. She can ask friends for recommended Web sites. Or she can turn to Web indexes, of which there are at present two kinds: manually constructed tables of contents that list Web sites by category and search engines that can rapidly scan an index of Web pages for certain key words.

Using dozens of employees who assign category labels to hundreds of Web sites a day, Yahoo compiles the best-known table of contents. To use Yahoo, one chooses from a menu [see illustration at far left] the category that seems most promising, then views either a more specialized submenu or a list of sites that Yahoo technicians thought belonged in that section. The interface can be awkward, however. The categories are not always mutually exclusive: Should Alice choose "Recreation," "Regional" or "Environment"? Whatever she selects, the previous menu will vanish from view, forcing her either to make a mental note of all the alternative paths she could have taken or to retrace her steps methodically and reread each menu. If Alice guesses wrong about which subcategory is most relevant (it is not "Environment"), she has to back up and try again. If the desired information is deep in the hierarchy, or is not available at all, this process can be time-consuming and aggravating.

### **The (Slow) Speed of Thought**

Research in the field of information visualization during the past decade has produced several useful techniques for transforming abstract data sets, such as Yahoo's categorized list, into displays that can be explored more intuitively. One strategy is to shift the user's mental load from slower, thought-intensive processes such as reading to faster, perceptual processes such as pattern recognition. It is

easier, for example, to compare bars in a graph than numbers in a list. Color is very useful for helping people quickly select one particular word or object from a sea of others.

Another strategy is to exploit the illusion of depth that is possible on a computer screen if one departs from the page model. When three-dimensional displays are animated, the perceptual clues offered by perspective, occlusion and shadows can help clarify relations among large groups of objects that would simply clutter a flat page. Items of greater interest can be moved to the foreground, pushing less interesting objects toward the rear or the periphery. In this way, the display can help the user preserve a sense of context.

Such awareness of one's virtual surroundings can make information access a more exploratory process. Users may find partial results that they would like to reuse later, hit on better ways to express their queries, go down paths they didn't think relevant at first--perhaps even think about their topic from a whole new perspective. Aunt Alice could accomplish a lot of this by jotting down notes as she pokes around Yahoo, but a prototype interface developed by my colleagues at the Xerox Palo Alto Research Center aims to make such sense-making activities more efficient.

Called the [Information Visualizer](#), the software draws an animated 3-D tree that links each category with all its subcategories. If Alice searches the Yahoo tree for "garden," all six areas of Yahoo in which "garden" or "gardening" is a subcategory will light up. She can then "spin" each of these categories to the front to explore where it leads. When one path hits a dead end, the roads not taken are just a click away.

When Alice finds useful documents, this interface allows her to store them, along with the search terms that took her to them, in a virtual book. She can place the book on a virtual bookshelf where it is readily visible and clearly labeled. Next weekend, Alice can pick up where she left off by reopening her book, tearing out a page and using it to resubmit her query.

Our interface does not offer much help to the Sisyphean attempt to organize the contents of the entire Web. Because new sites appear on the Web far faster than they can be

indexed by hand, the fraction listed by Yahoo (or any other service) is shrinking rapidly. And sites, such as Time magazine's, that contain articles on many topics often appear under only a few of the many relevant categories.

Search engines such as Excite and [AltaVista](#) are considerably more comprehensive--but this is their downfall. Poor Aunt Alice, entering the string of key words "garlic onion autumn fall garden grow" into Excite will, as of this writing, retrieve 583,430 Web pages, which (at two minutes per page) would take more than two years to browse through nonstop. Long lists littered with unwanted, irrelevant material are an unavoidable result of any search that strives to retrieve all relevant documents; conversely, a more discriminating search will almost certainly exclude many useful pages.

The short, necessarily vague queries that most Internet search services encourage with their cramped entry forms exacerbate this problem. One way to help users describe what they want more precisely is to let them use logical operators such as AND, OR and NOT to specify which words must (or must not) be present in retrieved pages. But many users find such Boolean notation intimidating, confusing or simply unhelpful. And even experts' queries are only as good as the terms they choose.

When thousands of documents match a query, giving more weight to those containing more search terms or uncommon key words (which tend to be more important) still does not guarantee that the most relevant pages will appear near the top of the list. Consequently, the user of a search engine often has no choice but to sift through the retrieved entries one by one.

### **Organizing Search Results**

A better solution is to design user interfaces that impose some order on the vast pools of information generated by Web searches. Algorithms exist that can automatically group pages into certain categories, as Yahoo technicians do. But that approach does not address the fact that most texts cannot be shoehorned into just one category. Real objects can often be assigned a single place in a taxonomy (an onion is a kind of vegetable), but it is a rare Web page indeed that is only about onions. Instead a typical text might discuss produce distributors, or soup recipes, or a

debate over planting imported versus indigenous vegetables. The tendency in building hierarchies is to create ever more specific categories to handle such cases ("onion distributors," for example, or "soup recipes with onion," or "agricultural debates about onions," and so on). A more manageable solution is to describe documents by whole sets of categories that apply to them, along with another set of attributes (such as source, date, genre and author). Researchers in Stanford University's digital library project are developing an interface called [SenseMaker](#) along these lines.

At [Xerox PARC](#), we have developed an alternative scheme for grouping the list of pages retrieved by a search engine. Called [Scatter/Gather](#), the technique creates a table of contents that changes along with a user's growing understanding of what kind of documents are available and which are most relevant.

Imagine that Aunt Alice runs her search using Excite and retrieves the first 500 Web pages it suggests. The Scatter/Gather system can then analyze those pages and divide them into groups based on their similarity to one another [see upper illustration on next page]. Alice can rapidly scan each cluster and select those groups that appear interesting.

Although evaluation of user behavior is an inexact process that is difficult to evaluate, preliminary experiments suggest that clustering often helps users zero in on documents of interest. Once Alice has decided, for example, that she is particularly keen on the cluster of 293 texts summarized by "bulb," "soil" and "gardener," she can run them through Scatter/Gather once again, rescattering them into a new set of more specific clusters. Within several iterations, she can whittle 500 mostly irrelevant pages down to a few dozen useful ones.

By itself, document grouping does not solve another common problem with Web-based search engines such as Excite: the mystery of why they list the documents they do. But if the entry form encourages users to break up their query into several groups of related key words, then a graphical interface can indicate which search topics occurred where in the retrieved documents. If hits on all topics overlap within a single passage, the document is more likely to be relevant, so the program ranks it higher.

Alice might have a hard time spelling out in advance which topics must occur in the document or how close together they must lie. But she is likely to recognize what she wants when she sees it and to be able to fine-tune her query in response. More important, the technique, which I call [TileBars](#), can help users decide which documents to view and can speed them directly to the most relevant passages.

The potential for innovative user interfaces and text analysis techniques has only begun to be tapped. Other techniques that combine statistical methods with rules of thumb can automatically summarize documents and place them within an existing category system. They can suggest synonyms for query words and answer simple questions. None of these advanced capabilities has yet been integrated into Web search engines, but they will be. In the future, user interfaces may well evolve even beyond two- and three-dimensional displays, drawing on such other senses as hearing to help Aunt Alices everywhere find their bearings and explore new vistas on the information frontier.

---

### Further Reading

*Rich Interaction in the Digital Library*. Ramana Rao, Jan O. Pedersen, Marti A. Hearst and Jock D. Mackinlay *et al.* in *Communications of the ACM*, Vol. 38, No. 4, pages 29-39; April 1995.

*The WebBook and the Web Forager: An Information Workspace for the World-Wide Web*. Stuart K. Card, George G. Robertson and William York in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, April 1996. Available on the [World Wide Web](#)

### [Selected publications by Marti Hearst](#)

["The WebBook and the Web Forager: An Information Workspace for the World-Wide Web."](#) Stuart K. Card, George G. Robertson and William York in *Proceedings of the ACM/SIGCHI Conference on Human Factors in Computing Systems*, Vancouver, April 1996.

["Reexamining the Cluster Hypothesis: Scatter/Gather on Retrieval Results."](#) Marti A. Hearst and Jan O. Pedersen in *Proceedings of the 19th Annual International ACM/SIGIR Conference*, Zurich, August 1996.

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

Michelle Q. Wang Baldonado and Terry Winograd in  
*Proceedings of the ACM/SIGCHI Conference on Human  
Factors in Computing Systems*, Atlanta, 1997 (in press).

[Research in Support of Digital Libraries at Xerox PARC](#)

---

**The Author**

[MARTI A. HEARST](#) has been a member of the research staff at the Xerox Palo Alto Research Center since 1994. She received her B.A., M.S. and Ph.D. degrees in computer science from the University of California, Berkeley. Hearst's Ph.D. dissertation, which she completed in 1994, examined context and structure in text documents and graphical interfaces for information access.

---

# A Scatter/Gather Example

Here we demonstrate the use of Scatter/Gather on a collection of encyclopedia articles. Our query is very simple:

**Retrieve the top 250 documents that contain the word *star* .**

Here we show that Scatter/Gather text clustering does a reasonably good job at organizing the documents into meaningful themes or topics.

We ask Scatter/Gather to place the 250 documents into 5 groups. Here is what results. (Bear in mind that encyclopedia articles are well-written and uniform format. The [next example](#) shows the results of a more complicated query on a more unruly text collection.)

<input type="checkbox"/> Cluster 1 Size: 8	key army war francis spangle banner air song scott word poem british
<input type="radio"/> Star-Spangled Banner, The <input type="radio"/> Key, Francis Scott <input type="radio"/> Fort McHenry <input type="radio"/> Arnold, Henry Harley <input type="radio"/> ...	
<input type="checkbox"/> Cluster 2 Size: 68	film play career win television role record award york popular stage p
<input type="radio"/> Burstyn, Ellen <input type="radio"/> Stanwyck, Barbara <input type="radio"/> Berle, Milton <input type="radio"/> Zukor, Adolph <input type="radio"/> ...	
<input type="checkbox"/> Cluster 3 Size: 97	bright magnitude cluster constellation line type contain period spectr
<input type="radio"/> star <input type="radio"/> Galaxy, The <input type="radio"/> extragalactic systems <input type="radio"/> interstellar matter <input type="radio"/> ...	
<input type="checkbox"/> Cluster 4 Size: 67	astronomer observatory astronomy position measure celestial telescop
<input type="radio"/> astronomy and astrophysics <input type="radio"/> astrometry <input type="radio"/> Agena <input type="radio"/> astronomical catalogs and atlases <input type="radio"/> ...	
<input type="checkbox"/> Cluster 5 Size: 10	family specie flower animal arm plant shape leaf brittle tube foot hor
<input type="radio"/> blazing star <input type="radio"/> brittle star <input type="radio"/> bishop's-cap	

☐ feather star

Shown here are the clusters' sizes (how many documents they contain), a list of topical terms, and a list of document titles. One can see from the topical terms of Cluster 1 that this cluster contains documents that involve stars as symbols, as in military rank and patriotic songs.

Cluster 2 has 68 documents that appear mainly to be about movie and tv stars.

Cluster 3 contains 97 documents that having to do with aspects of astrophysics.

Cluster 4 contains 67 documents also about astronomy and astrophysics. This cluster contains many articles about people who are astronomers (this is apparent when the list is scrolled down).

Cluster 5 contains all the articles that discuss animals or plants, and that happen to contain the word star, for example, star fish.

If we ask Scatter/Gather to re-cluster the 68 documents that appear in Cluster 2, the one that discusses movie and tv stars, and place the results into three clusters, we see the following clusters:

☐ Cluster 1 Size: 14    player league hit game national set bat average season history basebal

- ☐ Musial, Stan
- ☐ Bench, Johnny
- ☐ Carew, Rod
- ☐ Robertson, Oscar
- ☐ Beliveau, Jean
- ☐ Casper, Billy
- ☐ Chinese checkers
- ☐ Best, George
- ☐ Beamon, Bob

☐ Cluster 2 Size: 47    role stage broadway comedy performance actress production musical

- ☐ Burstyn, Ellen
- ☐ Stanwyck, Barbara
- ☐ Berle, Milton
- ☐ Bankhead, Tallulah
- ☐ Murphy, Eddie
- ☐ Walsh, Raoul
- ☐ Martin, Mary
- ☐ Zukor, Adolph
- ☐ Cosby, Bill

☐ Cluster 3 Size: 7    music country jazz folk pop paul cowboy leader williams hampton boy

- ☐ Williams, Hank
- ☐ Crosby, Bing
- ☐ Campbell, Glen
- ☐ DeLafonte, Henry

- ☐ Belafonte, Harry
- ☐ Shore, Dinah
- ☐ Denver, John
- ☐ Hampton, Lionel

This re-clustering reveals that in actuality this cluster had more kinds of documents than we originally thought, based on the topical terms. These three clusters can be rather neatly summarized as containing articles about (Cluster 1) people who are sports stars, (Cluster 2) stars of film, tv, and theatre, and (Cluster 3) musicians.

Now if we back up a step and re-cluster Cluster 3 from the original set, placing the results into four clusters, we see the following:

☐ Cluster 1 Size: 12    black white nuclear hole reaction helium neutron gravitational collap

- ☐ stellar evolution
- ☐ gravitational collapse
- ☐ black hole
- ☐ main sequence
- ☐ carbon cycle
- ☐ mass–luminosity relation

☐ Cluster 2 Size: 49    galaxy type distance stellar variable spectral interstellar brightness ga

- ☐ star
- ☐ extragalactic systems
- ☐ Galaxy, The
- ☐ interstellar matter
- ☐ cluster, star
- ☐ population, stellar

☐ Cluster 3 Size: 29    constellation northern hemisphere sky locate dipper celestial double r

- ☐ constellation (astronomy)
- ☐ Auriga
- ☐ Big Dipper
- ☐ Cassiopeia
- ☐ Cygnus
- ☐ Taurus

☐ Cluster 4 Size: 7    fraunhofer designate map joseph frown fur wollaston english von davi

- ☐ Fraunhofer lines
- ☐ Fraunhofer, Joseph von
- ☐ Star Carr
- ☐ Star of David

☐ **Star Chamber**☐ **Unsubstantiated**

The contents of these four clusters can be glossed as general astrophysics, galaxies and stars, constellations, and a cluster of leftover, or outlying documents.

This example suggests the potential power of the system for automatically grouping documents according to themes. It also shows some issues that remain to be addressed. First, we need to determine automatically what the best number of clusters is at each phase. Currently we have the user make the decision of how many clusters to show for each document subcollection. We are working on how to make this choice automatically, based on the characteristics of the subcollection. Second, sometimes the summary is misleading or incomplete in terms of what documents are to be found in the cluster. We saw this with the cluster about film and tv stars -- it also contained documents about sports and music stars, although these were in the minority. We are working on determining how to indicate to the user when there are hidden topic areas in the cluster.

Click [here](#) for another example on a more complex query.

[Back to Scatter/Gather Overview](#)

# References:

---

- [Courses](#): Digital Library and related courses being offered at various Universities.
- [Conferences/Workshops](#): Links to various conferences/workshops that have been held in the recent past or will be held in the near future.
- [Journals](#): Digital Library related journal information with links.
- [Repositories & Bibliographies](#): contains information and links to some of the repositories maintained by various organizations such as the [D-Lib Magazine](#).
- [Books](#): Some books that contain valuable information on Digital Libraries (along with links to some publishers)

---

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

---

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

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

# Topics:

---

- [Search, retrieval, resource discovery](#) (See Chapter 2 in Dr. Lesk's book.)
  - [Multimedia, representations](#) (See Chapter 4 in Dr. Lesk's book.)
  - [Architectures](#) (See Chapter 6 in Dr. Lesk's book.)
  - [Interfaces](#) (See Chapter 7 in Dr. Lesk's book.)
  - [Metadata](#)
  - [Electronic publishing, SGML](#)
  - [Database issues](#)
  - [Agents](#)
  - [Commerce, economics, publishers](#) (See Chapter 9 in Dr. Lesk's book.)
  - [Intellectual property rights, copyright laws & security](#) (See Chapter 10 in Dr. Lesk's book.)
  - [Social issues](#) (See Chapters 11, 12 in Dr. Lesk's book.)
- 

## Pedagogy:

We recommend that the topics be covered in the order given above, with the reader examining the material in the book by Dr. Lesk before visiting the online information. Topics that do not correspond to chapters in the book have been included as supplementary material that seemed to be of special interest to students at Virginia Tech, and/or where there is keen interest and progress by the digital library community. However, these can be skipped by novices interested in a general overview.

---

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

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

# Search, retrieval, resource discovery:

---

## Searching - LoC

- [LoC Home Page](#)
- Z39.50 [maintenance agency](#); [part 1](#)
- [The WWW Virtual Library arranged by LoC standards](#)
- [UNDERSTANDING AND COMPARING WEB SEARCH TOOLS](#)
- [Matrix of WWW Indices: A comparison of Internet indexing tools](#)

## **Federated search**

- [UIUC Federation Across Heter. DBs](#)
- [STARTS](#)
- [INFOSEEK patent](#)
- [TSIMMIS](#)
- [Virginia Tech Federated Search Demonstration for NDLTD \(theses, dissertations\)](#)
- [Emerge \(NCSA component architecture\)](#)

## **CyberStacks (WWW, Classification, Catalogs, Reviews/Clearinghouses)**

- [Home Page](#)
- [Net Projects](#)
- [Alphabetical topics vs. LC ranges](#)
- [Call for contributions](#)
- Question: Which efforts are far along? What demonstrations can you find that are the most informative / explanatory? How well does the Library of Congress classification system fit for WWW resources?
- Related work: [OCLC's Scorpion Project](#); [DDC](#); [Mantis](#); [CORC](#)

## **Columbia**

- [D-Lib Article on Images/Video](#)
- [WebSeek Home Page](#)

## Database Groups

## **Filtering**

- [Defn](#) from U. Md. [Information Filtering Project](#)
- Fast Data Finder: [Genetic sequence analysis](#)
- What is *information filtering*? How does it differ from information retrieval?

## [Cross-Language Information Retrieval Resources](#)

- [Eurospider](#) and [ISN LASE Search demo](#)
- [Readware Demo](#)
- [Mundial](#) - English and Spanish Demo
- Questions:
  - What languages are covered?
  - How well are phrases handled?

## [Stanford DL info finding projects](#)

[Berkeley documents and queries](#) (please study carefully, answering questions)

## [UCSB spatial indexing and retrieval](#)

---

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

---

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

**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**

International Standard  
Maintenance Agency

# Z39.50

The Library of Congress  
Network Development  
& MARC Standards Office

[Z39.50 Resources](#) - [Z39.50 Document](#) - [Related Specifications](#) - [Object Identifiers](#)  
[Implementor Register](#) - [Z39.50 Profiles](#) - [ZIG Meetings](#) - [Site Index](#)

This page provides links to information about Z39.50 resources and about the development and maintenance of Z39.50 (existing as well as future versions) and the implementation and use of the Z39.50 protocol.

"Z39.50" refers to the International Standard, ISO 23950: "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification", and to ANSI/NISO Z39.50. The Library of Congress is the Maintenance Agency and Registration Authority for both standards, which are technically identical (though with minor editorial differences).

The standard specifies a client/server-based protocol for searching and retrieving information from remote databases.

[Comments: z3950@loc.gov](mailto:z3950@loc.gov)  
[Maintenance Agency Procedures](#)

[Next ZIG Meeting:](#)  
[July -- Leuven, Belgium](#)

- [Registration Now Open!](#)

[ZIG Member Status Reports](#)

---

[Library of Congress Home](#) - [Other Standards Maintained by the Library](#) - [Z39.50 Gateway](#)

---



Library of Congress

General Comments: [lcweb@loc.gov](mailto:lcweb@loc.gov) Updated: May 30, 2000



## CONTENTS



[Minutes](#)

## FEDERATION ACROSS HETEROGENEOUS DATABASES



[Presentations](#)

April 3-4, 1997  
Grainger Engineering Library Information Center

University of Illinois at Urbana-Champaign  
1301 W. Springfield Ave., Urbana, IL



[AGENDA](#)

Welcome to the official site for the UIUC Digital Library  
Initiative Spring '97 Partners Workshop.

Please contact Susan Harum [dli@uiuc.edu](mailto:dli@uiuc.edu) for any questions  
or comments about the workshop.

[Go back to the DLI workshop page](#)



## ATTENDEES

# STARTS

## Stanford Protocol Proposal for Internet Retrieval and Search

STARTS is the result of an informal "standards" effort that we ([Luis Gravano](#), [Kevin Chang](#), [Hector Garcia-Molina](#), [Carl Lagoze](#), and [Andreas Paepcke](#)) coordinated at Stanford. This project developed a simple protocol that text search engines should follow to facilitate searching and indexing multiple collections of text documents.

[Final writeup](#) of the STARTS protocol ([PostScript version](#))

[A reference-implementation](#) of STARTS by Carl Lagoze

[A more readable description](#) of the STARTS protocol that appeared in Sigmod'97

[List of participants](#) of the STARTS Workshop, Stanford, August 1st, 1996

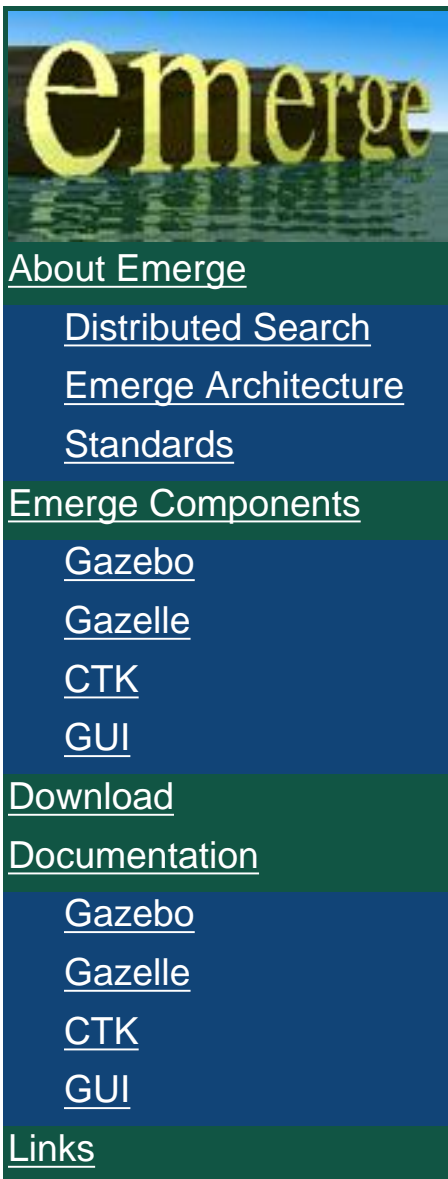
Slides of the talk that Prof. Hector Garcia-Molina gave at the STARTS workshop ([Powerpoint Version](#))

Slides of the talk that Luis Gravano gave at the STARTS workshop ([Powerpoint Version](#))

---

[Luis Gravano](#)

[gravano@cs.stanford.edu](mailto:gravano@cs.stanford.edu)



[emerge@ncsa.uiuc.edu](mailto:emerge@ncsa.uiuc.edu)

# About Emerge

Emerge is an NCSA effort to develop middleware components of a new distributed search infrastructure which addresses the scale and heterogeneity of scientific data. Our components enable search services to interoperate across scientific domains by providing user-configurable tools for mapping between metadata schemas, performing search queries against multiple data sources, and performing query pre- and post-processing. Access to our search services is through platform-neutral standard and emerging-standard tools such as [Z39.50](#), [XML](#), and [Java](#).

Here's a [slide show](#) with an overview of our research area and component architecture. And [here's one](#) which gives an overview of interoperability issues in distributed scientific information retrieval.

## Collaborations

Emerge is part of NCSA's Data Management and Visualization Division. Our components have been developed in collaboration with the [National Cancer Institute](#), the UIUC Digital Library Initiative and [CANIS](#), [NASA Project 30](#). We've also participated in panel discussions and advisory meetings with the [Committee for Institutional Cooperation](#) and the [UIUC Library Gateway](#) project.



## News

April 18, 2000: Alpha versions of our Java [XER](#) utilities are available. [Download](#) them or browse the online [documentation](#).

November 23, 1999: An alpha version of Gazelle has been released. Browse the new Gazelle [documentation](#) and [download](#) the latest version from our ftp server!

Emerge was featured in the August 27th, 1999 edition of [Science](#) magazine, in the NetWatch column (Vol. 285, number 5432). The article describes recent work to integrate NCSA's [Astronomy Digital Image Library](#) with diverse sources of astronomy data using Emerge components and a new XML format called [AML](#) (Astronomy Markup Language), developed by Damien Guillaume. This work was covered in a paper co-authored by Bob McGrath, Ray Plante,

Guillaume and Joe Futrelle, which was presented Aug. 13 at the Digital Libraries '99 meeting in Berkeley, CA.

Gazebo 0.9b1 is now available. New features include a revamped configuration file syntax, a Linux port, [better documentation](#), and improved error handling.

A slightly updated version of the demonstration GUI is also available. The GUI will be substantially re-worked and revisions will be released soon.

## Contact

Emerge can be reached at [emerge@ncsa.uiuc.edu](mailto:emerge@ncsa.uiuc.edu) and at [futrelle@ncsa.uiuc.edu](mailto:futrelle@ncsa.uiuc.edu).



# The Scorpion Project



[Scorpion](#) is a project of the [OCLC Office of Research](#) exploring the indexing and cataloging of electronic resources. Since subject information is key to advanced retrieval, browsing, and clustering, the primary focus of Scorpion is the building of tools for automatic subject recognition based on well known schemes like the [Dewey Decimal System](#).

---

## Scorpion Documentation

- [A brief introduction to Scorpion](#)
- [Evaluating Dewey Concepts](#)
- [Evaluating Scorpion Results](#)
- [Measures for Evaluating ...](#)
- [Clustering](#)
- [AMIGOS 97](#) (full image [version](#))
- [Scorpion helps catalog the Web](#)
- [Dewey Database Design](#)
- [ESS Field Label Descriptions](#)
- [Example ESS Record](#)
- [SMART Weighting Schemes](#)
- [Scorpion Usage Stats \(OCLC Internal Use Only\)](#)

## Automatic Subject Assignment

- [Simple URL Input Form](#)
- [Simple Text Input Form](#)
- [Advanced Input Form](#)

Thank you for your interest in the Scorpion project. The Research phase of this project that provided automatic subject assignment using the Dewey Decimal Classification (DDC) ended **November 2, 1999**. If you are an OCLC participating member, you can access the Scorpion automatic classifier through CORC. If you are a library or library school, you may also apply for CORC membership. For more information about CORC, send a message to [corc@oclc.org](mailto:corc@oclc.org), or visit the CORC Web site at <http://purl.oclc.org/corc>.

For more information about electronic access to the Dewey Decimal Classification, please visit the OCLC Forest Press Web site at <http://www.oclc.org/fp> or contact Dawn Lawson, OCLC Forest Press Electronic Products Manager ([dawn\\_lawson@oclc.org](mailto:dawn_lawson@oclc.org)).

## Related Work

- [Our "staging area" of related work.](#)
- [Online Classification: Implications for Classifying and Document\[-like Object\] Retrieval](#)
- [Using Library Classification Schemes for Internet Resources](#)
- [Dewey 2000](#)
- [Cataloguing Rules and Conceptual Models](#)
- [The Dublin Core](#)
- [Prototype Dublin Core Metadata System](#)
- [Electronic classification schemes](#)
- Pharos ([demo](#), [publications](#))

---

### About Scorpion

*The Scorpion service is Copyright 1996-1999 OCLC Online Computer Library Center, Incorporated. All and any portion thereof and all trademarks, copyrights, and other proprietary rights contained or existing therein are and shall remain the sole and exclusive property of OCLC*

*Scorpion uses a database based on the Dewey Decimal Classification (DDC) to assign DDC numbers. The Dewey Decimal Classification (DDC) system is Copyright 1996-1999 OCLC Online Computer Library Center, Incorporated. All and any portion thereof and all trademarks, copyrights, and other proprietary rights contained or existing therein are and shall remain the sole and exclusive property of OCLC. OCLC, Forest Press, Dewey, DDC, and Dewey Decimal Classification are registered trademarks of OCLC.*

### About the Dewey Decimal Classification (DDC) system

*The Dewey Decimal Classification (DDC) system is Copyright 1996-1999 OCLC Online Computer Library Center, Incorporated. All and any portion thereof and all trademarks, copyrights, and other proprietary rights contained or existing therein are and shall remain the sole and exclusive property of OCLC. OCLC, Forest Press, Dewey, DDC, and Dewey Decimal Classification are registered trademarks of OCLC.*

---



Comments/suggestions to [shafer@oclc.org](mailto:shafer@oclc.org)  
Scorpion [contributors](#)

# A Brief Introduction to Scorpion

[Keith Shafer](#)

*Last modified: Fri Dec 27 10:08:54 EST*

[Scorpion](#) is a [research](#) project at [OCLC](#) exploring the indexing and cataloging of electronic resources. Since subject information is key to advanced retrieval, browsing, and clustering, the primary focus of Scorpion is the building of tools for automatic subject recognition based on well known schemes like the [Dewey Decimal System](#).

---

## Electronic Chaos

The recent explosion of electronic information has far outpaced the availability of automated tools to effectively manage it. This phenomenon has gained visibility due to the popularity of the World Wide Web, but the problem is not restricted to the Web. Some organizations already have more electronic information available online than all of the publicly available Web pages combined. Furthermore, most printed material now exists in electronic format long before it is published.

As the Web has grown, there has been a general push to apply and develop techniques to make Web resources searchable and more widely accessible. For instance, there are now many free search services like Yahoo! and AltaVista. There is even wide acceptance of information retrieval techniques like ranked retrieval to help users sift through the volumes of Web information now available. Yet searching the raw content of every document still seems to be an unacceptable solution since it's not uncommon to retrieve hundreds of documents for a given search.

As the amount of accessible electronic information increases, the cost of accessing this information will increase. That is, even though users can now use free search services to find items of interest, they will increasingly spend their valuable time wading through masses of irrelevant documents to get the information they need. Accordingly, communities unfamiliar with library science are beginning to grapple with the problem of metadata and the organization of large collections of data.

## Order Makers

Historically, librarians have organized the world's information. For centuries, they have successfully managed, classified, and filtered information of many types. This has been accomplished via the creation of surrogates to manage the items of interest. For instance, to provide access to a book, a **catalog entry** is created to describe the book. Then, the potential reader need not have the book in hand to know what it is about, who wrote it, where it can be found, etc. This is a very effective scheme. Creating **metadata** about an object makes searching, filtering, organizing, and retrieving the item very efficient. This is true of traditional materials as well as new electronic resources.

These two worlds -- the seemingly unorganized Web and the organized world of libraries -- have much to offer one another. The Web can offer automated tools for searching raw information and the library world can offer experience organizing and understanding information of all types. By combining their talents and techniques, these two communities can bring powerful resources to bear on the problems of

accessing, maintaining, and supplying electronic information.

