Federated Search of Scientific Literature

A Retrospective on the Illinois Digital Library Project

Bruce Schatz, William Mischo, Timothy Cole, Ann Bishop, Susan Harum, Eric Johnson, Laura Neumann, Hsinchun Chen, and Dorbin Ng

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 before they were available in print. The DLI testbed focused on using the document structure to provide federated searches across publisher collections. Our sociology research included the evaluation of its effectiveness under use by over 1,000 UIUC faculty and students, a user community an order of magnitude bigger than the last generation of research projects centered on searching scientific literature. Our technology research developed indexing of the contents of text documents to enable a federated search across multiple sources, testing this on millions of documents for semantic federation.

This article will discuss the achievements and difficulties we experienced over the past four years. In section 1 (the DLI testbed and Structured Documents), we will review our experiences in building the DLI testbed, in which repositories (indexed collections) of full-text multiple-source documents have been built and federated (merged and mapped), so that they appear as a single virtual collection. Section 2 (Testbed Evaluation and Sociology Research) presents the results of our user studies and evaluation of the testbed and its user community in addition to our work in investigating the social practices of digital libraries. We then describe, in section 3 (Semantic Indexing and Technology Research), our research to improve information retrieval technology through the development of the Interspace, which focuses on statistical technologies for semantic in-

dexing that are scalable across subject domains. Section 4 (Multiple Views and Federated Search) describes our development of an Internet client for federated repositories, which allows the user to retrieve information from multiple servers by dynamically combining multiple indexes. Finally, in section 5 (Conclusion), we discuss our vision of the future of the Internet in the twenty-first century, where every community maintains its own repository of its own knowledge, and scalable semantics enables federation across repositories.

DLI TESTBED AND STRUCTURED DOCUMENTS

The overarching focus of the DLI testbed team has been on the design, development, and evaluation of mechanisms that provide effective access to full-text physics and engineering journal articles within an Internet environment. The primary goals of the testbed team were: (1) the construction and testing of a multi-publisher SGML-based full-text testbed employing flexible search and rendering capabilities and offering rich links to internal and external resources; (2) the integration of the testbed and other full-text repositories into the continuum of information resources offered to end-users within the library system; (3) determining the efficacy of full-text article searching vis-a-vis document surrogate searching, and exploring end-user full-text searching behavior in an attempt to identify user-searching needs; and (4) identifying models for effective publishing and retrieval of full-text articles within an Internet environment and employing these models in the testbed design and development.

Over the last four years, in conjunction with a number of professional societies, the testbed team has implemented a large-scale Web-based testbed of full-text journal articles featuring enhanced access and display capabilities. The Illinois DLI testbed is presently comprised of the article full-text in SGML format, the associated article metadata, and bit-mapped images of figures for sixty-three journal titles containing over 50,000 articles from five scholarly professional societies in physics and engineering. The full-text articles for the testbed have been contributed by the American Institute of Physics (AIP), the American Physical Society (APS), the American Society of Civil Engineers (ASCE), the Institute of Electrical and Electronics Engineers Computer Society (IEEE CS), and the Institution of Electrical Engineers (IEE).

The DLI testbed is based within the Grainger Engineering Library Information Center, a $22 million facility that opened in 1994 and is dedicated to the exploration of emerging information technologies. The Web-based retrieval system developed by the DLI testbed and evaluation teams is called DeLIver (Desktop Link to Virtual Engineering Resources). The DeLIver client, which replaced a Microsoft Windows-based custom client in use for the first two years of the project, has been in operation since

October 1997 and is being used by over 1,200 registered UIUC students and faculty and designated outside researchers. Detailed transaction log data of user search sessions (gathered and merged from database and Web servers) is being kept, and a preliminary analysis of user search patterns from 4,200 search sessions has been completed.

Figure 1 depicts a search session using the DLI Web client, DeLIver. The initial interface prompts the user that parts of the documents are searchable, and "plasma density" as a figure caption has been selected. The second interface displays the search results, showing four of the articles retrieved with "plasma density" in the figure caption. Note that the articles are from four different journals (federated repositories), and that "plasma density" is not found in the titles. The last interface displays an example of a figure retrieved in this manner. Note that SGML tags the complete structure of the document including figures and equations.

The cornerstones of the DLI testbed are the effective utilization of the article content and structure revealed by SGML and the production of the associated article-level metadata, which serves to normalize the heterogeneous SGML and provide short-entry display capability. The SGML is taken directly from the publisher's collections, then processed into a canonical format for federated search. The metadata also contain links to internal and external data, such as forward and backward links to other DLI testbed articles and links to A & I service databases and other repositories. The metadata and index files, which contain pointers to the full-text data, can be stored independently of, and separately from, the full text.

