Source Book on Digital Libraries

Edward A. Fox

TR 93-35

Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061

December 6, 1993

Source Book on Digital Libraries

Version 1.0
December 6, 1993

Prepared for and Sponsored by the
National Science Foundation
Directorate for Computer and Information Science and Engineering (CISE)
Division of Information, Robotics and Intelligent Systems (IRIS)

Covering a series of
NSF Invitational Workshops
and Related Information

Edward A. Fox, Editor
(fox@vt.edu)

Preface

This report has its origins in discussions Mike Lesk and I had in October 1991, in connection with an NSF-sponsored workshop. With Mike McGill we drafted a White Paper, calling for a National Electronic Science, Engineering, and Technology Library. Since then, many things have changed. Now, the term "Digital Library" is more in vogue than "Electronic Library," though one might also call it a hypermedia archive or a data/information/knowledge base.

Regardless of the name, the intent is clear. Step one is for NSF to play a lead role in launching a concerted R&D program in the area. This is essential, so that any large scale development efforts by the Federal Government, along with state and private agencies, will be sensibly coordinated and undertaken in the most cost-effective and timely manner possible. We hope this initiative will begin in 1993, and proceed into the 21st Century. We feel that it must focus on large scale prototypes, and on setting the stage for a new, open, task-oriented, computer-assisted, type of access to information.

Step two involves partnerships, cooperative ventures, and production conversion of back-archives. ARPA, NASA, NIST, Library of Congress, NLM, NAL, and many other groups must become involved if we are to serve the broad base of users, from kindergarten to Nobel laureate researchers. It will only be successful if supported by top-quality research on information storage and retrieval, hypertext, document processing, human-computer interaction, scaling up of information systems, networking, multimedia systems, visualization, education, training, etc.

It is hoped that this Source Book will help show the potential, progress, proposals and planning efforts underway in relation to Digital Libraries --- leading to a large and effective R&D initiative Toward this aim, please send in corrections and additions so we can keep this book up-to-date..

Acknowledgements

Keith Tomlin provided typing, editing, formating, FrameMaker expertise, and general secretarial assistance in assembling and organizing this report into a usable form from the myriad pieces we began with.

Thanks also go to the Department of Computer Science at Virginia Tech for making software and resources available to help in this effort, and to the many students, faculty, and staff who assisted in a variety of ways.

Finally, thanks go to the numerous people who provided their time in writing the various pieces included, who attended the related NSF workshops, or gave pointers to interesting work. Special thanks go to Mike Lesk and Mike McGill who have worked with me since 1991 on the topic of digital library initiatives, and to Chris Borgman and David Levy for providing lots of helpful texts and files, as well as for their support of the workshops.

CHAPTER 1

Future Directions in Text Analysis, Retrieval and Understanding

METHODOLOGIES FOR INTELLIGENT TEXT PROCESSING AND INFORMATION RESOURCES MANAGEMENT: ACHIEVEMENTS AND OPPORTUNITIES

Report of the NSF Invitational Workshop on Future Directions in Text Analysis, Retrieval and Understanding

The Workshop was funded by NSF grant IRI-9114210 and co-sponsored by The University of Nebraska Center for Communication and Information Sciences.

Edited by

Jitender S. Deogun, Edward Fox, and Vijay V. Raghavan

Program Committee:

Abraham Bookstein

Christine Borgman

Jitender S. Deogun (Chair)

