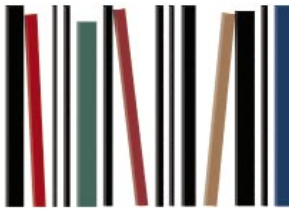
Cover Feature

Federated Search of Scientific Literature



The Illinois Digital Library Project has developed an infrastructure for federated repositories. The deployed testbed indexes articles from many scientific journals and publishers in a production stream that can be searched as though they form a single collection.

Bruce Schatz
William Mischo
Timothy Cole
Ann Bishop
Susan Harum
Eric Johnson
Laura Neumann
University of Illinois at Urbana-Champaign

Hsinchun Chen
Dorbin Ng
University of Arizona

The Digital Libraries Initiative (DLI) project at the University of Illinois at Urbana-Champaign (UIUC) was one of six sponsored by the NSF, DARPA, and NASA from 1994 through 1998. Our goal was to develop widely usable Web technology to effectively search technical documents on the Internet. We concentrated on building the experimental Illinois DLI Testbed with tens of thousands of full-text journal articles from physics, engineering, and computer science, and on making these articles available over the Internet before they are available in print.

Our DLI Testbed used document structure to provide federated search across publisher collections, by merging diverse tags from multiple publishers into a single uniform collection. Our sociology research evaluated the usage of the DLI Testbed by more than a thousand UIUC faculty and students. Our technology research moved beyond document structure to document semantics, testing contextual indexing of document content on millions of documents.

DLI TESTBED AND FEDERATED SEARCH

The DLI Testbed team designed, developed, and evaluated mechanisms to provide effective access to full-text physics and engineering journal articles within an Internet environment. The team, based in the Engineering Library at UIUC, had as its primary goals to

- construct and test a multipublisher, full-text DLI Testbed that employs flexible search and rendering capabilities and offers rich links to internal and external resources, with the sources tagged in Standard Generalized Markup Language (SGML);
- integrate the DLI Testbed and other full-text repositories into the continuum of information resources offered to end users within the Engineering Library system;
- determine the efficacy of full-text article searching compared to document surrogate searching and explore end-user full-text searching behavior, in order to identify user-searching needs; and
- identify models for effective publishing and retrieval of full-text articles within an Internet environment and employ these models in the DLI Testbed design and development.

Document collection and retrieval

The DLI Testbed supports full text in SGML format, associated article metadata, and bit-mapped figure images for scientific journal articles. At present, the collection includes 63 journals containing 66,000 articles from five professional societies:

- American Institute of Physics
- American Physical Society
- American Society of Civil Engineers
- Institution of Electrical Engineers
- IEEE Computer Society

Each publisher transmits electronic copies of their journals to us as they go to print, allowing the issues to appear in the DLI Testbed before the hard copies appear in the Engineering Library. The production stream is increasing at approximately 2,000 articles per month. We achieved a critical mass for useful search in 1997 when the SGML collection reached coverage of two years for each journal.

To support federated search across this collection, our DLI Testbed team developed a Web-based retrieval system called DELIVER (DEsktop LInk to Virtual Engineering Resources). In operation since October 1997, DELIVER has been used by more than 1,900