It is clear that a rich markup format such as XML (eXtensible Markup Language), which is a nearly complete instance of SGML, will become the language of open document systems. SGML permits documents to be treated as objects to be viewed, manipulated, and output. The major strength of SGML, in terms of its retrieval capabilities, lies in its ability to reveal the deep content and structure of a document. While SGML is becoming ubiquitous in the publishing world, it is still, for the most part, being generated by publishers as a byproduct rather than serving as an integral part of their production process.

The Document Type Definition (DTD) accompanying an individual publisher's SGML is the instrument that actually specifies the semantics and syntax of the tags to be used in the document markup. The DTD also specifies the rules that describe the manner in which the SGML tags may be applied to the documents. One of the major roadblocks in the successful deployment of the DLI testbed has been the processing involved with the heterogeneous DTDs directly from publishers. In the process of creating a viable DLI testbed, the Illinois testbed team developed a number of techniques to address problems and normalize SGML processing, indexing, storage, retrieval, and rendering.

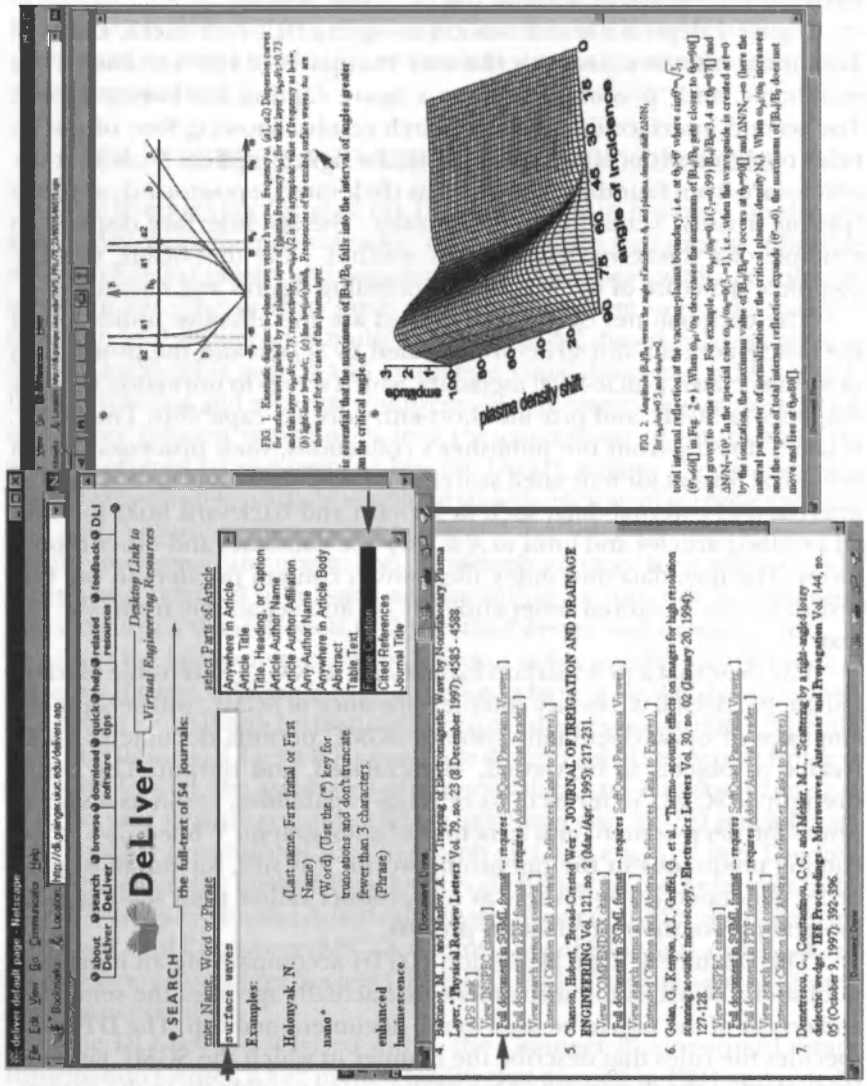

Figure 1. DLI Testbed Web-Based DeLIver Search Session

Another important concern of the DLI testbed group has been in exploring effective retrieval models for a Web-based electronic journal publishing system. It has become commonplace for both major and small-scale publishers to provide Internet (Web-based) access to their publications, including journal issues and articles. The DLI testbed team has proposed a distributed repository model which federates the individual publisher repositories of full-text documents. These distributed repositories are federated or connected by the extraction of normalized metadata and index data from the full-text which can then be searched via a parallel execution monitor. This model addresses the challenge of providing standardized and consistent subject, title, and author searching capabilities across these distributed and disparate repositories.