If every electronic item had a catalog entry or its equivalent, then interfaces could provide the best of both worlds -- access to the raw content (the free search services model) and access to the filtered information (the library model). However, librarians are so overburdened that they can barely keep up with their traditional workload, let alone begin to catalog and organize the vast amounts of information available electronically.

All electronic resources will never be humanly cataloged. It's just too expensive. Clearly, automated tools to apply library science ideas like classification and filtering to electronic resources at high speed and low cost are needed.

## Automation

While traditional catalog entries contain a wealth of metadata, the subject portion is arguably the most important when it comes to building advanced search and retrieval interfaces. If there were some way to automatically assign subject headings or concept domains to electronic items, then powerful filtering tools could be built. That's where [Scorpion](#) comes in.

Scorpion is a research project attempting to combine indexing and cataloging, based on the observation that these are complementary activities. Scorpion specifically focuses on building tools for automatic subject recognition by combining library science and information retrieval techniques. For instance, to assign subject codes to a document, the document can be treated as a query against a Dewey Decimal System database using ranked retrieval. The results of the search can then be treated as the subjects of the document. Subject assignment in this manner provides clear differentiation from the traditional computer indexing behind the currently available free search services.

Scorpion cannot replace human cataloging. There are many aspects of human cataloging that are difficult if not impossible to automate. However, Scorpion should produce tools that help reduce the cost of traditional cataloging by automating subject assignment when items are available electronically. For instance, a list of potential subjects could be presented by Scorpion to a human cataloger who could then choose the most appropriate subject.

---

For more information, please visit the Scorpion site at <URL:<http://purl.oclc.org/scorpion>> where additional documentation explaining the Scorpion tools and experimental results will be posted. Local clients or users with passwords and IDs will also be able to see the Scorpion tools in action. Anyone desiring a password and ID should contact [Keith Shafer](#).



Up



OCLC Home



Search



Site Map



What's New



Feedback



Site Help

[News](#)[About OCLC](#)[OCLC Services](#)[Support & User Doc.](#)[Contacts & Addresses](#)

[OCLC Cataloging Services](#) or [OCLC Reference Services](#)

## Considering CORC?

- [Service Overview](#)
- [About CORC](#)
- [View Participants List](#)
- [Sign up to Participate](#)  
Free to members until July 1, 2000

# Cooperative Online Resource Catalog

CORC is a state of the art, Web-based system that helps libraries provide well guided access to Web resources using new, automated tools and library cooperation. CORC empowers librarians with automated tools for the cooperative creation, selection, organization, and maintenance of Web based resources.

## Using CORC

**Patron** is the default authorization. Other access levels including Cataloging are available through assigned authorizations.

### [CORC Users Meeting](#)

**July 12 at the Chicago Public Library**

## Using CORC?

- [Log on to CORC](#)
- [Practice Area](#)
- [Frequently Asked Questions](#)
- [News](#)
- [Documentation](#)
- [Honor Roll](#)
- [Training Materials and PowerPoint Presentations](#)
- [Users Meeting Registration Form](#)



OCLC ONLINE COMPUTER LIBRARY CENTER, INC.

# Finding Images/Video in Large Archives

## Columbia's Content-Based Visual Query Project

Shih-Fu Chang, John R. Smith  
Horace J. Meng, Hualu Wang, and Di Zhong  
Department of Electrical Engineering and  
Center for Telecommunications Research  
Columbia University

*{sfchang,jrsmith,jmeng,hwang,dzhong}@ctr.columbia.edu*

**D-Lib Magazine**, February 1997

ISSN 1082-9873

---

### Table of Contents

- [An Application Driven Problem](#)
- [State of the Art](#)
- [Research Strategies](#)
- [Prototype Systems](#)
- [Testbed Support and User Evaluation](#)
- [Open Issues](#)
- [References](#)

---

## An Application Driven Problem

How do we find a photograph from a large archive which contains thousands or millions of pictures? How does a CNN video journalist find a specific clip from the myriad of video tapes, ranging from historical to contemporary, from sports to humanities? How do people organize and search the content of personal video tapes of family events, travel scenes, or social gatherings?

The era of "the information explosion" has brought about the wide dissemination and use of visual information, particularly, digital images and video, which we are also seeing in combination with text, audio, and graphics. The development of tools and systems that enhance image functionalities, such as searching and authoring, is critical to the effective use of visual information in the new media applications.

The current research and development of images and video search tools is driven by practical applications. We are seeing the establishment of large digital image and video archives, such as the Corbis catalog, which includes the Bettman Archive; the Picture Exchange, which is a joint venture between Kodak and Sprint; and many digital video libraries in various domains (e.g., environment, politics, arts), such as the on-line CNN news archives.

The systems for the search and retrieval of images and video from these archives require the development of efficient and effective image query tools.

## State of the Art

The use of comprehensive textual annotations provide one method for image and video search and retrieval. Today, text-based search techniques are the most direct, accurate, and efficient methods for finding "unconstrained" images and video. Text annotation is obtained by manual effort, transcripts, captions, embedded text, or hyperlinked documents. In these systems, keyword and full text searching may be enhanced by natural language processing techniques to provide great potential for categorizing and matching images.

The searching of images by their visual content complements the text-based approaches. Very often, textual information is not sufficient. Visual features of the images and video also provide a description of their content. By exploring the synergy between textual and visual features, these image search systems may be further improved. However, it is a significant challenge to automatically reconcile inconsistency between input from these features.

Many content-based image search systems have been developed for various applications. There has been substantial progress in developing powerful tools which allow users to specify image queries by giving examples, drawing sketches, selecting visual features (e.g., color and texture), and arranging spatial structure of features. Using these approaches, the greatest success is achieved in specific domains, such as remote sensing and medical applications. The reason is that in constrained domains, it is easier to model the users' needs and to restrict the automated analysis of the images, such as to a finite set of objects.

The integration of computer vision and image processing promises a wealth of techniques for solving the image and video search problems. But new challenges remain. In unconstrained images, the set of known object classes is not available. Also, use of the image search systems varies greatly. Users may want to find the most similar images, find an appropriate class of images, browse the image collection quickly, and so on. One unique aspect of image search systems is the active role played by users. By modeling the users and learning from them in the search process, we can better adapt to the users' subjectivity. In this way, we can adjust the search system to the fact that the perception of the image content varies between individuals, or over time.

The general system architecture for a content-based visual query system is included in Figure 1. The analysis of images and feature extraction plays important roles in both off-line and on-line processes. Other important aspects of the system include the closed interaction loop (including users), the supporting database components for retrieval and indexing, the integration with multimedia features, and the efficient user interfaces for query specification and image browsing.

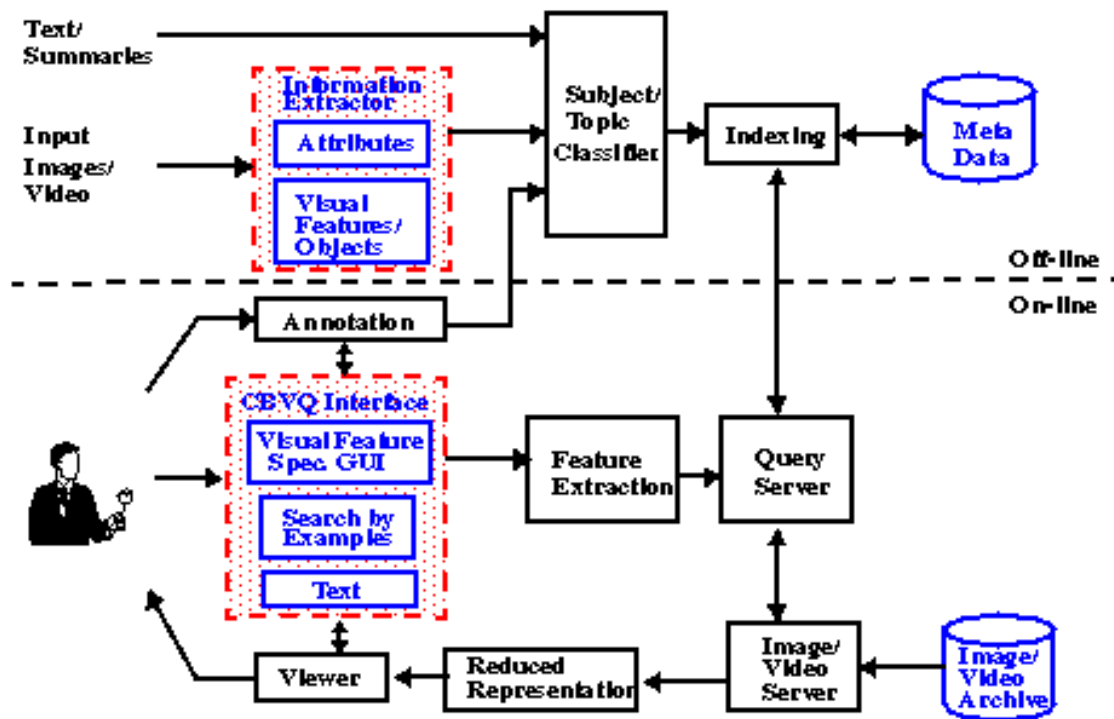


Figure 1. A general CBVQ system architecture.

The search of images is an emerging field with many exciting research challenges. The research tasks are practical, important, but not easy. In the following, we present our research strategies, prototype systems for image/video search, and our views on the important open research issues.

## Research Strategies

We present our strategies for tackling the above challenging issues in this section.

### Create a visual feature library by automatic image analysis

Although today's computer vision systems cannot recognize high-level objects in unconstrained images, we are finding that low-level visual features can be used to partially characterize image content. These features also provide a potential basis for abstraction of the image semantic content. The extraction of local region features (such as color, texture, face, contour, motion) and their spatial/temporal relationships is being achieved with success. We argue that the automated segmentation of images/video objects does not need to accurately identify real world objects contained in the imagery. Our goal is to extract the "salient" visual features and index them with efficient data structures for fast and powerful querying. Semi-automated region extraction processes and use of domain knowledge may further improve the extraction process.

In the later sections, we discuss the use of automatically extracted spatial/color regions for image search, and the integration of multiple visual features for video object indexing. We use a hierarchical object based schema for feature indexing and high-level object abstraction [4] (see Figure 2). The fusion of multiple visual features improves the region extraction process. We also show that the aggregation of regions into higher level objects is influenced by the spatial/temporal relationships of the regions. For example, Figure 3 shows the results of automatic video object segmentation and tracking. The visual features and spatial/temporal attributes of regions generate an index for searching for the video objects stored in the archive.

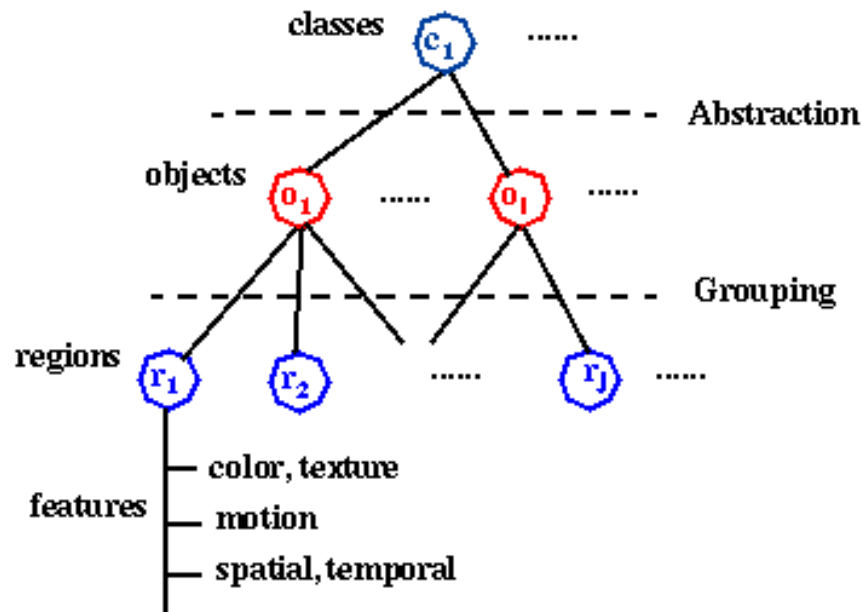


Figure 2. A hierarchical object based schema for images/video.

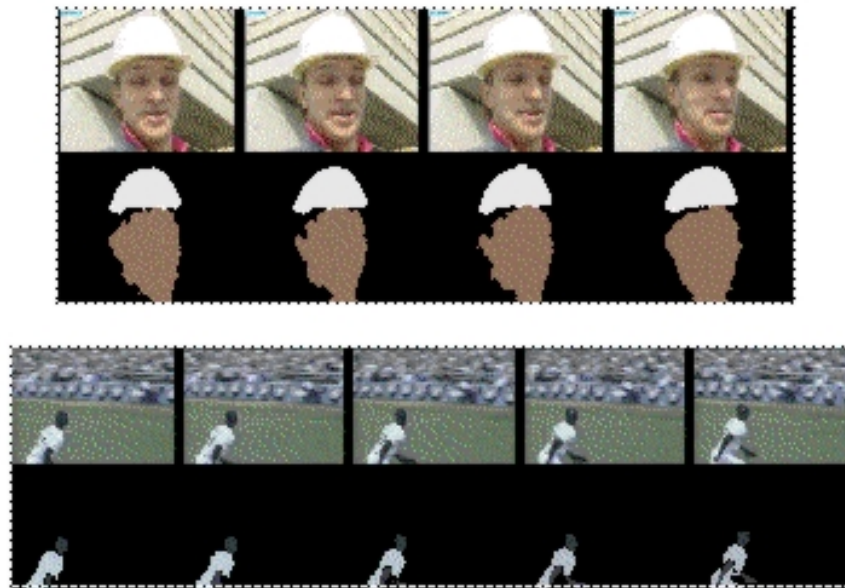


Figure 3. Examples of automatically segmented and tracked video objects.

## Explore the synergy between compression and functionalities

It's impossible to anticipate the users' needs completely at the feature extraction and indexing stage. The ideal solution is that images and video are represented (for compression also) in a way that is amenable to dynamic feature extraction. Today's compression standards (such as JPEG, MPEG-1, MPEG-2), are not suited to this need. The objective in the design of these compression standards was to reduce bandwidth and increase subjective quality. Although many interesting analysis and manipulation tasks can still be achieved in today's compression formats (as described later), the potential functionalities of the images were not considered. However, recent trends in

compression, such as MPEG-4 and object-based video, have shown interest and promise in this direction. The goal is to develop a system in which the video objects are extracted, then encoded, transmitted, manipulated, and indexed flexibly with efficient adaptation to users' preference and system conditions.

## Learn from users and domain ontologies

To break the barrier of decoding semantic content in images, user-interaction and domain knowledge is needed. These systems learn from the users' input as to how the low-level visual features are to be used in the matching of images at the semantic level. For example, the system may model the cases in which low-level feature search tools are successful in finding the images with the desired semantic content. In this way, the categories can be monitored and better analyzed by the system. Learning and other techniques in artificial intelligence provide great potential for these systems.

If the applications require the definition of specific semantic subjects, the feature models of images in these classes are constructed by hand and then used to match objects in the unknown images/video. This object recognition and subject classification method provides a system for on-line information filtering. We see great potential for improving image search systems to link the low-level visual features with high-level semantics. However, in unconstrained application domains, we expect only moderate success early on.

## Integrate visual and other multimedia features

Exploring the association of visual features with other multimedia features, such as text, speech, and audio, provides another potentially fruitful direction. Our experience indicates that it is more difficult to characterize the visual content of still images compared to video. Video often has text transcripts and audio that may also be analyzed, indexed, and searched. Also, images on the World Wide Web typically have text associated with them. In this domain, the use of all potential multimedia features enhances image retrieval performance.

## Prototype Systems

We have developed several content-based visual query prototype systems. WebSEEk and VisualSEEk explore the problem of efficiently searching large image archives. WebClip focuses on browsing, search, and content editing of networked video.

In WebSEEk, the images and video are analyzed in two separate automatic processes:

- (1) visual features (such as color histograms and color regions) are extracted and indexed off-line,
- (2) the associated text is parsed, and utilized to classify the images into subject classes in a customized image taxonomy (including more than 2000 classes).

More than 650,000 unconstrained images and video clips from various sources have been indexed in the initial prototype implementation. Users search for images by navigating through subject categories, or by using content-based search tools. The details of the system design and operation are described in [\[1\]](#).

One objective of WebSEEk is to explore the synergy between visual features and text. We also demonstrate the feasibility of image searching in a large scale testbed, the World Wide Web. We are developing more sophisticated content-based image search techniques in the VisualSEEk system [\[2\]](#). VisualSEEk enhances the search capability by integrating the spatial query (like those used in geographic information systems) and the visual feature query. Users ask the system to find images/video that include regions of matched features and spatial relationships. Figure 4 shows a query example in which two spatially arranged color patches were issued to find images with blue sky and open grass fields.

Back Forward Home Edit Reload Images Open Print Find Stop

Location: <http://disney.ctr.columbia.edu/SaFe/>

Query Clear Reset

Grid Paint Help

A B C D E F G  
H I J K L M N  
O P Q R S T U  
V W X Y Z 1 2

Spatial Query:  
☒ Absolute ☐ Relative

Query Weights:  
Spatial  10  
Feature  10  
Size  10  
Region  10



Photographs ☐

# SaFe

## Spatial and Feature query system

VisualSEEK Photographs Database: Spatial and Feature Query

2346 matches for REGION 1  
2171 matches for REGION 2

		
0 [364] ( 479.16)	1 [2841] ( 511.52)	2 [2788] ( 528.12)
		
3 [99] ( 541.52)	4 [1606] ( 558.76)	5 [372] ( 609.52)
		
6 [1141] ( 620.24)	7 [1138] ( 621.68)	8 [1774] ( 622.48)

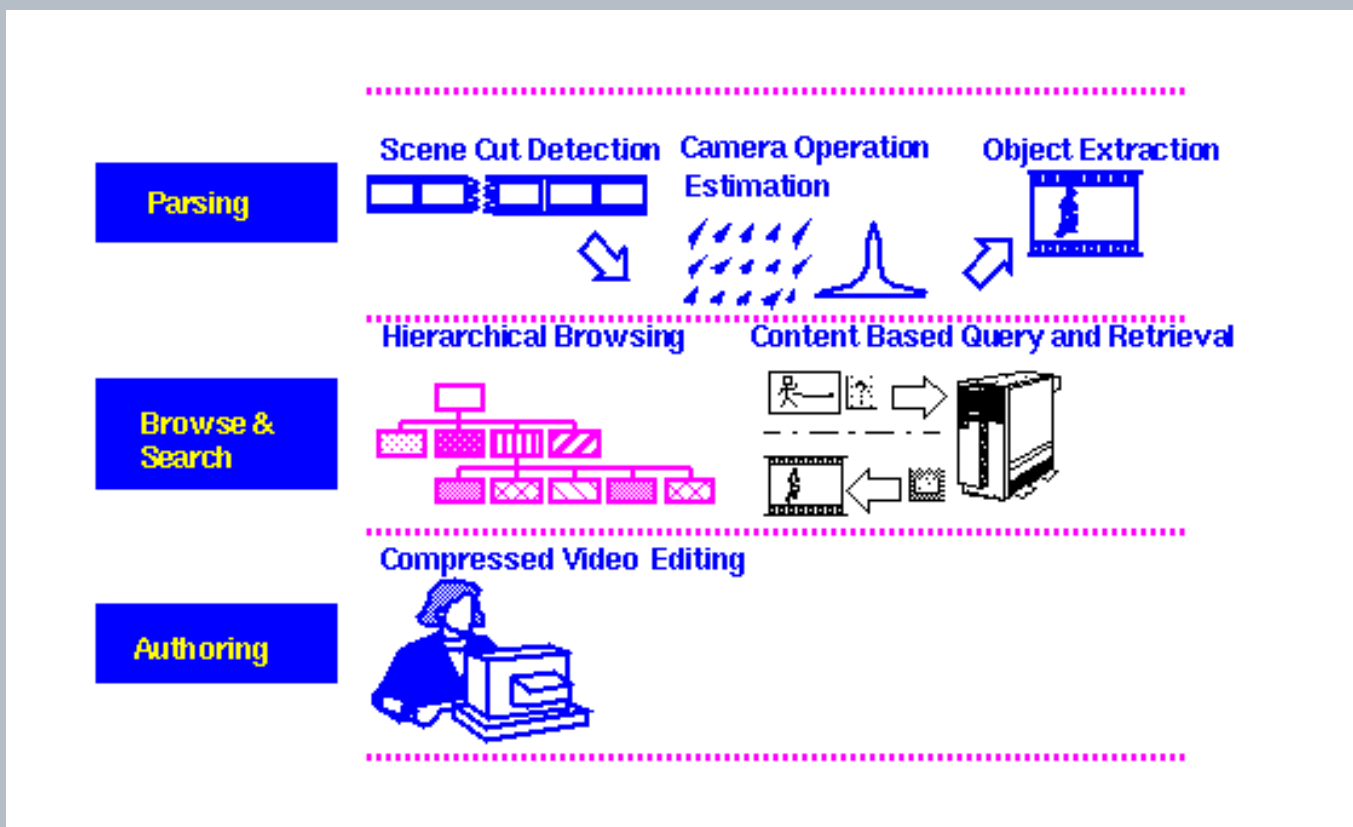
**Figure 4. An example query using VisualSEEK.**

For video, we have developed a system called WebClip [3], which allows for efficient browsing and editing of compressed video over the Web. One objective is to demonstrate the benefits of using compressed video without full decoding during the content analysis and manipulation stages. Visual features (like scene changes, foreground motion objects, and icon streams) can be extracted directly from the compressed video. Web users do not need high-end video decoding or processing facilities like those used in professional studios. Another objective of WebClip is to integrate the search and editing functionalities in the same environment. Tools developed in image search systems (like the above mentioned WebSEEK and VisualSEEK systems) are being ported to the video system. We are also adding new tools for searching by motion feature and temporal characteristics. After retrieval of matched video clips, the users use web-based tools to edit the video and compose new presentations with various video special effects.

Figure 5 shows the functionality components of WebClip. The compressed video sequences are parsed to obtain visual features and objects. The browsing and search interface provides a tree-structure hierarchical scene-based interface. This display can be adapted to different browsing modalities:

- (1) the time-based model,
- (2) the story-based model, and
- (3) the feature-based model.

The time-based model hierarchically lays out the icons of key frames from each video scene. This allows for rapid inspection of video content according to a sequential order of time. The story-based model recognizes (automatically or manually) the story structure within the video (e.g., a complete news story) and groups all video scenes belonging to the same story under a single node in the tree. The feature-based model clusters all video scenes to classes within each of which all video scenes have similar visual features. We have also undertaken new efforts to extend the joint spatial/feature query tools of the VisualSEEK system to the video domain. Video is indexed and searched by spatial/temporal relationships and visual features of video objects contained in the video sequence.

**Figure 5. Functionality components of WebClip.**

## Testbed Support and User Evaluation

Most of the test images and video in our testbed are collected from the public domain, including data on the Web, copyright free photograph stock from commercial CD's, MPEG simulation test video, and proprietary content from local research groups. Features extracted from these images are stored in our SGI ONYX-based server, which has 50GB storage space on disk arrays, and 50GB tertiary space on a tape archive.

Network facilities include standard Internet connections (via a T-3 line to outside), ATM connections within the campus and with external wide area networks (NYNET), and internal wireless networks running mobile IP. A video-on-demand (VoD) system which supports software-based video servers, MPEG-2 transport, and heterogeneous client terminals has been developed in the Image and Advanced TV lab. We envision the integration of our search systems with the VoD system soon to provide integrated image services.

An important work plan for the near future is the collaboration with faculty and students in the School of Journalism and at Teachers College, Columbia University. User studies and performance evaluation are being conducted in the news and education domains. One example is the Columbia Digital News Systems group [5], which integrates our efforts with others on information tracking, natural language processing, and multimedia briefing.

## Open Issues

Image/video searching is a relatively new field, but it has many exciting research issues. It requires close interaction between multiple technical disciplines and applications users. Researchers have made great progress in recent years, but a few critical issues have still not been addressed adequately. In particular, we believe that further breakthroughs need to be made in the following areas before image search systems can make significant impacts on real applications.

### Effective evaluation metrics and testset

Today, there are no satisfactory methods for measuring the effectiveness of image search techniques. Precision/recall types of metrics have been used in some of the literature but are impractical due to the tedious process of measuring image relevancies. There are no standard image corpus or benchmark procedures. We believe that resolution of this issue is of top priority for researchers and users in this field.

### Dynamic extraction and matching of visual features

As mentioned earlier, the image indexing and search schemes must adapt to dynamic user needs, resource conditions, and input data. In particular, the user needs and application requirements vary over time. A static set of features and matching schemes is limited. Efficient, if not real-time, methods should be developed to perform dynamic feature extraction, matching and abstraction. Real-time is defined in three different aspects:

- (1) fast enough to process live information (like live video),
- (2) fast enough to process a large amount of new information on-line (like on-line information filtering), and
- (3) fast enough to re-process existing data in the archive.

The degree of time urgency decreases in the same order. All these aspects demand breakthroughs in image/video representation and dynamic content analysis.

### Linking low-level features to high-level semantics

Today's content-based image search systems allow for image queries based on image examples, feature specification, and primitive text-based search. The WebSEEk system uses automatically extracted text in image subject classification. Other researchers have also shown some success in using newspaper photograph captions and

video transcripts to assist visual content analysis. Adaptive visual feature organization through user interaction has also been proposed. But the linkage between low-level visual features and high-level semantics is still very weak. Non-technical, general users tend to expect the same level of functionalities as those seen in today's text search systems. We admit that this is a difficult objective. But, as they are driven by critical application needs, image search systems will benefit from any breakthrough made in this direction.

## Acknowledgements

This project is supported in part by the ADVENT industry partnership project at the Image and Advanced TV Lab of CTR, Columbia University, Columbia Digital Library project, and National Science Foundation (IRI-9501266). We appreciate the research collaboration in this area with Dr. Chung-Sheng Li of IBM, Dr. Kenrick Mock of Intel, Dr. Harold Stone of NEC, Dr. HongJiang Zhang of Hewlett-Packard, and Mr. Jan Stanger.

## References

1. J. R. Smith and S.-F. Chang, "Searching for Images and Videos on the World-Wide Web," to appear in IEEE Multimedia Magazine, Summer, 1997. (also Columbia University CU/CTR Technical Report #459-96-25). Demo: <http://www.ctr.columbia.edu/webseek> <ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96e.ps>
2. J. R. Smith and S.-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System," ACM Multimedia Conference, Boston, MA, Nov. 1996. Demo: <http://www.ctr.columbia.edu/VisualSEEk> <ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/smith96f.ps>
3. J. Meng and S.-F. Chang, "CVEPS: A Compressed Video Editing and Parsing System," ACM Multimedia Conference, Boston, MA, Nov. 1996. Demo: <http://www.ctr.columbia.edu/WebClip> <ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/96/meng96c.ps>
4. D. Zhong and S.-F. Chang, "Video Object Model and Segmentation for Content-Based Video Indexing," IEEE Intern. Conf. on Circuits and Systems, June, 1997, Hong Kong. (special session on Networked Multimedia Technology & Application) <ftp://ftp.ctr.columbia.edu/CTR-Research/advent/public/papers/97/zhong97a.ps>
5. A. Aho, S.-F. Chang, K. McKeown, D. Radev, J. Smith, and K. Zaman, "Columbia Digital News Systems," to appear in Workshop on Advances in Digital Libraries, 1997.

*Approved for release, February 14, 1997.*

Copyright ©1997 Shih-Fu Chang, John R. Smith, Horace J. Meng, Hualu Wang, and Di Zhong



*hdl:cnri.dlib/february97-chang*

# Database Groups:

---

- [PENN](#)
- [Stanford](#)
- [Garlic - IBM Almaden](#)
- [U. Md.](#)
- [UCB database management](#)

---

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

---

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

**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**



# The Garlic Project

## Introduction

Garlic is a project being developed by members of the database group in Computer Science. The goal of Garlic is to enable large-scale multimedia information systems: large scale in that they involve lots of data with multimedia taken as broadly as possible to mean data of many types. We are particularly concerned about situations in which there is enough data of sufficiently specialized types that users have already made decisions about how to manage it, and have stored it in separate repositories that are specifically adapted to data of that type.

## The Need:

The bulk of the data in the world is not stored in database management systems. There are many specialized systems emerging to store and search for particular data types, including image management systems, etc. However, many applications can benefit from combining information from these various systems.

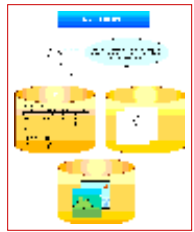


## In Medicine:

For example, in the medical field, hospitals often have separate information systems for each department. Radiology may store MRI scans, etc., in one system, Cardiology may store EKG's in another, the Lab may store lab reports in a document management system, and Administration may store its records in a relational DBMS. Doctors, however, need access to all of this information when treating a patient. Today, hard copies are made and collected in a folder, leading to delays and inconsistencies. In the future, hospitals would like to be able to store patient folders on-line, enabling doctors to search within and across folders (find all folders where the patient has symptoms



similar to this one'). However, they are unlikely to move all the data to a new, centralized system, or in fact, to any new system that disrupts their existing applications or threatens the autonomy of the various departments.



## In Kitchen Design:

Consider the interior designer of the future. He will need information on wallpapers, cabinets, appliances, floor tiles, etc., as well as information on his previous designs and a collection of powerful modeling tools. These different collections of information are likely to be owned by separate establishments, which will want to make independent decisions on what software and hardware to use to store them. Yet there are financial advantages to all concerned if they can share this data.