0018-9162/99/$10.00 © 1999 IEEE

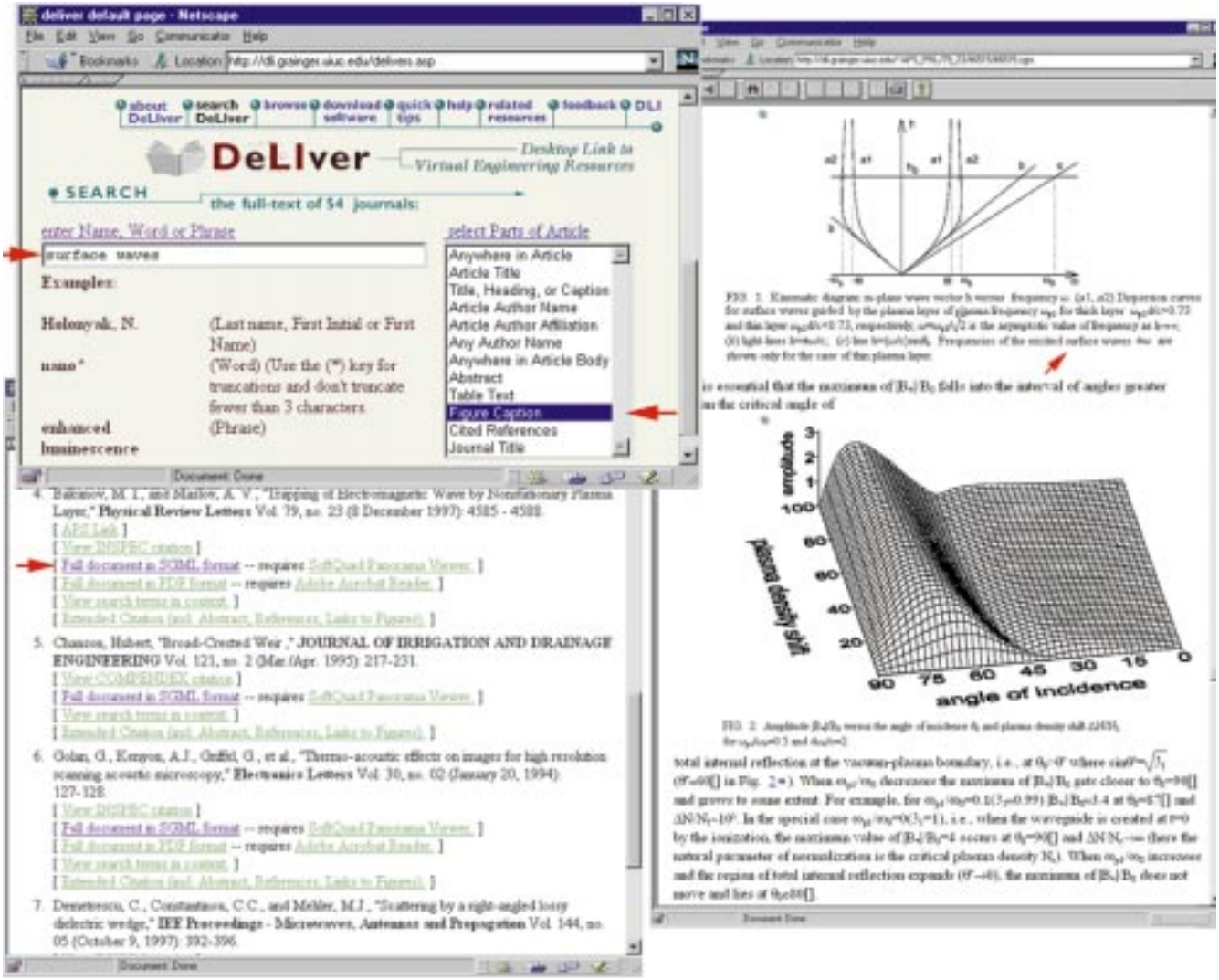


Figure 1. DLI testbed Web-based DELIVER search session, showing the query interface (upper left), the results interface (lower left), and full-text display (right).

registered UIUC students and faculty, plus designated outside researchers. We have recorded detailed transaction logs for more than 97,000 user search sessions.

Figure 1 shows a DELIVER search session. The initial screen, in the upper left, prompts the user that parts (structures) of the documents are searchable. The user has requested the term `surface waves` when it occurs as a figure caption. The search results appear in the lower left, showing four of the articles retrieved for the requested text phrase occurring in the requested document structure. Note that each article is from a different journal and that these journals span multiple publishers. The screen on the right gives a portion of the SGML display for the full text of the article. The red arrows show that `surface waves` occurs in a figure caption, but not in the title. Note that SGML tags the complete structure of the document, including figures and equations.

Federated structure search

A critical element of the DLI Testbed was the effective use of SGML to reveal document structure and produce associated article-level metadata, which homogenizes heterogeneous SGML and allows short-entry display. We take the SGML directly from the publishers' collections, converting it to a canonical format for federated searching and transforming tags into a standard set.

The metadata also contains links to internal and external data, such as other DLI Testbed articles and bibliographic abstract databases. The metadata and index files—which contain pointers to the full-text data—are stored independently and separately from the full text.

With SGML, documents can be treated as objects, allowing viewing, manipulation, and output. For retrieval purposes, SGML's major strength is its ability to reveal a document's component structure. While SGML is becoming ubiquitous in publishing, it is largely generated by publishers as a production by-product. The coming widespread availability of rich markup formats, such as XML (eXtensible Markup Language)—a nearly complete instance of SGML—will likely make such formats the standard for open document systems. Future versions of our DLI Testbed are planning to use XML to represent structure.

The DTD (Document Type Definition), which accompanies each publisher's SGML file, specifies the

semantics and syntax of the SGML tags. The DTD also specifies the rules for how SGML tags may be applied to the documents to identify where components occur.

One of the hardest problems in successfully deploying the DLI Testbed has been processing heterogeneous DTDs. We developed a number of techniques to address these problems and normalize processing, indexing, storage, retrieval, and rendering. For example, there is a standard canonical set of document tags, and all tags from all publishers are heuristically mapped into these.

