[47] S.K.M. Wong, W. Ziarko, V.V. Raghavan, P.C.N. Wong, Extended Boolean query processing in the generalized vector space model, *Information Systems*, v 14, pp 47-63, 1989.

[48] C.T. Yu, G. Salton, Precision weighting -- an effective automatic indexing method, *JACM*, v 23, pp 76-88, 1976.

CHAPTER 2 *July 1992 Workshop*

Chapter 2, Section A: List of Attendees

Bill Arms, Carnegie-Mellon University
Christine Borgman, UCLA
Y.T. Chien, IRIS Division, NSF
Steve Cisler, Apple Library of Tomorrow
Ed Fox, VPI & State University
John Garrett, Corp. Nat'l Research Initiatives
Hector Garcia-Molina, Stanford University
David Hartzband, DEC
Beverly Hunter, Education & Human Resources, NSF

Larry Hunter, National Library of Medicine
Hosagrahar Jagadish, AT&T Bell Labs
Paul Kantor, Rutgers University
Larry Lannom, Meridian Corp. (for DARPA)
John Leggett, Texas A&M University
Mike Lesk, Bellcore
Mark Liberman, UPenn
David Liston, Council on Library Resources
Clifford Lynch, UC Berkeley
Alexa McCray, National Library of Medicine
Mike McGill, University of Michigan
Judy Moline, NIST
Connie Moraff, U.S. Information Agency
Jan Olson, Cornell University
Dave Penniman, Council on Library Resources
Larry Rosenberg, IRIS Division, NSF
Dan Vanbelleghem, Networking/Communications, NSF
Maria Zemankova, IRIS Division, NSF
Robert Zich, Library of Congress

Chapter 2, Section B: Agenda

Workshop on Electronic Libraries
Agenda

Day 1 - Monday, July 20, 1992

8:30	Coffee
9:00 - 9:30	Introduction and Opening Remarks - YT Chien
9:30 - 10:00	Background and Vision - Mike McGill
10:00 - 10:30	Current effort (1) - Ed Fox
10:30 - 11:00	Break
11:00 - 11:30	Current effort (2) - Mike Lesk
11:30 - 12:00	Current effort (3) - Bill Arms
12:00 - 12:30	Current effort (4) - Dave Penniman
12:30	Lunch
2:00 - 5:30	Research & other issues - session 1

Assignments of topics and leaders

Networking/Infrastructure - Mike McGill
DB/Retrieval/Media - Ed Fox
Linkage to Industry - Mike Lesk
User Interface - Chris Borgman
Standards, IPR, etc. - Cliff Lynch

Day 2 - Tuesday, July 21, 1992

8:00	Coffee
8:30 - 10:00	Research & other issues -session 2 Summary by group leaders
10:00	Break
10:30 - 12:30	Discussion on priorities, actions, costs
12:30 - 1:30	Lunch
1:30 - 3:00	Future plans - next workshop; working groups, etc.
3:00	Adjournment

Chapter 2, Section C: Lesk Summary

Workshop on Electronic Libraries*

Michael Lesk

Bellcore
Morristown, NJ 07960

August 4, 1992

** This report is the personal opinion of the author and does not represent a position of either Bellcore or the National Science Foundation.*

On July 20-21 the National Science Foundation held a workshop in Washington to discuss a proposal for a National Electronic Library for Science, Engineering and Technology. The original proposal called for $50M over five years, half to be devoted to creation of material and half to be devoted to exploring the uses of the material. After discussions we made this considerably vaguer.

We have all watched as catalogs, abstract/index journals, and now full text has switched from paper to electronic form. For scientific and engineering information, however, the presence of graphics, the complex economic condition of technical publishers, and other constraints still leaves us without general availability of most scientific articles in electronic form. As the NREN develops and would make it possible to deliver such information, how can we facilitate the development of electronic information systems for the benefit of the whole nation? What would be the modern equivalent of the railroad land grants, the airline mail contracts, or the requirement for interchangeable parts in rifles that stimulated other important industries in the past? The success of the French with Minitel warns us that we can not just assume that technological advance will happen first in the United States regardless of what we do.

The intent of this meeting was to talk about how NSF might put together a new national initiative designed to encourage development of practical applications of electronic information, phrased in terms of a national electronic library for scientific and technical information at the undergraduate level. The original proposal devoted \$25M to library acquisition, with the intent of dividing this money among 5-10 discipline-specific projects, each of which would select the best material for their area and deal with making it available in machine-readable form.