## In Business: The Ad Agency Example:

## Publications

### General Garlic Papers

- [Data Engineering Bulletin '99: Transforming Heterogeneous Data with Database Middleware: Beyond Integration \(postscript ~364k\)](#)
- [RIDE-DOM '95: Towards Heterogeneous Multimedia Information Systems: The Garlic Approach](#)
- [Visual Database Systems '95: Querying Multimedia Data from Multiple Repositories by Content: The Garlic Project](#)
- [SIGMOD '96: Demo Announcement \(postscript ~50k\)](#)

### Query Optimization

- [VLDB '99: Cost Models DO Matter: Providing Cost Information for Diverse Data Sources in a Federated System \(postscript ~284k\)](#)
- [IBM Technical Report RJ10141 \(extended version of VLDB '99 paper\): Cost Models DO Matter: Providing Cost Information for Diverse Data Sources in a Federated System \(postscript ~316k\)](#)
- [VLDB '97: Optimizing Queries across Diverse Data Sources \(postscript ~293k\)](#)
- [Data Engineering Bulletin '96: An Optimizer for Heterogeneous Systems with Nonstandard Data and Search Capabilities \(postscript ~213k\)](#)

### Caching

- [VLDB '99: Loading a Cache with Query Results \(postscript ~194k\)](#)
- [IBM Technical Report RJ6291 \(extended version of VLDB '99 paper\): Loading a Cache with Query Results \(postscript ~300k\)](#)

#### Fagin's Algorithm for Merging Ranked Results

- [JCSS '99 \(extended version of PODS '96 paper\): Combining Fuzzy Information from Multiple Systems \(postscript ~636k\)](#)
- [COOPIS '99: Using Fagin's Algorithm for Merging Ranked Results in Multimedia Middleware \(postscript ~140k\)](#)
- [PODS '98:Fuzzy Queries in Multimedia Database Systems \(postscript ~247k\)](#)

#### Wrapper Architecture

- [VLDB '97: Don't Scrap It, Wrap It! An Architecture for Legacy Data Sources \(postscript ~186k\)](#)
- [IBM Technical Report RJ10077 \(extended version of VLDB'97 paper\): An Architecture for Legacy Data Sources \(postscript ~235k\)](#)

#### "Magic Formula" for Incorporating User Weights

- [ICDT '97:A Formula for Incorporating Weights into Scoring Rules \(postscript ~331k\)](#)

#### Query Browsing

- [\(extended version of VLDB '96 paper\): PESTO: An Integrated Query/Browser for Object Databases](#)

---

[ [IBM Almaden Computer Science](#) | [IBM Almaden](#) | [IBM Research](#) ]

[ [IBM home page](#) | [Order](#) | [Search](#) | [Contact IBM](#) | [Help](#) | [\(C\)](#) | [\(TM\)](#) ]

# Information Filtering Defined

A universally accepted definition of information filtering is, unfortunately, still lacking. So here is my personal definition, which I have used to build the Information Filtering Resources [web page](#). Generally, the goal of an information filtering system is to sort through large volumes of dynamically generated information and present to the user those which are likely to satisfy his or her information requirement.

In order to sharpen this definition, a distinction should be drawn between information collection and information filtering. In some domains (e.g. USENET News) the collection effort is minimal because the information comes to you. In other domains (e.g. the World Wide Web) the collection effort can be considerable because no mechanism exists to draw new information to the attention of a filtering system. The point to be made here, though, is that information collection is an interesting area in its own right, but I do not propose to include it in my definition of information filtering. In my view, the information filtering problem begins only after you have gained access to the new information.

Information filtering has been applied to a several domains using a variety of technical approaches. The original methods were manual alerting services that brought new information to the attention of users of research and special libraries. At the time this was referred to as Selective Dissemination of Information (SDI), a name which fell from favor about the time the Strategic Defense Initiative (SDI) was introduced in the United States :-). A few modern systems have adopted this remarkably descriptive name for the filtering process, however, and the interest in information filtering that has resulted from the present research thrusts in digital libraries arises at least in part from this tradition.

With the growth of the internet and other networked information, research in automatic filtering of networked information has exploded in recent years. Because of their low cost, large volume, and ease of recognizing new information, the most popular domains for research systems have been USENET News and electronic mail. The recent explosive growth of the World Wide Web has made this an interesting domain which has attracted some good research, although the information collection problem appears to make this a more difficult domain in which to conduct basic research on information filtering techniques. Another domain which has attracted considerable research interest is the annual Text REtrieval Conference (TREC) in which a standard text collection is used and a carefully controlled evaluation methodology is enforced. In TREC the information filtering task is referred to as "routing," adding somewhat to the confusion of terminology in this field. In fact, TREC recently adopted a special interest "filtering" track which adopts a different evaluation methodology but which conforms to the definition of filtering presented above. Commercial systems which filter newswire articles and other specialized information sources are becoming available as well. Filtering techniques will likely be applied to other domains such as images, sound and video in the future.

The distinction between information filtering and the more established field of information retrieval has proven to be the source of some confusion as well. Information retrieval broadly deals with the selection of information, and many of the features of information retrieval system design (e.g. representation, similarity measures or boolean selection, document space visualization) are present in information filtering systems as well. If one considers information retrieval from a very general "information selection" viewpoint, information filtering is simply a special case in which the information space is very dynamic. If, on the other hand, your personal definition of information retrieval involves selection of relatively static information in response to relatively dynamic queries, then information filtering is best

viewed as the dual problem to information retrieval. Regardless of which viewpoint you take, though, it is clear that researchers in information filtering will likely benefit from familiarity with the legacy of research in various aspects of information retrieval. For practical reasons I have not attempted to compile a comprehensive listing of network-accessible resources on information retrieval, however, so the interested researcher should refer to the Related Web Pages section of the Information Filtering Resources web page for some starting points on information Retrieval.

---

[Doug Oard](#) Last modified: Tue Dec 12 15:33:26 1995

# University of Maryland Information Filtering Project

The Information Filtering Project was a joint effort of the University of Maryland Electrical Engineering Department's [Medical Informatics and Computational Intelligence Laboratory](#), The Institute for Advanced Computer Studies' Computational Linguistics and Information Processing ([CLIP](#)) Lab and the College of Library and Information Services' [Digital Library Research Group](#), that extended from September 1993 through August 1996. Research on these topics is continuing, and information on the current work can be found [here](#).

## Our Web Pages

### [Information Filtering](#)

Links to what was at the time every known network-accessible resource on information filtering. New links are added as changes are noted, but this list is no longer comprehensive.

### [Cross-Language Text Retrieval](#)

Links to every known resource on cross-language text retrieval. Includes links to network accessible resources and a fairly comprehensive BibTeX file identifying published literature in the field. This page is still being maintained actively, and is fairly comprehensive.

## Papers and Talks

### [Alignment of Spanish and English TREC Topic Descriptions](#)

Poster paper presented at the Fifth Text REtrieval Conference (TREC-5), Gaithersburg MD, November 1996.

### [Evaluating Cross-Language Filtering Effectiveness](#)

Presented at the Cross-Linguistic Multilingual Information Retrieval Workshop at SIGIR-96, Zurich Switzerland, August 22, 1996.

### [Adaptive Vector Space Text Filtering for Monolingual and Cross-Language Applications](#)

A Ph.D. dissertation by Doug Oard that was completed in August 1996.

### [A Conceptual Framework for Text Filtering](#)

A selective survey of present practice in information filtering with an emphasis on defining the field and identifying significant research issues. The version linked above is HTML with links last verified in April 1997. The [postscript](#) version with the original URL's is also available. A greatly revised version will appear in the journal User Modeling and User Adapted Interaction in 1997.

### [A Survey of Multilingual Text Retrieval](#)

A survey of present practice in retrieval of texts in one language based on queries in another. More

recent papers on this subject are available [here](#).

### [Multilingual Information Filtering](#)

Some viewgraphs which provide a brief overview of the field, from a University of Maryland Digital Library Forum presentation on June 3, 1996.

### [Advanced User Models for Document Routing](#)

Viewgraphs from a Computational Linguistics Seminar presentation on April 25, 1996.

### [Experimental Investigation of High Performance Cognitive and Interactive Text Filtering](#)

Presented at the 1995 IEEE Conference on Systems, Man and Cybernetics, Vancouver, BC, October, 1995.

### [On Automatic Filtering of Multilingual Texts](#)

Presented at the 1994 IEEE Conference on Systems, Man and Cybernetics, San Antonio, TX, October 2-5, 1994.

### [A Survey of Information Retrieval and Filtering Methods](#)

A broad survey of recent research on techniques for information filtering and retrieval.

### [Filtering Networked Information Resources](#)

Viewgraphs from a presentation to the sixth annual meeting of the Special Interest Group on Networked Information Discovery and Retrieval in College Park, Maryland on March 24, 1995.

### [Information Filtering and Retrieval: Overview, Issues and Directions](#)

A background paper for a panel discussion at the 26th Annual Conference of the IEEE Engineering in Medicine and Biology Society, Baltimore, MD, November 3-6, 1994.

### [User Modeling for Information Filtering](#)

A position paper presented at the Fourth International Conference on User Modeling Special Interest Group on User Modeling in Information Retrieval, Hyannis, MA, August 17, 1994.

### [Neural Networks in Information Filtering and Retrieval](#)

An informal annotated bibliography of significant applications of connectionist networks to information filtering and retrieval.

## Project Members

- [Doug Oard](#)
- [Nicholas DeClaris](#)
- [Bonnie Dorr](#)
- [Christos Faloutsos](#)
- [Gary Marchionini](#)

# Related Web Pages at the University of Maryland

- [Digital Library Research Group](#)
- [Document Processing Group](#)

---

Last modified: Wed Jan 28 18:35:44 1998 [Doug Oard](#) oard@glue.umd.edu

# Cross-Language Information Retrieval Resources

This page is designed as a resource for people conducting research in [cross-language information retrieval](#). It is intended to collect references to all information on information retrieval systems which can accept queries in one language and return documents in another. It is maintained by the [Digital Library Research Group](#) of the [College of Library and Information Services](#) at the University of Maryland. If you are aware of resources that are within the scope of this page but do not appear here, please [send mail to Doug Oard](#).

## [December 1997 D-lib Magazine Article](#)

An introduction to cross-language information retrieval. A web page that was prepared for a [public lecture](#) here at Maryland provides another perspective on the topic that reflects some of my more recent thinking.

## [Conferences](#)

The best single source for information in the field. This page includes links to the full proceedings of every major cross-language information retrieval workshop as well as to a fairly complete list of upcoming conferences and workshops that include some treatment of cross-language information retrieval.

## [Cross-Language Information Retrieval Papers and Project Descriptions](#)

Another excellent place to look for information. Here you will find descriptions of experimental work on cross-language text retrieval that may not have been presented at one of the major workshops

## [Working Systems](#)

Here you will find links to experimental and commercial cross-language information retrieval systems that you can either obtain or use over the net. Some carry a fairly hefty price tag, others are free.

## [Bibliography](#)

A fairly comprehensive bibliography of published work on cross-language information retrieval in BibTeX form, last updated on July 3, 1997. The bibliography is also available in [postscript](#). Most of the references are described in at least one of my survey [papers](#) on cross-language information retrieval.

## [Related Resource Pages](#)

Web pages which collect links to resources that may be of interest to cross-language information retrieval researchers. None of these pages are devoted solely to cross-language information retrieval.

---

Last modified: Sat Apr 24 03:42:45 1999 [Doug Oard](#) oard@glue.umd.edu

# Multimedia, Representations:

---

## The Basics:

- [text file formats](#)
- [graphic file formats](#)
- [hypermedia & multimedia](#)

ACM DL'97 Tutorial: [Multimedia Information and Systems](#)

[ACM SIG on Information Retrieval](#) ; [ACM SIG on Multimedia](#) ; [IEEE-CS TC on Multimedia Computing](#) ; [Computing Curricula 2001](#)

## Digital Video

- [KRDL: Seamless Integration of Video Contents for Web-based Presentations over Different Devices](#)
- [KRDL: Video to SlideShow System \(ViSS\)](#)
- [CNN uses Quicktime for WWW daily news clips](#)

## MHIA Courseware and Curricula

- [Curriculum Resources in Interactive Multimedia \(CRIM\) Home Page](#)
- [MHIA Home Page](#)
- [SIGIR 96 Workshop](#)
- [Drexel 96 Workshop](#)
- [IR Courses](#)
- [Multimedia Courses](#) (Dublin, Ireland)
- [MM 1996 Workshop](#)
- [Lisbon 1997 Workshop](#)
- Questions:
  - What is the need for education related to information? What jobs?
  - What subjects should be covered in such education programs?
  - How should those subjects be ordered into each specific program?

---

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

---

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

**(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta**

# Architectures:

---

Core topics include:

- [D-Lib article on architecture](#)
- [Other CNRI activities](#)
- **Naming**
  - [PURL](#)
  - [Handles](#)
- [Networks](#): online notes of Dr. Lesk

Other topics of general interest, that are being studied by the [D-Lib Metrics Group](#) include:

- **Distributed processing (client/server)**
- **Interoperability** (see [IITA workshop on Interoperability](#) and some of work at [Stanford](#), as well as the [Open Archives initiative](#))
- **Performance**

---

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

---

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

(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta

# 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 architecture has been the subject of a series of useful discussions from which eight general principles have emerged; they are discussed in this introduction. These principles form the key issues in the transition to a true digital library from the network services that we have today. The Kahn/Wilensky paper also contains a comprehensive framework for resolving these issues.

## General Principles

- [1. The technical framework exists within a legal and social framework](#)
  - [2. Understanding of digital library concepts is hampered by terminology](#)
  - [3. The underlying architecture should be separate from the content stored in the library](#)
  - [4. Names and identifiers are the basic building block for the digital library](#)
  - [5. Digital library objects are more than collections of bits](#)
  - [6. The digital library object that is used is different from the stored object](#)
  - [7. Repositories must look after the information they hold](#)
  - [8. Users want intellectual works, not digital objects](#)
  - [Reference](#)
-

# General Principles

## 1. The technical framework exists within a legal and social framework

Early networked information systems were developed by technical and professional communities, concentrating on their own needs. The emphasis was on making information available to colleagues and the public, without charge. The digital library of the future will exist within a much larger economic, social and legal framework.

For example, musical works represent the livelihood of composers and musicians. Their artistic reputations depend on their work not being changed in storage or transmission. They require payment, as do recording studios and concert halls. Such work will only be part of the digital library, if the library supports their interests.

The legal system's task is to codify this rapidly changing economic and social framework. The relevant areas of law include copyright, performance, and other intellectual property, libel and obscenity, communications law, privacy, and international law.

The Kahn/Wilensky architecture can not write the law, but it provides a technical design that matches the legal structure that is expected to emerge. The architecture respects the creators and owners of intellectual property. It allows the preservation of rights that can last for more than one hundred years, and recognizes that digital works may include material from many sources, with separate property rights.

Society expects the creators of works to be responsible for their content, and for those who make decisions about content to behave responsibly. However, the digital library will not thrive if legal liability for content is placed upon parties whose only function is storage and transmission. Therefore, the architecture establishes clear boundaries between the areas of responsibility of the various parties.

## 2. Understanding of digital library concepts is hampered by terminology

Terminology proves to be a barrier in describing a digital library. Some words have such strong social, professional, legal, or technical connotations that they obstruct discussion between people of varying backgrounds. Simple words mean different things to different people. For example, the words "copy" and "publish" have different meanings to computing professionals, publishers, and lawyers. Common English usage is not the same as professional usage, and the versions of English around the world have subtle variations of meaning.

Certain words cause such misunderstandings that they are best expunged from any precise discussion of the digital library. The list includes "copy", "publish", "document", and "work". 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 stored in a digital object is called "content", which is divided into "data" and information about the data, 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 is the same. 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 servers is being added for increased local control.

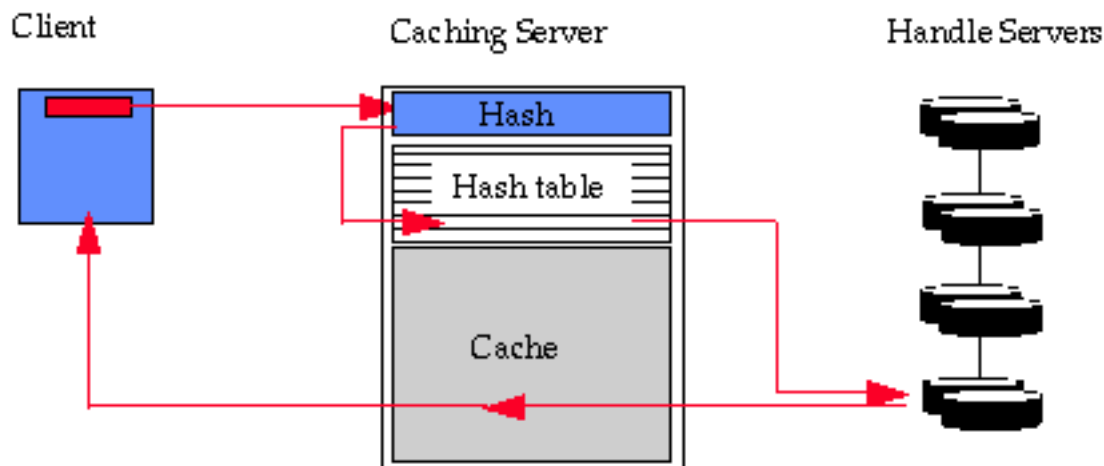


Figure 1. The CNRI handle system

## 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".

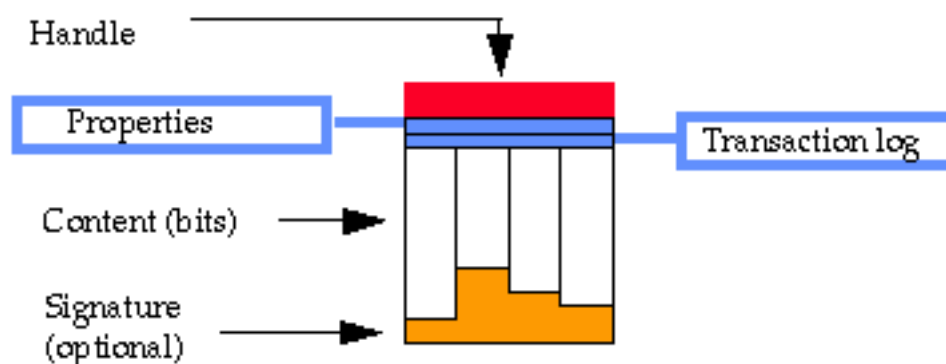


Figure 2. Parts of a digital object

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. 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 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.

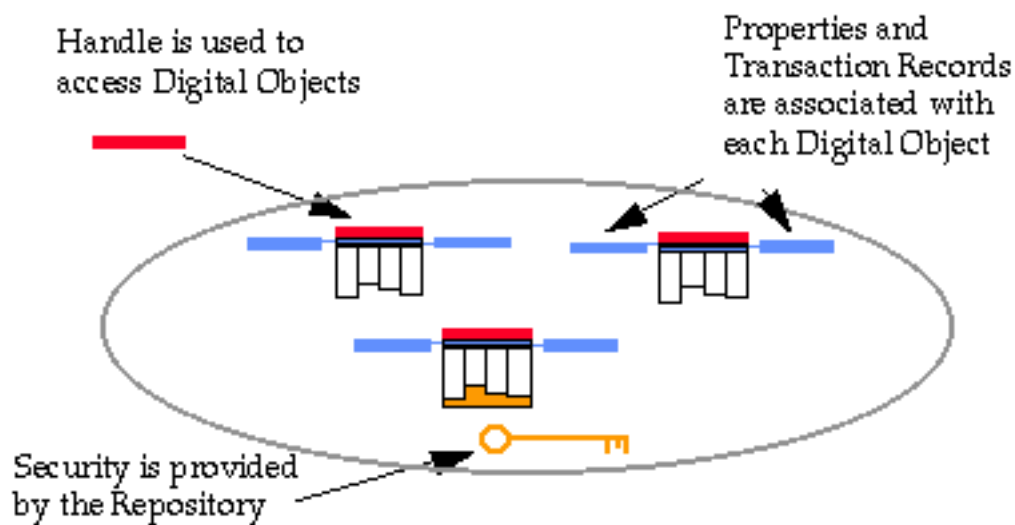


Figure 3. A repository

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 "report", "computer program", or "musical work", often refer to many digital objects that can be grouped together. The individual objects may have different formats, minor differences of content, different usage restrictions, and so on, but certain users are willing to consider them as equivalent.

Which digital objects should be grouped together can not be specified in a few dogmatic rules. The decision depends upon the context, the specific objects, their type of content and sometimes the actual content. The underlying architecture has to support two main needs. It must provide methods for grouping digital library objects and must provide means for retrieval.

The Kahn/Wilensky architecture supports these higher level ideas in several ways. One is to have a digital object containing several digital objects. Thus several formats of a text might be assemble into a single digital object. Another approach is to have these variants stored as separate digital objects, each with its own handle. These handles are contained in a digital object, known as a "meta-object", which acts like a catalog record. It contains a list of the variants with their handles and information about the differences amongst them.

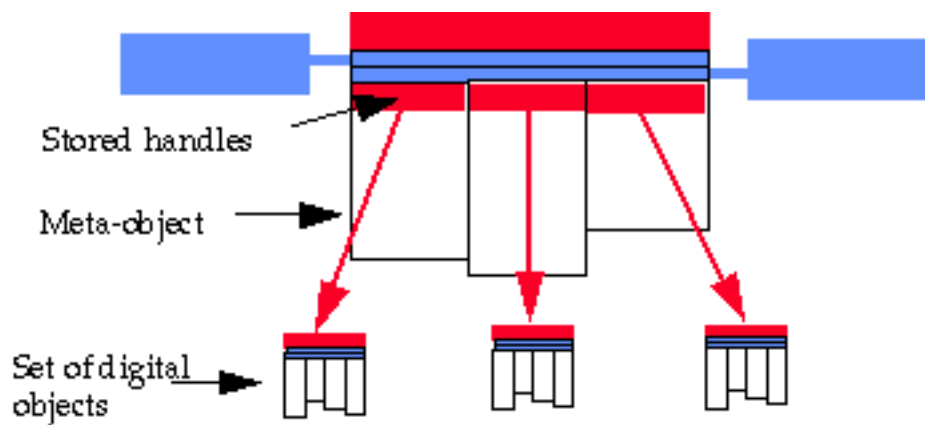


Figure 4. A digital object used as a catalog record

## Reference

[hdl:cnri.dlib/tn95-01](http://hdl:cnri.dlib/tn95-01) Kahn, Robert and Wilensky, Robert. "A framework for distributed digital object services". May, 1995. (<http://WWW.CNRI.Reston.VA.US/home/cstr/arch/k-w.html>)

Copyright © 1995 Corporation for National Research Initiatives

---

d-Lib forum

d-Lib magazine

---

*[hdl:cnri.dlib/july95-arms](http://hdl:cnri.dlib/july95-arms)*



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 Internet Engineering Task Force Uniform Resource Identifier working groups. There is nothing incompatible between PURLs and the ongoing URN (Uniform Resource Name) 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
- [Validate](#) PURLs NEW
- [Power](#) user's page (all features)

---

As of *Thu Jun 1 04:09:59 EDT 2000* : PURLs Created = **26981** , PURLs Resolved = **0** and Unique Client Systems = **0**

---

[The PURL Team](#)  
[purl@purl.lib.vt.edu](mailto:purl@purl.lib.vt.edu)

Corporation for National Research Initiatives

# HANDLE SYSTEM<sup>®</sup>

Home

Introduction

Software

Documentation

Support

Handle Resolver

**A general-purpose global  
name service enabling  
secure name resolution  
over the Internet.**

**Introducing the**

**JAVA<sup>™</sup> Version**

[Get the details▶](#)

| [introduction](#) | [software](#) | [documentation](#) |  
| [support](#) | [resolver](#) |

Updated: 11 Apr 2000  
[Corporation for National Research Initiatives](#)  
Contact: [hdladmin@cnri.reston.va.us](mailto:hdladmin@cnri.reston.va.us)

# *D-Lib* Working Group on Digital Library Metrics

---

## D-Lib Working Group on Digital Library Metrics

This Working Group is aimed at developing a consensus on an appropriate set of metrics to evaluate and compare the effectiveness of digital libraries and component technologies in a distributed environment. Initial emphasis will be on (a) information discovery with a human in the loop, and (b) retrieval in a heterogeneous world.

[Working Group Charter](#)

[Other Working Group Documents](#)

[Working Group Private Area](#)

This is an open working group, and anyone interested in the subject and in contributing to the work of the group is encouraged to join. For further information or to join the group, contact Barry Leiner <[BLEiner@riacs.edu](mailto:BLEiner@riacs.edu)>.

The Working Group sponsored a [Workshop on Digital Library Metrics](#), organized by [Bill Pottenger](#) and [Bob McGrath](#), and held 27 June 1998, just after the [DL'98 conference](#).

---

**D-Lib**

*The D-Lib Program is based at the [Corporation for National Research Initiatives](#) and is sponsored by the [Defense Advanced Research Projects Agency](#) (DARPA) on behalf of the Digital Libraries Initiative under Grant No. N66001-98-1-8908.*

*prepared by [Barry Leiner](#)*

*last modified 3/21/00*

*bl/bw*

# 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](mailto:clifford.lynch@ucop.edu))

Hector Garcia-Molina ([hector@db.stanford.edu](mailto:hector@db.stanford.edu))

*Converted to HTML using GradStudentWare 2.2*

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

[Introduction](#)

[Definitions and Roles of Digital Libraries](#)

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

[Infrastructure Requirements for Digital Library Research](#)

[Research Issues and Priorities](#)

[1. Interoperability](#)

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

[3. Collection Management and Organization](#)

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

[Conclusions](#)

[Executive Summary](#)

[Appendix 1 - List of Participants](#)

[Appendix 2 - Strawman Report](#)

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

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

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

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

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

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

# Introduction

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

While there have been a number of workshops and other meetings examining the broader questions of support for applications in the National Information Infrastructure (NII), we believe this was the first workshop that focused specifically on Digital Libraries in this context. In the past year, Digital Libraries have emerged as one of the central and most compelling applications enabled by the NII; numerous digital library research projects are underway, including six large-scale pilot projects that have been funded jointly by ARPA, NASA, and NSF. While Digital Libraries are now a vibrant research area, and also a field in which considerable commercial development is taking place (presaging the future economic importance of Digital Library technology to the United States), many new questions are emerging as a result of this flowering of research activity. Informed by insights gained from current research, this workshop offered an opportunity to consider questions such as interoperability objectives that might be defined among projects now underway.

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

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

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

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

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

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

Michael Lesk,  
Bellcore:  
The Commercial Perspective

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

Mike Schwartz,  
University of Colorado:  
The Internet Perspective

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

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

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

## Definitions and Roles of Digital Libraries

Considerable work has already been done on operational definitions of Digital Libraries and their relationship to traditional library institutions, as well as to the broader systems of scholarly and

commercial publishing (see, for example, Communications of the ACM, April 1995). Much of the discussion in this workshop was motivated by questions of scaling, interoperability and needed support infrastructure.

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

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

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

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

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

Libraries -- digital or traditional -- exist to serve diverse purposes and constituencies. To some extent,

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

## Defining Interoperability in the Digital Library Environment

Defining interoperability proved difficult. It is clear that this is still a central research problem in its own right, and one that merits continued attention. Discussions of infrastructure focused on common tools, enabling technologies and standards that would provide a basis for further exploration of interoperability issues, particularly by encouraging and facilitating the growth of digital libraries on the Internet. Considerable effort was spent on identifying infrastructure that was either unique or particularly critical to progress in digital libraries, as opposed to more general-purpose infrastructure that a range of NII applications, including digital libraries, might share. One clear theme was that an understanding of interoperability issues required operational experience which could only be gained by large-scale deployment of digital library systems. Speculation about interoperability in the abstract is of very limited value.

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

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

# Infrastructure Requirements for Digital Library Research

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

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

## Research Issues and Priorities

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

### 1. Interoperability

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

The more technical interoperability research involve protocol design that supports a broad range of interaction types, inter-repository protocols, distributed search protocols and technologies (including the ability to search across heterogeneous databases with some level of semantic consistency), and object

interchange protocols. Interoperability is not simply a matter of providing coherence among passive object repositories. Digital library systems offer a range of services, and these services must be projected in an interoperable fashion as well. One particular issue that emerged was that existing Internet protocols (such as HTTP, the basis of the World Wide Web) are clearly inadequate. Research must move beyond the current base of deployed protocols and systems. This raises complex questions about how to deploy prototype systems and the tradeoffs between advanced capabilities and ubiquity of access.

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

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

## 2. Description of Objects and Repositories

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

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

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

approaches.

### 3. Collection Management and Organization

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

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

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

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

### 4. User Interfaces and Human-Computer Interaction

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

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

## 5. Economic, Social, and Legal Issues

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

## Conclusions

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

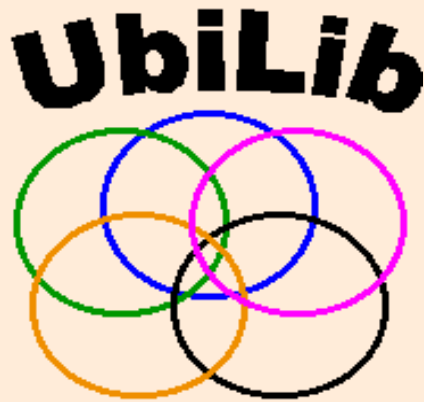
Scaling was identified as a major area of concern. The common vision is one of tens of thousands of repositories of digital information that are autonomously managed yet integrated into what users view as a coherent digital library system. Accommodating this very large number of repositories -- a very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

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

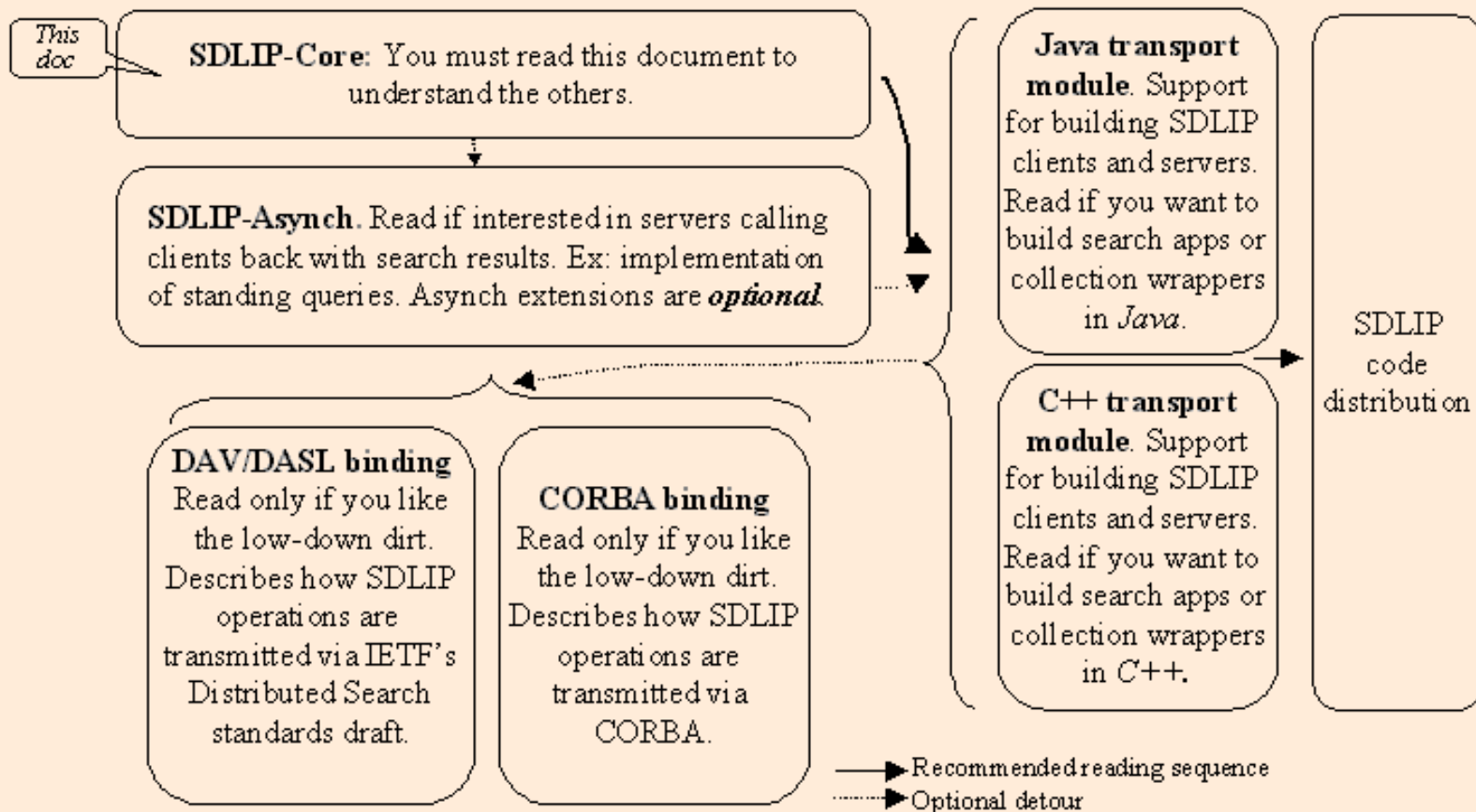
library system. Establishing infrastructure and tools to facilitate experimentation with large-scale systems is essential, as is funding to study use and behavior of large-scale systems once deployed through this infrastructure. The Internet as a context for deploying digital library systems offers an 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.