Distributed repositories and links

An important concern was developing effective retrieval models for journals published on the Web. We designed a distributed repository architecture that federates individual publisher repositories of full-text documents. Normalized metadata and index data are extracted from the full text, allowing searches via a parallel execution monitor. This architecture enables standardized and canonical searches of subject and author that are consistent across distributed and disparate repositories.

The DLI Testbed team successfully demonstrated the efficacy of the distributed repository model. We produced cross-DTD metadata, providing parallel database querying and distributed retrieval techniques across a distinguished subset of the full-text repositories. We then installed an off-site repository by cloning the testbed environment at the actual site of a publisher partner (the American Institute of Physics, in New York City).

We made significant progress developing a metadata specification to support standardized retrieval across repositories. This allowed for short-entry display independent of the full-text document repositories and links to associated testbed items and bibliographic databases. We used SGML tag aliasing for normalization to accommodate heterogeneous DTDs. The DELIVER client supports searching, retrieval, and display across multiple repositories, providing cross-repository retrieval with single searches.

Our innovations include integration of DELIVER with other retrieval services. We implemented Inspec and Compendex proxies for the Ovid retrieval system, with links to the DELIVER Testbed. These proxies enable those databases to be searched with comprehensive coverage for journal abstracts, with transparent links following into the full-text SGML documents when the article is covered by the DLI Testbed. The DLI Testbed also provides links from the bibliographies of retrieved DELIVER articles to other items in the testbed, citation links to previous testbed articles, and links from bibliographic references in retrieved DELIVER articles to Inspec and Compendex database records.

Multiple-view interfaces

Complete search sessions across multiple sources are necessary to effectively handle scientific literature. Our DLI Testbed provides support for federated searching across document structures from different publisher repositories. The user can use a single high-level structure, such as *author* or *caption*, and have it automatically translated into the appropriate SGML tags for each document.

Our experimental user interfaces, which ran in the Engineering Library prior to DELIVER, showed that effective information retrieval for full-text structure search is greatly facilitated by multiple views. Traditional information retrieval has supported only a single view, which sends a query to an index and returns a result. This is the model currently supported within commercial online systems and within Web search systems. A multiple-view interface supports sessions, with combinations across the results of different search/queries.

We developed an experimental multiple-view interface, called IODYNE, which seamlessly integrates many different kinds of indexes with drag-and-drop between screen windows for search indexes.[1] Such a client is the paradigm for the next generation of search systems on the Web, where multiple indexes for different purposes can be easily combined within an entire session. IODYNE supports text search of full-text SGML via the DLI Testbed and of bibliographic abstracts via Ovid.

This text search can be boosted by term suggestion, where the user specifies a broad query and the system returns related terms to be interactively selected for future queries. IODYNE supports term suggestion via subject thesauri (such as the Inspec thesaurus) and via concept spaces (automatically generated thesauri provided by our technology research, as described below in the "Semantic Indexing and Technology Research" section).

Testbed partners and continuance

The collaborative relationship between the DLI Testbed team and its publishing partners was particularly strong—they grew to refer to and rely upon us as their "R&D arm." The strong partnering relationship is evidenced by the agreement between the DLI project and the publisher partners to initiate a Collaborative Partners Program, whose funding is enabling the continuation of the DLI Testbed beyond the DLI grant period. The Engineering Library is also a recipient of a three-year grant from the US Defense Advanced Research Projects Agency (DARPA) to continue the SGML testbed for evaluation purposes.

These new funds will allow the DLI Testbed team to continue investigating issues connected with full-text article indexing, interface design, retrieval, and

The DELIVER client supports search, retrieval, and display across multiple repositories, providing cross-repository retrieval with single searches.

We explored how researchers use journal components in their work.

rendering. Continued contributions of materials from the publishing partners will allow for the increase of both the depth and breadth of the digital collection. Plans are also underway to extend DLI Testbed access to the Big Ten University Consortium throughout the US Midwest, to enlarge the user population, and further develop the distributed repository model.

TESTBED EVALUATION AND SOCIOLOGY RESEARCH

Our social science team pursued an integrated investigation of the social practices of digital libraries.[2] Throughout the project, we carried out user studies and evaluations aimed at improving the DLI Testbed. We also documented and analyzed the extent and nature of DLI Testbed use, satisfaction, and impacts regarding engineering work and communication. These efforts informed our broader contributions to knowledge about engineering work, the use of scientific and engineering journals, and the changing information infrastructure.