At the beginning of the meeting Michael McGill (Univ. of Michigan) presented this original proposal, which argued for creating electronic information resources to permit use of the NREN to improve U.S. science education. The U. S. information companies devoted to scientific and scholarly information are relatively small compared, for example, to those involved in financial information. A Federal role would greatly accelerate the availability of scientific and technical information, particularly for educational applications. In addition to creating an electronic library, the proposal also called for funding four major research centers and a dozen additional research projects on the uses and exploitation of electronic libraries. Two oversight boards, one to handle technical compatibility and one to handle inter-discipline cooperation, were proposed. Although this sounded ambitious, it was very small compared to the Information Infrastructure and Technology Act of 1992 (Sen. Gore) which called specifically for building digital libraries and which suggested (over seven years) \$150M for NSF-funded research relating to libraries and, adding up over all agencies and applications, a total of \$1.15B.

Ed Fox (Va. Tech) then discussed his ENVISION project, which will use ACM's electronic data (journals, algorithms, hypertext, and so on) and will combine in its architecture SGML, HyTime, multimedia, objects, links, user-centered design, Z39.50 and open systems. He also mentioned several other electronic library projects, including Elsevier's TULIP (three years of 42 journals), and a joint graduate school effort on dissertations.

I described the CORE project, a joint effort of Cornell, Bellcore, the American Chemical Society and Chemical Abstracts Service, and OCLC. We work with chemical journal data, and have run experiments showing that for any task involving searching, access to electronic data is much faster and better than to paper data. For reading, paper and screens are now comparable. Both image and ASCII data are available for our collection, and experiments are being run to compare them.

Bill Arms (CMU) talked about his MERCURY project, a follow-on to MIT's Project Intrex of the 1960s. Everyone, he pointed out, wants lots of well-managed information, available rapidly over a distributed network, easy to navigate around, and reasonably priced. He felt the important problems were what kinds of objects to store, how to navigate in a large on-line file, dealing with the scale of really useful libraries, and educating the many disciplines involved to work together.

Dave Penniman (Council on Library Resources) had attended the NENGIS conference in June, devoted to an ultimate goal of a "virtual" national engineering library in electronic form. This conference proposed focussing on improving the information gathering skills of engineers, assessing the needs of the community, and building coalitions. To work with the embedded base, focus its effort on engineering, Larry Lannom (DARPA) talked about getting CS tech reports on the desktop; the plan is that all reports will be in image format with bibliographic descriptions in MARC compatible form.

Dave Hartzband (DEC) described a collaboration with a major multinational corporation on office/factory automation. They found just technology is not enough, but that social/anthropological knowledge is also needed to effect change in such an organization (which he described as a "cornered animal").

We then broke into four groups to address (1) what tools are needed, (2) what content an electronic library should have, (3) what the applications are, and (4) economic/social/legal questions. The results of these groups were not directly useful – as always, everyone asked for everything. The tool group wanted open architecture, including distributed heterogeneous extensible modular systems (plus I assume motherhood and apple pie), plus both research and prototypes, short and long-term work, and a basic structure including a network and peer review. The content panel wanted a cross product of traditional scientific publishing, grey literature, and multimedia with various subject disciplines and with various technological solutions (interfaces, systems, and so on). The applications group talked about IR (Information Retrieval), text generation, CSCW (Computer Supported Cooperative Work), and so on; with various requests for navigation, ease of use, and leveraging. The social/economic panel wanted studies of social and cultural implications of libraries, and also suggested a specific organization in NSF to deal with electronic libraries. None of the groups did much about pruning the list of demands or giving NSF good reasons for much of what they recommended.

There was a good discussion in the full workshop of facilitating access to copyrighted material. John Garrett suggested a compulsory licensing scheme with a procedure for publishers to opt out, but requiring them to provide a reasonable alternative. Bill Arms suggested that NSF, DARPA, and NIH require that work done on government funding should have the copyright transferred to the government. Another proposal was that first drafts of papers should be available electronically. Unfortunately, the workshop did not have enough representation from the publishing community to hear reaction to such proposals.