Susanne M. Humphrey
Alexa T. McCray
Vijay V. Raghavan
Clement Yu

Workshop Participants

Richard Belew, Univ. of California, San Diego, La Jolla, CA
Nicholas J. Belkin, Rutgers University, New Brunswick, NJ
Abraham Bookstein, University of Chicago, Chicago, IL
Peter Bollmann-Sdorra, Technical University of Berlin, Berlin, Germany
Christine Borgman, Univ. of California, Los Angeles, CA
W. Bruce Croft, Univ. of Massachusetts, Amherst, MA
Jitender S. Deogun, Univ. of Nebraska, Lincoln, NE
Edward A. Fox, Virginia Polytechnic Institute & State U., Blacksburg, VA
Mark E. Frisse, Washington University School of Medicine, St. Louis, MO
Ralph Grishman, New York U., New York, NY
Rick Hemphill, Naval Ocean Systems Center, San Diego, CA
Susanne M. Humphrey, National Library of Medicine, Bethesda, MD
Michael Lesk, Bellcore, Morristown, NJ
David Lewis, Univ. of Chicago, Chicago, IL
Alexa T. McCray, National Library of Medicine, Bethesda, MD
Michael McGill, Ameritech Information Systems, Dublin OH (Later U.Mich., Ann Arbor, MI)
Christopher C. Owens, Univ. of Chicago, Chicago, IL
Vijay V. Raghavan, U. of SW Louisiana, Lafayette, LA
Lisa F. Rau, GE Research & Development, Schenectady, NY
Gerard Salton, Cornell University, Ithaca, NY
Phil Smith, Ohio State University, Columbus, OH
Craig Stanfill, Thinking Machines Corp., Cambridge, MA
Tomek Strzalkowski, New York U., New York, NY
Beth Sundheim, Naval Ocean Systems Center, San Diego, CA
Kazem Taghva, Univ. of Nevada, Las Vegas, NV
Richard Tong, Advanced Decision Systems, Mountain View, CA
Irene Travis, The World Bank, Washington, D.C.
Carl Weir, Unisys, Paoli, PA
Michael Wong, Univ. of Regina, Regina, Canada
William A. Woods, Sun Microsystems Laboratories Inc., Billerica, MA
Clement Yu, Univ. of Illinois at Chicago, Chicago, IL

National Science Foundation Observer:
Maria Zemankova

The Report of the NSF invitational workshop on *Future Directions in Text Analysis, Retrieval and Understanding* is published in two volumes. Volume I, this document, summarizes discussions at the workshop and presents recommendations resulting from the workshop as well as related follow up activities. This volume consists of two parts. Part I presents the *Workshop Report* and Part II presents a proposal for a *National Electronic Library,* the ideas for which originated during the workshop. However, the details presented in this part were not actually discussed at the workshop, but were worked out via post-workshop meetings and discussions.

Volume II, a separate document, contains the collection of *Position Papers* prepared for and presented at the workshop. However, there are many places in Volume I where reference is made to papers that appear in Volume II. The two volumes are published as technical reports in the Department of Computer Science and Engineering at the University of Nebraska. Copies of both volumes can be obtained free of charge by ordering technical report number **UNL-CSE-93-006** for Volume I and technical report number **UNL-CSE-93-007** for Volume II.

Editor's Note: Part II, the proposal for a ***National Electronic Library*** *appears first here in Section A, followed by excerpts from Part I in Section B, the* ***Workshop Report****. Minor editing has enhanced the original versions that are available from University of Nebraska.*

Address for ordering Technical Reports:
Ms. Linda Endres
Department of Computer Science & Engineering
University of Nebraska,
Lincoln, NE 68588-0115

email: tech-reports@cse.unl.edu

Chapter 1, Section A: National Electronic Library

A National Electronic Science, Engineering, and Technology Library

Edited by

M. Lesk, E. Fox, and M. McGill

Executive Summary

The NREN is a new high-speed data communications network for the United States, and a key element of the High Performance Computing and Communications initiative HPCC). This initiative will bring significant benefits to the research and education communities of the U.S.; D. Allan Bromley, Director of the Office of Science and Technology Policy wrote that it "will lay the foundation for changes in education at all levels." In order to achieve these goals we must expand beyond the currently planned physical resources by adding to them a knowledge component -- providing in particular a basic scientific and engineering using high performance systems, extending applications of software algorithms, and supporting basic research and human resources training.

It is strongly recommended that the NSF should also support research on information and document management in the sciences and on the exploitation of electronic information resources. Research on scientific and technical information resources management would maximize value achieved from HPCC in general and this new on-line information resource in particular.

A prototype electronic library, accessible everywhere in the country, would protect our national leadership in the information industry, industrial competitiveness, and science and technology education. It is estimated that such a library could be created, and important related research funded, for a total of $50M over five years. As Vannevar Bush, the Director of the Office of Scientific Research and Development, wrote in 1945, "instruments are at hand which ... will give man access to and command over the inherited knowledge of the ages. The perfection of these pacific instruments should be the first objective of our scientists as they emerge from their war work." As the Cold War ends, perhaps we should think again about this job, whose time has certainly come.