We pursued several research threads that are relevant to understanding social practices associated with the development and use of federated repositories of structured documents: article disaggregation during knowledge construction, user understanding of newly encountered digital libraries, convergence of communities of practice with information artifacts, and resolution of digital library visions held by different stakeholders.

In our research, we carefully adapted traditional social science methods to the study of social phenomena involving information systems. We employed a variety of qualitative and quantitative techniques for collecting and analyzing data, including

- observing engineering work and learning activities,
- conducting focus groups with potential system users,
- conducting interviews with actual system users,
- performing usability testing of system prototypes,
- recording transaction logs of system sessions, and
- conducting large-scale user surveys.

In addition, we initiated computer-mediated data-gathering techniques, such as user registration and exit polls after sessions. We have considered results from all these methods to triangulate our findings and provide a deeper understanding of the nature of digital library use and social phenomena involved.

Analysis of DELIVER users and use

Users are required to fill out an online demographic questionnaire to register for a DELIVER login. These questionnaires were analyzed when the registrations reached a total of 1,200 UIUC faculty, staff, and students. Half of these users are graduate students, who also account for the most searches. About 75 percent of users are men, most between 23 and 29 years old. Faculty members are a small, but intense, segment of users.

DELIVER users cover a wide spectrum, representing all campus engineering disciplines, science-related fields (such as ecological modeling and biology), and fields such as communications and psychology. We found, however, that most users' backgrounds reflect the DLI Testbed's contents, which concentrates on journals from physics, civil engineering, electrical engineering, and computer science.

A preliminary analysis of 226 recently completed user surveys suggests that people are generally satisfied with our system. The mean responses to three separate questions meant to gauge people's reaction to DELIVER was 3.5 (where 1 corresponded to "terrible," "frustrating," and "inadequate search power," and 5 corresponded to "wonderful," "satisfying," and "adequate search power").

DELIVER transaction logs reveal the use of various system features. Analysis of more than 4,200 sessions indicates that about 20 percent of sessions used the extended citation screen, while 38 percent of sessions viewed the full text of the article. In usability interviews, we found that users' ability to view full text was limited by the fact that they had to first download additional software (an SGML plug-in) to view it.

Use of document structure

Given the nature of searching and display made possible through the use of SGML, we explored how researchers use journal components—such as abstracts, figures, equations, or bibliographic citations—in their work.[3] We identified five basic purposes for article components:

- To identify documents of interest.
- To assess the relevance of an article before retrieving and reading the full text.
- To create a customized document surrogate after retrieval that includes a combination of bibliographic and other elements (for example, author's name, article title, tables).
- To provide specific pieces of information, such as an equation, a fact, or a diagram.
- To convey knowledge not easily rendered by words, especially through figures and tables.

Engineers describe a common pattern for utilizing document components by zooming in on and filtering information in their initial reading of an article. They tend to first read the title and abstract, then skim section headings. Next, they look at lists, summary state-

ments, definitions, and illustrations, before zeroing in on key sections, reading conclusions, and skimming references.

But engineers pursue unique practices after this initial reading, as they disaggregate and reaggregate article components for use in their own work. Everyone takes scraps or reusable pieces of information from the article, but they do this differently—perhaps by using a marker to highlight text portions of interest or by making a mental register of key ideas.

Engineers then create some kind of transitory compilation of reusable pieces, such as a personal bibliographic database, folders containing the first page of an article stapled to handwritten notes, or a pile of journal issues with key sections bookmarked. These intellectual and physical practices associated with component use seem to be based on a combination of tenure in the field, the nature of the task at hand, personal work habits, and cognitive style.

Use of digital libraries

Our digital library also allowed us to step back and take a broader look at the use of online digital collections and how people attempt to make sense of them. In analyzing results from several different data collection efforts, we found that users can be confused by a newly encountered digital library, and that it takes some time and interaction for them to figure out what a particular system, like DELIVER, is.

In usability tests, we identified patterns of user actions designed to uncover what sort of system the DLI Testbed was and what it could do. What first appeared to be random trial-and-error use of the interface was actually structured exploration, which occurred frequently across sessions. Such exploration is a cut-and-try approach.[4] Situating usage in the real world forced us to think about who our most likely audience was, what they were probably most interested in using our system for, and how best to reach them.

SEMANTIC INDEXING AND TECHNOLOGY RESEARCH

Improving Web searching beyond full-text retrieval requires using document structure in the short term and document semantics in the long term. Our technology research team focused on developing new infrastructure for our vision of the future Internet, termed the *Interspace*, where each community indexes its own repository of its own knowledge.[5] For community amateurs to provide classification comparable to today's trained professionals, information infrastructure must provide substantial support for semantic indexing and retrieval.