Y. T. Chien (NSF) then talked about what is involved in a research initiative. He described the way different government agencies had cooperated to develop the HPCC and Biotechnology initiatives, and the kind of public arguments that are needed. It is clear that we need to collect support from some of the groups represented at the meeting (e.g., NIH, LOC, NIST, DARPA, and the National Archives), and to show arguments, with graphic demonstrations, of the value of what we propose; and also to get testimonials from respected scientists.

After a discussion about the philosophical role of libraries, we wound up with a summary by Michael McGill. He gave as the justifications for this work:

- leverage to transform research and education
- promotion of synergy and multidisciplinary research
- improving manufacturing competitiveness

- widening access to information resources
- following the logical development of the NREN and its applications
- providing more cost effective access to information

What is needed for an electronic library service is an architecture including the network, servers, and standards, plus information resources that will meet a discipline's needs.

The ''goal statement'' was edited for some time and came out as

> To realize a national service supporting research, education, and commerce by providing ubiquitous access to relevant, high-quality, usable information in all media for all people.

Among the steps to be done for this were to:

(a) interconnect, catalog and enhance existing and emerging services and resources;

(b) support development of electronic collections;

(c) do research and development in IR, CSCW, navigation, education, and so on;

(d) address the social and economic issues;

Finally, all members of the audience were asked to say what should $1M in 1993 be spent on. Some of us (me, Paul Kantor, Bill Arms) wanted all or most of the money to go for the most realistic prototype or demonstration project that could be done. Others wanted scenarios of the future, building relationships, and so on. Cliff Lynch, John Garrett, and Chris Borgman wanted some on collection of content, some on social studies, and some on visualization of the future.

On balance, this workshop was useful for extending the number of people who know about the proposed initiative, but provided relatively little useful guidance on what specifically to do.

We struggled, through much of the workshop, with issues relating to publishers, authors, and other aspects of the existing scientific publishing system. We have no good answers, but we should remember that the most likely challenge to the existing system will be a totally outside effort – something like WAIS (Wide Area Information Server) or WWB (World-Wide Web) coming forward, developing into a general information system, and being adopted as a substitute for conventional publication. There is, after all, no general reason to think that the organizations that dominated one industry will dominate its technological successor: canal companies did not build the railways, steam locomotive builders disappeared when diesels came along, Kodak did not do well in instant photography, and Winsor & Newton doesn't make photographic film. Both libraries and publishers are in danger of being bypassed, and I believe aggressive action on their part is needed to preserve their pre-eminent role in information handling.

Similarly, there is increasing development in Europe on information infrastructure. It is not just Minitel -- imaging projects, for example, are more ambitious in Europe than here. The economies of scale in the electronic business are so large, with essentially no per-copy costs, that the U.S. domination of the information business is at risk. We are not talking about research here -- many of the systems are practical and can be put forward immediately.

Perhaps the key lack in this area is a set of purposes to which electronic scientific information could be directly applied. The advantages of the electronic format are clear -- Y. T. Chien pointed to its ubiquity, speed and convenience of access, searchability, and so on, while Maria Zemankova talked about the ability to reformat, e.g., for the disabled (large type, or text to speech conversion). But few people addressed the specific applications, so that we did not make progress on deciding whether we should focus on undergraduate education, exchange of information between practicing scientists and engineers, research support, technology transfer to manufacturing groups, or what. In practice, we still must survey the possible application areas and decide where the best return on NSF investment can be obtained. Nor was there serious discussion of the trade-off between expensive new multi-media concoctions and straightforward conversion of existing printed materials.

As one choice, I feel science education is a serious problem in the U.S. and we need to try whatever new tools we can. I realize that the education community has lived through programmed instruction and computer-aided instruction and is very skeptical, but research in this area is important and should be stepped up. Thus, I would focus an effort on material aimed at undergraduate science education; and I would focus on conventional material, in bulk, rather than a few multi-media shows.

I would hope that NSF can move forward on some kind of electronic library projects. I regret that the meeting went towards more general statements rather than focussing on what might actually be done. Personally, I would prefer to go back to the original initiative, and get to work on actually building an electronic library.