# The Simple Digital Library Interoperability Protocol (SDLIP-Core)

**SDLIP document map and recommended reading sequence (click to navigate):**



## Contents:

### [1. Introduction and Overview](#)

#### [1.1 Grouping of Operations Into Interfaces](#)

#### [1.2 Different Ways of Using SDLIP](#)

#### [1.3 When Can Servers Discard State?](#)

#### [1.4 Implementation Architecture](#)

### [2. SDLIP Operations in Detail](#)

#### [2.1 Search Interface](#)

#### [2.2 The Result Access Interface](#)

[2.3 The Source Metadata Interface](#)[3. XML Formats Used in SDLIP](#)[3.1 Property Lists](#)[3.2 Exceptions](#)[3.3 SearchResult](#)[3.4 Subcollection Specifications](#)[3.5 Source Metadata](#)[3.6 Server Delegates](#)[4. Implementing SDLIP With IETF's DASL](#) (details in separate document!)[5. Implementing SDLIP With CORBA](#) (details in separate document!)[Appendix A: Error codes and their meanings](#)

# 1. Introduction and Overview

This document describes the Simple Digital Library Interoperability Protocol (SDLIP; pronounced S-D-Lip). Clients use SDLIP to request searches to be performed over information sources. The result documents are returned synchronously, or they are streamed from service to client as they become available. Implementations can be constructed over HTTP or CORBA based transports. In fact, any search service can be accessible through both kinds of transports at the same time. Implementations for IETF's HTTP based DASL protocol, and for CORBA are available.

Figure 1 shows a typical example of where SDLIP is relevant.

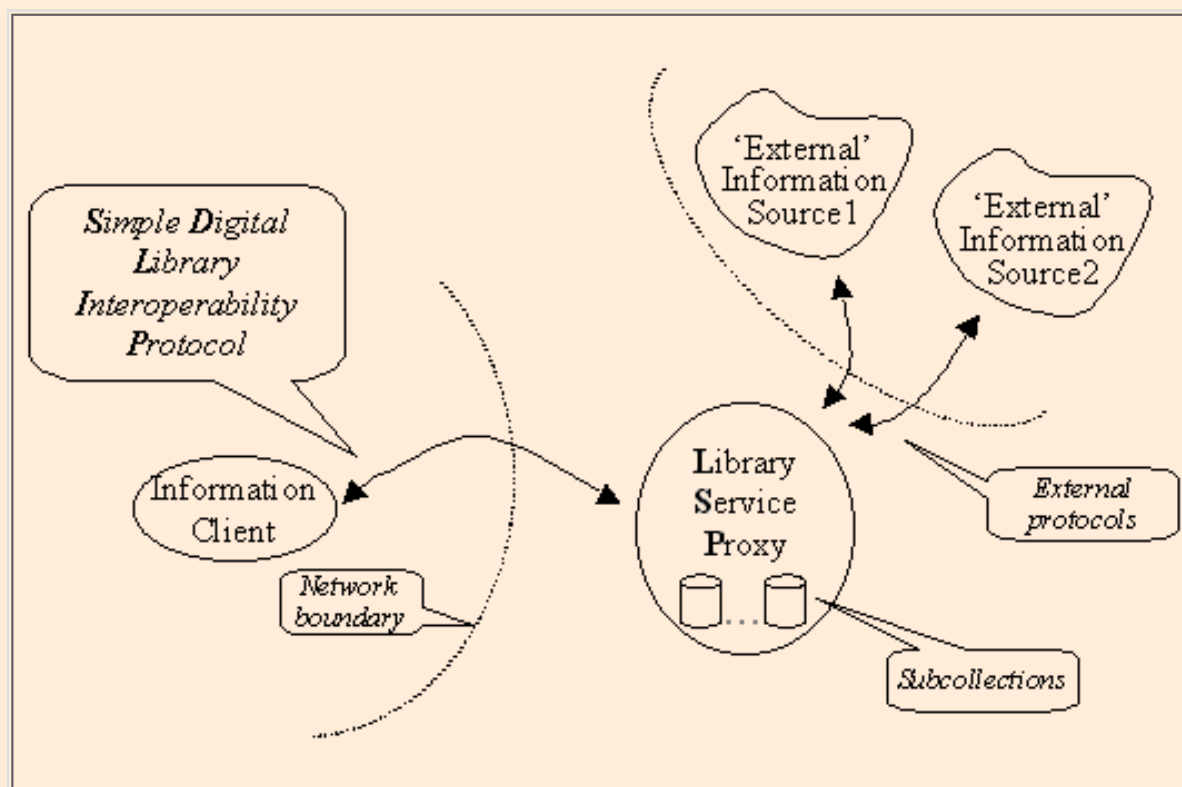


Figure 1: The Role of SDLIP in a Digital Library Architecture With Autonomous Sources and Wrappers

The dotted line in Figure 1 indicates a network boundary: entities on the same side of the line are assumed to be in the same address space. Note in Figure 1 that the information to be served is stored in repositories that do not (necessarily) implement SDLIP. This is a typical scenario, because information sources are often autonomously maintained, and do not present uniform

interfaces to programs trying to extract information from them. Examples for external, non-conforming information sources are Web search engines, library catalogs, and commercial information providers, such as Nexus-Lexus, or the Dialog Corporation.

The 'Library Service Proxy' (LSP) in Figure 1 wraps two external sources. Through its back end, the LSP interacts with the external services via the transport and higher-level protocols required for these services. One LSP may thus serve out multiple 'subcollections'. At the front end, the proxy supports SDLIP. Of course, an information source may itself provide SDLIP access. In that case, the client can interact directly with the source.

The basic interaction is for the client to request a search across the network. Part of the request specifies how many documents are to be returned initially, once the search will be complete. The request also specifies which portion of each document is to be returned. For example, the client might ask for authors and titles of the first 10 documents to be returned right away. The client may later request more documents of the result, or it may request additional portions of the documents already delivered.

We define two levels of SDLIP capabilities: SDLIP-Core implements synchronous operations only. Clients invoke search operations on servers, and 'hang' until the operations return with the result. This document focuses on SDLIP-Core. The second level, SDLIP-Asynch adds the ability for clients to invoke search operations that return immediately. Services then deliver result information back to the client through one or more callbacks. SDLIP-Asynch thus subsumes SDLIP-Core. SDLIP-Asynch's additional capabilities are described in the separate [SDLIP-Asynch document](#).

SDLIP has the following goals:

- Simplicity for both client and server side implementations
- Implementations possible via both distributed object technology, such as CORBA, and via HTTP
- Support for stateful and stateless operation at the server side
- Support for dynamic load balancing in server implementations
- Support for thin clients, such as handheld devices

## 1.1 Grouping of Operations Into Interfaces

Figure 2 shows how the SDLIP operations are divided into three interfaces.

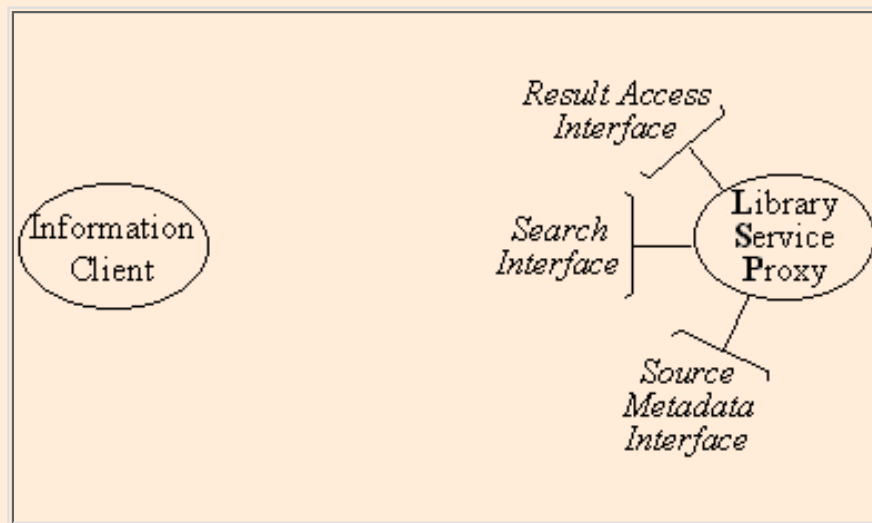


Figure 2: SDLIP-Core Consists of Operations Grouped into Three Interfaces

The search interface on the service contains the operations needed for submitting a search request to the service.

The result access interface allows client applications to access the set of result documents, wherever that set is maintained. The source metadata interface, finally, allows clients or services such as metasearch engines to question a library service proxy about its capabilities. This might include a list of the subcollections served by the LSP, or the attributes that may be searched.

The partitioning into interfaces has three advantages. First, the interfaces make it clear which role each operation plays, and for which participants of the search transaction the operation needs to be implemented. Second, the interface notion enables clean expansions to the protocol in the future. One can subclass the existing interfaces to accommodate more elaborate facilities, or

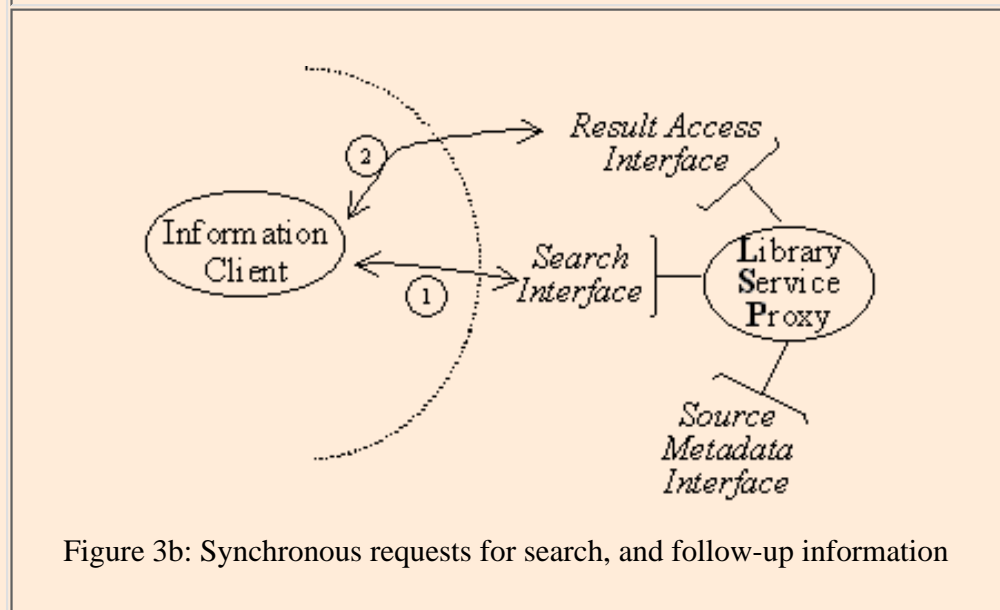
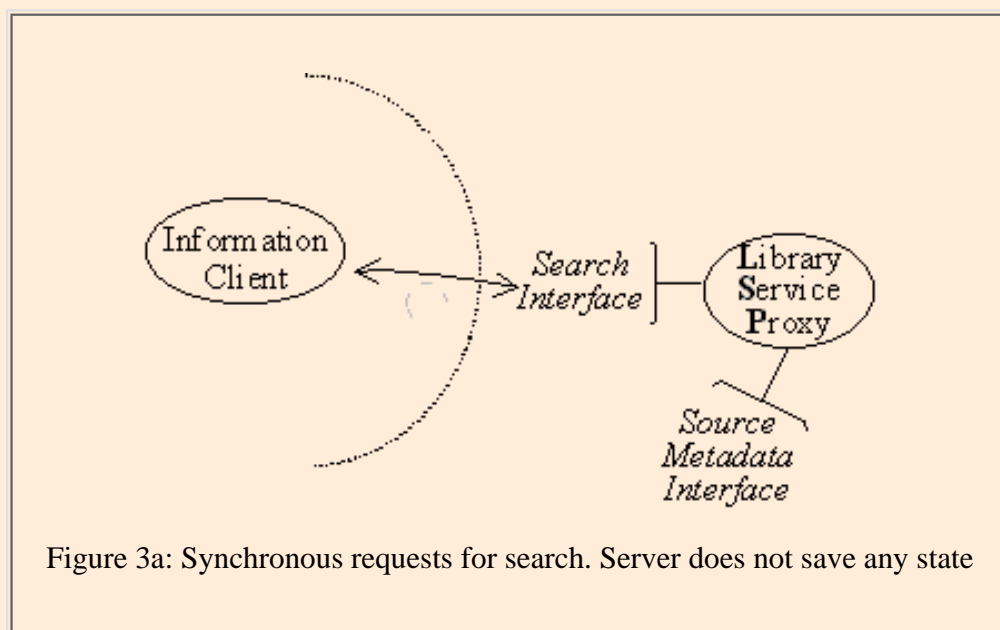
one can add additional interfaces. For example, one could use interface inheritance to add operations to the source metadata interface, if in the future some LSPs wish to export additional metadata, or wish to export that data in some new format. Or maybe one might want to add a whole new interface for financial transactions. Neither of these expansions would impact the existing core protocol. A third advantage of organizing SDLIP's operations into functionally coherent interfaces is that for some scenarios, or 'configurations', some of the interfaces are not needed. Rather than having to list various operations to be dropped for these cases, we can then simply say that interface X is not needed. For example, if a server is stateless, it does not need to implement a result access interface, because all results are returned in the operations of the search interface.

The minimum a stateless SDLIP server needs to implement is the search interface. Clients can rely on it being present. If a server maintains result sets which clients can access, then the server also needs to implement the result access interface. Though not required, all servers should implement the source metadata interface. As documented below, this is a very simple interface to provide.

## 1.2 Different Ways of Using SDLIP

Figure 3 illustrates how SDLIP-Core can be used in three configurations. The simplest is the configuration of Figure 3a. It features one library service proxy serving the information, and a single client application object. The client submits the search request synchronously via the service's search interface. The results are returned as part of that call.

Figure 3b shows a somewhat more sophisticated usage in which the server maintains the result set of the search, at least for a while. Later, the client might, again synchronously, ask for more documents of the same result set (2).



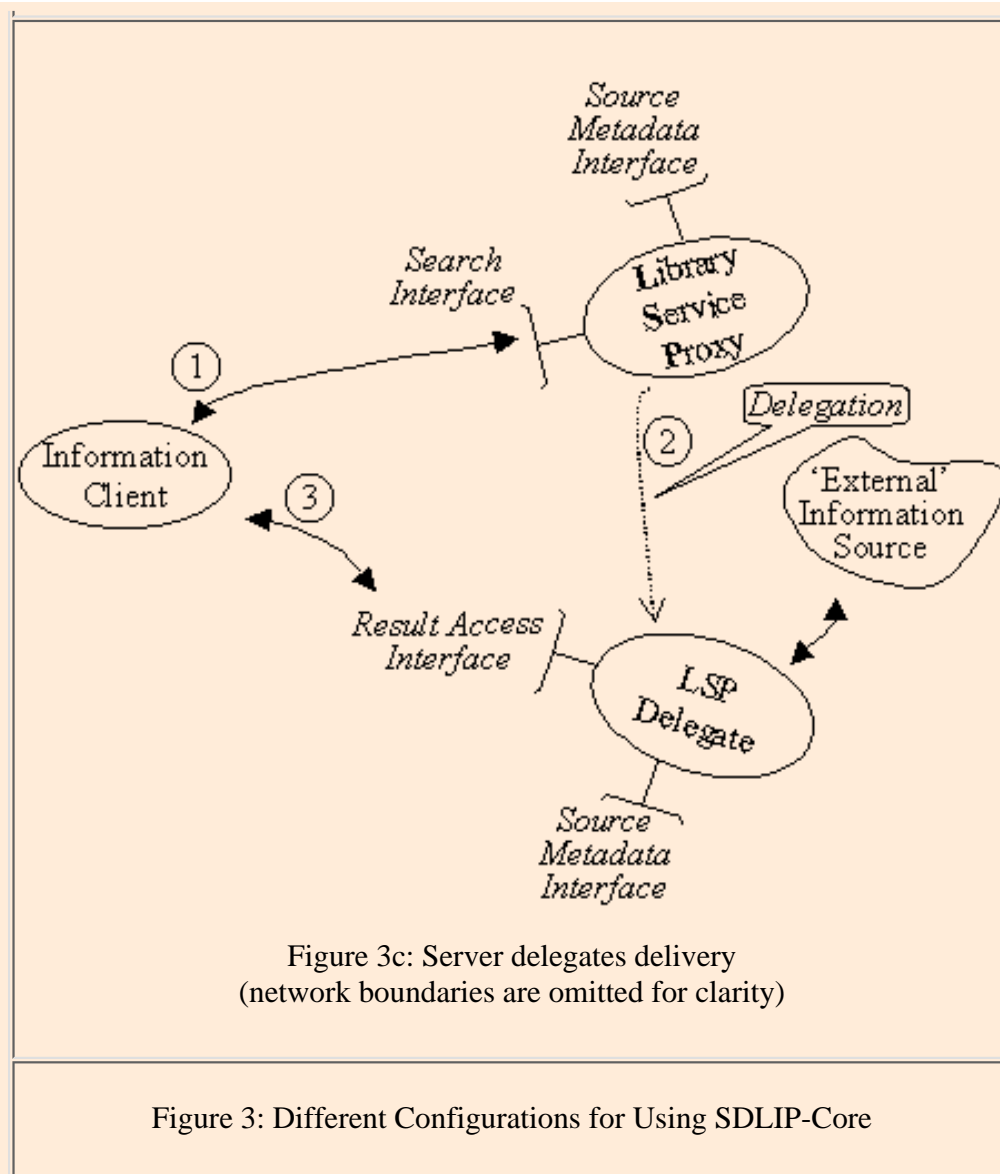


Figure 3c, finally, illustrates how services can delegate interactions with clients if the service object gets overloaded (2), yet wishes to maintain state for the client. When servers return (partial) results from a search operation, they can also specify a future contact address (3). All future interactions regarding the result set are made by the delegate. If the client wishes, for example, to recall more documents of the result set than it asked for in its initial search request, then it will use the delegate's address, rather than the main address (4). More configurations are possible if the [asynchronous SDLIP extensions](#) are also supported.

### 1.3 When Can Servers Discard State?

If SDLIP claims that it enables both stateful and stateless implementations of servers, how do clients and servers agree on whether or not there is state at the server? The notion of a 'state parking meter' takes care of this. It is a very simple notion.

When clients submit a search request to a server, they include the amount of time they would like the service to retain the state associated with the session. The server returns the actual time it is willing to maintain the result set and related state. For a completely stateless server, this time could be zero. The clock starts ticking right after the search call returns. Once the time has expired, the server is free to discard all state associated with the search. The client, meanwhile, has the option of invoke an 'extend state timeout' operation on the server to add additional time. The server has the option of granting or refusing the request. This is rather like the client feeding a parking meter. Note that this scheme ignores some uncertainties in that the server and client clocks might not be synchronized. Also, the server starts its clock when it returns from the search call, while the client starts counting down when the return process is complete. Since state maintenance times are expected to be large compared to these differences, the issue is ignored in favor of simplicity.

Of course, if clients and servers are to converse about a result set, they need the ability to refer to the set, and to (potentially multiple, parallel) operations the client invokes on the service. These references are provided through a server session ID, and client request IDs, respectively: One of the values a server returns with every new search request is a 'cookie' that uniquely identifies the result set within the server. The client must pass this session ID to the server whenever the client requests more information from the result access interface.

In addition to referencing a result set, clients may need to reference operations they have invoked, and that have not returned. For example, the client might use multiple threads to invoke several follow-up requests to an existing result set. The client might then wish to use an additional thread to cancel one of the hanging invocations. In order to identify multiple requests, many of the SDLIP operations include a client-side request ID. Like the server session ID, client request IDs are cookies. The server may compare request IDs for equality, but other than that, these IDs are opaque to the server. In addition to the server request ID, clients pass their own request ID with every result access request. When cancelling an operation, clients use these two IDs to uniquely identify which operation they want to stop.

Notice that this scheme does not require server or client to generate any globally unique identifiers. Server session IDs must be unique only within the server. Similarly, client request IDs need to be unique only within the client.

## 1.4 Implementation Architecture

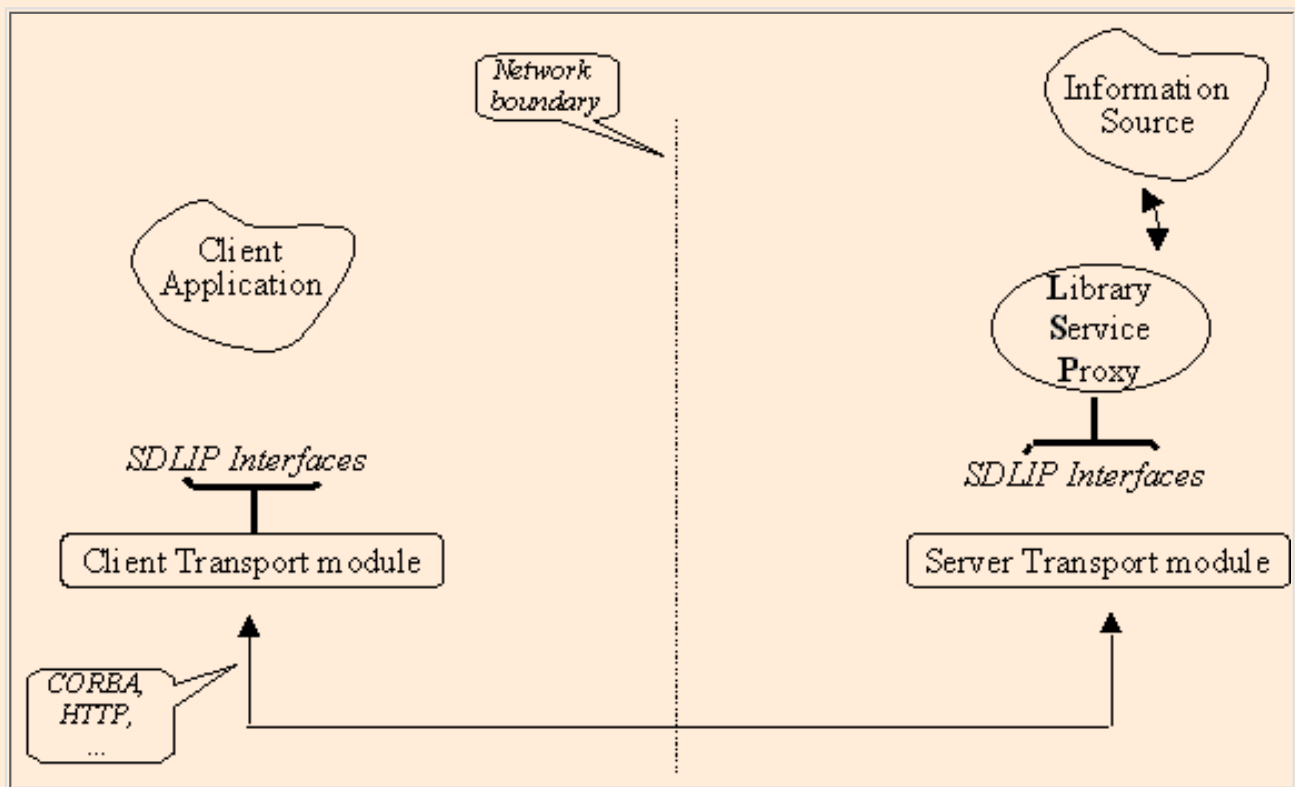


Figure 4: SDLIP Implementation Architecture

One of SDLIP's key goals is to make it very easy to build SDLIP clients, and to construct library service proxies (LSPs) that wrap arbitrary sources. Figure 4 shows how SDLIP implementations accomplish this. Implementors need to produce only the client application and/or the library service proxy in figure 4. Everything else is taken care of by standard libraries. An important point: Client applications and services need not be aware of the methods used for transporting operation requests and replies. The transmission of requests and replies might be accomplished through different 'transport bindings': CORBA, HTTP, or, maybe in the future, some other means. Client applications are unaffected by the transport binding. A client application merely creates a *client transport module* object in its local address space. This module implements the SDLIP interface. The client then invokes SDLIP operations on this local module. The module packages the operations for transport via one of the supported SDLIP transport bindings. Any given client transport module instance uses one particular transport binding. If a different binding is to be used, the client application simply instantiates a different class of transport module.

Implementation of an LSP is analogous. The LSP is an object that implements one or more of the SDLIP interfaces. It is similar in spirit to Web servlets. The LSP's operations are invoked locally by the server transport module object. Again, the

details of transport are of no concern to the LSP. One LSP could provide service via different transport modes simply by instantiating two kinds of server transport modules. That is, an LSP could make itself available via both CORBA and [\[DASL\]](#) transports at the same time, simply by instantiating a CORBA server transport module and a DASL transport module.

Please see the documentation for the transport module libraries ([Java](#) or [C++](#)) for examples of code.

Let's get to specifics. [Section 2](#) will describe SDLIP operations in detail. [Section 3](#) explains the XML structures used with SDLIP. [Section 4](#) sketches how the IETF Distributed Authoring, Searching and Locating (DASL) facility can be used as an SDLIP transport. [Section 5](#), finally, summarizes SDLIP's mapping to a CORBA transport. The [Appendix A](#) lists the error codes used in SDLIP.

## 2. SDLIP Operations in Detail

We describe each interface in turn. For each interface, we list the associated operations, and explanations for each parameter. Whenever a parameter is a specially encoded XML string, we just state its purpose. [Section 3](#) defines the XML formats in more detail. For clarity, we do summarize the simple XML property list format ahead of time below. A concise Interface Definition Language (IDL) specification is available in the combination of [SDLIPCore.idl](#) and [SDLIPLocal.idl](#).

Some technologies that might be used to implement an SDLIP transport allow parameter defaulting. For CORBA transports, defaults are NULL values. For HTTP based transports, a defaulted parameter is simply left out. SDLIP enables implementations to make use of such defaulting facilities by specifying default values for as many parameters as possible. Apart from saving on required bandwidth, these defaults also make it easier to gracefully degrade interactions in which either the client or the service are not truly SDLIP compliant. For example, DASL clients might interact with SDLIP services. These clients will not include some of the parameters in their search requests. Defaulting these parameters ensures that the SDLIP servers can still provide a reasonable level of service. Whenever a parameter can be left out in the specifications below, its default value is specified in curly braces in the parameter explanations.

First, a couple of conventions we follow, and some very brief preview hints to set the reader at ease about how these operations work in an implementation.

**Entity Addresses:** When we use the word 'address' to specify the target of a method invocation, we could use the term 'object identifier' (OID). We instead use the term 'address', so that the mapping to HTTP based implementations is more obvious. The realization of 'address' in that case is, of course, a URL.

**Property Lists:** Some of the method parameters below are property lists. These are XML-encoded lists of attribute value pairs. For example:

```
<propList>
  <QualityOfService>Fastest</QualityOfService>
  <UserID>Miller</UserID>
</propList>
```

We use property lists as a catch-all expansion facility. Since property lists can be as large as needed, they are a great way to take care of special needs that arise in the future, and cannot be included in a core protocol, such as SDLIP. For the formal DTD of property lists, see [Section 3.1](#).

**XMLObject:** We introduce the following type that is used when an operation's parameter is supposed to be an XML-encoded string:

```
interface XMLObject {
  string getString();
  void   setString(in string XMLStr);
};
```

All XML-encoded strings are packaged into an XMLObject. So, in the specifications below, you might see:

```
...
void search() {
```

```
XMLObject subcols, // Choice of collections to search w/in LSP.
...
}
```

This parameter specification calls an `XMLObject` to be passed into the call. This object is to contain an XML string with the subcollections to be searched. Again, the formats of all XML strings are explained in [Section 3](#). Implementations of this interface come with the standard SDLIP transport module libraries.

In its simplest form, `XMLObject` just stores an XML string and provides the two operations for setting and getting that string. Individual implementations may subclass `XMLObject` and provide richer services to SDLIP client/server applications. For example, the [transport modules for DASL](#) and [CORBA](#) provide some of the facilities specified in the Document Object Model (DOM). This makes it very easy for clients to extract values from the XML they receive as results of SDLIP operations, and to construct valid XML to pass into the calls.

If an XML element is to contain binary data, such as images, maps, or other digital objects, it needs to be encoded in base64 [\[BASE64\]](#). Base64 character data must be wrapped into an `SDLIP:base64` element:

```
<my:bindata>
  <SDLIP:base64>SGV5LCB5b3UglzIG1lc3Mh</SDLIP:base64>
</my:bindata>
```

The syntax of the tag '`SDLIP:base64`' uses the notion of XML name spaces. The tag denotes the identifier 'base64' in the 'SDLIP' name space. Tags without name space specifications are understood to be in the SDLIP name space. For example, the `<propList>` tag that introduces a property list really stands for `<SDLIP:propList>`. Clients and server programs do not need to include the SDLIP name space specification when constructing the XML tags that are used for the built-in SDLIP XML structures.