The Interspace focuses on scalable technologies for semantic indexing that work generically across all subject domains.[6] We can automatically generate analogues of concepts and categories. We can use concept spaces—collections of abstract concepts generated from concrete objects—to boost searches by interactively suggesting alternative terms.[1,7] We can use category maps to boost navigation by interactively browsing clusters of related documents.[8]

Scalable semantics

Scalable semantics is our term for the new technologies that can index the semantics of document contents on large collections. These algorithms rely on statistical techniques, which correlate the context of phrases within the documents. For example, concept spaces use text documents as the objects and noun phrases as the concepts. The concept spaces are then the co-occurrence frequencies between related terms within the documents of a collection.

Situating usage in the real world forced us to think about who our most likely audience was, what they were probably most interested in using our system for, and how best to reach them.

Over the past several years, using DLI materials, we have used the supercomputers at the National Center for Supercomputing Applications (NCSA) to compute concept spaces for progressively larger collections,[9] until the scale of entire disciplines, such as engineering, has been reached. By partitioning a large existing collection into discipline subcollections, which are the equivalent of community repositories, we use supercomputers to simulate the future world of a billion repositories.

In 1995 we generated concept spaces for 400,000 abstracts from Inspec (deep coverage in physics, electrical engineering, and computer science), and in 1996 we generated concept spaces for 4 million abstracts from Compendex (broad coverage across all of engineering, some 38 subject disciplines). The first computation took one day of supercomputer time, and the second took 10 days of high-end time on the HP Convex Exemplar. The second computation provided a comprehensive simulation of community repositories for 1,000 collections across all of engineering, generated by partitioning the abstracts along the subject classification hierarchy.[5]

Concept spaces and document search

The Interspace consists of multiple spaces at the category, concept, and object levels. Within the course of an interaction session, users will move across different spaces at different levels of abstraction and across different subject domains. For example, the system enables users to locate desired terms in the concept space by starting from broad terms, then traversing into narrow terms specific to that document collection. They can then move across into document space to perform full-text searches by dragging the concept term into the document space search window.