Chapter 2, Section D: Statements of Participants

Christine L. Borgman, UCLA

A. Statement of Work Relevant to Electronic Libraries:

A.1 Research:

- Interface design for information retrieval systems, both for adults and children. Research environments studied include Boolean-based on-line catalogs, Boolean-based scientific and technical databases, and hypertext, browsing-oriented systems for children

- User behavior in seeking information, both of adults and children

- User characteristics in seeking information and the use of information retrieval systems, including mental models and individual differences

- Bibliometrics and scholarly communication, studying the characteristics of the fields of library/information science and communication, as reflected in published literature, and the broad context of scholars' use of published information

A.2 Teaching:

- Human-computer interaction, including the social context of the use of computers, behavioral aspects of computer usage, and the design of interfaces, with special attention to the design of information retrieval systems

- Library automation, including the internal library operations (circulation, acquisitions, cataloging), data management, access to external resources (local, regional, national networks, etc.), and relevant standards and technologies

- Information systems analysis and design

- On-line searching of textual databases, including technical skills, resources, and service management

- Scholarly communication and bibliometrics, including bibliometric methodology, multi-disciplinary perspectives on scholarly communication, and application to social theory, library management, and information retrieval

- Social implications of information technologies, including privacy, the changing nature of work, and other technology policy issues

A.3 Professional Experience:

- Systems analysis and design of information retrieval systems for libraries and for industry

- Design of personal name matching algorithms for information retrieval

B. NSF Workshop in Electronic Libraries - Research Agenda Statement

This is the abstract of a talk I have given (in varying forms) around the U.K. on behalf of the British Library, in Ireland, and at several universities and research organizations around the U.S. from 1990 to date, which summarizes the current state of affairs in the electronic library area and proposes directions for the future. I will bring to the meeting a long outline of specific research issues extracted from this talk.

B.1 The Library Of The Future: Intelligence, Integration, Interconnections

Libraries have always been the primary and central repositories of scholarly and other forms of public knowledge in print form. As electronic formats have become an increasingly important part of the information environment, libraries have changed rapidly, along with their users. People now expect their information to come to them over electronic networks, in the form of newsletters, journals, discussion groups, and databases. This evolution presents both challenges to the professions and research questions to the disciplines at the nexus of information, computing, and behavior. We need a better understanding of why and when people seek information, how they search for information, and how we can model their behavior. We need standard data structures beyond the bibliographic and flat text structures now in place, and we need more sophisticated retrieval algorithms tailored to the models of information seeking and models of data. Several factors will affect the directions taken by future information systems. The hardware and software technology for these developments will be less of a problem than the intellectual activities to describe, access, and retrieve the volume and variety of data already extant. Social and cultural factors will drive specific implementations, including organizational attitudes toward the support of work tasks, training and education in the use of information technologies, and standards developments. Economic and political factors may determine the balance of local vs. remote resources employed and will result in debates over social equity, privacy, copyright, etc. Libraries will face ever more difficult choices in resource allocation between acquiring materials and providing remote access, as well as managing an ever-increasing variety of organization-wide information resources. The talk will examine these issues and research questions in the context of the current computing and communications environment in which the educational community operates.

Steve Cisler, Apple Library of Tomorrow

My background is in public library work, so I have a strong belief that we must understand the needs of a spectrum of user capabilities and information requirements.

I manage the Apple Library of Tomorrow program which has provided equipment grants to a wide variety of small to large libraries of all types. The types of projects include easy-to-use children's catalogs, multi-media databases, provision of Internet connectivity to rural libraries, experimentation with spread spectrum radio to establish a metropolitan area network in San Diego, and a trial project to mount Smithsonian images in JPEG compressed formats for Internet users.

I have also participated in a research project called Rosebud to develop a non-boolean searching tool that makes use of software agents.

I am involved in information policy formulation for the Computer Systems Policy Project which is comprised of 12 CEO's of various computer companies. Their present focus is on the National Information Infrastructure, or what happens after the NREN.

Ed Fox, Virginia Tech

A. Background:

In the late 60's I worked with J.C.R. Licklider of Project MAC, and Mike Kessler of Project TIP, both at MIT, both interested in electronic libraries. In the late 70's I worked at Cornell with Gerry Salton: implementing the SMART system in C for UNIX, showing that searches with citation data added to keywords was more effective, and proving that extended Boolean schemes outperform standard Boolean methods.