The DLI testbed team has succeeded in demonstrating the efficacy of the distributed repository model by producing cross-DTD metadata, providing parallel database querying and distributed retrieval techniques across a distinguished subset of the full-text repositories, and by establishing and employing an off-site repository at the site of an actual publisher (AIP).

In the four years of the grant, the testbed team has made significant progress in the development of a metadata specification to support standardized retrieval across repositories. This allowed for a short-entry display independent of the discrete full-text document repositories and links to associated testbed items, A & I service databases, and other repositories. SGML tag aliasing or normalization of the system was done to accommodate heterogeneous DTDs. Development was needed of the Web DeLIver and custom Windows clients for search, retrieval, and display across multiple discrete repositories to provide cross-repository retrieval from single-search command arguments.

Issues that came to the forefront during the project were connected with the rendering of SGML within the Softquad Panorama viewer and the rendering mathematics in particular. As a result of these difficulties, an international mathematics rendering conference was organized and held at the Grainger Library in May 1996.

Innovations by the testbed team include the integration of DeLIver with other retrieval services. An Ovid INSPEC and Compendex proxy with links to the DeLIver Testbed and other remote publisher repositories were implemented. Links from the bibliographies of retrieved DeLIver articles to other items contained in the testbed were incorporated as well as citation links from testbed articles that cite previous testbed articles. Links from the retrieved DeLIver articles and references in the bibliographies of retrieved DeLIver articles to INSPEC and Compendex database records in an Ovid system were also incorporated. DeLIver users may also re-execute the DeLIver search command arguments in the INSPEC/Compendex/CurrentContents periodical index databases.

This work has been accomplished with the cooperation and support of our publisher partners and through the use of commercial software from OpenText, Hewlett-Packard, SoftQuad, and Microsoft. The partnering relationship between the testbed team and its publishing partners was particularly strong, as evidenced by their assertion that the DLI was their "R&D" arm.

Technology transfer was fostered through frequent formal and informal meetings, quarterly newsletters updating research results, frequently updated Web pages, and an annual expense-paid workshop. DLI partners had access to DeLIver, the UIUC Web client, and hands-on consulting for SGML specifications. In addition, they received versions of the Windows custom client and processing code along with copies of research results, including statistics and user evaluation.

The strong partnering relationship is also evidenced by the agreement between the DLI and their partners to initiate a Collaborative Partners Program, which will provide for the continuation of the DLI testbed beyond the grant period. The Grainger Library is also a recipient of a three-year grant from DARPA to continue the SGML testbed.

The Collaborative Partners Program and the DARPA grant will allow the testbed team to continue researching issues connected with full-text article indexing, interface design, retrieval, and 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 testbed access to the Big 10 university consortium throughout the Midwest so as to enlarge the user population and further develop the distributed repository model.

TESTBED EVALUATION AND SOCIOLOGY RESEARCH

The DLI Social Science Team has pursued an integrated interdisciplinary research program that investigates the social practices of digital libraries (Bishop & Star, 1996). Throughout the course of the project, we have carried out user studies and evaluation work aimed at improving the DLI testbed. We are also working on documenting and analyzing extent and nature of testbed use, satisfaction, and impacts within the context of engineering work and communication. Both of these lines of work also inform our broader interests in contributing to existing knowledge about engineering work, use of scientific and engineering journals, and the changing information infrastructure.

We have pursued several specific research threads that are of particular relevance to understanding social practices associated with the development and use of federated online repositories of full-text documents. These include studies of the disaggregation of journal articles in the course of knowledge construction; how people make sense of new DLs they

encounter; the convergence of communities of practice with information artifacts and infrastructure; and negotiating among multiple visions of a DL held by different stakeholders.

The DL Initiative has afforded us the opportunity to share ideas on human-centered design and evaluation with colleagues across the six projects, most notably those at Berkeley and Santa Barbara. We feel that our research has been greatly enriched by this association. Our experiences will be pooled in an edited monograph on research related to social aspects of digital library design and use planned for publication in 1999. The book aims to identify and discuss challenging issues that arise in socially-grounded approaches to studying digital libraries based on the recent work of leading researchers in the field. We feel the book will be of interest to digital library policy-makers, designers, implementers, and evaluators.

In conducting our research, we paid particular attention to the adaptation and application of traditional social science methods to studying social phenomena associated with information systems. We have employed a variety of qualitative and quantitative techniques for collecting and analyzing data. These include observation of engineering work and learning activities, interviews and focus groups with a range of potential and actual system users, usability testing, and large-scale user surveys. In addition, we have initiated a number of computer-mediated data-gathering techniques, such as user registration, exit polls displayed after an individual's DeLIver session, and system instrumentation (the creation of transaction logs). We are bringing the results of each of these methods together in order to triangulate our findings and provide a deeper understanding of the nature of digital library use and the social phenomena involved.