**Error reporting (Exceptions):** SDLIP's approach to exceptions is to be as robust as possible. This means that if a request can be filled at least partially, the operation proceeds. Of course, if something really does go wrong, LSPs may raise exceptions in response to operations invoked on them. The information returned with an exception is always an XML-encoded string as defined in [Section 3.2](#). It includes information about the type of exception, a short description, and possibly additional debug information. Like all other return information, exceptions travel from the server-side LSP to the client via the transport modules. These modules are responsible for transmitting the exceptions across the wire. When the client transport module receives an exception it raises an error condition for the client application to catch. This is done using the error signaling facility of the programming language common to the transport module and the client application. Examples are the C++ throw/catch mechanism, or the analogous facility in Java. Exception objects have the following interface:

```
interface SDLIPException {
    int getCode();
    int getReason();
    XMLObject getDetails();
}
```

## 2.1 Search Interface

The `search()` method is the basic way of submitting a search to a server. We distinguish between IN parameters and OUT parameters. IN parameters hold information passed to the callee. All information that operations return to callers are passed back in OUT parameters. Consequently, the return type of all operations is `void`. For distributed object implementations, OUT parameters are a familiar notion. The SDLIP DASL binding specifies how OUT parameters are returned via HTTP.

```
void search(

    Long clientSID,           // {0} Client-side session ID (unique within client)
    XMLObject subcols,        // {service's default (or sole) subcollection}
                               // Choice of collections to search w/in LSP.
    XMLObject query,          // The query.
                               // (e.g. <ADL:GazeteerLang>Lake Tahoe</ADL:GazeteerLang>)
    Long numDocs,             // {-1} Number of documents to return right away (-1: all)
```

```

XMLObject docProps,          // {all possible properties} Properties to return
                                // for each result doc
                                // (e.g. <propList><Abstract/><Title/>, ...)
Long stateTimeoutReq,        // {0} Request for number of seconds to
                                // maintain state at server. -1: request unlimited time
XMLObject queryOptions,      // {none} Additional info for the LSP.
OUT Long expectedTotal       // {0} -1 if unknowable. -2 if not yet known.
OUT Long stateTimeout,       // {0} Time server is willing to maintain state
OUT Long serverSID,          // {0} ID by which server identifies this session
OUT XMLObject serverDelegate, // {same as for original query} For followup requests
OUT XMLObject result         // XML-Encoded result list.
)

```

The client invents a `clientSID`, which allows the service proxy to refer to this query request later on. This ID only needs to be as unique as the client requires. LSP implementations use their own internal mechanisms to separate sessions with different clients.

The `subcols` parameter is used when one LSP serves out many collections. This is sometimes true for commercial information providers, or for Z39.50 sites. If an LSP only serves one source of information, this parameter can be empty. A special case of subcollection are the result sets of previous searches: For the purpose of query refinement, clients must be able to specify such existing result sets within the LSP. The subcollection string is formatted like this (for details on server `sessionIDs`, see later in this document):

```

<subcols>
  <subcolName>[subcollection name]</subcolName>
  <resSet>[server sessionID]</resSet>
  <resSet>[server sessionID]</resSet>
  <subcolName>[subcollection name]</subcolName>
  <resSet>[server sessionID]</resSet>
</subcols>

```

One or more result sets or subcollections may be specified in any order. Of course, the result sets referenced must still be 'alive', that is clients must have fed the state timeout parking meter. See [Section 3.4](#) for details on the format of this string.

The query parameter contains the query itself. This is an XML string whose outermost tag names the query language being used. The value inside depends on the query language. For example, a query for a Dialog Corporation database might look like this:

```

<Dialog:StandardQuery>
  au=Miller and py=1994
</Dialog:StandardQuery>

```

The same query issued using the DASL basicsearch query language might look like this:

```

<basicsearch xmlns="DAV:" xmlns:Dialog="http://dialog.com/">
  <where>
    <and>
      <eq>
        <prop><Dialog:au/></prop>
        <literal>Miller</literal>
      </eq>
      <eq>
        <prop><Dialog:py/></prop>
        <literal>1994</literal>
      </eq>
    </and>
  </where>
</basicsearch>

```

The `numDocs` parameter specifies how many documents the LSP is to return initially. A value of '-1' means 'return all documents that are found'. Remember that the client may use the result access interface later on to request additional documents.

The `docProps` parameter is a property list. Properties are the names of document properties the LSP is to return for each of the result documents. One example value is:

```
<propList>
  <Title/>
  <Author/>
</propList>
```

This is a somewhat funny looking property list: none of the properties have values. This lack of values is indicated by the trailing '/'. This syntax is standard XML.

A more involved example for a document property specification is:

```
<propList>
  <USMARC:245/>
  <DublinCore:Creator/>
</propList>
```

The detailed format of the property names is not part of SDLIP. But SDLIP does assume that an XML namespace notation may be used to introduce the 'attribute model' within which the name of the respective attribute should be interpreted. For example, USMARC:245 is assumed to denote 'author' in the Library of Congress' USMARC attribute model.

The `stateTimeoutReq` is the number of seconds the client would like the server to hold on to the result set. After that time, the server may discard the result state. A value of -1 requests that the server hold state indefinitely, or until the client calls `cancelRequest()`.

The `queryOptions` parameter is a property list (i.e. a `<propList>` XML structure) that is not further defined by SDLIP. Clients and services may use properties to hold additional information regarding the query being transmitted. For example, the property list might be used to pass authorization information, financial arrangements, or quality of service specifications to the LSP.

The remaining parameters are all OUT parameters. They contain the following information.

The `expectedTotal` is the total number of hits found. Sometimes, servers cannot tell right away how many hits will be found. In this case, `expectedTotal` is set to -2. If, for example, the client asks for only the first 10 hits to be returned right away, then a stateful server may return right away, once the first 10 hits have been retrieved from the underlying collection. The server would then continue to build its result set while the client processes the initial results. The client can later use the result access interface's `getSessionInfo()` to find out the final number of hits. Sometimes, a total number of hits will never be known. Servers indicate this by setting `expectedTotal` to -1. Example: a service that takes a single query and keeps filling a result set forever. A news subscription service might do this. Clients would pull information from the result set at their convenience.

The `stateTimeout` parameter is the number of seconds the server is willing to hold the state. Recall that this number may be different from the number of seconds requested in the `stateTimeoutReq`. In particular, it may be zero if the server is stateless.

The `serverSID` is the session ID the server uses to identify this session. All correspondence with the server regarding this session must include that ID. Using a separate session identifier for the client and the server avoids having to invent globally unique IDs.

The `serverDelegate` is important only for servers that are stateful *and* are performing load balancing. This parameter is an XML structure listing the addresses of server-side delegates that are willing to serve follow-on requests over the result set. Much of the time, this parameter will be defaulted. The default means that follow-on requests are to be directed to the same address as the original query. When not defaulted, the format of this return parameter looks like this:

```
<redirect>
  <serverDelegate>[URL_or_IOR_1]</serverDelegate>
```

```
<serverDelegate>[URL_or_IOR_2]</serverDelegate>
...
</redirect>
```

The client can pick one of the delegates and use it for follow-on requests: each of the delegates implements the result access interface. Client applications can use the transport module facilities to make this switch easy: The module contains a static method that takes one of the 'URL\_or\_IOR' strings and returns a transport object of the right kind. The client application can then invoke the result access methods on that object.

The `search()` operation blocks until the LSP has finished setting up the result set. Then the result is returned in the `result` OUT parameter. This return type is an XML string that contains a list of `numDocs` documents (if that many were indeed found). For each document, only the attributes specified in `docProps` are returned. The format of the search result type is explained in [Section 3.3](#). Note that if some of the requested properties are not available, the server still returns the other requested properties that can be retrieved.

## 2.2 The Result Access Interface

Once some of the documents have been returned, clients might want to get more documents than they had originally requested, or they might need to request additional properties of documents they already have. This is accomplished through the result access interface.

The `getSessionInfo()` allows clients to find out about the result set, especially if the initial return parameters of the `search()` operation could not return the total number of hits:

```
void getSessionInfo()
    Long serverSID,           // {0}
    OUT Long expectedTotal, // {0} -1 if unknowable. -2 if not yet known.
    OUT Long stateTimeout   // {0} Total number of seconds server is willing to
                           // hold state. -1 if forever.
)
```

The `expectedTotal` return parameter is -1 if the remote LSP has indicated that the total number of documents cannot be determined. If the callee simply does not know yet, but there is a chance that it will find out later, a -2 is returned.

The `stateTimeout` is the total number of seconds that the server is willing to hold the state without receiving an `extendStateTimeout()` call.

The `getDocs()` operation is the means by which clients ask for more documents, or for additional portions of partially transferred documents in a result set. The key notion is that client and server think of the result documents as being arranged in order within a result array. Documents are referenced by their index into that array.

```
void getDocs(           // Returns a SearchResult XML string
    Long serverSID,      // {0} ... of the original query
    Long reqID,          // {0} new each time
    XMLObject docProps,  // {all possible properties} properties to get.
                        // A property list.
    String docsToGet,    // {1-} document indexes to retrieve. 1- for all.
    OUT XMLObject result // the XML-encoded search result.
)
```

The `serverSID` is the session number that was established during the original search request. The `reqID` is an identifier for this particular request within the overall session. It can be used to cancel this particular delivery request via a separate thread (see `cancelRequest()`).

The `docProps` is the same kind of parameter that is used in the `search()` call: it specifies which document properties should be included with each document.

The `docsToGet` parameter specifies which documents to get. Documents are identified by their index into the result array.

The array is one-based. We make the array start with the index '1' to allow document range descriptions that are familiar from document print dialogs in common user interfaces. For example, "1,3,5-7" will retrieve documents 1, 3, 5, 6, and 7. A dangling "-" after a number denotes an open range. For example: "3-" means "all documents beginning with the third one". Similarly: "1-" means "all documents in the result set".

When the originally agreed upon time limit for result state maintenance is about to expire, and the client wants the server to maintain the result set for an additional amount of time, the client needs to call `extendStateTimeout()`:

```
void extendStateTimeout(// Request more time for server to
    Long serverSID,      // {0} maintain search result state.
    Long additionalTime,
    OUT Long timeAllotted // Num of secs server agrees to maintain state
)
```

The LSP returns the amount of additional time it is actually willing to maintain the state for the client.

Alternatively, sometimes, a client may want to release server state prematurely, or it may have changed its mind about a request it delivered earlier. It can use `cancelRequest()` for these situations. (Even though the `getDocs()` call is synchronous, a multi-threaded client might use a separate thread to contact the server and cancel a 'hanging' `getDocs()`):

```
void cancelRequest(
    Long serverSID,          // {0}
    Long reqID               // {0}
)
```

The `reqID` is 0 if all outstanding requests for this session are to be canceled. This has the semantics of closing the session, and allowing the server to free its resources. If `reqID` is not zero, only the corresponding request within the session is canceled.

## 2.3 The Source Metadata Interface

By source metadata we mean information about the service itself, and about the collections that are being served out. In this context, one could ask many complicated questions. Example: 'Which of your subcollections has a searchable DublinCore:Creator property?' While such questions are certainly of interest, SDLIP takes a minimalist approach. The emphasis of the source metadata interface is to provide the most important information easily, but to require almost no implementation effort on the server side. This improves the chance that services will actually provide this interface, which is traditionally the most neglected facility in server implementations. In particular, the interface is defined in such a way that services can return one constant XML string for each request. If services are more ambitious, they may provide a subcollection called 'SourceMetadata' which can be queried like other subcollections.

The source metadata interface includes three operations:

- `getInterface()` returns the names and versions of every SDLIP interface that is supported.
- `getSubcollectionInfo()` returns a list of subcollections with their names, descriptions, and supported query languages.
- `getPropertyInfo()` takes a subcollection name as parameter. It returns a description of all the document properties that are acceptable for that subcollection (author, title, etc.). This returned information also includes information on how clients may work with each property: search over it, retrieve it, etc.

The operation for asking sources about their interface versions:

```
void getInterface(          // Get info about LSP's interfaces
    OUT XMLObject version   // XML-encoded info about the versions of each supported
interface.
    // Ex:
    // <SDLIPInterface>
    //   <SearchInterface>
    //     <version>1.0</version>
    //   </SearchInterface>
    //   <ResultAccessInterface>
    //     <version>1.1</version>
```

```
//    </ResultAccessInterface>
//    <MetadataInterface>
//        <version>1.0</version>
//    </MetadataInterface>
// </SDLIPInterface>
```

```
)
```

The returned value contains information about all of the supported interfaces. If more than one version is supported, the `<version>` element may be repeated. See [Section 3.5](#) for details on the information that is returned.

To get the names and supported query languages of all the subcollections a service makes accessible:

```
void getSubcollectionInfo(
    OUT XMLObject subcolInfo // XML list of subcollections and their supported query
    languages
)
```

The `getSubcollectionInfo()` operation returns something like this:

```
<subcolInfo>
  <subcol>
    <subcolName>New York Times</subcolName>
    <defaultSubcol/>
    <queryLangs>
      <DAV:basicsearch/>
      <DialogCorp:standard/>
      <Z3950:RPN>
    </queryLangs>
  </subcol>
  <subcol>
    <subcolName>StockQuotes</subcolName>
    <subcolDesc> Current stock market values. Delayed by at least 15
minutes</subcolDesc>
    <queryLangs>
      <DAV:basicsearch/>
    </queryLangs>
  </subcol>
</subcolInfo>
```

Three pieces of information are provided for each subcollection: Its name, an optional human-readable description of the subcollection's contents, a list of query languages that may be used to query the subcollection, and whether this is the service's default subcollection. Knowing the default subcollection is, of course, important when operation parameters that specify a service's subcollection are left out during operation invocations.

Finally, to get information about a service's supported document properties and attribute models:

```
void getPropertyInfo(          // Returns a propList XML string
    String subcolName,         // {null} If not supplied, request for default
    subcollection.             //
    OUT XMLObject propInfo     // Acceptable attribute models and document properties
    for the                    //
                                // specified subcollection, and whether they are
                                // searchable/retrievable
)
```

This operation allows clients to retrieve information about which document properties may be searched or retrieved for the specified subcollection. Here is an example of what is returned in `propInfo`.

```
<propList>
  <DublinCore:creator>
```

```

    <searchable/>
    <retrievable/>
</DublinCore:creator>
<USMARC:245>
    <searchable/>
    <retrievable/>
    <phraseSearch/>
</USMARC:245>
<USMARC:711c>
    <retrievable/>
    <accessPermission>ALL</accessPermission>
</USMARC:711c>
</propList>

```

## 3. XML Formats Used in SDLIP

In order to keep SDLIP's datatypes simple, parameters that contain multiple pieces of information are encoded as XML. This section describes these XML formats. We use DTD-like syntax to describe the structures. Note, however, that we are taking some liberties in that we assume extensibility. It is expected that more entities might be added within SDLIP's XML structures over time, even though a strict adherence to the DTDs would not allow that.

XML has some primitive data types that are difficult to remember. Here are the ones that are used below:

- **ANY:** Unicode text that must be valid XML. If there are tags, they must match, etc. You must escape XML-reserved characters, such as '>' with standard escapes, such as '&gt;'.
- **#PCDATA:** Character data. Must not include XML-reserved characters.

### 3.1 Property Lists

An example of a property list is this:

```

<propList>
  <QualityOfService>Fastest</QualityOfService>
  <UserID>Miller</UserID>
</propList>

```

Think of this as a dictionary, or list of key/value pairs. You may have empty elements in a `propList`, as in the following list of delivered groceries:

```

<propList>
  <Ham/>
  <Eggs/>
  <Bacon/>
</propList>

```

In summary, `propList` is: `<!ELEMENT propList (ANY)>.`

### 3.2 Exceptions

Exceptions are delivered in whatever form the underlying transport allows. Once the SDLIP transport module receives the exception information over the wire, it raises an exception for the client or server implementation on the local machine. The exception will have the following interface:

```

interface SDLIPException {
    unsigned short getCode();
    string getReason();
    XMLObject getDetails();
}

```

Either two or three pieces of information are provided in an exception: an error code, a human-readable message, and optionally a property list with additional information.

Here is an example:

```
catch (SDLIPException e) {
    e.getCode();    // returns 451
    e.getReason();  // returns "Bad Query"
    e.getDetails().getString(); /* returns "<propList>
                                <remedy>Read the manual, you
                                <stacktrace>...</stacktrace>
                                </propList>"
                                */
}
```

The error codes follow a subset of HTTP conventions. All error codes are three-digit numbers. In SDLIP there are two classes of error codes. Codes of the form 4xx indicate errors in the information supplied by the client. Errors of the form 5xx indicate errors the server encountered, even though the client has supplied correct information. See [Appendix A](#) for the full sets of codes.

### 3.3 SearchResult

This kind of XML string is used to return result documents in response to a search. An explanation follows:

```
<SearchResult>
  <doc>
    <DID>1</DID>
    <propList>
      <author>Bill Smith</author>
      <author>Frank Miller</author>
      <title>This is My Life</title>
      <abstract>It's been great so far.</abstract>
    </propList>
  </doc>
  <doc>
    ...
  </doc>
</SearchResult>
```

For each document in a search result we communicate its document ID (DID) which is a numeric index into the result set (one-based), and the document's properties as requested by the client (i.e. author, title, etc). Note that exactly which properties are delivered depends on what the client asked for in its request (e.g. in the property list parameter of the `search()` call).

Here is the DTD for search results:

```
<!ELEMENT SearchResult (doc*)>
<!ELEMENT doc (DID, propList)>
<!ELEMENT DID (#PCDATA)>
```

Notice that a server could generate a more elaborate structure for the documents. For example, the author field could be subdivided into first name initials and last name portions, and a 'pict' attribute could be added, which points to a gif image of the author:

```
<SearchResult>
  <doc>
    <DID>1</DID>
    <propList>
      <DublinCore:Creator>
        <initials>A.</initials>
```

```

        <lastName>Miller</lastName>
        <authorPict>http://peopleServer.org/~miller/icon.jpg</authorPict>
    </DublinCore:Creator>
    <title>How I Did It</title>
    <abstract>With lots of effort.</abstract>
</propList>
</doc>
<doc>
    ...
</doc>
</SearchResult>

```

### 3.4 Subcollection Specifications

When requesting a search, clients specify a set of subcollections and/or result sets to run the search over. This set is expressed as an XML string. For example:

```

<subcols>
  <subcolName>New York Times</subcolName>
  <resSet>3<resSet>
  <resSet>5<resSet>
  <subcolName>Washington Post</subcolName>
</subcols>

```

This instructs the LSP to search over the New York Times, the Washington Post, and result sets produced in session IDs 3 and 5.

The DTD for this is:

```

<!ELEMENT subcols (subcolName | resSet)*>
<!ELEMENT subcolName (#PCDATA)>
<!ELEMENT resSet (#PCDATA)>

```

### 3.5 Source Metadata

Version information about supported SDLIP interfaces are returned by a simple XML element:

```

<SDLIPInterface>
  <SearchInterface>
    <version>1.0</version>
  </SearchInterface>
  <ResultAccessInterface>
    <version>1.1</version>
  </ResultAccessInterface>
  <MetadataInterface>
    <version>1.0</version>
  </MetadataInterface>
</SDLIPInterface>

```

The DTD is:

```

<!Element SDLIPInterface (SearchInterface, ResultAccessInterface?,
                           MetadataInterface?, SearchAsynchInterface?,
DeliveryInterface?, ANY)>
<!Element SearchInterface (version?, ANY)>
<!Element version (#PCDATA)>
<!Element ResultAccessInterface (version?, ANY)>
<!Element MetadataInterface (version?, ANY)>

```

```
<!Element SearchAsynchInterface (version?, ANY)>
<!Element DeliveryInterface (version?, ANY)>
```

The searchAsynch and delivery interfaces are part of the optional SDLIP-Asynch extension. The ANY at the end of the SDLIPInterface element definition above allows services to provide additional interfaces, such as authorization, payment, etc. Similarly, the ANY in the interface definitions allow additional details about each interface to be included. If any of the interface elements are empty, version 1.0 is assumed. Example:

```
<SDLIPInterface>
  <SearchInterface/>
</SDLIPInterface>
```

Information about a service's subcollections has the following DTD (for an example, see [Section 2.3](#)):

```
<!Element subcolInfo (subcolName, subcolDesc?, defaultSubcol?, queryLangs)>
<!Element subcolName (#PCDATA)>
<!Element subcolDesc (#PCDATA)>
<!Element defaultSubcol EMPTY>
<!Element queryLangs ANY>
```

When clients ask LSPs for information about the properties that are available for the documents in the collections the LSPs serve out, an attribute list is returned. The following example indicates that a source supports a small portion of Dublin Core and USMARC:

```
<propList>
  <DublinCore:creator>
    <searchable/>
    <retrievable/>
  </DublinCore:creator>
  <USMARC:245>
    <searchable/>
    <retrievable/>
    <phraseSearch/>
  </USMARC:245>
  <USMARC:711c>
    <retrievable/>
    <accessPermission>ALL</accessPermission>
  </USMARC:711c>
</propList>
```

The properties in the Dublin Core and USMARC name spaces each contain one or more subelements which describe what may be done with the respective property. Standard capabilities for a property are <searchable/> and <retrievable/>. Services may add others, as exemplified by the <phraseSearch/> and <accessPermission> examples.

Notice that we again use empty XML fields for Boolean conditions. For example, if the (empty) <searchable/> attribute is present, then the attribute being described may be searched in queries. The absence of the <searchable/> element indicates that the source does not maintain a search index for this attribute.

## 3.6 Server Delegates

Server delegate specifications have a very simple DTD:

```
<!Element redirect (serverDelegate+)>
<!Element serverDelegate (#PCDATA)>
```

## 4. Implementing SDLIP With IETF's DASL

The Internet Engineering Task Force (IETF) is defining a standard for searching document repositories over the Web. The effort is part of the Web-based Distributed Authoring and Versioning ([WebDAV](#)) initiative, and is called Distributed Authoring, Searching and Locating ([DASL](#)). DASL is an HTTP-based protocol. It defines how search requests are delivered, and how results are returned. It also defines a basic query language that every DASL server must support. In contrast to SDLIP, the DASL protocol is very Web-centric.

SDLIP defines a mapping of SDLIP-Core operations onto DASL. The goal of the mapping is (i) to provide a coherent transport between SDLIP clients and services over HTTP, (ii) to allow SDLIP clients to search DASL servers, and (iii) to allow DASL clients a minimum of search capabilities over SDLIP servers.

SDLIP's DASL binding is described in a [separate document](#).

## 5. Implementing SDLIP With CORBA

Client and server transport modules may use CORBA to communicate search requests and results. With this transport binding, client applications still communicate with their local client transport module through local SDLIP calls, and library service proxies are still called by their local server transport modules. The CORBA based interactions between the transport modules are specified with CORBA IDL ([SDLIPCorba.idl](#) and [SDLIPCore.idl](#)). The interactions are very similar to the SDLIP specifications of this document.

SDLIP's CORBA binding is described in a [separate document](#).

## Appendix A: Error Codes and their Meanings

Error conventions follow a subset of HTTP conventions: 4xx are errors in information supplied by the client. Error codes of the form 5xx signal problems at the server.

Code	Error Name	Meaning
400	eInvalidRequest	Use if none of the more specific error codes fits
401	eUnauthorized	Operation invocation may be correct, but it requires authorization
402	ePaymentRequired	Client needs to supply payment
404	eNotFound	The requested document is not served by this server
405	eIllegalMethod	Specified operation is not part of SDLIP
408	eRequestTimeout	Server has discarded state
450	eQueryLanguageUnknown	Server does not support the specified query language
451	eBadQuery	Query malformed
452	eInvalidProperty	One or more specified document properties are not supported
453	eInvalidSessionID	Session ID specified unknown to this server
454	eInvalidSubcollection	Specified subcollection not supported on this server
455	eMalformedXML	An XML parameter is not parsable
500	eServerError	Use if none of the more specific errors fits
501	eNotImplemented	Operation is legal SDLIP, but server doesn't support it
503	eServiceUnavailable	Service or requested subcollection is supported in principle, but is currently down

# References

- [BASE64]  
N. Borenstein, N. Freed. Base64 Content Transfer Encoding, in MIME Part One, RFC 1521, Sep 1993  
<http://src.doc.ic.ac.uk/packages/rfc/rfc1521.txt>
- [DASL]  
Saveen Reddy, Dale Lowry, Surenda Reddy, Rick Henderson, Jim Davis, Alan Babich: *DAV Searching & Locating*, Internet Draft, June 3, 1999  
<http://www.webdav.org/dasl/protocol/draft-dasl-protocol-00.html>



## the Open Archives initiative

---

The Open Archives initiative is a forum to discuss and solve matters of interoperability between author self-archiving solutions, as a way to promote their global acceptance.

---

## 2nd meeting of the Open Archives initiative

- June 3rd 2000, San Antonio, Texas
- At the occasion of the ACM Hypertext'2000 and ACM Digital Libraries'2000 conferences.
- [Agenda & participants](#): Extending Interoperability of Digital Libraries: Building on the Open Archives Initiative.

## The Santa Fe Convention

The convention is the result of the Santa Fe meeting of the initiative which has taken place on October 21-22 1999. The meeting was supported by [the Council on Library and Information Resources](#), [the Digital Library Federation](#), [the Scholarly Publishing & Academic Resources Coalition](#), [the Association of Research Libraries](#) and [the Los Alamos National Laboratory](#).

- The [Santa Fe Convention](#) for the Open Archives initiative presents a simple technical and organizational framework to support basic interoperability among e-print archives.
  - Read the companion paper in [D-Lib Magazine](#), February 2000.
  - Check out the e-print [archives](#) that comply with the Santa Fe Convention, or that are working towards compliancy.
  - The [press release](#) of the Santa Fe meeting of the Open Archives initiative.
-

Read about the goals of the Open Archives initiative in [the call](#) sent out by [Paul Ginsparg, Rick Luce and Herbert Van de Sompel](#) to a group of experts, inviting them to join the first meeting

The public archive of the closed [discussion list](#)

Have a look at the **VERY** experimental [UPS prototype](#) that was created to facilitate discussions at the Santa Fe meeting. It has a cross-archive search engine as well as a linking service.

---

*last updated April 30 2000*

*get in touch with the Open Archives initiative by contacting [openarchives@openarchives.org](mailto:openarchives@openarchives.org)*



# Interfaces:

---

## [Stanford DL user interface projects](#)

### Xerox Interfaces for Information Access

- [Home Page](#)
- [Scientific American article](#)
- [Cat-a-Cone figures](#)
- [Scatter/Gather examples](#)
- Questions:
  - Compare
    - What are the various interfaces built? How do they compare? What is the best use of each?
  - Scatter/gather
    - Explain clustering, relate it to scatter/gather.
    - What are special problems with large category systems and how can they be solved?

[Envision](#) project at Virginia Tech, [MARIAN](#) sequel

[Berkeley](#): TileBars, Multivalent documents

---

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

---

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

(c) Copyright 1998-2000, Edward A. Fox, Rajat Gupta

# Metadata:

---

- [IMS Metadata](#)
- [Metadata: the Foundations of Resource Description](#)
- [OCLC/NCSA Metadata Workshop Report](#)
- [RFC-1807](#)
- [TEI](#)
- [BASIS article](#)
- [D-Lib Working Group on Metadata](#)
- [STARTS](#)
- [Dublin Core Metadata Initiative](#)
- [Alliance Metadata Standards Working Group at NCSA](#)

---

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

---

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

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

**Specifications****Meta-Data**[Enterprise](#)[Content Packaging](#)[Question & Test](#)[XML, Bindings and  
Examples](#)[Toolkits  
Requirements](#)[Home](#)

IMS is pleased to release the IMS Meta-data Specification - Version 1, to the public. The creation of this document would not have been possible without the outstanding efforts of the worldwide IMS community plus active collaboration and participation by organizations around the world including: IEEE, ARIADNE, and GESTALT.

The IMS Meta-data Specification was approved unanimously by the IMS Technical Board, which is comprised of one representative from each of the [IMS Contributing member](#) organizations.

The IMS Meta-data Specification is comprised of 3 documents. Links are provided below to the relevant materials as html and PDF files.

1. [IMS Learning Resource Meta-data Information Model](#)
  - [Information Model PDF](#) - (23K)
2. [IMS Learning Resource Meta-data XML Binding Specification](#)
  - [Binding Specification PDF](#) - (72K)
3. [IMS Learning Resource Meta-data Best Practices and Implementation Guide](#)
  - [Best Practices and Implementation Guide PDF](#) - (109K)

IMS XML Bindings, DTDs, and Examples may be found at:  
<http://www.imsproject.org/xml/>.

Direct your questions or comments on the IMS Meta-data Specification to: [md-question@imsproject.org](mailto:md-question@imsproject.org)

Check out Sun Microsystems [Toolkit](#) for creating IMS-compliant Meta-data



# Metadata: The Foundations of Resource Description

Stuart Weibel

Office of Research, OCLC Online Computer Library Center, Inc.

*weibel@oclc.org*

**D-Lib Magazine**, July 1995

---

This paper is an abbreviated version of the [Summary Report of the OCLC/NCSA Metadata Workshop](#). It sets forth a proposal for the content of a simple resource description record (the Dublin Core Metadata Element Set) and outlines a series of further steps to advance the standards for the description of networked information resources.

- [Introduction](#)
- [Underlying Assumptions](#)
- [Implementations](#)
- [Next Steps](#)
- [References](#)

---

**d-lib forum**

**d-lib magazine**

---

## Introduction

The explosive growth of interest in the Internet in recent years has created a digital extension of the academic research library for certain kinds of materials. Valuable collections of texts, images and sounds from many scholarly communities -- collections that may even be the subject of state-of-the-art discussions in these communities--now exist only in electronic form and may be accessible from the Internet. Knowledge regarding the whereabouts and status of this material is often passed on by word of mouth among members of a given community. For outsiders, however, much of this material is so difficult to locate that it is effectively unavailable.

Why is it so difficult to find items of interest on the Internet or the World Wide Web? A number of well-designed locator services, such as Lycos [\[http://lycos.cs.cmu.edu/\]](http://lycos.cs.cmu.edu/), are now available that automatically index many of the resources available on the Web and maintain up-to-date databases of

locations. But indexes are most useful in small collections within a given domain. As the scope of their coverage expands, indexes succumb to problems of large retrieval sets and problems of cross disciplinary semantic drift. Richer records, created by content experts, are necessary to improve search and retrieval. Formal standards such as the [TEI Header](#) and [MARC](#) cataloging) will provide the necessary richness, but such records are time consuming to create and maintain, and hence may be created for only the most important resources.

An alternative solution that promises to mediate these extremes involves the creation of a record that is more informative than an index entry but is less complete than a formal cataloging record. If only a small amount of human effort were required to create such records, more objects could be described, especially if the author of the resource could be encouraged to create the description. And if the description followed an established standard, only the creation of the record would require human intervention; automated tools could discover these descriptions and collect them.

Can a simple metadata record be defined that sufficiently describes a wide range of electronic objects? The Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) convened the invitational Metadata Workshop on March 1-3, 1995, in Dublin, Ohio to address this issue. Fifty-two librarians, archivists, humanities scholars and geographers, as well as standards makers in the Internet, Z39.50 and Standard Generalized Markup Language (SGML) communities, met to identify the scope of the problem, to achieve consensus on a list of metadata elements that would yield simple descriptions of data in a wide range of subject areas, and to lay the groundwork for achieving further progress in the definition of metadata elements that describe electronic information.

## Goals

Goals of the workshop included fostering a common understanding of the problems and potential solutions among the stakeholders and promoting a consensus on a core set of metadata elements to describe networked resources.