At Virginia Tech we have worked on multimedia (especially DVI -- digital video interactive), testing the utility of advanced retrieval methods for library catalog access, building a next generation library catalog system based on that experience, hashing algorithms, building an English lexicon from machine readable dictionaries, and recently on data fusion in the TREC project (a DARPA supported large scale test of retrieval methods). The LEND (Large object-oriented External Network Database) effort was designed to show how our hashing algorithms and viewing information, data, and knowledge as all stored in a large graph, would be better for information retrieval and hypertext. The CODER (COmposite Document Expert/extended/effective Retrieval) system was designed to show the use of AI in information retrieval, and has been applied to electronic mail, Navy messages, and a large collection of materials on Cardiology (including a book and the MeSH thesaurus.)

As chair of the ACM Special Interest Group on Information Retrieval (SIGIR) I've worked with Mike McGill and Mike Lesk to encourage a federal initiative on electronic libraries. My view of SIGIR encompasses not only traditional Information Storage and Retrieval, but also certain issues of multimedia information systems, along with algorithms, computational linguistics, experimental as well as theoretical studies, and, of course, this matter of electronic libraries.

Project Envision, in its first year out of 3, is building an electronic library in Computer Science. We are exploring issues regarding collection and conversion of items, document description and markup, preparation of guidelines to simplify the work of authors and publishers, construction of a prototype system (including work on human-computer and Mathematica interfaces, user task analysis and testing, efficient and effective retrieval, working with objects instead of text strings or bitmaps, etc. In some areas we are dealing with the key problems of building a National Electronic Archive, albeit on a smaller scale.

B. Issues and Ideas to Contribute:

- How can an archive of sufficient size be developed to allow research to proceed? Where will funding come from for development work?

- How can a coherent scheme of support for standards for a National Electronic Library emerge, be accepted by authors, professional societies, universities, corporations, publishers, readers?

- How can we model users and their tasks, so as to best serve them as they make use of this very large information collection?

- How can we use the archive to leverage personalized computer aided instruction, perhaps using hypertext and multimedia, so it will be less labor intensive to improve education?

Hector Garcia-Molina, Stanford University

A. My areas of interest include database management systems and distributed computing. In particular, I have focused on synchronization and crash recovery problems.

Recently, I have worked (or am working) on three projects that are related in some way to Electronic Libraries:

1. Distributed Indexing.
Indexing in an electronic library is typically through inverted lists (e.g., the inverted list for the word “cat” contains pointers to all documents where the word occurs). As the number of documents and queries grows, there is a need for parallel searching and for distributing the lists across multiple processors and disks. We are evaluating the options for such distribution, initially via detailed simulations. A preliminary technical report is available (STAN-CS-92-1434).

2. Unstructured Document Management.
We have constructed an information management system for unstructured office

documents such as electronic mail, files, and faxes. The interface to the system is a mailer (e.g., MH) that has been enhanced with library style filing and retrieval facilities. The system can be linked to other information sources (including more traditional electronic libraries) to form a distributed repository. Through the mailer interface, one can easily exchange with others information that has been retrieved. (This work is being done at the Matsushita Information Technology Laboratory in Princeton, and a Technical Report is available.)

3. Disaster Recovery
When important information is stored in a database system, disaster recovery techniques are employed to safeguard the data. Typically, a remote facility maintains a "hot standby" copy. As the primary site updates the data, the log of the changes is propagated to the backup. We have been investigating algorithms for efficient propagation of updates and for management of the backup copies. (Various papers are available.) In an electronic library one must also cope with disasters and preserve documents. The operational parameters may be different from those of traditional database applications (e.g., no updates to existing documents), but still, some sort of disaster recovery will be needed.

B. There are a wide variety of important issues that must be investigated in the area of electronic libraries. Here is one way to classify the issues:

I. SYSTEMS ISSUES

1. Single Site — Tertiary storage, Scalability, ...
2. Distributed Systems — Information finding, distributed index structures
3. Providing index terms for documents — Optical character recognition, ...

II. DOCUMENT ISSUES

1. Electronic Journals — Can we incorporate journal processing into the library?
2. Annotation and Collaboration — Instead of just submitting papers to the library, can papers be developed within the library, in a cooperative way?
3. Copyright and Privacy — Copyrights and royalties, what can we do?
   Can we make additions to the electronic library documents, that are only to be seen by selected groups?

III. INTELLIGENCE ISSUES

1. Intelligent Finding — Mediators, Agents, Know-bots, and the like...
2. Integration — How can documents from heterogeneous sources be found?
3. Document Modeling and Understanding — How do we represent a document? [What are the] operations?