We currently have over 1,200 registered patrons of DeLIver representing University of Illinois faculty, students, and staff. About half of our users are graduate students, who also do the highest average number of searches. Approximately 75 percent of DeLIver patrons are men, mostly in the 23-29 age bracket. The relatively small number of faculty members who use the system seem to be intense users. There is a surprisingly wide audience for DeLIver, representing all campus engineering disciplines, science-related fields such as ecological modeling and biology, and other fields such as communications and psychology. We have found, however, that our heaviest users closely reflect the content of our testbed, which concentrates its holdings on journals from civil engineering, electrical and computer engineering, and computer science.

A preliminary analysis of recently completed user surveys ($N = 226$) suggests that people are generally satisfied with our system. The mean overall responses to three separate questions meant to gauge people's reaction to DeLIver was 3.5 (where one corresponded to "terrible," "frustrating," "inadequate search power" and five corresponded to "wonderful,"

"satisfying," and "adequate search power"). DeLIver transaction logs reveal the extent to which various system features have been used.

Analysis of over 4,200 sessions indicates that about 20 percent of sessions invoked the Extended Citation screen, while 38 percent of sessions resulted in viewing the full text of an article. In situated usability interviews, we found that the extent to which people use the available full text is compromised by the fact that they must download additional software in order to view it. Comments made by DeLIver users in interviews also suggest that new system features—like the Extended Citation screen—are bypassed, at least initially, in favor of familiar ways of handling print journal articles. While the Extended Citation screen allows users to identify and browse material in a manner that potential users said would be desirable, actual use requires an initial learning effort not demanded by following one's habitual manner of reading print material.

Given the nature of searching and display that is made possible through the use of SGML and the layered means of displaying search results, we have explored how researchers use journal components—such as abstracts, figures, equations, or bibliographic citations—in their work (Bishop, 1998). We have identified five basic purposes for use of article components: (1) to identify documents of interest; (2) to assess the relevance of an article before retrieving and reading the full text; (3) to create a customized document surrogate after retrieval that includes a combination of bibliographic and other elements—e.g., author's name, article title, tables; (4) to provide specific pieces of information such as an equation, a fact, or a diagram; and (5) to convey knowledge not easily rendered by words, especially through figures and tables.

Engineers describe a common pattern of utilizing document components to focus on and filter information in their initial reading of an article. They tend to read the title and abstract first and then skim section headings. Next, they look at lists, summary statements, definitions, and illustrations before focusing 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 everyone does this differently—e.g., by using a marker to highlight text portions of interest or making a mental register of key ideas. People 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.

Our digital library also provides an opportunity 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 have found that users can be confused by a newly encountered DL, and that it takes some time and interaction for them to decide what a particular system, like DeLIver, is. In usability tests, we identified patterns of user actions designed to uncover what sort of system our testbed was and what it could do. What first appeared to be a random trial and error use of the interface was actually structured exploration that occurred frequently across sessions.

We argue that users take a "cut and try approach" (Neumann & Ignacio, 1998) to help them differentiate our system from other genres of online systems, such as a general Web search engine or an online library catalog. In addition, users look for cues that indicate which conventions different platforms and interfaces hold. Because our DeLIver interface, in particular, draws on many different information system genres without carrying anyone's conventions through entirely, users are confused. For example, in one version of the system, all underlined terms did not represent hypertext links and, in the current version, not all the links are easily identifiable. Because there are no consistently followed conventions for interfaces to Web-based digital libraries, we need to find a way to signal to users what is and is not different about the individual systems they encounter.

Finally, one other area of general research deals with the larger implications of our changing information infrastructure. Communities of practice converge with information artifacts and information infrastructure to produce the "ready-to-hand-ness" of particular resources. Transparency is created and maintained through access to, and participation in, communities of practice and their associated information worlds. We have investigated this through three case studies: (1) of academic researchers, (2) of a profession creating a classification of work practices, and (3) of a large-scale classification system (Star, Bowker, & Neumann, 1998).

There have been many hurdles that we have overcome and challenges in our work that we are still dealing with. Many of these stem from the difficulties of resolving the multiple facets of this large and distributed project. All the different teams on the project bring different expertise, interests, and assumptions about how everything should work. We have often found ourselves at the crux of these differences in our roles of eliciting user feedback, running usability tests, meeting with reference librarians charged with incorporating our DLI testbed into existing library services, and taking broader theoretical perspectives. Negotiating these multiple and sometimes competing visions was the subject of one study in which we focused on understanding the ways in which potential use, new and old infrastructure, and large project organization interact.