## Scope

Since the Internet contains more information than professional abstractors, indexers and catalogers can manage using existing methods and systems, it was agreed that a reasonable alternative way to obtain usable metadata for electronic resources is to give authors and information providers a means to describe the resources themselves. The major task of the Metadata Workshop was to identify and define a simple set of elements for describing networked electronic resources. To make this task manageable, it was limited in two ways. First, only those elements necessary for the discovery of the resource were considered. It was believed that resource discovery is the most pressing need that metadata can satisfy, and one that would have to be satisfied regardless of the subject matter or internal complexity of the object.

Secondly, the discussion was further restricted to the metadata elements required for the discovery of what were called **document-like objects**, or **DLOs** by the workshop participants. It was believed that DLOs are still the most common type of resource sought in the Internet and that whatever solution could be proposed for DLOs could be extended to other kinds of resources. More importantly, the likelihood of making progress on this challenging problem would be increased if attention could initially be restricted to something familiar.

DLOs were not rigorously defined, but were understood by example. For example, an electronic version of a newspaper article or a dictionary is a DLO, while an unannotated collection of slides is not. Of course, the crux of the problem is that in a networked environment, DLOs can be arbitrarily complex because they can consist of text with callouts to images, audio or video clips, or to other hypertext documents. The Metadata Workshop participants made no attempt to limit the complexity of DLOs, except to say that the intellectual content of a DLO is primarily text, and that the metadata required for describing DLOs will bear a strong resemblance to the metadata that describes traditional printed texts.

As a result of the restricted focus of the workshop, certain issues required for a complete description of DLOs, such as cost, archival status and copyright information, were eliminated from the scope of the discussion. Elements required for the description of objects other than DLOs, such as the elements required for the description of complex geological strata in a geospatial resource, were also beyond the scope of the discussion. The goal was to define a core set of metadata elements that would allow authors and information providers to describe their work and to facilitate interoperability among resource discovery tools. But because the core elements do not yield a complete description of objects in a networked environment, careful consideration was also given to mechanisms for extending the element set.

The primary deliverable from the workshop was a set of thirteen metadata elements, named the **Dublin Core Metadata Element Set** (or Dublin Core, for short). The Dublin Core was proposed as the minimum number of metadata elements required to facilitate the discovery of document-like objects in a networked environment such as the Internet. The syntax was deliberately left unspecified as an implementation detail. The semantics of these elements was intended to be clear enough to be understood by a wide range of users.

Below is a brief description of the elements in the Dublin Core **Dublin Core Element Description**

- **Subject:** The topic addressed by the work
- **Title:** The name of the object
- **Author:** The person(s) primarily responsible for the intellectual content of the object
- **Publisher:** The agent or agency responsible for making the object available
- **OtherAgent:** The person(s), such as editors and transcribers, who have made other significant intellectual contributions to the work
- **Date:** The date of publication
- **ObjectType:** The genre of the object, such as novel, poem, or dictionary
- **Form:** The physical manifestation of the object, such as Postscript file or Windows executable file
- **Identifier:** String or number used to uniquely identify the object
- **Relation:** Relationship to other objects
- **Source:** Objects, either print or electronic, from which this object is derived, if applicable
- **Language:** Language of the intellectual content
- **Coverage:** The spatial locations and temporal durations characteristic of the object

To make this discussion concrete, consider an electronic a record created with the relevant portions of the Dublin Core, and a sample syntax, that describes an electronic version of Maya Angelou's poem "On the Pulse of Morning". This description is based on a record created by the University of Virginia Library's

Electronic Text Center. (For a description of that project, see Gaynor [\[Gaynor\]](#).)

- **Subject:** Poetry
- **Title:** On the Pulse of Morning
- **Author:** Maya Angelou
- **Publisher:** University of Virginia Library Electronic Text Center
- **OtherAgent:** Transcribed by the University of Virginia Electronic Text Center
- **Date:** 1993
- **Object:** Poem
- **Form:** 1 ASCII file
- **Identifier:** AngPuls1
- **Source:** Newspaper stories and oral performance of text at the presidential inauguration of Bill Clinton
- **Language:** English

## Underlying Assumptions

The discussions at the Metadata Workshop revealed several principles that should guide the further development of the element set. Adherence to these principles increases the likelihood that the core element set will be kept as small as possible, that the meanings of the elements will be understood by most users, and that the element set will be flexible enough for the description of resources in a wide range of subject areas. These principles are intrinsicality, extensibility, syntax independence, optionality, repeatability, and modifiability.

### Intrinsicality

The Dublin Core concentrates on describing intrinsic properties of the object. Intrinsic data refer to the properties of the work that could be discovered by having the work in hand, such as its intellectual content and physical form. This is distinguished from extrinsic data, which describe the context in which the work is used. For example, the "Subject" element is intrinsic data, while transaction information such as cost and access considerations are extrinsic data. The focus on intrinsic data in no way demeans the importance of other varieties of data, but simply reflects the need to keep the scope of deliberations narrowly focussed.

### Extensibility

In addition to its use in dealing with extrinsic data, extension mechanisms will allow the inclusion of intrinsic data for objects that cannot be adequately described by a small set of elements.

Extensibility is important because users may wish to add extra descriptive material for site-specific purposes or specialized fields. In addition, the specification of the Dublin Core itself will change over time, and the extension mechanism will allow revisions while maintaining some backward compatibility with the originally defined element set.

## Syntax Independence

Syntactic bindings are avoided because it is too early to propose formal definitions and because the Dublin Core is intended to be eventually used in a range of disciplines and application programs.

## Optionality

All the elements are optional. The Dublin Core may eventually be applied to objects for which some elements have no meaning (who is the author of a satellite image?). It also seems counterproductive to mandate complex descriptions if the creators of the content are expected to provide the descriptive material. A simple description is better than no description at all.

## Repeatability

All elements in the Dublin Core are repeatable. For example, multiple author elements would be used when a resource has multiple authors.

## Modifiability

Each element in the Dublin Core has a definition that is intended to be self-explanatory. However, it is also necessary that the definitions of the elements satisfy the needs of different communities. This goal is accomplished by allowing each element to be modified by an optional qualifier. If no qualifier is present, the element has its common-sense meaning; otherwise, the definition of the element is modified by the value of the qualifier.

Qualifiers will be typically derived from well-known conventions in the library community or from the field of knowledge appropriate to the resource. Qualifiers are important because they give the Dublin Core a mechanism for bridging the gap between casual and sophisticated users. For example, the data in the **Subject** element consists of any word or phrase that describes the object's content. However, a professional cataloger may wish to supply the name of the authoritative source from which the subject terms are taken. In such a case, the element may be written as **Subject (scheme=LCSH)**, indicating that the subject terms are taken from the Library of Congress Subject Headings.

## Implementations

One of the goals of the OCLC/NCSA Metadata Workshop was to promote prototype resource description projects based on a common model of resource description. A number of Metadata Workshop conferees represent organizations that have ongoing activities or are starting activities that will be influenced by the results of the workshop. These include:

- The OCLC Spectrum Project  
Contact:Diane Vizine-Goetz, [vizine@oclc.org](mailto:vizine@oclc.org)
- [The OCLC Internet Resources Cataloging Project](#)  
Contact:Erik Jul, [jul@oclc.org](mailto:jul@oclc.org)
- Library of Congress

Contact:Rebecca Guenther, [rgue@loc.gov](mailto:rgue@loc.gov)

- O'Reilly Associates

Contact:Terry Allen, [terry@ora.com](mailto:terry@ora.com)

- Los Alamos National Laboratory and Indiana University

Contact:Ron Daniel Jr.,[rdaniel@acl.lanl.gov](mailto:rdaniel@acl.lanl.gov)

Contact:Pete Percival,[percival@bronze.ucs.indiana.edu](mailto:percival@bronze.ucs.indiana.edu)

- Bunyip Systems

Contact:Chris Weider,[clw@bunyip.com](mailto:clw@bunyip.com)

- Georgia Institute of Technology

Contact:Michael Mealling, [michael.mealling@oit.gatech.edu](mailto:michael.mealling@oit.gatech.edu) , <http://www.gatech.edu/iiir>

- SoftQuad

Contact: Yuri Rubinsky,[yuri@sq.com](mailto:yuri@sq.com)

- Concordia University

Contact:Bipin Desai, [bcdesai@cs.concordia.ca](mailto:bcdesai@cs.concordia.ca),

<http://www.cs.concordia.ca/~faculty/bcdesai/cindi-system-1.0.html>

## Next Steps

Refinement and standardization of the metadata element set defined in this document will be an ongoing, dynamic process involving many stakeholder communities. No single forum will suffice to air all concerns and no single standard can be expected to accommodate the needs of all communities. The problem must be divided into manageable chunks and the process must engage the relevant stakeholder communities. Implicit in the present activity is the proposition that there are core elements common to many object types, and that a simple, extensible framework of such elements can be defined to support more complete resource descriptions.

The initial objective--the specification of elements for the discovery of document-like objects--can be extended in a variety of directions:

- Expansion of the Dublin Core to include other object types, such as services or collections.
- Expansion of the Dublin Core to embrace functionality other than resource discovery, such as archival control and the authentication of users and charging mechanisms.
- Establishing standardized methods for extensibility.
- Refinement of existing work. The Dublin Core is an untested approach to the description of resources that will need to be modified with experience.

OCLC and NCSA will establish a workshop series to address aspects of this agenda. A Metadata Workshop Steering Committee will be established to define topics and assure appropriate representation of stakeholders. Design groups of perhaps a dozen or fewer individuals will be solicited to prepare discussion papers to focus workshop activities. Participants will be invited based on their publicly evident accomplishments in relevant areas or by reviewed application. Workshops will be limited to 50 or fewer participants and conducted in roughly the style of the March 1995 Workshop.

Other work will be done in coordination with IETF working group on Uniform Resource Identifiers

(URIs) to assure that the results can be integrated into the emerging protocols for resource location and persistent naming.

Finally, active promotion of results will be carried out by establishing liaison with formal associations of stakeholders. In the library community, MARC standards evolve under the guidance of the Machine-Readable Bibliographic Information Committee (MARBI), composed of representatives of the Library of Congress and other stakeholders in the library community. A close relationship should be sustained between this committee and the Metadata Work Group. Relationships should also be established with publishers, document vendors, SGML vendors and theoreticians working on the problem of text encoding. Other communities also have requirements that must be accommodated in any framework for resource description. These communities include the GIS community, government information providers and business communication groups.

---

## References

### [MARC]

Network Development and MARC Standards, Office, ed. 1994. USMARC Format for Bibliographic data. 1994. Washington, DC: Cataloging Distribution Service, Library of Congress.

### [TEI]

Sperberg-McQueen, C. M., and Leu Burnard, ed. 1994. Guidelines for Electronic Text Encoding and Interchange. Chicago and Oxford: Text Encoding Initiative.

### [Gaynor]

Gaynor, Edward. 1994. "Cataloging Electronic Texts: The University of Virginia Library Experience." Library Resources and Technical Services 38(4): 403-413 (October 1994).

**Copyright © 1995 OCLC**

---

**d-Lib forum**

**d-Lib magazine**

---

*hdl:cnri.dlib/july95-weibel*

# rfc1807

Press [here](#) to go to the top of the rfc 'tree'.

Network Working Group  
Request For Comments: [1807](#)  
Obsoletes: [1357](#)  
Category: Informational

R. Lasher  
Stanford  
D. Cohen  
Myricom  
June 1995

## A Format for Bibliographic Records

### Status of this Memo

This memo provides information for the Internet community. This memo does not specify an Internet standard of any kind. Distribution of this memo is unlimited.

### Abstract

This RFC defines a format for bibliographic records describing technical reports. This format is used by the Cornell University Dienst protocol and the Stanford University SIFT system. The original RFC (RFC [1357](#)) was written by D. Cohen, ISI, July 1992. This is a revision of RFC [1357](#). New fields include handle, other\_access, keyword, and withdraw.

### Introduction

Many universities and other R&D organizations routinely announce new technical reports by mailing (via the postal services) the bibliographic records of these reports.

These mailings have non-trivial cost and delay. In addition, their recipients cannot conveniently file them, electronically, for later retrieval and searches.

Publishing organizations that wish to use e-mail or file transfer to obtain these announcements can do so by using the following format.

Organizations may automate to any degree (or not at all) both the creation of these records (about their own publications) and the handling of the records received from other organizations.

This format is designed to be simple, for people and for machines, to be easy to read ("human readable") and create without any special programs.

This RFC defines the format of bibliographic records, not how to process them.

Lasher &amp; Cohen

Informational

[Page 1]

RFC [1807](#)

A Format for Bibliographic Records

June 1995

This format is a "tagged" format with self-explaining alphabetic tags. It should be possible to prepare and to read bibliographic records using any text editor, without any special programs.

This RFC includes the CR-CATEGORY, a field useful for Computer Science publications. It is expected that similar fields will be added for other domains.

This format, as described in RFC [1357](#), was implemented as part of the Dienst system and has been in use by the five ARPA-funded computer science institutions to exchange bibliographic records (Cornell, SU, UC, MIT, and CMU). Programs have been written to map between this RFC and structured USMARC (format developed at the Library of Congress) cataloging records, also from USMARC to the RFC.

The focus of this ARPA-funded research has been into many aspects of digital libraries including searching and accessing techniques that do not necessarily use bibliographic records (for example, natural language processing, automatic and full-text indexing). However, the continued use of bibliographic records is expected to remain an important part of the library system environment of the future and its use is an important link between the physical world of scientific works and the on-line world of digital objects. The format described in this paper allows a link between these two worlds to be created.

This format was developed with considerable help and involvement of Computer Science and Library personnel from several organizations, including Carnegie Mellon University, Corporation for National Research Initiatives (CNRI), Cornell University, University of Southern California/Information Sciences Institute (ISI), Meridian (now called DynCorp), Massachusetts Institute of Technology, Stanford University, and the University of California. Key contributions were provided by Jerry Saltzer of MIT, and Larry Lannom of DynCorp. The initial draft was prepared by Danny Cohen and Larry Miller of ISI. The revision was done by Rebecca Lasher from Stanford with assistance from the CS-TR participants.

This RFC does not place any limitations on the dissemination of the bibliographic records. If there are limitations on the dissemination of the publication, it should be protected by some means such as passwords. This RFC does not address this protection.

The use of this format is encouraged. There are no limitations on its use.

Lasher & Cohen	Informational	[Page 2]
RFC <a href="#">1807</a>	A Format for Bibliographic Records	June 1995

## The Information Fields

The various fields should follow the format described below.

<M> means Mandatory; a record without it is invalid.

<O> means Optional.

The tags (aka Field-IDs) are shown in upper case.

<M>	BIB-VERSION of this bibliographic records format
<M>	ID
<M>	ENTRY date
<O>	ORGANIZATION
<O>	TITLE

<O> TYPE  
<O> REVISION  
<O> WITHDRAW  
<O> AUTHOR  
<O> CORP-AUTHOR  
<O> CONTACT for the author(s)  
<O> DATE of publication  
<O> PAGES count  
<O> COPYRIGHT, permissions and disclaimers  
<O> HANDLE  
<O> OTHER\_ACCESS  
<O> RETRIEVAL  
<O> KEYWORD  
<O> CR-CATEGORY  
<O> PERIOD  
<O> SERIES  
<O> MONITORING organization(s)  
<O> FUNDING organization(s)  
<O> CONTRACT number(s)  
<O> GRANT number(s)  
<O> LANGUAGE name  
<O> NOTES  
<O> ABSTRACT  
<M> END

- \* One bibliographic record for each publication, where a "publication" is whatever the publishing institution defines as such.
- \* A record contains several fields.
- \* Each field starts with its tag (aka the field-ID) which is a reserved identifier (containing no separators) at the beginning of a new line with or without spaces before it), followed by two colons ("::"), followed by the field data.
- \* Continuation lines: Lines are limited to 79 characters. When needed, fields may continue over several lines, with an implied space in between. In order to simplify the use no special marking is used to indicate continuation line. Hence, fields are terminated by a line that starts (apart from white space) with a word followed by two colons. Except for the "END::" that is terminated by the end of line.) For improved human readability it is suggested to start continuation lines with some spaces.
- \* Several fields are mandatory and must appear in the record. All fields (unless specifically not permitted to) may be in any order and may be repeated as needed (e.g., the AUTHOR field). The order of the repeated fields is always preserved.
- \* Only printable ASCII characters are to be used. The permissible characters are ASCII codes 040 (Space) through 176(~) and line breaks which are \012 (LF) or \012\015 (CRLF). Empty lines indicate paragraph break. \009 (tab) must be replaced by spaces. This specifically forbids tabs, null characters, DEL, backspaces, etc. (i.e., if used, the record is invalid.)

However full 8 bit ASCII may be used. WARNING: some electronic mailers cannot handle 8 bit ASCII and these records may need to be transported via other mechanisms.

Throughout this document the word "publisher" means the publishing organization of a report (e.g., a university or a department thereof), not necessarily an organization authorized to issue ISBN numbers.

## EXAMPLE

-----  
BIB-VERSION:: CS-TR-v2.1  
ID:: OUKS//CS-TR-91-123  
ENTRY:: January 15, 1992  
ORGANIZATION:: Oceanview University, Kansas, Computer Science  
TYPE:: Technical Report  
REVISION:: January 5, 1995; FTP access information added  
TITLE:: Scientific Communication must be timely  
AUTHOR:: Finnegan, James A.  
CONTACT:: Prof. J. A. Finnegan, CS Dept, Oceanview Univ,  
Oceanview, KS 54321 Tel: 913-456-7890  
<Finnegan@cs.ouks.edu>  
AUTHOR:: Pooh, Winnie The  
CONTACT:: 100 Aker Wood  
DATE:: December 1991  
PAGES:: 48  
COPYRIGHT:: Copyright for the report (c) 1991, by J. A.  
Finnegan. All rights reserved. Permission is granted  
for any academic use of the report.  
HANDLE:: hdl:oceanview.electr/CS-TR-91-123  
OTHER\_ACCESS:: url:http://electr.oceanview.edu/CS-TR-91-123  
OTHER\_ACCESS:: url:ftp://electr.oceanview.edu/CS-TR-91-123  
RETRIEVAL:: send email to Finnegan@cs.ouks.edu with fax number  
KEYWORD:: Scientific Communication  
CR-CATEGORY:: D.0  
CR-CATEGORY:: C.2.2 Computer Sys Org, Communication nets, Net  
Protocols  
SERIES:: Communication  
FUNDING:: FAS  
CONTRACT:: FAS-91-C-1234  
MONITORING:: FNBO  
LANGUAGE:: English  
NOTES:: This report is the full version of the paper with  
the same title in IEEE Trans ASSP Dec 1976  
ABSTRACT::

Many alchemists in the country work on important fusion problems.  
All of them cooperate and interact with each other through the  
scientific literature. This scientific communication methodology



record and is used in management of these records. Its format is "ID:: XXX//YYY", where XXX is the publisher-ID (the controlled symbol of the publisher) and YYY is the ID (e.g., report number) of the publication as assigned by the publisher. This ID is typically printed on the cover, and may contain slashes.

The organization symbols "DUMMY" and "TEST" (case independent) are reserved for test records that should NOT be incorporated in the permanent database of the recipients.

Format: ID:: <publisher-ID>//<free-text>

Example: ID:: OUKS//CS-TR-91-123

\*\*\*\* See the note at the end regarding the \*\*\*\*  
\*\*\*\* controlled symbols of the publishers \*\*\*\*\*

Lasher & Cohen	Informational	[Page 6]
RFC <a href="#">1807</a>	A Format for Bibliographic Records	June 1995

ENTRY (M) -- This is a mandatory field. It is the date of creating this bibliographic record.

The format for ENTRY date is "Month Day, Year". The month must be alphabetic (spelled out). The "Day" is a 1- or 2-digit number. The "Year" is a 4-digit number.

Format: ENTRY:: <date>

Example: ENTRY:: January 15, 1992

ORGANIZATION (O) -- It is the full name spelled out (no acronyms, please) of the publishing organization. The use of this name is controlled together with the controlled symbol of the publisher (as discussed above for the ID field).

Avoid acronyms because there are many common acronyms, such as ISI and USC. Please provide it in ascending order, such as "X University, Y Department" (not "Y

Department, X University").

Format: ORGANIZATION:: <free-text>

Example: ORGANIZATION:: Stanford University, Department of  
Computer Science

TITLE (O) -- This is the title of the work as assigned by the author. This field should include the complete title with all the subtitles, if any.

Format: TITLE:: <free-text>

Example: TITLE:: The Computerization of Oceanview with  
High Speed Fiber Optics Communication

TYPE (O) -- Indicates the type of publication (summary, final project report, etc.) as assigned by the issuing organization.

Format: TYPE:: <free-text>

Example: TYPE:: Technical Report

REVISION (O) -- Indicates that the current bibliographic record is a revision of a previously issued record and is intended

to replace it. Revision information consists of a date and/or followed by a semicolon and by text in an open ended format. The revised bibliographic record should contain a complete record for the publication, not just a list of changes to the old record. If revision is omitted, the record is assumed to be a new record and not a revision. If the revision date is specified as 0, this is assumed to be January 1, 1900 (the previous RFC, used revision data of 0, 1, 2, 3, etc. this specification is for programs that might process records from RFC[1357](#)).

The text before the semicolon in this field is a date of the form month day, year. Any record with a more recent revision date replaces completely any record with an earlier revision date (supplied either explicitly or by default). Use the text to describe the revision. Reasons to send out a revised record include an error in the original, or change in the access information.

Format: REVISION:: January 1, 1995; <free-text>

Example: REVISION:: January 1, 1995; FTP information  
added

WITHDRAW (O) Withdraw means the document is no longer available. Some Institutions choose to delete the record others remove some of the fields. It is up to each institution to decide how to process withdraw records.

A withdraw record has all of the mandatory fields plus the withdraw field and a mandatory revision field.

The Withdraw field should indicate the reason for the withdraw in free text.

Example for withdrawing a bibliographic record::

```

BIB-VERSION:: CS-TR-v2.1
ID::          OUKS//CS-TR-91-123
ENTRY::       January 21, 1995
ORGANIZATION:: Oceanview University, Kansas, Computer
                Science
TITLE::       The Computerization of Oceanview with
                High Speed Fiber Optics Communication
REVISION::    January 21, 1995
WITHDRAW::    Withdrawn, found to be irrelevant
END::         OUKS//CS-TR-91-123

```

AUTHOR (O) -- Personal names only. Normal last name first inversion. Editors should be listed here as well, identified with the usual "(ed.)" as shown below in the last example.

If the report was not authored by a person (e.g., it was authored by a committee or a panel) use CORP-AUTHOR (see below) instead of AUTHOR.

Multiple authors are entered by using multiple lines, each in the form of "AUTHOR:: <free-text>".

The system preserves the order of the authors.

Format: AUTHOR:: <free-text>

Example: AUTHOR:: Finnegan, James A.  
 AUTHOR:: Pooh, Winnie The  
 AUTHOR:: Lastname, Firstname (ed.)

CORP-AUTHOR (O) -- The corporate author (e.g., a committee or a panel) that authored the report, which may be different from the ORGANIZATION issuing the report.

In entering the corporate name please omit initial "the" or "a". If it is really part of the name, please invert it.

Format: CORP-AUTHOR:: <free-text>

Example: CORP-AUTHOR:: Committee on long-range computing

CONTACT (O) -- The contact for the author(s).  
 Open-ended, most likely E-mail and postal addresses.

A CONTACT field for each author should be provided, separately, or for all the AUTHOR fields.  
 E-mail addresses should always be in "pointy brackets" (as in the example below).

Format: CONTACT:: <free-text>

Example: CONTACT:: Prof. J. A. Finnegan, CS Dept,  
 Oceanview Univ., Oceanview, Kansas, 54321  
 Tel: 913-456-7890 <Finnegan@cs.ouks.edu>

DATE (O) -- The publication date. The formats are "Month Year" and "Month Day, Year". The month must be alphabetic (spelled out). The "Day" is a 1- or 2-digit number. The "Year" is a 4- digit number.

Format: DATE:: <date>

Example: DATE:: January 1992

Example: DATE:: January 15, 1992

PAGES (O) -- Total number of pages, without being too picky about it. Final numbered page is actually preferred, if it is a reasonable approximation to the total number of pages.

Format: PAGES:: <number>

Example: PAGES:: 48

COPYRIGHT (O) -- Copyright information. Open ended format. The COPYRIGHT field applies to the cited report, rather than to the current bibliographic record.

Format: COPYRIGHT:: <free-text>

Example: COPYRIGHT:: Copyright for the report (c) 1991,  
by J. A. Finnegan. All rights  
reserved.

Permission is granted for any academic  
use of the report.

HANDLE (O) -- Handles are unique permanent identifiers that are used in the Handle Management System to retrieve location data. A handle is a printable string which when given to

Handles are used to identify digital objects stored within a digital library. If the technical report is available in electronic form, the Handle MUST be supplied in the bibliographic record.

Format is "HANDLE:: hdl:<naming authority>/string of characters". The string of characters can be the report number of the technical report as assigned by the publisher. For more information on handles and handle servers see the CNRI WEB page at

Lasher & Cohen                      Informational                      [Page 10]

---

RFC [1807](#)      A Format for Bibliographic Records      June 1995

<http://www.cnri.reston.va.us>.

\*\*\*\* NOTE: White space in HANDLE due to line wrap is ignored.

Format: HANDLE:: hdl:<naming authority>/string of characters

Example: HANDLE:: hdl:oceanview.electr/CS-TR-91-123

OTHER\_ACCESS (0) -- For URLs, URNs, and other yet to be invented formatted retrieval systems.

Only one URL or URN per occurrence of the field.

URL and URN information is available in the internet drafts from the IETF (Internet Engineering Task Force). The most recent drafts can be found on the CNRI WEB page at <http://www.cnri.reston.va.us>.

\*\*\*\* NOTE: White space in a URL or URN due to line wrap is ignored.

```
Format:  OTHER_ACCESS:: URL:<URL>
        OTHER_ACCESS:: URN:<URN>
```

Example: OTHER\_ACCESS:: URL:<http://elib.stanford.edu/Docume>

nt /STANFORD.CS:CS-TN-94-1

Example: OTHER\_ACCESS:: URL:ftp://JUPITER.CS.OUKS.EDU/PUBS/  
computerization.txt.

When the URN standard is finalized naming authorities will be registered and URNs will be viable unique identifiers. Until then this is a place holder. For the latest URN drafts see CNRI WEB page at <http://www.cnri.reston.va.us>.

RETRIEVAL (0) -- Open-ended format describing how to get a copy of the full text. This is an optional, repeatable field.

No limitations are placed on the dissemination of the bibliographic records. If there are limitations on the dissemination of the publication, it should be protected by some means such as passwords. This format does not address this protection.

Format: RETRIEVAL:: <free-text>

[illegible]

RFC [1807](#)      A Format for Bibliographic Records      June 1995

RETRIEVAL:: for full text with color pictures  
send a self-addressed stamped envelope to  
Prof. J.A. Finnegan, CS Dept,  
Oceanview University, Oceanview, KS 54321

KEYWORD (0) -- Specify any keywords, controlled or uncontrolled. This is an optional, repeatable field. Multiple keywords are entered using multiple lines in the form of "KEYWORD:: <free-text>.

Format: KEYWORD:: <free-text>

Example: KEYWORD:: Scientific Communication  
KEYWORD:: Communication Theory

CR-CATEGORY (0) -- Specify the CR-category. The CR-category (the Computer Reviews Category) index (e.g., "B.3") should always be included, optionally followed by the name of that category. If the name is specified it should be fully specified with parent levels as needed to clarify it, as in the second example below. Use multiple lines for multiple categories.

Every year, the January issue of CR has the full list of these categories, with a detailed discussion of the CR Classification System, and a full index. Typically the full index appears in every January issue, and the top two levels in every issue.

Format: CR-CATEGORY:: <free-text>

Example: CR-CATEGORY:: D.1

Example: CR-CATEGORY:: B.3 Hardware, Memory Structures

PERIOD (O) -- Time period covered (date range). Applicable primarily to progress reports, etc. Any format is acceptable, as long as the two dates are separated with " to " (the word "to" surrounded by spaces) and each date is in the format allowed for dates, as described above for the date field.

Format: PERIOD:: <date> to <date>

Example: PERIOD:: January 1990 to March 1990

[illegible]

RFC [1807](#)      A Format for Bibliographic Records      June 1995

SERIES (0) -- Series title, including volume number within series. Open-ended format, with producing institution strongly encouraged to be internally consistent.

Format: SERIES:: <free-text>

Example: SERIES:: Communication

FUNDING (O) -- The name(s) of the funding organization(s).

Format: FUNDING:: <free-text>

Example: FUNDING:: ARPA

MONITORING (O) -- The name(s) of the monitoring organization(s).

Format: MONITORING:: <free-text>

Example: MONITORING:: ONR

CONTRACT (O) -- The contract number(s).

Format: CONTRACT:: <free-text>

Example: CONTRACT:: MMA-90-23-456

GRANT (O) -- The grant number(s).

Format: GRANT:: <free-text>

Example: GRANT:: NASA-91-2345

LANGUAGE (O) -- The language in which the report is written.  
Please use the full English name of that language.

Please include the Abstract in English, if possible.

If the language is not specified, English is assumed.

Format: LANGUAGE:: <free-text>

Example: LANGUAGE:: English

Example: LANGUAGE:: French

NOTES (O) -- Miscellaneous free text.

Format:   NOTES:: <free-text>

Example:   NOTES:: This report is the full version of the  
                  paper with the same title in IEEE Trans ASSP  
                  Dec 1976

ABSTRACT (O) -- Highly recommended, but not mandatory. Even though no limit is defined for its length, it is suggested not to expect applications to be able to handle more than 10,000 characters.

The ABSTRACT is expected to be used for subject searching since titles are not enough. Even if the report is not in English, an English ABSTRACT is preferable. If no formal abstract appears on document, the producers of the bibliographic records are encouraged to use pieces of the introduction, first paragraph, etc.

Format:   ABSTRACT:: xxxx ..... xxxxxxxxx  
                  xxxx ..... xxxxxxxxx  
  
                  xxxx ..... xxxxxxxxx  
                  xxxx ..... xxxxxxxxx

END (M) -- This is a mandatory field. It must be the last entry of a record, identifying the record that it ends, by stating the same ID that was used at the beginning of the records, in its "ID::".

Format:    END:: XXX//YYY

Example:   END:: OUKS//CS-TR-91-123

>>>>>>   [END OF FORMAT DEFINITION]   <<<<<<<

A Note Regarding the Controlled Symbols of the Publishers

In order to avoid conflicts among the symbols of the publishing organizations (the XXX part of the "ID:: XXX//YYY") it is suggested that the various organizations that publish reports (such as universities, departments, and laboratories) register their <publisher-ID> symbols and names, in a way similar to the registration of other key parameters and names in the Internet.