Information Retrieval

The inspiration for desktop access to large information resources was described in 1945 by President Roosevelt's science advisor, Vannevar Bush. His proposal for the "Memex" machine was ingeniously designed around bar-coded microfilm. The development of digital computers changed our conception of such machines, and beginning in the 1950s and 1960s a new field emerged, called "information retrieval," a term invented by Calvin Mooers and popularized by Cyril Cleverdon. Computer typesetting of reference books made it possible to develop on-line systems as a by-product of conventional publishing. On-line systems now dominate some information use, for example, use of abstracts by professional literature searchers in libraries.

The information resources available include those from universities. For example, the University of California provides free access to its library book catalog through its MELVYL system, which handles 350K searches/week and is available on the Internet. Other resources come from governments, e.g., the National Library of Medicine's MEDLINE system of medical documents (with 70K searches made each week) or the catalog of publications from HMSO (Her Majesty's Stationery Office) in the United Kingdom. Many electronic resources now are privately developed and funded, such as the BRS/Saunders on-line full-text medical library.

The National Science Foundation played a key role in the development of this industry. Research funded in the 1960s by the Office of Science Information Services effectively started electronic information use. One of the largest projects was the computerization of Chemical Abstracts and the resulting development of CA files on-line and the STN (Science and Technology Network), with files relating to chemistry and other disciplines, which were funded originally by the NSF (although long since maintained, in this case, by funding from a professional society).

Other government agencies have been important to the development of the electronic information area. For example software developed by NIST (then called the National Bureau of Standards) was a model for commercial on-line retrieval systems. The National Technical Information Service (NTIS), the National Library of Medicine (NLM), National Aeronautics and Space Administration (NASA), the National Endowment for the Humanities (NEH), the National Agricultural Library (NAL), the Library of Congress (LC), and many other agencies have developed on-line information systems or machine-readable information resources. For example, none of the on-line library catalogs would have developed without the MARC tapes from the Library of Congress. These made possible a complete change in U. S.

library cataloging practice and stimulated organizations like the On-line Computer Library Center (OCLC) and the Research Libraries Group (RLG), and more recently the development of companies like Bibliofile. However, the NSF has a special responsibility to coordinate educational and human resources applications of the HPCC, and so logically should take the lead in any new science and technical information resources management programs using the NREN.

Information Technology Today

Until now the United States has led this industry. Most retrieval systems are based in the U.S., most technology was developed here, and the U.S. based industry here is larger than anywhere else. As we move further into the Information Age, the U.S.A. must retain its leadership in the information industries and must learn how to make better use of information resources.

Scientific and technical information resources management is critically tied to the HPCC. For example, parallel computers are now being used in advanced new text retrieval systems (such as the DowQuest service offered by Dow Jones that is implemented on two Connection Machines). New software algorithms for clustering and knowledge handling are part of modern retrieval systems. Networked and remote access is now typical, with many on-line catalogs, for example, available on the Internet both here and abroad. And, of course, neither basic research nor advanced education is possible without good information access, today increasingly meaning remote access to electronic databases.

Technology for Information Handling

The technology of information systems is diverse and changing rapidly. In the last thirty years the standard forms for distribution of machine-readable information have changed from paper tape, to punched cards, to magnetic tape, to optical disk (CD-ROM). Gigabyte stores (i.e., able to record 10^9 characters of text or other types of data) are affordable for workstations, and on-line terabyte stores (i.e., recording 10^{12} bytes) are commercially available. The information includes ASCII files, databases in various formats, images, video, sound, and combinations of all of these. The compactness of the data formats, in bytes per unit of weight, has increased by perhaps 300,000 fold since the days of punched cards.

Similarly, communications technology has greatly improved. Instead of the 110 baud teletypes of twenty years ago, we now have fibers carrying gigabits per second and experimental switches capable of the same speeds. The Arpanet of 1969 evolved into the NREN, and went from slow speed links to the present 45 Mbits/second and the future multi-Gbit backbone. Instead of only a few major research institutions, we now find tens of thousands of worldwide sites on the Internet.

Computational power has grown almost as fast. Even ten years ago a computer that performed one million operations per second was something to be proud of. There are now two manufacturers offering machines which promise one trillion operations per second. Supercomputers are routinely accessed over the Internet for remote operation. RISC architectures have made very fast workstations available for under $5000.