IV. USER INTERFACE ISSUES

1. User interface to an Electronic Library— How does a user submit queries, view documents? do updates?
2. Multimedia — How are sound, video, pictures displayed?
3. Preservation — How does one preserve look and feel of documents that will no longer be available in their original form?
4. Librarian of the Future — What will the librarian of the year 2010 look like?
5. Social Impact — What will be the social impact of Electronic Libraries?

John Garrett, Corporation for National Research Initiatives

A. I am currently Director of Information Resources, Corporation for National Research Initiatives. Until July, 1991, I was Director of Market Development for the Copyright Clearance Center, with responsibility for building electronic pilot projects (among other things). In my current position, I am responsible for building structural relationships and designing programs which would enhance CNRI's goal of supporting the creation of a national digital information infrastructure for the United States. One of our major programs involves designing and building digital library systems in pursuit of this goal. Therefore, I am actively involved in a number of different forums and contexts in working with other individuals and organizations who seek to create digital libraries, and to address the many issues and problems which are involved in the process.

My particular interests involve addressing rights and permissions issues in the digital library framework, and building sound programs to assess the impact of digitally accessible information on users, information creation and dissemination in general, and the larger community. I have written, published and spoken widely on digital libraries, copyright management, the economics of digital information, and related issues. At CNRI, we are working on several technologies which are important to building digital libraries, including further expansion of our Knowbot software agents, and the problems of security and confidentiality in networked environments.

At the end of the meeting, we were asked to respond to several general questions. On the first of these, "What do electronic libraries look like to....", I don't feel qualified to respond because we don't fit any of the categories. I would suggest that, for a couple of them, you might want to ask the questions to key people who weren't at the meeting: specifically, publishers and information providers (and perhaps authors). I think the other categories were pretty well covered in the group. If you'd like suggestions of publishers and providers who might comment, let me know and I'll recommend some names.

B. The second charge was to identify specifics in eight areas, beginning with problems. Here's an initial response.

PROBLEMS

- Inadequate long term funding
- Economics of electronic libraries are unclear
- Technologies for sorting and screening complex information are poorly developed (or poorly known)
- No clear way for existing projects to coordinate, and learn from each other
- Too little comprehensive, mission-critical information in electronic libraries for anyone except lawyers to do real work

EFFECT OF DELAY

- More and more uncoordinated, slipshod, narrowly focused efforts to build individual electronic libraries; more failures
- Increasing risk of control by largest players (including government)
- Further decline of U.S. competitive posture versus successful information-intensifying economies

APPROACHES

- Support intelligent electronic library construction and dissemination
- No one way: let all flowers bloom
- Move aggressively to identify and coordinate existing efforts

- Move aggressively to involve commercial providers (Dialog, West, Mead Data Central, Dow Jones, etc.)
- Support evaluation models
- Support standards development and dissemination
- Support development of alternative approaches to copyright

PROJECTS

Rather than identify particular projects, it would make more sense (I think) to build a process for continuing identification and coordination of related efforts. I'd be glad to participate in that, but it's a significant effort.

These seem to me to be the key items: let me know if more ideas on any of these, or others, would be useful.

C. In no particular order, these are the issues around this topic I'm worrying about today:

- Too much information, undigested, unusable: I want active, tuneable filtering agents
- No higher order integration of existing projects -- how can the various efforts learn from each other?
- Divergent, slowly moving standards efforts
- Insufficient body of important digitally available material for real world experimentation. Only the lawyers can do real work based on digital library materials (a frightening thought).
- Too much emphasis on technology, too little interest in the social, economic, legal, political, consensus-building activities
- Is there a marketplace? If there is, how would I know? (88% of the on-line business goes to three companies; more than 50% is for legal information)
- Limited involvement of rightsholders, authors, creators, librarians (e.g., on the list for this meeting)
- Automated rights and permissions

- Lack of linkages to other major industry players (computer and software manufacturers, in particular -- not to mention the largest producers of information, like the insurance companies)
- What are the real costs of maintaining and sustaining the networks? Where will the money come from? Where should it come from?
- Who is going to step up and assert leadership? NSF?

H. V. Jagadish, AT&T Bell Labs

A. Research Work

1. RIGHTPAGES: This is an electronic journal browsing and delivery system. Currently it has over 100 users and covers most major journals in Electrical Engineering. and Computer Science. The guiding philosophy is to maintain the look-and-feel of the paper original as far as the screen display is concerned, while exploiting the electronic representation to providing facilities for searching, alerting, annotating, etc.