Just as our social science research can be different things to different people, so can our digital library. DeLIver is a hybrid system, something of a research system, a system to demonstrate to various stakeholder groups, and also a production system upon whose stability users rely. This mix was particularly apparent during DeLIver's roll-out in October 1997. Our team, in conjunction with the Testbed Team, struggled to make sense of, and deal with, initial user access barriers in the form of authentication and registration procedures. Situating (potential) use 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 World Wide Web searching beyond full-text retrieval requires using document structure in the short-term and document semantics in the long-term. As the testbed team made progress in research with SGML, our Technology Research team focused on the development of the Interspace. The Interspace is a vision of the future Internet where each community maintains its own repository of its own knowledge (Schatz, 1997). For amateur classifiers to be comparable to today's professionals, information infrastructure must provide substantial support for semantic indexing and semantic retrieval.

The focus of the Interspace is on scalable technologies for semantic indexing that work generically across all subject domains (Schatz, Johnson, Cochrane, & Chen, 1996). Analogues of concepts and categories are automatically generated. Concept spaces can be used to boost a search by interactively suggesting alternative terms (Chen, Yim, Fye, & Schatz, 1995; Schatz, Johnson, Cochrane, & Chen, 1996). Category maps can be used to boost navigation by interactively browsing clusters of related documents (Chen, Yim, Fye, & Schatz, 1998). Collectively we refer to these techniques as semantic indexing.

The scalable semantics algorithms rely on statistical techniques, which correlate the context of phrases within the documents. Over the past several years using DLI materials, we have used NCSA supercomputers to compute progressively larger collections until the scale of entire disciplines, such as engineering, has been reached. We use supercomputers as time machines to simulate the world of a billion repositories by partitioning a large existing collection into discipline subcollections, which are the equivalent of community repositories.

Concept spaces were generated in 1995 for 400K abstracts from INSPEC (electrical engineering and computer science) and in 1996 for 4M abstracts from Compendex (all of engineering, some thirty-eight broad subjects). The first computation took one day of supercomputer time (Chen, Schatz, Ng, Martinez, Kirchoff, & Lin, 1996) and the second took

ten days of high-end time on the HP Convex Exemplar (Schatz, 1997). 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.

Concept spaces are collections of abstract concepts that are generated from concrete objects. Traditionally, the objects have been text documents and the concepts all canonical noun phrases. The concept spaces are then the co-occurrence frequencies between related terms with the documents of the collection.

Figure 2 depicts an example of the use of concept spaces for engineering literature (Chen, Martinez, Kirchoff, Ng, & Schatz, 1998). The upper window displays abstract indexes for categories and concepts, while the lower window displays concrete indexes for document collections. The pane in the upper left of the figure shows an integrated list of abstract indexes over the INSPEC, Compendex, and Patterns collections. INSPEC and Compendex are standard commercial bibliographic databases, and Patterns is a Software Engineering community repository.

The upper left-hand pane is a snapshot of a query session in Software Engineering incorporating all three indexes: Computers and Data Processing from the Compendex index, Software engineering techniques from the INSPEC index, and Design Patterns from the Patterns index. The two panes to the right of the upper window show portions of an automatically generated concept space for the INSPEC categories "Software Engineering Techniques" and "Object Oriented Programming." The concept space allows the user to interactively refine a search by selecting concepts that have been automatically generated and presented to the user. In the example below, the user has specified "complex object" and the system returned a list of related concepts such as "configuration management."

If the concept space is further navigated (not shown), additional related concepts, such as "revision control system," can be found. Such concepts may be used in conducting a full-text search as portrayed in the "Full Text Search" pane below the concept space display. This allows the user to descend to the level of actual objects in a collection at any time. In the lower left pane, the user has performed a full-text search on the concept space term "revision control system" and the system has identified several abstracts containing this term with the one selected by the user displayed in the lower right pane.

The Interspace consists of multiple spaces at the category, the concept, and the object levels. Within the course of an interaction session, a user 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.