Fortunately, standards development has not slowed down this growth. The most important standards have been in static data coding: ASCII for character coding, and SGML for document representation. In interactive messaging, there is X.400 for email, X.500 for directories, and Z39.50 for formatting retrieval requests (at the heart of the Wide Area Information Servers network involving over 100 servers on the Internet). Z39.58 is the proposed Common Command Language to provide a consistent communication language for information systems and X.12 is the family of standards that provide for Electronic Data Interchange. For multimedia, two new standards will play an increasingly important role: the JPEG standard for gray scale and color images, and the MPEG standard for motion video and audio. All these standards are issued by ANSI (or its predecessors ASA and USASI) or by NISO or by international groups like ISO, IEC and CCITT.

A major change is coming from the need to include images and other forms of multimedia information. Image processing requires much larger storage, processing capacity, and communications speeds than does text handling. One page of text, for example, might contain 500 words or 3000 bytes.

A picture of this text (black and white, 300 dpi, Group IV FAX compressed) might be 50,000 bytes. If the 500 words were read aloud, which would take about 4 minutes, the corresponding sound signal, digitized and compressed (8KHz, 8 bits/sample) would contain 2,000,000 bytes. If the 4 minutes were stored as digital video, even with extreme compression to DS-1 (previously known as T-1) transmission rates (1.5 Mbit/second), more than 45,000,000 bytes would be needed. HDTV (high definition television) without compression is likely to require perhaps 45 Mbits/sec., another factor of 30 in information demand. Thus, the approximate expansion from text to image is 1:20; from text to speech is about 1:700, from text to compressed video is more than 1:15,000, and from text to HDTV 1:450,000.

The Information Industries

Electronic information access is an essential and rapidly growing industry. All "knowledge industries" (this includes all education) yield a third of the GNP; "information services," broadly defined (including, for example, consultants and advertising agencies) contribute directly about 8% of the GNP. On-line information services now generate directly $6-7B per year and are growing at 20% a year [32].

The U.S. information industry in 1991 is estimated at about $10B, growing at 13% per year; the industry in Europe is $3.4B, growing at 19%; and in Japan it is $0.8B, growing at 30%. The largest company in the business is Dun & Bradstreet, with over $1B in revenues; other U.S. companies with more than $100M in revenues are ADP, Dow Jones, CDC, TRW, Mead Data Central, Citicorp, McGraw-Hill, Knight-Ridder, Trans-Union Credit, Equifax, and Computer Sciences Corp. Although the largest parts of the industry at the moment are financial services, scientific-technical-medical information in the U.S. is now a $600M industry, and is growing faster: 19% compared to 13% in financial and 9% in credit [11].

The largest general bibliographic on-line vendor in this area, Dialog Information Services, had revenues of around $100M/yr in 1989. By contrast, King Research estimates that the average scientist spends the equivalent of $8000 per year (in time and expenses) getting and reading information in all forms; the total market is estimated at $21B per year. Carnegie-Mellon University personnel have projected that an electronic offprint service would attract 6 million users, and represent a $7B market over the next twenty years.

There are now about one million CD-ROM readers in the world, and the number is doubling every year. About $3B worth of CD-ROM disks have been sold. In 1996 it is estimated that the amount of money the U.S. spends on electronic information will exceed the amount spent on paper information; at present publishers only get 10-20% of their revenue from electronic products. Multi-media information, it is predicted, will be another $10B business in 1996. As an example of how quickly these industries can grow, the 900 number business, which did not exist a few years ago, is now $700M per year.

Today the United States dominates the large text database industry. Vendors such as Dialog Information Services, Mead Data Central, Chemical Abstracts Service, the On-line Computer Library Center, and others are used by customers throughout the world. These databases grew out of a tradition of bibliographic and textual searching. Originally, they were intended for, and are still largely used by, professional searchers. Now end-user access is growing, which greatly increases the importance of information systems.