2. PAGE LAYOUT ANALYSIS: Two-dimensional grammars can be defined for page layouts and special parse techniques used to match physical blocks on a page against logical entities in a parse tree. This sort of work is crucial to translation of documents from paper to electronic form.

3. INDEXING: Index techniques for document retrieval have been studied much. It is important to see how these may change for electronic documents. One issue in particular is indexing on other than alphanumeric strings.

B. Discussion

There are two aspects -- one is research topics, the other is utility to the scientific and research community from a working electronic library system. These two should be considered separately.

As far as a working practical system is concerned, most of the basic technology is in place, at least for a first-cut (but definitely useful) system, provided the source material is available in electronic form. The reason we do not have such systems in place today is simply a question of cost and return on investment. NSF should select a commercial vendor(s) and award a contract to build and maintain such an electronic library system. This will provide the seed impetus needed commercially, just as ARPA's funding of Arpanet through a contract to BBN worked successfully several years ago.

While we know enough to build a useful system today, there still remain vast gaps in our knowledge, and there is ample scope for research in related areas. Two areas in particular that I believe are both important and worth attacking are (1) the conversion of documents from paper to electronic form, and (2) the development of storage and retrieval techniques appropriate for large collections of electronic documents. NSF can encourage research in these and other related areas through judicious funding of appropriate research projects at universities.

Paul B. Kantor, Rutgers

Needs Assessment by Development and Evaluation of Prototypes.
July 16, 1992

Statement of Principle

At Rutgers' Alexandria Project we are convinced that the best path to full information access for the nation's scientists and engineers is the systematic development and evaluation of prototype tools for access and organization. We believe that the events of the past two decades have shown that new developments in information and communication technology are so transformative that the need for them cannot be assessed by conventional market survey techniques. Development of prototypes makes the potential concrete, while providing the base of experience that will be needed to develop the instruments whose prototypes prove to be most used and most valuable.

Present and Planned Work at Rutgers

Work presently underway includes the development of an Adaptive Network Interface, which permits users who belong to a common interest group to share with each other pointers from book to book in a library's on-line catalog. [Supported by the U.S. Dept of Education. KANTOR, RICE, CONTRACTOR]. On the drawing board [proposal submitted to U.S. Department of Education] in collaboration with Waksman Institute, SCILS and Rutgers University Libraries is a project to prototype "Biotell" interfaces which permit seamless transition among the separate components of a typical on-line “data chase” in genetic engineering. [KESSELMAN, DAY, SOFER, HAMM, SARACEVIC, KANTOR].

In the planning stages are efforts to prototype semi-intelligent searchers [“know-bots”, “cobots”] which can conduct searches for engineering and scientific information across the Internet. [KANTOR, RATZAN, SCHWARTZ].

A more theoretical venture is an investigation of the degree to which retrieval enhancement schemes can be accommodated in a common framework represented by data fusion applied to the inverse image of indexing maps. The hope is that new developments in the management of parallel database structures will enhance retrieval effectiveness when problems are formulated in this way [KANTOR, FRIEDER at GMU]. Network optimization in general is a major research interest of several Rutcor Fellows [AVI-YITCHAK et al.]

Other specific research projects underway or on the drawing board include: (1) designed experiment to test whether the signal detection model can fully characterize the results of the TREC, or whether question-system interaction effects are present [KANTOR]; (2) development of alternate search strategies for use with INQUERY for study of the implications of combining strategies and/or fusion of retrieval status data [BELKIN, KANTOR]; (3) investigation of reverse storyboarding of prepared video broadcasts, as a step towards image indexing of continuous image materials [ATWATER, KANTOR, BURNETT]; (4) potential of Kohonen feature maps as a tool for the organization of bibliographic materials [SCHWARTZ, KANTOR]; (5) conjoint analysis as a tool for assessing feature significance in the provision of reference support [CRAWFORD, KANTOR].