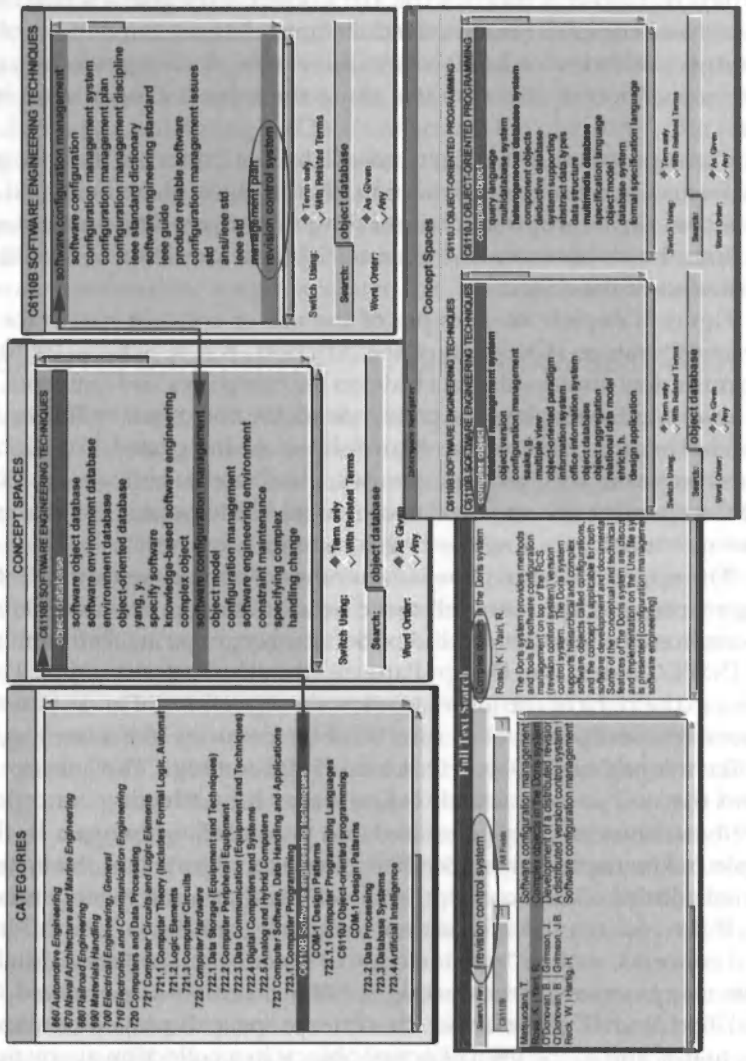

Figure 2. Semantic Indexing for Engineering Community Repositories

They can then move into document space to perform a full-text search by dragging the concept term into the document space search window. This sequence is shown for "revision control system."

Finally, to search a subject domain they are less familiar 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 has been depicted as follows: first, the user has identified "complex object" as a desirable search term

and refined the search by locating the related term "revision control system." Next, the user has determined to pursue the object-oriented theme in greater detail and wishes to switch from the "Software Engineering Techniques" subject domain into the "Object Oriented Programming" domain.

The result of this concept switching is depicted in Figure 2, where the portion of the corresponding concept space for "Object Oriented Programming" has been displayed in the upper right pane. The user can now deepen the search by browsing related terms of "complex object" in the "Object Oriented Programming" subject domain. Such a fluid flow across levels and subjects supports semantic interoperability and is our approach toward vocabulary switching (Chen, Martinez, Ng, & Schatz, 1997). This form of interactive concept switching by space navigation is a key reason for naming the system the Interspace.

MULTIPLE VIEWS AND FEDERATED SEARCH

Complete search sessions across multiple sources are necessary to handle effectively scientific literature. The DLI testbed efforts provided support for federated search across the document structures from different publisher repositories. The user could then use a single high-level structure, such as author or caption, and have it automatically translated into the appropriate SGML tags for each document. The research efforts provided support for federated search across the document contents from different publisher repositories. Higher-level indexes for term suggestion were automatically generated; these enabled the user to provide a general term and choose, from a list of suggestions, a specific term actually useful for a search.

These results make it clear that general network information retrieval systems must integrate multiple views. Traditional information retrieval has supported only a single view—i.e., it sends a query to an index and returns a result. This is the model currently supported within the commercial online systems and within the World Wide Web. A multiple view interface supports complete sessions with a federated search of multiple indexes and with dynamic combination across the results of different searches.

We have developed a multiple-view, multiple-source information retrieval client and tested it on a prototype basis with the sources available within our DLI project. IODyne is custom software that runs on the PC Windows 95/NT platform (Schatz, Johnson, Cochrane, & Chen, 1996). The prototype can retrieve records from various kinds of text sources (SQL, Z39.50, Opentext) and provides search term suggestion from specially prepared subject thesauri and concept spaces. The user can simultaneously display and compare results from multiple bibliographic sources, as well as multiple suggestion sources, and can drag and drop objects from any

window into any other window to create new queries and present different views of data.