The information industry is extremely diverse. It includes government agencies (the National Library of Medicine and the National Agricultural Library), nonprofit private sector organizations (OCLC and RLG), and for-profit private firms (such as Dialog and Mead Data Central). Revenues for even the largest such firms are in the range of $400M per year. None of these organizations, inside or outside the government, compare in budget with, for example, the $6B printing/publishing companies in Japan or the almost as large companies in Europe. (Recall Robert Maxwell, of U.K., who was acclaimed for his "empire" in this area, which included a multi-media company and CD-ROM pressing plants.) Although the U.S. has thrived on this diversity, the size of some foreign competitors makes our situation dangerous.

The Information Challenge

Although the U. S. leads in the current on-line information industry, the changes in the industry leave future dominance at risk.

Foreign Initiatives

The French, for example, are very active in image processing and storage. At the Centre Pompidou a collection of 150,000 images is to be made available via ISDN throughout France. The French national library is proposing to scan 1,000,000 books, dwarfing any project elsewhere in the world. The first large distributed scientific image bank was the Adonis project (CD-ROMs of biomedical journals), run by a consortium of publishers out of Amsterdam. Information handling is a business with high start-up costs and low incremental costs; it is not going to be easy for competitors to enter this business if other countries get too much of a head start. Action in the U.S. is required to maintain our position in the information industry.

Foreign enterprises are also moving strongly into the U.S. information business. Foreign publishers, recording companies, and motion picture studios are buying U.S. companies in these businesses. The four largest printers in the world, for example, are all foreign. BRS and Orbit (two U.S. on-line system vendors) now belong to Maxwell (or a British bankruptcy court). Consumer electronics companies such as Philips and Sony are investing heavily in new electronic products and dominate these markets. For example, although the phonograph and videotape recorders were invented in the U.S. and the various formats of traditional records and tapes developed here, the VHS player market was taken over by the Far East, the CD was developed and introduced outside the U.S., and future products of this sort (the Sony Data Discman) continue to be developed by foreign companies. Major electronic information programs such as Adonis (a consortium of biomedical publishers) and TULIP (an Elsevier program for electronic archives in the crucial area of material science and engineering) are being run by foreign groups. Even in traditional areas, the increasing prices of journals and books, and the drop in the value of the dollar, have made U.S. research libraries unable to keep up their subscriptions to foreign journals [31].

U.S. technology in the information areas is in danger of becoming obsolete. Just as we have fuzzier television than Europe because we introduced the service at an early stage of technology, we have old software and systems in the information area. We focus on text when other media are becoming important; we have old fashioned intermediary-tailored search systems when end-user systems are becoming popular; and we have not integrated searching systems with information delivery systems. We rely on low-speed dial-up for most commercial services as medium speed (ISDN) and high-speed (NREN) networks are being built. Even in retrieval technology, commercial systems rely on simple Boolean search many years after the research community has demonstrated the value of term weighting, relevance feedback, and other techniques.

Information Competitiveness

Perhaps even more important than the status of the information industry itself, however, are the consequences to the U.S. of not having access to current scientific, engineering, and other technical information. With global competition in the marketplace, rapid technology transfer is essential to allow new products and services to be produced first by American corporations, and for maturing products to benefit as quickly as possible from new advances. It is critical for national competitiveness that technical information be made available rapidly and to the widespread domestic community. Presently, there are many places where the best technology is not used, because of failures in either education of staff or in technology transfer. Computer programmers use old languages, scientists use old measuring devices, and engineers lack the latest tools. The steel, automotive, and consumer electronics industries are well known examples where technology transfer failures have allowed other nations to supersede U.S. corporations as world leaders.

Delays in getting the best information to all workers arise from many reasons. Publications and patents are often delayed, sometimes for three or more years; electronic processing could yield significant speed-up. Access to publications is often too hard and/or too expensive for individuals or libraries. Indeed, the entire scientific publication process has been called into question regarding finances and effectiveness. Education is viewed as something that happens in school, where people simply retain information after graduation.

Even then, many students are not aware of the latest developments, and rarely have time or encouragement to browse or see how findings in disparate disciplines are inter-related. Further, we need to improve the ability of workers in industry to retrieve new information and keep at the cutting edge of new technology, i.e., engage in "lifelong learning."[18]

Other countries recognize these issues. In Japan, for example, a great many scientific conferences are run for the purpose of technology transfer. Academic conferences in the United States are normally to present new results; many corporations don't pay for their staff to attend meetings unless they are giving a paper. By contrast, in Japan much less emphasis is placed on whether a conference describes new science, and more on whether many corporate employees are attending it. Often, major corporations in the U.S. recognize this problem, of course: IBM spends as much on training each year as the budget of Harvard.