Lasher &amp; Cohen

Informational

[Page 14]

RFC [1807](#)

A Format for Bibliographic Records

June 1995

Rebecca Lasher (RLASHER@Forsythe.stanford.edu), of Stanford working with CNRI has agreed to coordinate this registration with the IANA for the publishers of Computer Science technical reports. It is suggested that before using this format the publishing organizations would coordinate with her (by e-mail) their symbols and the names of their organizations.

In order to help automated handling of the received bibliographic records, it is expected that the producers of bibliographic records will always use the same name, exactly, in the ORGANIZATION field.

## Security Considerations

Security issues are not discussed in this memo.

## Acknowledgements

This work was supported by the Advanced Research Projects Agency under Grant No. MDA-972-92-J-1029 with the Corporation for National Research Initiatives (CNRI). Its content does not necessarily reflect the position or the policy of the Government or CNRI, and no official endorsement should be inferred.

## Authors' Addresses

Rebecca Lasher  
Mathematical and Computer Sciences Library  
M.S. 2125  
Stanford University  
Stanford, CA, USA 94305

Phone: +1 415 723 0864

EMail: rlasher@forsythe.stanford.edu

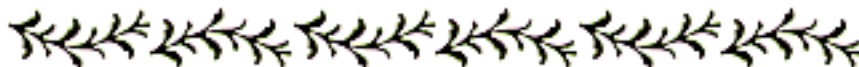
Danny Cohen  
Myricom  
325 N. Santa Anita Ave.  
Arcadia, CA 91006  
USA

Phone: +1 818 821 5555  
EMail: Cohen@myri.com

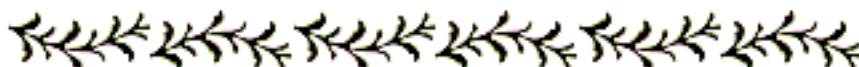


# TEI Guidelines for Electronic Text Encoding and Interchange (P3)

Made available by the Electronic Text Center at the University of Virginia

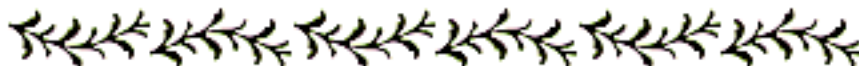


## [Search the entire \*TEI Guidelines\*](#)



New! [Quick Tag Usage Look-up](#):

search the *Alphabetical Reference List of Tags and Attributes* **only**



## Browse the *TEI Guidelines*

- [Bibliographic header of the TEI Guidelines](#)
- [Preface](#)
- [Acknowledgments](#)
  - TEI Working Committees (1990-1993)
  - Advisory Board
  - Steering Committee Membership
- [Changes from TEI P1 to TEI P3](#)
- [Part 1: Introduction](#)
- [Part 2: Core Tags and General Rules](#)
- [Part 3: Base Tag Sets](#)
- [Part 4: Additional Tag Sets](#)
- [Part 5: Auxiliary Document Types](#)

- [Part 6: Technical Topics](#)
  - [Part 7: Alphabetical Reference List of Tags and Attributes](#)
  - [Part 8: Reference Material](#)
- 

## Resources of Related Interest

- [The Text Encoding Initiative Home Page](#)
  - [Other Electronic Versions of the TEI Guidelines](#)
  - [The Electronic Text Center Introduction to TEI and Guide to Document Preparation.](#)
  - [The Electronic Text Center SGML page.](#)
- 



# Notes on Metadata and the Web

For an overview paper on related areas, read about the [Warwick Framework](#), a container architecture for aggregating metadata.

These notes are based on the articles that appear in the Oct./Nov. 1997 issue (v. 24 no. 1) of the *Bulletin of the American Society for Information Science* (ASIS). The issue title is *Organizing Internet Resources: Metadata and the Web*.

Some of the key topics considered are:

- Dublin Core, its evolution, its adaptations
- Cataloging, MARC, and their extension to Internet
- Automatic classification: Scorpion
- Naming: URL, URN, URI, URC, DOI

## Useful Links by Topic - Alphabetical

The following links are either taken from the articles in the *Bulletin* issue or relate closely and fill in helpful information.

- [InterCat Project](#)- proof-of-concept database, made of records extracted from OCLC's WorldCat, demonstrating catalog services plus Web access to resources of the Internet
- [International Conf. on Principles and Future Development of AACR](#)- related papers, on Anglo-American Cataloging Rules, and their revision
- [Persistent URLs](#)- PURLs
- [Dublin Core Home Page](#)
- [Dublin Core Elements](#)
- [Dublin Core element Coverage](#) - proposed standard
- [Center for Electronic Text in the Humanities](#)
- [EAD \(Encoded Archival Description\): SGML for Archival Finding Aids - LoC](#)
- [EAD \(Encoded Archival Description\): SGML for Archival Finding Aids - Berkeley](#)
- [UC Berkeley Finding Aids](#)
- [Cataloging Internet Resources: Manual and Practical Guide, by Nancy B. Olson](#)
- [RDF Home Page](#)- Resource Description Framework, on metadata architecture on the Web
- [UKOLN Metadata Home Page](#)- summary of pubs, projects, metadata resources from UK and beyond, definitions
- [metadata element sets crosswalks](#)- mappings and relationships between various metadata sets, including Dublin Core
- [OCLC](#) and its [Research Department](#)
- [Stuart Weibel](#)- senior research scientist at OCLC, leader of Dublin Core efforts

- [Workshops on Metadata](#)
- [Dublin Core Workshop, 4th, official report](#) - held at National Library of Australia - and a [light-hearted account](#)
- [Resource Discovery project in Australia](#)
- [National Library of Australia PANDORA Project](#) (Preserving and Accessing Networked Documentary Resources of Australia)
- [In the Company of Strangers: Challenges and Opportunities in Metadata Implementation](#) paper by Maxine Brodie, policy level issues which impact on metadata implementation at the State Library of New South Wales, Sydney, Australia
- [Architecture for Access to Government Information](#) : report, Australia, 1996
- [ERIN - Environmental Resources Information Network](#), Australia - also runs a metadata listserv
- [Core Data Elements for Land and Geographic Directories in Australia and New Zealand](#)
- [Dataset Publishing - A Means to Motivate Metadata Entry](#), by S.D. Callahan, B.D. Johnson, and E.P. Shelley - Australian Resources, NPI Theory (choice behavior)
- [meta-searcher called HotOIL that accesses both HTTP and Z39.50 servers - demo](#) - translates user requests, merges results, displays summary
- [MetaWeb project](#) - develop and disseminate metadata tools
- [GEM](#) - educational resources - which calls for adding elements like Resource Needed, Standard, Audience, Pedagogy, Quality - see [elements](#)
- [NetFirst](#) - database/directory, cataloging of Internet (uses Dewey)
- [Canadian Information by Subject](#) - info on Canada in Internet (uses Dewey)
- [BUBL Information Service, Scotland, higher education, with subject tree](#) (uses Dewey)
- [Internet Public Library Youth Division](#) (uses Dewey)
- [Blue Web'n, by Pacific Bell, to organize Web sites for students, educators, ...](#) (uses Dewey)
- [Enhancing the indexing vocabulary of DDC by C.J. Godby](#)
- [Scorpion project at OCLC](#)

## Acknowledgements

Thanks are given to the authors of the respective articles, from whose contributions the notes above are derived. All distortions of their content and intention are the fault of E. Fox, who apologizes for any misrepresentation inadvertently resulting from this attempt to summarize a valuable set of interesting articles.

- Guest editors' intro. to Special Section, by Efthimis N. Efthimiadis and Allyson Carlyle
- Cataloging Internet Resources: Survey and Prospectus, by Erik Jul
- The Dublin Core: A Simple Content Description Model for Electronic Resources, by Stuart Weibel
- Uniform Resource Identifiers and the Effort to Bring "Bibliographic" Control" to the Web: An

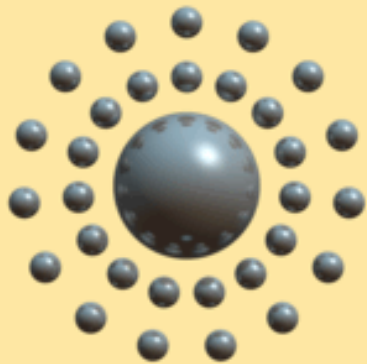
## Overview of Current Progress, by Ray Schwartz

- Options for Organizing Electronic Resources: The Coexistence of Metadata, by Sherry L. Vellucci
  - Metadata in Australia, by Carmel Maguire
  - GEM: Using Metadata to Enhance Internet Retrieval by K-12 Teachers, by Stuart Sutton and Sam G. Oh
  - From Book Classification to Knowledge Organization: Improving Internet Resource Description and Discovery, by Diane Vizine-Goetz
  - Scorpion Helps Catalog the Web, by Keith Shafer
- 

Please follow the above mentioned links to find answers to the following questions:

- What is metadata?
- How many elements are in the Dublin Core?
- What are some new elements added for educators in GEM?
- Describe TEI briefly and explain how it relates to Dublin Core work.
- Explain *finding aid*.
- Describe EAD briefly and explain how it relates to cataloging archival collections.
- Where are their detailed instructions on how to catalog the internet?
- What is RDF?
- What is happening in UK re metadata?
- What mappings are their between metadata representations?
- What is the Resource Discovery project in Australia?
- What happened at the Australian metadata meeting?
- What is covered by the Dublin Core *coverage* element?
- What metadata is needed for geographic information?
- When you search on "digital library" with HotOIL, what refinements are suggested? What are the results of the default processing of your query and what sources were used? Can you find the abstract of a talk on archiving the Internet?
- What WWW search/browse services use Dewey?
- What systems are available to automatically catalog WWW pages?

# DUBLIN CORE METADATA INITIATIVE

[Home](#)[Search](#)[Site Map](#)[What's New](#)[Feedback](#)[Home :](#)

## CONTENTS

- [Dublin Core Element Set](#)
- [About the Dublin Core Metadata Initiative](#)
- [News and Publications](#)
- [Documents](#)
- [DC in Multiple Languages](#)
- [Questions and Answers](#)
- [Projects](#)
- [Tools](#)
- [Working Groups](#)
- [Workshop Series](#)

## MIRRORS

- [Official DCMI Site](#)
- [Australian mirror](#)
- [UK mirror](#)

## Latest Important Information:

- 2000-05-06: [French translation](#) of the Dublin Core Element Set, v.1.1 is now available [[More Information](#)]
- 2000-05-04: New Project: [Francois Rabelais University Libraries](#) web site is indexed with Dublin Core. [[More Information](#)]
- 2000-04-17: [Approved Dublin Core Interoperability Qualifiers Announced](#) [[More Information](#)]
- 2000-03-28: [CEN Press Release](#) on the acceptance of Dublin Core as a CEN Workshop Agreement. [[More Information](#)]

## The Dublin Core: A Simple Content Description Model for Electronic Resources

### Metadata for Electronic Resources

The Dublin Core is a metadata element set intended to facilitate discovery of electronic resources. Originally conceived for author-generated description of Web resources, it has attracted the attention of formal resource description communities such as museums, libraries, government agencies, and commercial organizations.

The Dublin Core Workshop Series has gathered experts from the library world, the networking and digital library research communities, and a variety of content specialties in a series of invitational workshops. The building of an interdisciplinary, international consensus around a core element set is the central feature of the Dublin Core. The progress represents the emergent wisdom and collective experience of many stakeholders in the resource description arena. An open mailing list supports ongoing work.

The characteristics of the Dublin Core that distinguish it as a prominent candidate for description of electronic resources fall into several categories:

### Simplicity

The Dublin Core is intended to be usable by non-catalogers as well as resource description specialists. Most of the elements have a commonly understood semantics of roughly the complexity of a library catalog card.

### Semantic Interoperability

In the Internet Commons, disparate description models interfere with the ability to search across discipline boundaries. Promoting a commonly

understood set of descriptors that helps to unify other data content standards increases the possibility of semantic interoperability across disciplines.

### **International Consensus**

Recognition of the international scope of resource discovery on the Web is critical to the development of effective discovery infrastructure. The Dublin Core benefits from active participation and promotion in some 20 countries in North America, Europe, Australia, and Asia.

### **Extensibility**

The Dublin Core provides an economical alternative to more elaborate description models such as the full MARC cataloging of the library world. Additionally, it includes sufficient flexibility and extensibility to encode the structure and more elaborate semantics inherent in richer description standards

### **Metadata Modularity on the Web**

The diversity of metadata needs on the Web requires an infrastructure that supports the coexistence of complementary, independently maintained metadata packages. The World Wide Web Consortium (W3C) has begun implementing an architecture for metadata for the Web. The Resource Description Framework, or RDF, is designed to support the many different metadata needs of vendors and information providers. Representatives of the Dublin Core effort are actively involved in the development of this architecture, bringing the digital library perspective to bear on this important component of the Web infrastructure.



For questions or  
comments regarding  
the Dublin Core  
contact [dc@oclc.org](mailto:dc@oclc.org)

---

[Home](#) | [Search](#) | [Site Map](#) | [What's New](#) | [Feedback](#) |  
[About the Dublin Core](#) | [News and Publications](#) | [Documents](#) |  
[Questions and Answers](#) | [Schemas](#) | [Projects](#) | [Tools](#) | [Working Groups](#) |  
[Workshop Series](#)

# Alliance Metadata Standards Working Group



The Alliance Metadata Standards Working Group is an NCSA effort to develop metadata interoperability standards for use with scientific data collections on the Grid. Its work is tightly integrated with the [Grid Forum's Grid Information Service Working Group](#).

## More Information

- Some slides outlining the mission of the WG. [[.ppt](#)]
- Meeting reports and summaries
  - [Meeting of Alliance Science Portals Technical Working Group](#), Oct. 22-23, 1999.
  - [Meeting with AT Teams](#), Dec 10, 1999
  - [Meeting with Sun Portal Technologies Representatives](#), Dec 14, 1999
- Relevant Working Documents
  - The Data Grid (Ann Chervenak, Ian Foster, Carl Kesselman, Charles Salisbury, Steve Tuecke) [[.pdf](#)]
  - Chemical Engineering Scenarios:
    - Electrochemical Deposition and Dissolution Including Corrosion (Dick Alkire) [[.doc](#) [.html](#)]
    - Modeling and Control of Multidimensional Crystal Growth (Richard Braatz [[.pdf](#)])
- Useful [Links](#)

# The Data Grid: Towards an Architecture for the Distributed Management and Analysis of Large Scientific Datasets

Ann Chervenak\*   Ian Foster<sup>§+</sup>   Carl Kesselman\*   Charles Salisbury<sup>§</sup>   Steven Tuecke<sup>§</sup>

\* Information Sciences Institute, University of Southern California

<sup>§</sup> Mathematics and Computer Science Division, Argonne National Laboratory

<sup>+</sup> Department of Computer Science, The University of Chicago

## 1 Introduction

In an increasing number of scientific disciplines, large data collections are emerging as important community resources. In domains as diverse as global climate change, high energy physics, and computational genomics, the volume of interesting data is already measured in terabytes and will soon total petabytes. The communities of researchers that need to access and analyze this data (often using sophisticated and computationally expensive techniques) are often large and are almost always geographically distributed, as are the computing and storage resources that these communities rely upon to store and analyze their data [17].

This combination of large dataset size, geographic distribution of users and resources, and computationally intensive analysis results in complex and stringent performance demands that are not satisfied by any existing data management infrastructure. A large scientific collaboration may generate many queries, each involving access to—or supercomputer-class computations on—gigabytes or terabytes of data. Efficient and reliable execution of these queries may require careful management of terabyte caches, gigabit/s data transfer over wide area networks, coscheduling of data transfers and supercomputer computation, accurate performance estimations to guide the selection of dataset replicas, and other advanced techniques that collectively maximize use of scarce storage, networking, and computing resources.

The literature offers numerous point solutions that address these issues (e.g., see [17, 14, 19, 3]). But no integrating architecture exists that allows us to identify requirements and components common to different systems and hence apply different technologies in a coordinated fashion to a range of data-intensive petabyte-scale application domains.

Motivated by these considerations, we have launched a collaborative effort to design and produce such an integrating architecture. We call this architecture the *data grid*, to emphasize its role as a specialization and extension of the “Grid” that has emerged recently as an integrating infrastructure for distributed computation [10, 20, 15]. Our goal in this effort is to define the requirements that a data grid must satisfy and the components and APIs that will be required in its implementation. We hope that the definition of such an architecture will accelerate progress on petascale data-intensive computing by enabling the integration of currently disjoint approaches, encouraging the deployment of basic enabling technologies, and revealing technology gaps that require further research and development. In addition, we plan to construct a reference implementation for this architecture so as to enable large-scale experimentation.

This work complements other activities in data-intensive computing. Work on high-speed disk caches [21] and on tertiary storage and cache management [5, 19] provides basic building blocks. Work within the digital library community is developing relevant metadata standards and metadata-driven retrieval mechanisms [16, 6, 1] but has focused less on high-speed movement of large data objects, a particular focus of our work. The Storage Resource Broker (SRB) [2] shows how diverse storage systems can be integrated under uniform metadata-driven access mechanisms; it provides a valuable building block for our architecture but should also benefit from the basic services described here. The High Performance Storage System (HPSS) [24] addresses enterprise-level concerns (e.g., it assumes that all accesses occur within the same DCE cell); our work addresses new issues associated with wide area access from multiple administrative domains.

In this paper, we first review the principles that we are following in developing a design for a data grid architecture. Then, we describe two basic services that we believe are fundamental to the design of a data grid, namely, storage systems and metadata management. Next, we explain how these services can be used to develop various higher-level services for replica management and replica selection. We conclude by describing our initial implementation of data grid functionality.

## 2 Data Grid Design

The following four principles drive the design of our data grid architecture. These principles derive from the fact that data grid applications must frequently operate in wide area, multi-institutional, heterogeneous environments, in which we cannot typically assume spatial or temporal uniformity of behavior or policy.

*Mechanism neutrality.* The data grid architecture is designed to be as independent as possible of the low-level mechanisms used to store data, store metadata, transfer data, and so forth. This goal is achieved by defining data access, third-party data mover, catalog access, and other interfaces that encapsulate peculiarities of specific storage systems, catalogs, data transfer algorithms, and the like.

*Policy neutrality.* The data grid architecture is structured so that, as far as possible, design decisions with significant performance implications are exposed to the user, rather than encapsulated in “black box” implementations. Thus, while data movement and replica cataloging are provided as basic operations, replication policies are implemented via higher-level procedures, for which defaults are provided but that can easily be substituted with application-specific code.

*Compatibility with Grid infrastructure.* We attempt to overcome the difficulties of wide area, multi-institutional operation by exploiting underlying Grid infrastructure [10, 20, 15] (e.g., Globus [9]) that provides basic services such as authentication, resource management, and information. To this end, we structure the data grid architecture so that more specialized data grid tools are compatible with lower-level Grid mechanisms. This approach also simplifies the implementation of strategies that integrate, for example, storage and computation.

*Uniformity of information infrastructure.* As in the underlying Grid, uniform and convenient access to information about resource structure and state is emphasized as a means of enabling runtime adaptation to system conditions. In practice, this means that we use the same data model and interface to access the data grid’s metadata, replica, and instance catalogs as are used in the underlying Grid information infrastructure.

These four principles lead us to develop a layered architecture (Figure 1), in which the lowest layers provide high-performance access to an orthogonal set of basic mechanisms, but do not enforce specific usage policies. For example, we define high-speed data movement functions with rich error interfaces as a low-level mechanism, but do not encode within these functions how to respond to storage system failure. Rather, such policies are implemented in higher layers of the architecture, which build on the

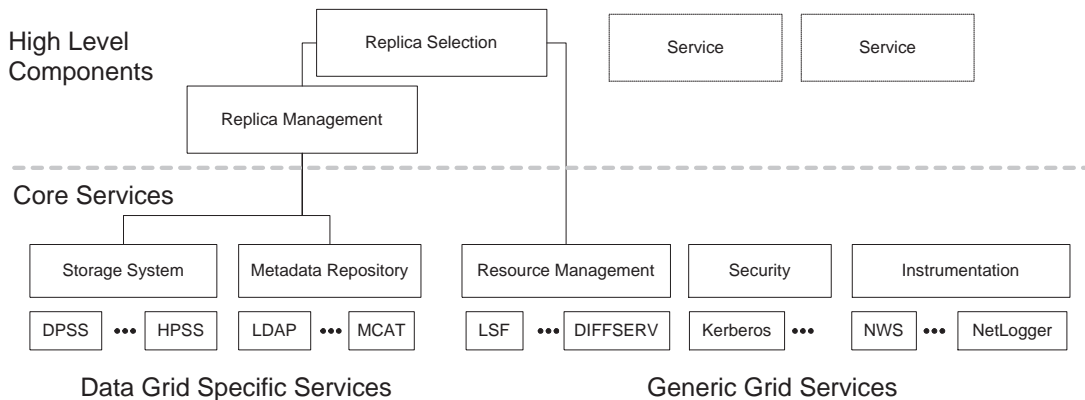


Figure 1: Major components and structure of the data grid architecture

mechanisms provided by the basic components.

This approach is motivated by the observation that achieving high performance in specific applications often requires that an implementation exploit domain-specific or application-specific knowledge. In data grids, as in other Grid systems, this focus on simple, policy-independent mechanisms will encourage and enable broad deployment without limiting the range of applications that can be implemented. By limiting application specific behaviors to the upper layers of the architecture, we can promote reuse of the basic mechanisms while delivering high-performance and specialized capabilities to the end user and application.

### 3 Core Data Grid Services

We now turn our attention to the basic services required in a data grid architecture. We focus in particular on two services that we view as fundamental: data access and metadata access. The data access service provides mechanisms for accessing, managing, and initiating third-party transfers of data stored in storage systems. The metadata access service provides mechanisms for accessing and managing information about data stored in storage systems. This explicit distinction between storage and metadata is worth discussing briefly. In some circumstances, for example when data is being stored in a database system, there are advantages to combining metadata and storage into the same abstraction. However, we believe that keeping these concepts separate at the architectural level enhances flexibility in storage system implementation while having minimal impact on the implementation of behaviors that combine metadata access with storage access.

#### 3.1 Storage Systems and the Grid Storage API

In a Grid environment, data may be stored in different locations and on different devices with different characteristics. As we discussed above, mechanism neutrality implies that applications should not need to be aware of the specific low-level mechanisms required to access data at a particular location. Instead, applications should be presented with a uniform view of data and with uniform mechanisms for accessing that data. These requirements are met by the storage system abstraction and our grid storage API. Together, these define our data access service.

### 3.1.1 Data Abstraction: Storage Systems

We introduce as a basic data grid component what we call a *storage system*, which we define as an entity that can be manipulated with a set of functions for creating, destroying, reading, writing, and manipulating the attributes of named sequences of bytes called *file instances*.

Notice that our definition of a storage system is a logical one: a storage system can be implemented by any storage technology that can support the required access functions. Implementations that target Unix file systems, HTTP servers, hierarchical storage systems such as HPSS, and network caches such as the Distributed Parallel Storage System (DPSS) are certainly envisioned. In fact, a storage system need not map directly to a single low-level storage device. For example, a distributed file system that manages files distributed over multiple storage devices or even sites can serve as a storage system, as can an SRB system that serves requests by mapping to multiple storage systems of different types.

Our definition of a file instance is also logical rather than physical. A storage system holds data, which may actually be stored in a file system, database, or other system; we do not care about how data is stored but specify simply that the basic unit that we deal with is a named sequences of uninterpreted bytes. The use of the term “file instance” for this basic unit is not intended to imply that the data must live in a conventional file system. For example, a data grid implementation might use a system such as SRB to access data stored within a database management system.

A storage system will associate with each of the file instances that it contains a set of properties, including a name and attributes such as its size and access restrictions. The name assigned to a file instance by a particular storage system is arbitrary and has meaning only to that storage system. In many storage systems, a name will be a hierarchical directory path. In other systems such as SRB, it may be a set of application metadata that the storage system maps internally to a physical file instance.

### 3.1.2 Grid Storage API

The behavior of a storage system as seen by a data grid user is defined by the data grid storage API, which defines a variety of operations on storage systems and file instances. Our understanding of the functionality required in this API is still evolving, but it certainly should include support for remote requests to read and/or write named file instances and to determine file instance attributes such as size. In addition, to support optimized implementation of replica management services (discussed below) we require a third party transfer operation used to transfer the entire contents of a file instance from one storage system to another.

While the basic storage system functions just listed are relatively simple, various data grid considerations can increase the complexity of an implementation. For example, storage system access functions must be integrated with the security environment of each site to which remote access is required [12]. Robust performance within higher-level functions requires reservation capabilities within storage systems and network interfaces [11]. Applications should be able to provide storage systems with hints concerning access patterns, network performance, and so forth that the storage system can use to optimize its behavior. Similarly, storage systems should be capable of characterizing and monitoring their own performance; this information, when made available to storage system clients, allows them to optimize their behavior. Finally, data movement functions must be able to detect and report errors. While it may be possible to recover from some errors within the storage system, other errors may need to be reported back to the remote application that initiated the movement.

### 3.2 The Metadata Service

The second set of basic machinery that we require is concerned with the management of information about the data grid itself, including information about file instances, the contents of file instances, and the various storage systems contained in the data grid. We refer to this information as *metadata*. The *metadata service* provides a means for publishing and accessing this metadata.

Various types of metadata can be distinguished. It has become common practice to associate with scientific datasets metadata that describes the contents and structure of that data. The metadata may describe the information content represented by the file, the circumstances under which the data was obtained, and/or other information useful to applications that process the data. We refer to this as *application metadata*. Such metadata can be viewed as defining the logical structure or semantics that should apply to the uninterpreted bytes that make up a file instance or a set of file instances. A second type of metadata is used to describe the fabric of the data grid itself: for example, details about storage systems, such as their capacity and usage policy, as well as information about file instances stored within a given storage system.

The metadata service provides a uniform means for naming, publishing, and accessing these different types of metadata. Each type of metadata has its own characteristics in terms of frequency and mechanism of update and its logical relationship to other grid components and data items. Interesting data management applications are likely to use several kinds of metadata. Although we have referred to several different sources of metadata, we propose that a single interface be used for accessing all types of metadata.

The difficulty of specifying a general structure for all metadata is apparent when one considers the variety of approaches used to describe application metadata. Some applications build a metadata repository from a specified list of file instances based on data stored in a self-describing format (e.g., NetCDF, HDF). High energy physics applications are successfully using a specialized indexing structure. The Digital Library community is developing sets of metadata for different fields (e.g., citedli3). Other user communities are pursuing the use of eXtended Markup Language (XML) [4] to represent application metadata.

The situation is further complicated when one considers the additional requirements imposed by large-scale data grid environments. Besides providing a means of integrating the different approaches to metadata storage and representation, the service must operate efficiently in a distributed environment. It must be scalable, supporting metadata about large number of entities being contributed by large numbers of information sources located in large numbers of organizations. The service must be robust in the face of failure, and organizations should be able to assert local control over their information.

Analysis of these requirements leads us to conclude that the metadata service must be structured as a hierarchical and distributed system. This approach allows us to achieve scalability, avoid any single point of failure, and facilitate local control over data. Distribution does complicate efficient retrieval, but this difficulty can be overcome by having data organization exploit the hierarchical nature of the metadata service.

This analysis leads us to propose that the metadata service be treated as a distributed directory service, such as that provided by the Lightweight Directory Access Protocol (LDAP) [23]. Such systems support a hierarchical naming structure and rich data models and are designed to enable distribution. Mechanisms defined by LDAP include a means for naming objects, a data model based on named collections of attributes, and a protocol for performing attribute-based searching and writing of data elements. We have had extensive experience in using distributed directory services to represent general Grid metadata [8], and we believe that they will be well suited to the metadata requirements of data grids as well.

The directory hierarchy associated with LDAP provides a structure for organizing, replicating,

and distributing catalog information. However, the directory service does not specify how the data is stored or where it is stored. Queries may be referred between servers, and the LDAP protocol can be placed in front of a wide range of alternative information and metadata services. This capability can provide a mechanism for the data grid to support a wide variety of approaches to providing application metadata, while retaining a consistent overall approach to accessing that metadata.

### 3.3 Other Basic Services

The data grid architecture also assumes the existence of a number of other basic services, including the following:

- An authorization and authentication infrastructure that supports multi-institutional operation. The public key-based Grid Security Infrastructure (GSI) [12] meets our requirements.
- Resource reservation and co-allocation mechanisms for both storage systems and other resources such as networks, to support the end-to-end performance guarantees required for predictable transfers (e.g., [11]).
- Performance measurements and estimation techniques for key resources involved in data grid operation, including storage systems, networks, and computers (e.g., the Network Weather Service [25]).
- Instrumentation services that enable the end-to-end instrumentation of storage transfers and other operations (e.g., NetLogger [22], Pablo [18], and Paradyn [13]).

## 4 Higher-Level Data Grid Components

A potentially unlimited number of components can exist in the upper layer of the data grid architecture. Consequently, we will limit our discussion to two representative components: replica management and replica selection.

### 4.1 Replica and Cache Management

A *replica manager* is a data grid service whose functionality can be defined in terms of that provided by the storage system and metadata repository services. The role of a replica manager is to create (or delete) copies of file instances, or replicas, within specified storage systems. Typically, a replica is created because the new storage location offers better performance or availability for accesses to or from a particular location. (In this section, we use the terms *replica* and *file instance* interchangeably.) A replica might be deleted because storage space is required for another purpose.

In this discussion, we assume that replicated files are read only; we are not concerned with issues of file update and coherency. Thus, replicas are primarily useful for access to “published” data sets. While this read only model is sufficient for many uses of scientific data sets, we intend to investigate support for modifying the contents of file instances in the future.

For convenience, we group all of the replicas of a file instance along with any associated metadata into a single entry in the metadata repository. We call this entry a *logical file*, since it represents the logical structure (i.e. metadata) associated with the referenced file instances. In most situations, the file instances contained in a logical file will be byte-for-byte copies of one another, but this is not required.