Figure 2 is a composite of a session with an experi-

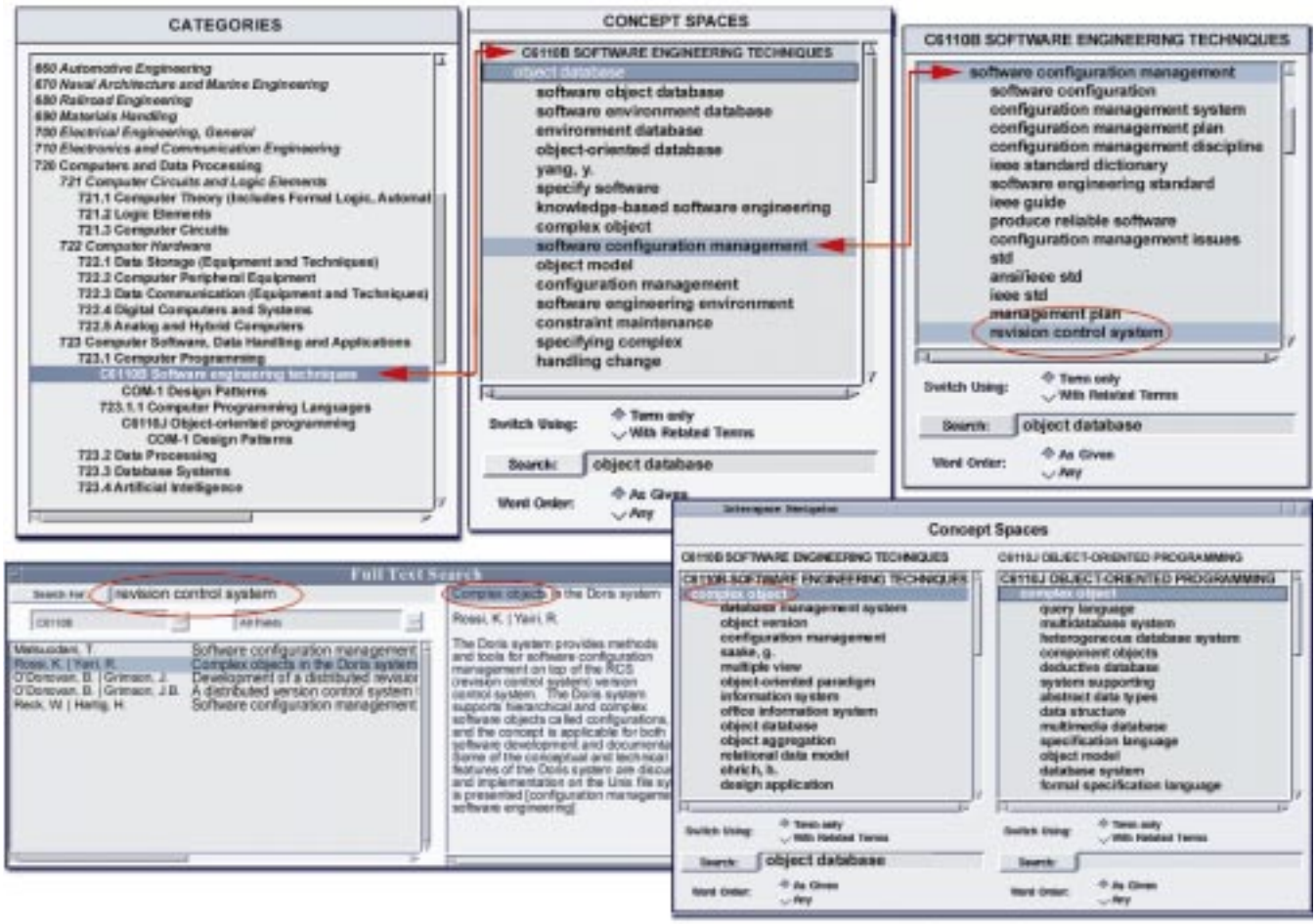


Figure 2. An Interspace Navigator for engineering literature, showing concept spaces for community repositories, with upper windows illustrating term suggestion using concept spaces and lower windows illustrating document search leading the user to concept switching across subject domains.

mental Interspace containing concept spaces for engineering literature.[10] The different windows sampled in this session illustrate abstract indexes for categories and concepts, plus indexes for documents within collections.

The upper-left window shows an integrated list of categories for the Inspec, Compendex, and Patterns collections. Inspec and Compendex are standard commercial bibliographic databases, and Patterns is a software engineering community repository. These categories can be selected to retrieve a concept space for a specified subject domain. The upper-middle window shows portions of a concept space for the Inspec category `Software Engineering Techniques`. The concept space allows the user to interactively refine a search by selecting from related concepts. The user specified the general term `object database`, and the system returned a list of specific terms, such as `software configuration management`.

The user wants to locate and search for an even more specific term related to the use of "object-oriented databases in software engineering." The upper-right window shows a further navigation of the concept space, listing related terms of `software configuration management`, which was selected from the previous related list. The very specific term `revision control system` is located and used immediately by dragging it into the Full-Text Search window.

This window at the lower left has two panes showing the results of the query for `revision control system`. The left pane lists the articles from the Software Engineering Techniques collection mentioning the term, while the right pane displays the abstract of a selected article. Note a specific article about `software objects` in `configuration management` has been found, by navigating the concept space of terms starting from the broad `object database`, without being required to ever type a specific search term from memory.

Vocabulary switching across concept spaces

Finally, to search a subject domain they are unfamiliar with, users can begin within the concept space for a familiar subject domain, then choose another concept space for the unfamiliar domain and navigate across spaces based on common terms. This interactive vocabulary exploration is our approach to vocabulary switching, the classic information retrieval problem of different terms for the same concepts across different subject domains.[11]

In Figure 2, for example, the user has navigated across concept spaces in the upper windows to locate a specific document on `revision control` in `configuration management` that discusses `software objects`. To further investigate the use of object-oriented techniques, the user selects `complex object`

from the title of the article displayed in the lower-left window and drags this term back into the concept space window for `Software Engineering`. This action switches the level of abstraction from documents (objects) to terms (concepts).

The lower-right window displays the related terms for `complex object` within `Software Engineering` in the left pane. Scanning this term list, the user now wants even more detailed information about object-oriented techniques than seem to be available in this collection. So the user moves up another level of abstraction (from concepts to categories) and selects the category for `Object-Oriented Programming` from the Categories window.

A vocabulary switch is now performed between the domains of `Software Engineering Techniques` and `Object-Oriented Programming`. The user drags `complex object` from one domain pane to another—the lower-right window is the result. The right pane (at the lower rightmost side) shows the related terms for `complex object` within `Object-Oriented Programming`. Note this related term list is different from that for the same term in `Software Engineering` since the collections are different (so the commonly occurring terms will be different). This new term list can then be scanned to select specific terms related to `software objects and configuration management` to use in searching any of the available collections.