Remarks on the Economic Picture

We have explored the basic economic picture [for the CLR/Engineering Foundation Conference] and have concluded that, across the full range of engineering and science activity - including computer and technical personnel, time worth $33 Billion per year is spent accessing and using information. Consider an economic model in which Authors are compensated for making their information available over the network. We estimate that the total compensation, paid by information users, will be of the same order of magnitude as the cost of the time. If 5% of that total is directed towards the organization and transmission of the information, more than 1.5 Billion dollars per year will be expended on the development and maintenance of archival files, pathways, navigators, servers and clients.

Summary

We believe that this industry will grow eventually. However, an organized array of intelligently designed research prototypes can have the same beneficial effect that NSF's early support of on-line bibliographic databases provided. We believe strongly that the research emphasis should be on prototype projects which provide: (1) an arena for developing new ideas, (2) systematic testing and evaluation of these ideas, (3) a head start in the implementation of the best of those ideas. As some ideas become successful they will provide the foundation for de facto standards which will supplement and enlarge the standards now under development.

Larry Lannom, DARPA

A. DARPA Workshops on an Electronic Library for CS Technical Reports

Prepared for the NSF Workshop on Electronic Libraries

7/15/92

DARPA has sponsored two workshops on the topic of increased availability of Computer Science Technical Reports. The workshops were attended by faculty from Libraries and CS departments of major university producers and collectors of CS TRs. The meetings, and subsequent e-mail exchanges, generated a consensus on how this might best be accomplished in the short term.

The goals were identified as 1) real time access to CS TRs, immediately upon publication and completely independent of location, i.e., one virtual library on your desktop; 2) a display format at least as good as paper for graphics, equations, etc.; 3) relatively low barriers to entry, both in terms of publishing reports and in accessing published reports; and 4) scalability, for both numbers of participants and for types of material.

The following general scheme was evolved to meet these goals:

Each TR publisher, as part of the publication process, agrees to 1) create a bibliographic record, in an agreed upon format, for each TR and broadcast that record to all other participants, probably through a validator of some sort; and 2) place the full text of the report on a local server, again in an agreed upon format. The server would be guaranteed to respond with the full document to requests that arrived according to a given protocol. The broadcast records would contain pointers back to the full documents.

This would allow, although not require, the participants to build local databases of the combined bibliographic records from all participants and to build and maintain local search, retrieval, and display mechanisms which would query the remote servers holding the complete documents.

A great deal of discussion went into the most appropriate standards and protocols. The rationales can't be summarized in this space, but the selections were 1) text based structured records for the bibliographic component; 2) bit-mapped images, e.g., G4, as a minimum for the full document, with additional formats allowed as multiple copies; and 3) e-mail for bibliographic record broadcast. Z39.50 for retrieval and Kerberos for authentication were also identified as likely candidates. Note that all other choices, e.g., what to do with the bibliographic database and received page images or other full document formats, are purely local.

B. Possible Issues to Address at the NSF Workshop on Electronic Libraries

The general topic of electronic libraries sits in the middle of a wonderfully interesting collection of technical problems, including distributed computing, document representation formats, and information retrieval. For such libraries to work at a national level there has to be tight agreement on certain issues, e.g., access methods and representation standards, while other issues, e.g., user interface development of selection, can be left as local decisions. A workshop such as this one must address these technical issues, but it may well turn out that the non-technical issues are more difficult to solve. These include:

- What constitutes an electronic library? A collection of electronic documents? A collection plus access protocols? A collection plus access plus interface?
- Who are the intended users? Anyone with a workstation? Anyone with a terminal? Anyone on the Internet? Anyone with a modem? Are there any access restrictions?
- How are rights and permissions handled in an electronic environment?
- What is the role of traditional publishers in an electronic environment?

- Who owns the library? Who manages it? Who decides on its content? Who controls access? Who pays for it?

Judi Moline, NIST

A) Work that's most relevant to the area of electronic libraries

> Standards for document creation, document presentation, and document publishing; prototypes using hypermedia; concepts such as open systems environments, electronic commerce, intellectual property rights.

B) The kinds of issues important for us to address

—What is the basic model of the electronic library:

- Distributed nodes acting as publishers of information
- Reference centers providing expert assistance and/or indices
  (Where will search be done?)
- Standard window environment for user interface

—What is a document? a database? an image? etc.

—How much of the "document" is important, i.e.

- Content
- Content + Generic Document Structure
- Image Structure
- Content + Image
- Content + Image + Generic Document Structure

—Pricing of information:

- As commodity?
- If based on projected use, how to deal with valuable information in small markets?
- Pay by use?