But education in schools also needs help. Computer networks are being put into a great many colleges and universities, but all too often there is not very much in the way of information resources that is available. Generally an academic network mostly serves to transmit news, electronic mail and printer commands. An effort to provide technical information in electronic form will exploit networking to help students.

There have been many criticisms of the U.S. educational system. Our international competitiveness is slipping; and we are not able to make it all up with corporate libraries and training programs, which mostly exist only in large companies. We should be using electronic techniques to compensate for the inadequate resources of many public libraries, and to supplement what is now being done in college and university training. As students in K-12 (Kindergarten through 12th grade) gain access to computer networks at an increasingly rapid pace, education during those important years will also benefit from investment in making information available in electronic forms. Indeed, innovative efforts in Texas and other locations have begun to replace textbooks with information resources.

The capabilities being developed by the HPCC offer capacities to deal with these problems. Most obviously, the NREN would permit access throughout the country to advanced resources. In addition, the more powerful machines would allow greater flexibility and speed in answering user queries, and in extending information formats to new media. New software algorithms are needed to deal with the new media, particularly graphics, and to exploit better retrieval techniques. And, of course, human resources training is both the goal of this activity, and a supporting component to provide the people to do the work of building it. But the overall message is that we can not have an HPCC initiative that deals entirely with machines and networks; we can not build the systems without information. We will certainly use them to distribute information, and we definitely should use them in the most effective way to support science and technology in the United States.

Proposal for an Initiative in IR & DM Resources and Document Management

We recommend an initiative to build a basic science, engineering, and technology library on-line, available over the Internet/NREN, and for research regarding building and exploiting it. This would accelerate the development of the libraries of the next century, and improve the rate at which basic scientific knowledge is disseminated.

Our belief is that the library of the future will be based on electronic data. It will contain both text and graphics and be widely accessible via electronic networks. It is likely to be decentralized; even today, we are finding more and more libraries grouping into consortia to deal with the financial pressures of increasing prices for materials, and moving books to storage facilities in remote areas to deal with the pressures of increasing prices for land and buildings adjacent to today's facilities. As a result, it is already necessary for users to make use of collections which are in several places and owned by diverse managements; in the future, these problems can be alleviated by the ease with which information can be transmitted to the users' locations.

Which topic areas should be transformed into electronic form first? The Army, realizing the importance of optical digital information systems, has recently awarded contracts to scan over 100,000,000 pages of records; this may serve as a model for larger future DoD (Department of Defense) efforts of this kind. There are many non-military subjects to be treated. Also, we must support much more than just old records: the extension of electronic libraries, as soon as practicable, into all intellectual areas. Initially, however, we recommend scientific and engineering materials, for a variety of reasons:

- The users of such material are most likely to already possess computer workstations suitable for reading on-line, or to purchase such systems as the first generations of PCs are retired;
- This material has important impact on industry in the U.S.;
- Frequent updating, which electronics makes possible, is important in fast-moving areas such as science and engineering; and
- Science and engineering teaching is an area of U.S. education which has faced particular criticism, and in which a new initiative is thus appropriate.

We must use existing infrastructure to start with, wherever possible. The on-line resources of NLM in medicine, CAS in chemistry, UMI in dissertations, and OCLC and RLG in catalogs should be used as a base for whatever is built. Many major publishers are collecting their typesetting tapes; where possible, these should be used, with appropriate financial arrangements to safeguard authors, publishers and readers.

Similarly, existing technology must be exploited where possible. The most important resource is the NREN, and the U.S. base of installed computer workstations in corporations and colleges is also critical. New technology (cheaper disk drives, processors, parallel processors, and other areas) also is particularly important.

Collection Plan

We have the opportunity to make a major step forward in document handling, and to improve national scientific and educational productivity. This can be done by providing technical information in electronic form over the NREN. It will require putting together technologies in high performance computing, high performance networking, artificial intelligence, information retrieval, computer-human interaction, and other areas. The NSF should start an initiative to do it.