Figure 3 illustrates a multiple view session with indexes and protocols from the DLI project. For text search, the testbed SGML repository is accessed via the Opentext search engine and the INSPEC bibliographic database via the Ovid Z39.50 proxy. For term suggestion, the INSPEC thesaurus is handled within a built-in IODyne browser, and the concept spaces for INSPEC come from the Research semantic index computations.

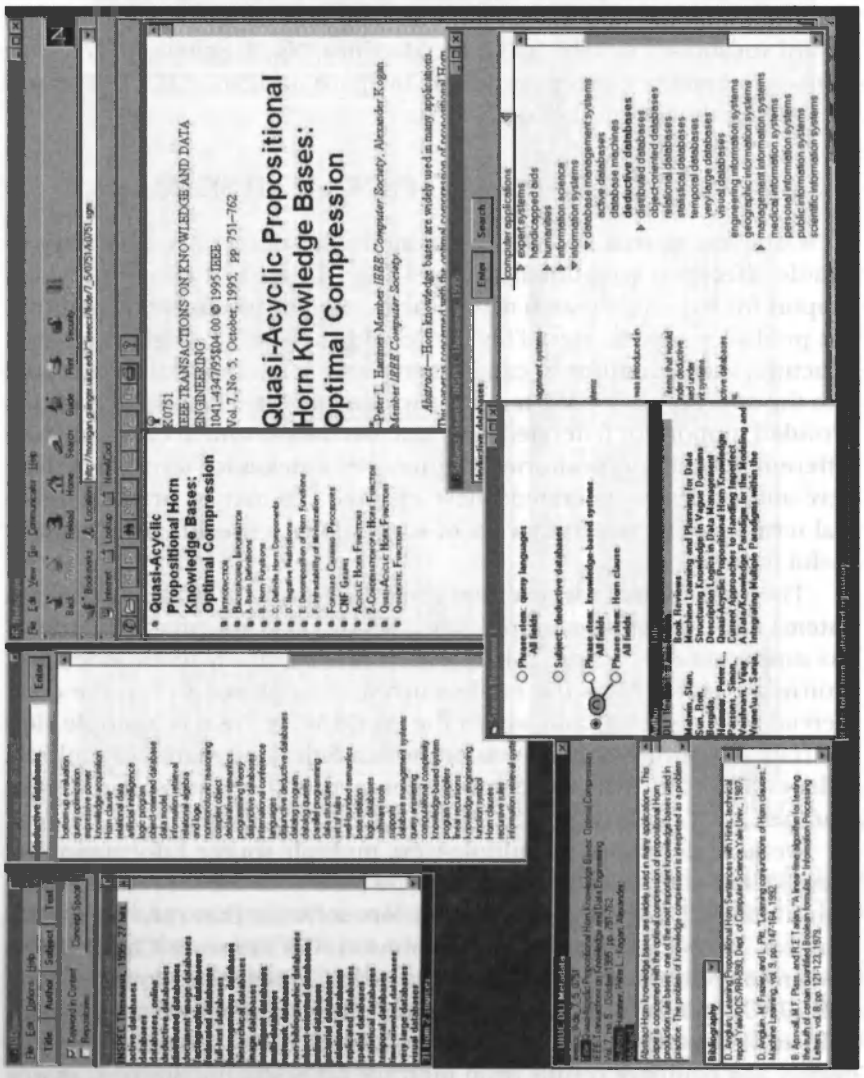

Figure 3. Multiple View Interface for Term Suggestion and Full-Text Search

In a typical session, a person would use the term suggestion sources to identify relevant terms and then the text search sources in order to retrieve relevant documents. The window at lower center is a search document with an SGML engineering repository attached to it. It contains several queries including a Boolean query. The Boolean query is selected, displaying its results in the lower half of the search document. The bibliographic record for one of the retrieved articles is at the lower left, and the full SGML document is displayed in the Netscape browser window at right, behind the search document and INSPEC thesaurus display.

The thesaurus display navigates subject hierarchies in thesaurus and classification systems. To perform a bibliographic search with any subject identifier in the thesaurus display, you drag it and drop it into the search document. Here "deductive databases" were used for a search. Often, the human-indexer subject thesaurus terms are too general for effective search directly but are used to navigate the machine-indexer concept spaces which contain all of the collection terms.

The concept space display, at top center, serves this detailed term suggestion. Search terms listed in it can be dropped onto the search document to perform bibliographic searches; the two terms in the Boolean search, "knowledge-based systems" and "Horn clause," were obtained from the concept space after dragging in "deductive databases" from the thesaurus. The Keyword in Context window, along the upper lefthand edge, shows terms matching a typed entry.