A logical file exists in the metadata repository, since it describes attributes of a set of file instances, and its position in the repository provides a logical file with a globally unique name. To facilitate data

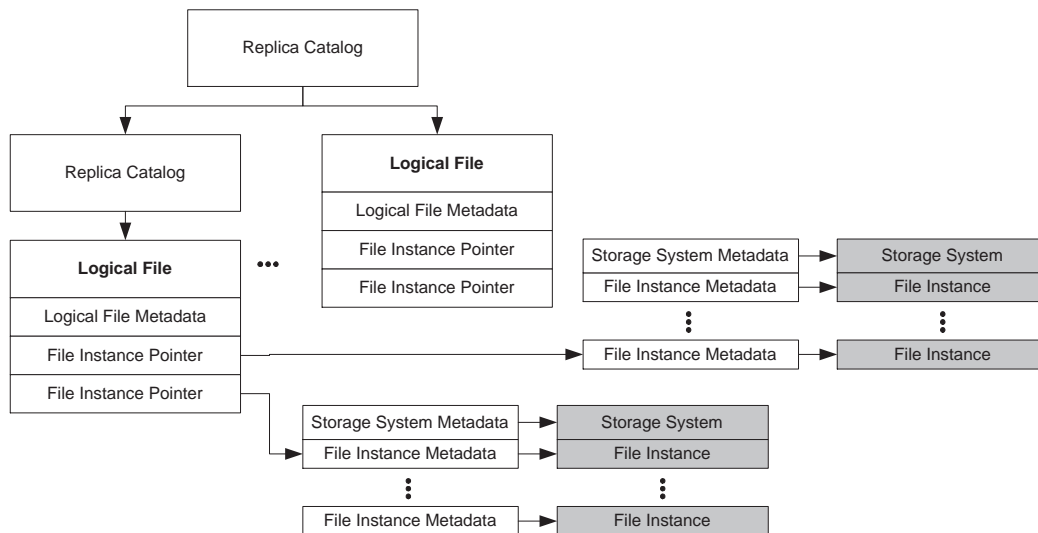


Figure 2: The structure of a replica catalog. Boxes shaded in gray represent storage system entities, all other boxes represent entity in the metadata repository

discovery, related logical files are grouped into collections, called replica catalogs, that are stored at well-known locations in the metadata repository. The relationship between file instances, logical files, and replica catalogs is shown in Figure 2.

A data grid may (and indeed typically will) contain multiple replica catalogs. For example, a community of researchers interested in a particular research topic might maintain a replica catalog for a collection of data sets of mutual interest. Replica catalogs can thus provide the functionality of logical collections, grouping logical files on related topics. It is possible to create hierarchies of replica catalogs to impose a directory-like structure on related logical collections.

A replica manager can perform access control on entire catalogs as well as on individual logical files. By combining the functionality provided by the storage system and metadata repository, the replica manager also can perform a number of basic operations, including creation and deletion of replicas, logical files, and replica catalogs.

Note that the existence of a replica manager does not determine when or where replicas are created, or which replicas are to be used by an application, nor does it even require that every file instance be entered into a replica catalog. In keeping policy out of the definition of the replica manager, we maximize the types of situations in which the replica manager will be useful. For example, a file instance that is not entered into the catalog may be considered to be in a local “cache” and available for local use only. Designing this as a policy rather than coupling file movement with catalog registration in a single atomic operation explicitly acknowledges that there may be good, user-defined reasons for satisfying application needs by using files that are not registered in a replica catalog.

## 4.2 Replica Creation and Replica Selection

The second representative high-level service provided in the upper level of the data grid is replica selection. Replica selection is interesting because it does not build on top of the core services, but rather relies on the functions provided by the replica management component described in the preceding section. Replica selection is the process of choosing a replica that will provide an application with data

access characteristics that optimize a desired performance criterion, such as absolute performance (i.e. speed), cost, or security. The selected file instance may be local or accessed remotely. Alternatively the selection process may initiate the creation of a new replica whose performance will be superior to the existing ones.

Where replicas are to be selected based on access time, Grid information services can provide information about network performance, and perhaps the ability to reserve network bandwidth, while the metadata repository can provide information about the size of the file. Based on this, the selector can rank all of the existing replicas to determine which one will yield the fastest data access time. Alternatively, the selector can consult the same information sources to determine whether there is a storage system that would result in better performance if a replica was created on it.

A more general selection service may consider access to subsets of a file instance. Scientific experiments often produce large files containing data for many variables, time steps, or events, and some application processing may require only a subset of this data. In this case, the selection function may provide an application with a file instance that contains only the needed subset of the data found in the original file instance. This can obviously reduce the amount of data that must be accessed or moved.

This type of replica management has been implemented in other data-management systems. For example, STACS is often capable of satisfying requests from High Energy Physics applications by extracting a subset of data from a file instance. It does this using a complex indexing scheme that represents application metadata for the events contained within the file. Other mechanisms for providing similar function may be built on application metadata obtainable from self-describing file formats such as NetCDF or HDF.

Providing this capability requires the ability to invoke filtering or extraction programs that understand the structure of the file and produce the required subset of data. This subset becomes a file instance with its own metadata and physical characteristics, which are provided to the replica manager. Replication policies determine whether this subset is recognized as a new logical file (with an entry in the metadata repository and a file instance recorded in the replica catalog), or whether the file should be known only locally, to the selection manager.

Data selection with subsetting may exploit Grid-enabled servers, whose capabilities involve common operations such as reformatting data, extracting a subset, converting data for storage in a different type of system, or transferring data directly to another storage system in the Grid. The utility of this approach has been demonstrated as part of the Active Data Repository [7]. The subsetting function could also exploit the more general capabilities of a computational Grid such as that provided by Globus. This offers the ability to support arbitrary extraction and processing operations on files as part of a data management activity.

## 5 Status of the Data Grid Implementation

We have made progress on several fronts in our effort to identify the basic low-level services for a data grid architecture. In particular, we have defined a Grid Storage API, a standard interface to storage systems that includes create, delete, open, close, read and write operations on file instances. This interface also supports storage to storage transfers. We have implemented the interface for several storage systems, including local file access, HTTP servers, and DPSS network disk caches.

We also have defined simple replica management and metadata services. These services use the MDS information infrastructure to store attribute information about file instances, storage systems, logical files, and replica catalogs. Using these attributes, we can query the information system to find the replicas associated with a logical file, estimate their performance, and select among replicas

according to particular performance metrics.

This work represents the first steps in our effort to create an *integrating architecture* for data-intensive petabyte-scale application domains. Performance studies of the Grid Storage API are under way, and we plan to further explore basic services such as instrumentation.

## Acknowledgments

We gratefully acknowledge helpful discussions with Steve Fitzgerald, Bill Johnston, Reagan Moore, Richard Mount, Harvey Newman, Arie Shoshani, Brian Tierney and other participants in the DOE Next Generation Internet “Earth System Grid” and “Particle Physics Data Grid” projects. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Computational and Technology Research, U.S. Department of Energy, under Contract W-31-109-Eng-38.

## References

- [1] M. Baldonado, C. Chang, L. Gravano, and A. Paepcke. The Stanford digital library metadata architecture. *Intl J. Digital Libraries*, 1(2):108–121, 1997.
- [2] Chaitanya Baru, Reagan Moore, Arcot Rajasekar, and Michael Wan. The SDSC storage resource broker. In *Proceedings of CASCON’98 Conference*. 1998.
- [3] M. Beck and T. Moore. The Internet2 distributed storage infrastructure project: An architecture for internet content channels. *Computer Networking and ISDN Systems*, 30(22-23):2141–2148, 1998.
- [4] Tim Bray, Jean Paoli, and C. M. Sperberg-McQueen. The extensible markup language (xml) 1.0. W3C recommendation, World Wide Web Consortium, February 1998. See <http://www.w3.org/TR/1998/REC-xml-19980210>.
- [5] L.T. Chen, R. Drach, M. Keating, S. Louis, D. Rotem, and A. Shoshani. Efficient organization and access of multi-dimensional datasets on tertiary storage systems. *Information Systems Special Issue on Scientific Databases*, 20(2):155–83, 1995.
- [6] S. Cousins, H. Garcia-Molina, S. Hassan, S. Ketchpel, M. Roscheisen, and T. Winograd. Towards interoperability in digital libraries. *IEEE Computer*, 29(5), 1996.
- [7] Renato Ferreira, Tahsin Kurc, Michael Beynon, Chialin Chang, Alan Sussman, and Joel Saltz. Object-relational queries into multidimensional databases with the active data repository. *International Journal of Supercomputer Applications*, 1999.
- [8] S. Fitzgerald, I. Foster, C. Kesselman, G. von Laszewski, W. Smith, and S. Tuecke. A directory service for configuring high-performance distributed computations. In *Proc. 6th IEEE Symp. on High Performance Distributed Computing*, pages 365–375. IEEE Computer Society Press, 1997.
- [9] I. Foster and C. Kesselman. Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications*, 11(2):115–128, 1997.
- [10] I. Foster and C. Kesselman, editors. *The Grid: Blueprint for a Future Computing Infrastructure*. Morgan Kaufmann Publishers, 1999.

- [11] I. Foster, C. Kesselman, C. Lee, R. Lindell, K. Nahrstedt, and A. Roy. A distributed resource management architecture that supports advance reservations and co-allocation. In *Proceedings of the International Workshop on Quality of Service*, pages 27–36, 1999.
- [12] I. Foster, C. Kesselman, G. Tsudik, and S. Tuecke. A security architecture for computational grids. In *ACM Conference on Computers and Security*, pages 83–91. ACM Press, 1998.
- [13] Jeffrey Hollingsworth and Bart Miller. Instrumentation and measurement. In [10], pages 339–365.
- [14] William Johnston. Realtime widely distributed instrumentation systems. In [10], pages 75–103.
- [15] William E. Johnston, Dennis Gannon, and Bill Nitzberg. Grids as production computing environments: The engineering aspects of NASA’s Information Power Grid. In *Proc. 8th IEEE Symp. on High Performance Distributed Computing*. IEEE Computer Society Press, 1999.
- [16] M. Lesk. *Practical Digital Libraries: Books, Bytes, and Bucks*. Morgan Kaufmann Publishers, 1997.
- [17] Reagan Moore, Chaitanya Baru, Richard Marciano, Arcot Rajasekar, and Michael Wan. Data-intensive computing. In [10], pages 105–129.
- [18] Daniel Reed and Randy Ribler. Performance analysis and visualization. In [10], pages 367–393.
- [19] A. Shoshani, L. M. Bernardo, H. Nordberg, D. Rotem, and A. Sim. Storage management for high energy physics applications. In *Computing in High Energy Physics 1998 (CHEP 98)*. 1998. <http://www.lbl.gov/arie/papers/proc-CHEP98.ps>.
- [20] R. Stevens, P. Woodward, T. DeFanti, and C. Catlett. From the I-WAY to the National Technology Grid. *Communications of the ACM*, 40(11):50–61, 1997.
- [21] B. Tierney, W. Johnston, L. Chen, H. Herzog, G. Hoo, G. Jin, and J. Lee. Distributed parallel data storage systems: A scalable approach to high speed image servers. In *Proc. ACM Multimedia 94*. ACM Press, 1994.
- [22] B. Tierney, W. Johnston, B. Crowley, G. Hoo, C. Brooks, and D. Gunter. The NetLogger methodology for high performance distributed systems performance analysis. In *Proc. 7th IEEE Symp. on High Performance Distributed Computing*. IEEE Computer Society Press, 1998.
- [23] M. Wahl, T. Howes, and S. Kille. Lightweight directory access protocol (v3). RFC 2251, Internet Engineering Task Force, 1997.
- [24] Richard W. Watson and Robert A. Coyne. The parallel I/O architecture of the High-Performance Storage System (HPSS). In *IEEE MSS Symposium*, 1995.
- [25] R. Wolski. Forecasting network performance to support dynamic scheduling using the network weather service. In *Proc. 6th IEEE Symp. on High Performance Distributed Computing*, Portland, Oregon, 1997. IEEE Press.

# Electronic Publishing:

---

- [The SGML/XML Web Page](#)
  - [CS5604 unit on SGML](#): check out the related course notes offered at Virginia Tech.
  - [Elsevier](#)  
[TULIP](#)
- 

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

---

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

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

# The SGML/XML Web Page has moved

Either click on the new location below, or wait 30 seconds and you will be automatically redirected.

<http://www.oasis-open.org/cover/sgml-xml.html>

Please take this opportunity to change your personal bookmarks and any links on public Web pages under your authority. Thank you!

Information about the site move is provided in a [note](#).

---

[\[SIL Home\]](#)

- Site Index
- News
- Applications
- Articles
- Software
- Biblio
- Events
- XML
- XSL
- XLink
- DSSSL
- CSS
- HyTime
- Search

Support for  
The XML Cover Pages  
is provided by:



# the XML COVER PAGES

Robin Cover, Managing Editor

Hosted by:

The XML Cover Pages is a comprehensive online reference work for the Extensible Markup Language (XML) and its parent, the Standard Generalized Markup Language (SGML). The reference collection features extensive documentation on the application of the open, interoperable "markup language" standards, including XSL, XSLT, XPath, XLink, XPointer, HyTime, DSSSL, CSS, SPDL, CGM, ISO-HTML, and others.

The XML Cover Pages is currently [sponsored](#) by [OASIS](#) (Organization for the Advancement of Structured Information Standards) and four OASIS Members: [ISOGEN International Corp](#), [Software AG](#), [Sun Microsystems](#), and [webMethods](#).

## What's New...

Read the [most recent SGML/XML news . . .](#)

## Overview



- [The XML Cover Pages](#)
- [News](#)
- [Introductions](#)
- [XML, XSL, XLink](#)
- [Related Standards](#)
- [Application Standards](#)

- [Publications](#)
- [Software](#)
- [Support](#)
- [Events](#)
- [Special Topics](#)
- [Contacts](#)

▲ The XML Cover Pages	<ul style="list-style-type: none"><li>● <a href="#">Site Index</a></li><li>● <a href="#">Site Description</a></li><li>● <a href="#">Site Search</a></li></ul>
▲ News	<ul style="list-style-type: none"><li>● <a href="#">What's New in the XML Cover Pages?</a></li><li>● <a href="#">XML News Articles</a></li><li>● <a href="#">XML Press News</a></li><li>● Earlier News: [<a href="#">1999 Q3</a>] - [<a href="#">1999 Q2</a>] - [<a href="#">1999 Q1</a>] - [<a href="#">1998</a>] - [<a href="#">1997</a>] - [<a href="#">1996</a>] - [<a href="#">1995</a>]</li></ul>

Web site [sponsorship opportunities...](#)

▲ Introductions	<ul style="list-style-type: none"> <li>● <a href="#">General Introduction to SGML</a></li> <li>● <a href="#">General Introduction to XML</a></li> <li>● <a href="#">SGML Frequently Asked Questions (FAQs)</a></li> <li>● <a href="#">XML Frequently Asked Questions (FAQs)</a></li> </ul>
▲ XML, XSL, XLink	<ul style="list-style-type: none"> <li>● <a href="#">XML (Extensible Markup Language)</a></li> <li>● <a href="#">XSL (Extensible Stylesheet Language)</a></li> <li>● <a href="#">XLink (XLink, XPath and XPointer)</a></li> <li>● <a href="#">XML Schemas</a></li> </ul>
▲ Related Standards	<ul style="list-style-type: none"> <li>● <a href="#">Style - CSS</a></li> <li>● <a href="#">Style - DSSSL</a></li> <li>● <a href="#">Hypermedia - HyTime</a></li> <li>● <a href="#">Other Standards Related to SGML/XML</a></li> </ul>
▲ Applications	<ul style="list-style-type: none"> <li>● <a href="#">General SGML/XML Applications</a></li> <li>● <a href="#">Academic Applications</a></li> <li>● <a href="#">Government and Industry Applications</a></li> <li>● <a href="#">Proposed XML Applications</a></li> </ul>
▲ Publications	<ul style="list-style-type: none"> <li>● <a href="#">Essential SGML/XML Books</a></li> <li>● <a href="#">Comprehensive SGML/XML Bibliography</a></li> <li>● <a href="#">Journals, Newsletters and other Serials</a></li> <li>● <a href="#">XML Books</a></li> <li>● <a href="#">XML Articles</a></li> <li>● XML Article Archive: [<a href="#">1999</a>] [<a href="#">1998</a>] [<a href="#">1997</a>]</li> </ul>
▲ Software	<ul style="list-style-type: none"> <li>● <a href="#">Public Software Tools for SGML/XML/DSSSL</a></li> <li>● <a href="#">XML Software Tools</a></li> <li>● <a href="#">XSL Software Tools</a></li> <li>● <a href="#">Commercial SGML/XML Software</a></li> </ul>
▲ Support	<ul style="list-style-type: none"> <li>● <a href="#">Industry Consortia, SIGS, Working Groups</a></li> <li>● <a href="#">SGML/XML Mailing Lists and Discussion Groups</a></li> <li>● Special Lists and Groups for <a href="#">XML</a> and <a href="#">XSL</a></li> <li>● <a href="#">Commercial XML Support</a></li> </ul>
▲ Events	<ul style="list-style-type: none"> <li>● <a href="#">Conferences, Seminars, Tutorials, Workshops</a></li> </ul>

 Special Topics	<ul style="list-style-type: none"><li>● <a href="#">SGML/XML Grammar</a></li><li>● <a href="#">Architectural Forms and SGML/XML Architectures</a></li><li>● <a href="#">Groves, Grove Plans, Property Sets</a></li><li>● <a href="#">SGML/XML and (La)TeX</a></li><li>● <a href="#">Miscellaneous</a></li></ul>
 Contacts	<ul style="list-style-type: none"><li>● <a href="#">Contact Addresses - Corporate Entities</a></li><li>● <a href="#">Personal Home Pages - Some SGML/XML Experts</a></li></ul>

 [Top](#)

Copyright © Robin Cover and OASIS, 1994-2000. [Other legal notices](#).

Document URL: <http://www.oasis-open.org/cover/sgml-xml.html>.

Please send comments and corrections to: [robin@isogen.com](mailto:robin@isogen.com)

# Ontologies and Agents in Digital Libraries

Key topics about *Ontology* adapted from *AI Magazine*, Fall 1997, 18(3), include:

- Defn
- Comparison criteria
- Top level categories, taxonomy. categories, realtions, axioms
- Comparison chart

URLs related include:

- [Ontologies](#)
  - [Indented list diagrams of important ontologies](#)
  - [CYC Home Page](#) and [ontology](#) and [table of contents](#)
  - [WordNet Home Page](#) and [online demo](#)
  - Generalized Upper Model: [model](#), [overall organization](#), [concept hierarchy](#), [relational hierarchy](#)
  - [UMLS Home Page](#) and [fact sheets](#), [MeSH](#), [Grateful Med](#) and [demo](#)
  - [TOVE - Toronto Virtual Enterprise](#)
  - [KIF](#) - Knowledge Interchange Format and [brief intro](#)
  - [Stanford Knowledge Modeling Group](#) and [Layout Editor](#)
  - [Ontolingua](#)
  - [EUROKNOWLEDGE Glossary etc.](#)
  - [Stanford DLI](#) and [agents](#), especially for Web browsing
    - [InterPay : Shopping Models](#), [Secure Electronic Marketplace for Europe](#)
  - [ILU](#) and [Stanford testbed use](#)
  - [Agents '97 Conf.](#)
  - [CHI '97 Software Agents Tutorial](#) by Pattie Maes and her [Software Agents Group](#)
  - [My Yahoo](#) (successor to Webdoggie from MIT)
  - [IBM Agents](#), [and the Agent Building Environment \(ABE\): A toolkit for building intelligent agent applications](#)
  - [Machine Learning software and datasets](#) - naive Bayes classifier - see *AI Magazine* Fall 1997 p. 18
  - [IBM DL: QBIC](#), [watermarking](#)
  - Hal Berghel: [CACM Nov. 1997 40\(11\): Watermarking Cyberspace](#), and [IEEE Computer 29:7 article](#)
  - [DigiCash](#) (Ch. 11)
- 
- Agents: people and places
    - [iimam@site.gmu.edu](#) adaptatation, intelligence

- yves.Kodratoff@Iri.Iri.fr
- Brian Gaines, U. Calgary: society of agents
- Haynes, Sen : U. Tulsa: cases
- Rus, Dartmouth: gather info
- Decker, Sycara, Williamson: CMU: multiagent society, planning, matchmaker info agent

Questions:

- Try WordNet on "library" and look for coordinate terms on senses 1,2,3
- Try Grateful Med and find MeSH / Meta Terms for "diabetes"

# Commerce, Economics, Publishers:

---

## NetBill

- [Home Page](#)
- [Demo](#)
- [Overview article on payment systems from IEEE Spectrum](#)
- Questions: How would this work with ETDs? What are the advantages and disadvantages relative to other approaches?

## Commerce part of CS6604 lecture

- [Workshop on Tech. of Terms and Conditions](#) and [Final Report to NSF](#) - including Breakout Group Reports
- [EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce](#), Hamburg, Germany, June 4-5, 1998

[Projections for Making Money on the Web](#) (Michael Lesk, Harvard Infrastructure Conference, 23-25 January 1997)

---

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

---

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

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

# Intellectual property rights, copyright laws and legal issues:

---

(Chapter 10, page 223, "Books, Bucks and Bytes", Michael Lesk)

- [Cyberspace Law for Non-Lawyers](#): This is an electronic course : a "real" course in the "real world" This site includes a discussion function which will allow you, if you are so inclined, to post your own comments and reactions to the individual messages that the instructors have mailed out.
- [Overview of Copyright Laws in the Digital Domain](#) and [References](#) : Check out the references for some very good links and information on copyright laws and related issues.
- [Pamela Samuelson](#) and pointers based on her pages and recommendations
- [Electronic Commerce](#)
- [Workshop on Tech. of Terms and Conditions](#) and [Final Report to NSF](#) - including Breakout Group Reports
- [EC98, International IFIP Working Conference on Distributed Systems for Electronic Commerce](#), Hamburg, Germany, June 4-5, 1998
- [Stanford U. work on electronic commerce, legal pointers](#)
- Copyright law in Netherlands (in Dutch): [background home page](#), [page on intellectual property and copyright](#)

## Other related references:

- Digital Copyright Protection - Peter Wayner - AP Professional - Boston, 1997
- Scholarly Publishing: The Electronic Frontier - ed. Robin P. Peek and Gregory B. Newby - The MIT Press, Cambridge, MA, 1996
- The Network Nation - Starr Roxanne Hiltz and Murray Turoff - The MIT Press, Cambridge, MA, 1994
- Ubiquitous Email ...

---

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

---

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

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

# Pamela Samuelson Plus Recommendations on Law and Digital Libraries

[Professor Pamela Samuelson](#) is one of the leading authorities on legal issues in the area of intellectual property rights (IPR). A new [MacArthur Fellow](#), a Fellow of the [Electronic Frontier Foundation](#), a Fellow of the [Cyberspace Law Institute](#), she is a Professor at the [University of California at Berkeley](#) with a joint appointment in the [School of Information Management and Systems](#) and the [School of Law](#).

For more information on this and related topics, see

- [Selected Papers by Pamela Samuelson](#)
- [Law 276: Cyberlaw](#) - by Pamela Samuelson, University CA, Berkeley
- [Infosys 296A: Future of the Information Society, Copyright & Community](#) - by Peter Lyman and Pamela Samuelson, University CA, Berkeley
- [Cyberspace Law for Non-Lawyers](#), which attracted over 20,000 subscribers, by [David Post, Temple U. School of Law](#); Lawrence Lessig, [Harvard Law School](#); [Eugene Volokh, UCLA School of Law](#)
- [Crash Course in Copyright](#) from UT system, including the [Digital Library](#)
- [Copyright Management Center](#) of IUPUI, directed by [Kenneth Crews](#)
- [The ILTguide to Copyright](#) at Columbia, for educators
- [Copyright Law Materials](#) at Cornell Legal Info. Institute
- [Copyright & Fair Use](#) site of Stanford University Libraries
- [Copyright Basics Circular from the U.S. Copyright Office](#)
- [Copyright Clearance Center \(CCC\) Online](#)
- [Digital Future Coalition \(DFC\)](#)
- [IIP Policy Gateway, Harvard Information Infrastructure Project](#)
  - [Bibliography](#)
  - [Policy resources in the area of Internet governance](#), supplement to MIT Press [book](#)
  - The Impact of the Internet on Communications Policy [conference](#)
- [ALAWON](#) - ALA (American Library Association) Washington Office Newsline providing urgent and late breaking news
- [ARL Federal Relations and Information Policy Program](#), Prue Adler



## DFC Front Page

**Current Issues**

**Past Issues**

**Hill Hotline**

**DFC Members**

**Links**

**Contact Us**

### HOT TOPICS

È **H.R. 354 is marked-up and the Commerce Cmte offers an alternative.**

On Thursday, 5/20/99, H.R. 354 was approved for consideration by the House Judiciary Committee. On Wednesday, 5/19/99 Commerce Committee Chairman Bliley offered his own alternative, the Consumer and Investor Access to Information Act of 1999.

È **DFC signs position statement on "database" legislation.**

The DFC and 130 others have signed a statement outlining a position on H.R. 354 which is now in the Judiciary Committee. [Click here to read the statement.](#)

È **The Digital Millennium Copyright Act is now Public Law No: 105-304.**

È **The site is changing**

The DFC has completed its work on the WIPO implementing legislation and we look forward to providing you with a COMPLETE history of the Digital Millennium Copyright Act. Stay tuned for major changes and more documents.

È **Timeline now available**

[The DFC has now completed an event timeline for passage of the DMCA.](#) Stop by and have a look or print it out for reference. Shortly, the timeline will be a fully clickable index of events and documents.

È **No more letters please!**

Thank you for writing to preserve balance in the DMCA. Your efforts made a difference. At this time we have ended our letter writing

### New Additions:

The Digital Future Coalition, along with other stakeholders in the "database" legislation, issued a press release Thursday applauding the introduction of H.R. 1858. This bill, the Consumer and Investor Access to Information Act of 1999, is a competitor to H.R. 354, "The Collections of Information Antipiracy Act." [Click here to see the press release and other materials that have just been released.](#)

-##-

With the start of the 106th Congress, a new crop of issues involving intellectual property and technology has sprouted. The DFC is involved with many of these issues.

Currently, the coalition is working on H.R. 354, "The Collections of Information Antipiracy Act." This piece of legislation will affect everything from sports scores and financial data to scientific research. A new web site, ["Database Data"](#) has the low-down on this potentially harmful law.

Later in the year, the DFC will be preparing for a federal rule making regarding section 1201 of the Digital Millennium Copyright Act. This section deals with the "fair use" of information. Specifically, the government will be deciding which, if any, types of information should be exempted from the anti-circumvention provisions of the DMCA. The DFC will participate in this process wherever possible to assure that the fair use doctrine is protected.

Also of interest, the Register of Copyrights will release a report on distance education, digital media, and copyright after completing its study and hearings.

Visit often because the DFC will have the most recent documents; action alerts on legislative activities and answers to your questions about the DMCA.

### Instant Info:

**Feel like you don't know enough about copyright law to talk to your representative? Just want to know more about these issues? Click on the topics below to get short, easy to understand descriptions:**

- [Fair Use](#)
- [Temporary Copies](#)
- [First Sale](#)
- [Preemption](#)
- [Distance Learning](#)
- [Library Exemptions](#)
- [Anti-Circumvention](#)

campaign.

- [Copyright Management Information](#)

[\[Main Page\]](#) [\[Current Issues\]](#) [\[Past Issues\]](#) [\[Hill Hotline\]](#) [\[DFC Members\]](#) [\[Links\]](#) [\[Contact Us\]](#)

# Social Issues:

---

- Social Aspects [D-Lib Working Group](#)
  - UCLA Workshop, Social Aspects of Digital Libraries, Feb. 16-17, 1996  
<http://www-lis.gseis.ucla.edu/DL/>
  - Life Cycle [http://www-lis.gseis.ucla.edu/DL/UCLA\\_DL\\_model.gif](http://www-lis.gseis.ucla.edu/DL/UCLA_DL_model.gif)
- 

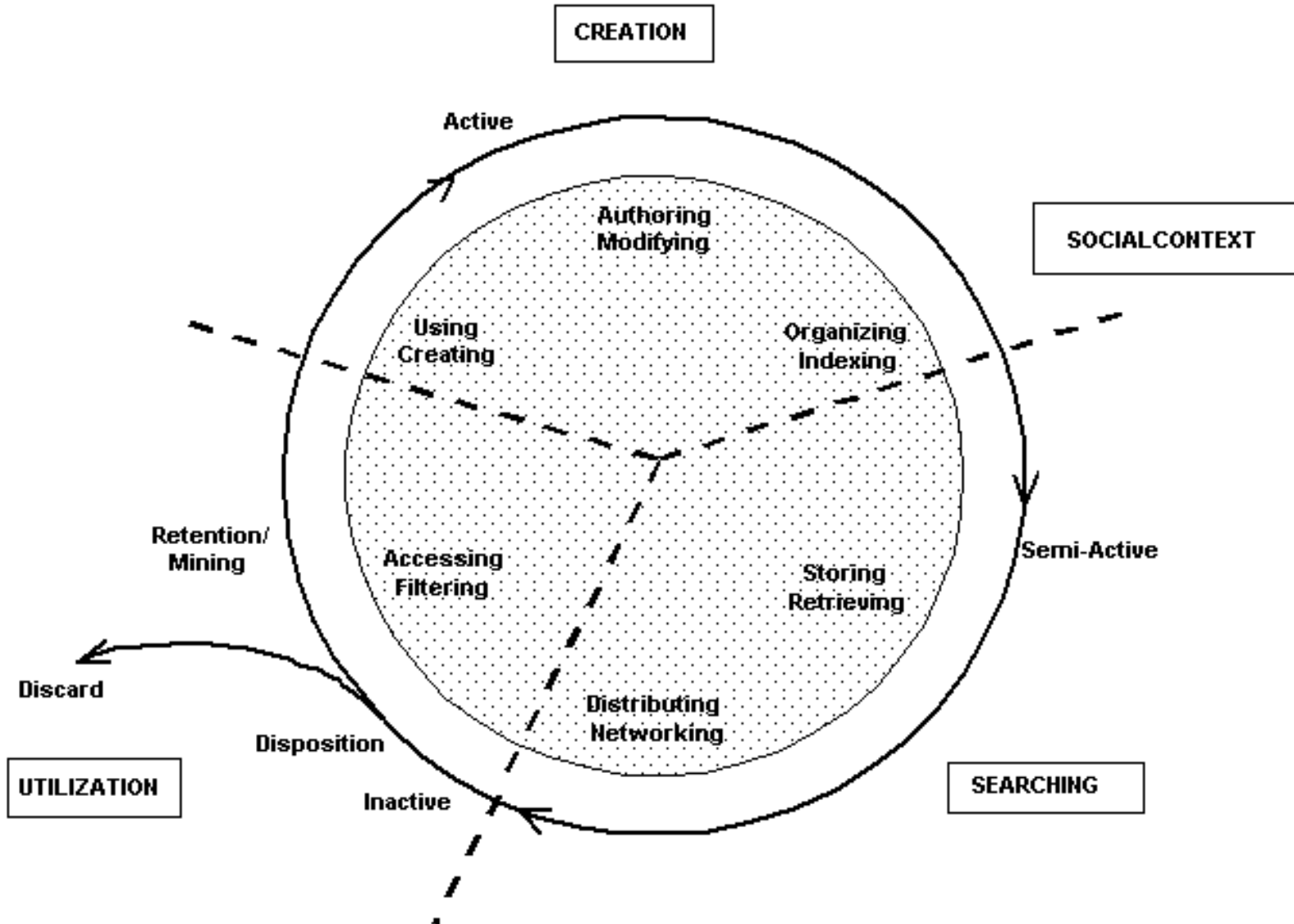
[\[Main\]](#) [\[Contents\]](#) [\[Topics\]](#)

---

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

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

## 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.