Such a fluid flow across levels and subjects supports semantic interoperability. This vocabulary switching by interactive navigation across concept spaces illustrates why the system is named the Interspace. Interspace navigation enables location of documents with specific concepts without previous knowledge of the terms within the documents. We are constructing and using a full-fledged Interspace prototype with semantic indexing and space navigation for community repositories in engineering and medicine.[12]

CONCLUSIONS AND IMPLICATIONS

We believe that both the DLI Testbed and the research efforts of the UIUC DLI project were major successes.

The DLI Testbed efforts built a production system with federated search across structured documents. The articles arrive in a production stream directly from major scientific publishers in full-text SGML and are fully federated at the DTD level with a Web interface. The DLI Testbed collection is currently the largest federated repository of SGML articles from scientific literature anywhere.

The DLI Testbed users represent an order of magnitude bigger population than the last-generation research system for search of scientific literature. The DLI Testbed evaluation performed comprehensive fine-grained methodologies such as user interviews and large-scale methodologies such as transaction logs. Our results will shortly lead to commercial technologies for federating structured documents across the Internet.

Our research efforts built an experimental system with semantic indexes from document content. Concept spaces are generated for term suggestion and integrated with text search via a multiple view interface. Vocabulary switching is supported by interactive navigation across concept spaces.

Our research computations are the largest ever in information science. They represent the first time that semantic indexes using generic technology have been generated on discipline-scale collections with millions of documents. They are the first large step toward scalable semantics, statistical indexes with domain-independent computations.

The knowledge networks of the Interspace will be connected via switching machines that switch concepts.

The Internet of the 21st century will radically transform how we interact with knowledge. Traditionally, online information has been dominated by data centers with large collections indexed by trained professionals. The rise of the World Wide Web and the information infrastructure of distributed personal computing have rapidly developed the technologies of collections for independent communities. In the future, online information will be dominated by small collections maintained and indexed by community members themselves.

The information infrastructure must similarly be radically different to support indexing of community collections and searching across such small collections. The base infrastructure will be knowledge networks rather than transmission networks. Users will consider themselves to be navigating in the Interspace, across logical spaces of semantic indexes, rather than in the Internet, across physical networks of computer servers.

Future knowledge networks will rely on scalable semantics, on automatically indexing the community collections so that users can effectively search within the Interspace of a billion repositories. The most important feature of the infrastructure is therefore support of semantic correlation across the indexed collections. Just as the transmission networks of the Internet are connected via switching machines that switch packets, the knowledge networks of the Interspace will be connected via switching machines that switch concepts. ❖

Acknowledgments

This work was supported by the NSF, DARPA, and NASA under Cooperative Agreement No. NSF-IRI-94-11318COOP. We thank the American Institute of

Physics, the American Physical Society, the American Society of Civil Engineers, the IEEE Computer Society, and the Institution of Electrical Engineers for making their SGML materials available to us on an experimental basis. Engineering Index and IEE kindly provided Compendex and Inspec, respectively. Indexing was done on Hewlett-Packard servers, obtained through an educational grant program. Many people have contributed to the research discussed here. In particular, we thank Robert Wedgeworth, Kevin Powell, Ben Gross, Donal O'Connor, Robert Ferrer, Tom Habing, Hanwen Hsiao, Heidi Kellner, Emily Ignacio, Cecelia Merkel, Bob Sandusky, Eric Larson, S. Leigh Star, Pauline Cochrane, Andrea Houston, Melanie Loots, Larry Jackson, Mike Folk, Kevin Gamiel, Joseph Futrelle, Roy Campbell, Robert McGrath, Duncan Lawrie, Leigh Estabrook.

References

1. B. Schatz et al., "Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval," *Proc. First ACM Int'l Conf. Digital Libraries,* ACM Press, New York, 1996, pp. 126-133.
2. A. Bishop and S. Star, "Social Informatics for Digital Library Use and Infrastructure," *Ann. Rev. Information Science and Technology,* Vol. 31, 1996, pp. 301-401.
3. A. Bishop, "Digital Libraries and Knowledge Disaggregation: The Use of Journal Article Components," *Proc. Third ACM Int'l Conf. Digital Libraries,* ACM Press, New York, 1998, pp. 29-39.
4. L. Neumann and E. Ignacio, "Trial and Error as a Learning Strategy in System Use," *Proc. 61st American Soc. Information Science Ann. Meeting,* Information Today, Medford, N.J., 1998, pp. 243-252.
5. B. Schatz, "Information Retrieval in Digital Libraries: Bringing Search to the Net," *Science,* Jan. 1997, pp. 327-334.
6. B. Schatz et al., "Federating Diverse Collections of Scientific Literature," *Computer,* May 1996, pp. 28-36.
7. H. Chen et al., "Automatic Thesaurus Construction for an Electronic Community System," *J. American Soc. Information Science,* Mar. 1995, pp. 175-193.
8. H. Chen et al., "Internet Browsing and Searching: User Evaluations of Category Map and Concept Space Techniques," *J. American Soc. Information Science,* July 1998, pp. 582-603.
9. H. Chen et al., "A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Project," *IEEE Trans. Pattern Analysis Machine Intelligence,* Aug. 1996, pp. 771-782.
10. H. Chen et al., "Alleviating Search Uncertainty through Concept Associations: Automatic Indexing, Co-occurrence Analysis, and Parallel Computing," *J. American Soc. Information Science,* Mar. 1998, pp. 206-216.
11. H. Chen et al., "A Concept Space Approach to Addressing the Vocabulary Problem in Scientific Information Retrieval: An Experiment on the Worm Community System," *J. American Soc. Information Science,* Jan. 1997, pp. 17-31.
12. B. Schatz et al., "The Interspace Prototype," http://www.canis.uiuc.edu.