CONCLUSION

Both the DLI testbed and the research efforts of the UIUC DLI project achieved major success. The 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 testbed collection is currently the largest existing federated repository of SGML articles from scientific literature. The DLI testbed users represent a population an order of magnitude bigger than the last generation research system for search of scientific literature. The testbed evaluation performed comprehensive methodologies at both a fine-grain level with user interviews and a large-scale level with transaction logs. Such results will lead shortly to practical commercial technologies for federating structured documents across the Internet.

The 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. The research computations are the largest ever in information science.

They represent the first time that semantic indexes using generic technology have been generated on collections with millions of documents. They are the first large-scale step toward scalable semantics and statistical indexes with domain-independent computations.

The Internet of the twenty-first century will radically transform the interaction 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 the 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 small collections so that they can effectively be searched within the Interspace of a billion repositories. The most important feature of the infrastructure is therefore support of 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 National Science Foundation (NSF), the Defense Advanced Research Projects Agency (DARPA), and the National Aeronautics and Space Administration (NASA) under Cooperative Agreement No. IRI-94-11318COOP. We thank the American Institute of Physics (AIP), the American Physical Society (APS), the American Society of Civil Engineers (ASCE), the Institute of Electrical and Electronics Engineers Computer Society (IEEE CS), and the Institution of Electrical Engineers (IEE) for making their SGML materials available to us on an experimental basis. Engineering Index (EI) and IEE kindly provided Compendex and INSPEC respectively.

Many people have contributed to the research discussed here. In particular, we thank Robert Wedgeworth, Kevin Powell, Ben Gross, William Pottenger, Donal O'Connor, Robert Ferrer, Tom Habing, Hanwen Hsiao, Emily Ignacio, Cecelia Merkel, Bob Sandusky, Eric Larson, S. Leigh Star,

Andrea Houston, Pauline Cochrane, Larry Jackson, Mike Folk, Kevin Gamiel, Joseph Futrelle, William Wendling, Roy Campbell, Robert McGrath, Duncan Lawrie, and Leigh Estabrook.

REFERENCES

Bishop, A., & Star, S. L. (1996). Social informatics for digital library use and infrastructure. In M. E. Williams (Ed.), *Annual review of information science and technology* (vol. 31, pp. 301-401). Medford, NJ: Information Today.

Bishop, A. (1998). Digital libraries and knowledge disaggregation: The use of journal article components. In I. H. Witten, R. M. Akscyn & F. M. Shipman (Eds.), *Digital libraries '98* (The third ACM International Conference Digital Libraries, June). New York: Association for Computing Machinery.

Chen, H.; Yim, T.; Fye, D.; & Schatz, B. (1995). Automatic thesaurus construction for an electronic community system. *Journal of the American Society for Information Science, 46*(3), 175-193.

Chen, H.; Schatz, B.; Ng, T. D.; Martinez, J.; Kirchhoff, A.; & Lin, C. (1996). A parallel computing approach to creating engineering concept spaces for semantic retrieval: The Illinois digital library project. *IEEE Transaction Pattern Analysis and Machine Intelligence, 18*(8), 771-782.

Chen, H.; Martinez, J.; Ng, T. D.; & Schatz, B. (1997). 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.

Chen, H.; Houston, A.; Sewell, R.; & Schatz, B. (1998). Internet browsing and searching: User evaluations of category map and concept space techniques. *Journal of the American Society for Information Science, 49*(7), 582-603.

Chen, H.; Martinez, J.; Kirchhoff, A.; Ng, T. D.; & Schatz, B. (1998). Alleviating search uncertainty through concept associations: Automatic indexing, co-occurrence analysis, and parallel computing. *Journal of the American Society for Information Science, 49*(3), 206-216.

Neumann, L., & Ignacio, E. (1998). Trial and error as a learning strategy in system use. In C. M. Preston (Ed.), *ASIS '98* (Proceedings of the 61st American Society Information Science Annual Meeting, Pittsburgh, PA, October). Medford, NJ: Information Today.

Schatz, B.; Johnson, E.; Cochrane, P.; & Chen, H. (1996). Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In E. A. Fox & G. Marchionini (Eds.), *Proceedings of the first ACM International Conference Digital Libraries* (Bethesda, MD, March) (pp 126-133). New York: Association for Computing Machinery.

Schatz, B.; Mischo, W.; Cole, T.; Hardin, J.; Bishop, A.; & Chen, H. (1996). Federating diverse collections of scientific literature. *Computer, 29*(5), 28-36.

Schatz, B. (1997). Information retrieval in digital libraries: Bringing search to the Net. *Science, 275*(January), 327-334.

Star, S. L.; Bowker, G.; & Neumann, L. (1998). *Transparency beyond the individual level of scale: Convergence between information artifacts and communities of practice.* Unpublished manuscript.