Bruce Schatz is director of the Community Architectures for Network Information Systems (CANIS) Laboratory at UIUC and professor in the Graduate School of Library and Information Science, with joint appointments in computer science, neuroscience, and health information sciences. He was the principal investigator of the Illinois Digital Library Project. He is senior research scientist at the National Center for Supercomputing Applications (NCSA). He served as the scientific advisor on information systems when NCSA developed Mosaic, which was inspired by his earlier network information systems and spawned the World Wide Web. His current research is building analysis environments to support community repositories (Interspace), and performing large-scale experiments in semantic retrieval for vocabulary switching. He received a BA in mathematical sciences from Rice University, an MS in artificial intelligence from MIT, an MS in computer science from Carnegie Mellon University, and a PhD in computer science from the University of Arizona.

William Mischo is director of the Grainger Engineering Library Information Center, the Engineering Librarian, and a professor of library administration at UIUC. His current interests include expanding the DLI project under the auspices of a CNRI grant to investigate the use of XML, cascading style sheets, and dynamic HTML in retrieving and rendering article full text. He received a BA in mathematics from Carthage College and an MA in library and information science from the University of Wisconsin.

Timothy Cole is system librarian for digital projects and associate professor of library administration at the UIUC library. His interests include information retrieval interfaces, processing of document structure, and the use of the Web in libraries. He received a BS in aeronautical and astronautical engineering and an MS in library and information science from UIUC.

Ann Bishop is assistant professor in the Graduate School of Library and Information Science at UIUC and principal investigator of a US Department of Commerce grant on the introduction of computers in low-income neighborhoods. Her primary research interest is social aspects of information system design, evaluation, and use. She received a BA in Russian literature from Cornell University and an MLS and a PhD in information transfer from Syracuse University.

Susan Harum was the external relations coordinator for the University of Illinois Digital Libraries Initia-

tive at UIUC from 1994 to 1998; she continues to assist the program director for digital libraries at the NSF. Her interests include collaborative research tools for international digital libraries. She received a BS in anthropology from Michigan State University, an MA in eastern Asian languages, and an MS in library and information science from UIUC.

Eric Johnson is a PhD candidate in the UIUC Graduate School of Library and Information Science. His research interests include hypermedia systems and hypertext use in thesaurus navigation and bibliographic retrieval. He received a BS in computer science and science and technology studies from Michigan State University, an MS in computer science from Northwestern University, and an MA in sociology and an MS in library and information science from UIUC.

Laura Neumann is a PhD candidate in the UIUC Graduate School of Library and Information Science. Her research interests include work practice and social issues surrounding the digitization of information, the use of digital libraries and other information systems, and the automation of work tasks. She received a BA in sociology and anthropology from the University of Minnesota.

Hsinchun Chen is a full professor in the Department of Management Information systems at the University of Arizona and director of the UA/MIS Artificial Intelligence Lab. His research interests include semantic retrieval, search algorithms, knowledge discovery, and collaborative computing. He received a BA from National Chiao-Tung University in Taiwan, an MBA from SUNY Buffalo, and an MS and a PhD in information systems from New York University.

Dorbin Ng is a PhD candidate in the Department of Management Information Systems at the University of Arizona. His research interests include multimedia information retrieval, knowledge discovery using supercomputers, and user interfaces for collaborative computing. He received a BS in business administration and an MS in management information systems from the University of Arizona.

For more information on the Illinois Digital Library Project, see http://dli.grainger.uiuc.edu or e-mail dli@uiuc.edu. Readers can contact the authors at CANIS Laboratory, University of Illinois, 704 S. Sixth St., Champaign, IL 61820; www.canis.uiuc.edu or schatz@canis.uiuc.edu.