Why should we expect that electronics will help? Results with Bellcore's SuperBook and OWL's Guide program, for example, show that people using computers for text handling can get better answers, and find them faster than people using paper. Hypertext and hypermedia systems link related items to avoid the time-consuming labor involved in looking at related books or other media types. Over the last thirty years, research in information retrieval has shown how to improve systems to let people find information in text; and, in library after library, electronic search systems are replacing card catalogs and books of abstracts and indexes. The average U.S. scientist and engineer spends half time searching for and acquiring information; if, as in some SuperBook experiments, this could be done 25% faster with a gain in accuracy, the consequences for national engineering productivity would be immense. In addition, new modes of information access are possible only with an "active" electronic library.

A traditional library is passive; users come to it and search for what they want. In an electronic library it is possible to have pro-active elements, which send users messages when agents discover items that, based on earlier commands, it is known the users are looking for. In the electronic network of the future, we must compensate for the inability of the users to look around rooms of physical books. To facilitate this, we can build either asynchronously running programs which sit in the background, waiting for new information to arrive, as in SDI (selective dissemination of information) programs; or we can provide sophisticated "agent" programs, such as the "knowbots" described by Vinton Cerf, which are dispatched by users to search out and return desired documents. To truly serve a community of users these must learn about the needs, preferences, and capabilities of those users and consider that information during both search and presentation phases.

The central focus of the initiative should be the creation of on-line resources in a variety of scientific areas, which then could be used in new and exciting ways.

Recommendation:

The NSF should solicit proposals from groups of researchers to create on-line libraries in key scientific areas, in accordance with the following guidelines:

- Each group should explain how they will create an archive that will include materials for at least basic scientific activities, at the level of a completed undergraduate degree, in their area. They must agree to do this in a way that the end result will be compatible with the archives for other disciplines.

- Each group should involve participants who are (a) scientific users of information; (b) computer and information researchers; (c) publishers and other information creators and distributors.

- Each proposal should look to national, rather than local, availability of on-line information.

- Each group should address how their information will be used.

- In each case, the economic basis by which copyright permissions will be managed (whether a royalty, license, or other scheme) should be explained.

- The sum of all the proposals should provide reasonable coverage of basic science and engineering, without duplication of effort, and should be inclusive rather than exclusive of information sources (e.g., all of textbook, monograph, technical report and journal literature are candidates for the libraries).

We do not believe that these proposals should be constrained in terms of the technology to be used. Each group may propose the use of scanning, keying, OCR, re-use of typesetting data, or other technology. However, an effort must be made to maintain compatibility between different systems, regarding the representations used in the archives.

Since the intent is to provide discipline-oriented projects, it is likely that the larger publishers and other information creators may be involved in more than one library creation group. For this reason, the number of projects is being kept relatively small, to minimize the effort in coordinating the information conversion work.

Recommendation:

The NSF should create two oversight boards.

A. The principal investigators of each project should form a board to manage document standards, interchange principles, and other necessary agreements to produce a national science and engineering network rather than just local or discipline-specific networks.

B. A second board should consist of one representative from each of the scientific agencies in the U. S. Government concerned with these areas (e.g., the National Science Foundation, the Department of Energy, the National Aeronautics and Space Administration, DARPA, NIST, and so on). These representatives should ensure that the right information is available, and that shared funding can be arranged where appropriate.

These two boards will determine at planning and coordination conferences the "standards" for coordinated development of systems, ensuring that the results of the development and research activities are available in their entirety to the research and educational communities.

We anticipate that the sum of the necessary collections would amount to perhaps 100,000 volumes, or 50,000,000 pages. If this were converted to machine-readable form by scanning and OCR, the cost might approach $10M. Correcting errors, coordinating figures with text, marking up text and other elements, storing and allowing access, and other management costs might bring the total to $25M. That seems like a relatively small price to pay for creating a valuable national resource in science and technology. For example, if the existence of this collection permitted a 1% reduction in teaching time spent with undergraduates, there would be on the order of $100M (i.e., 1M students * $10,000 tuition/student X .01 reduction) saved in equivalent tuition costs. In fact, the real expenses (which should be borne largely by the users, not the NSF) are likely to be the royalty or license payments to the creators or publishers of the material.

With an archive developed, authors would prepare future articles and manuscripts in a compatible format, so minimal expense would be incurred in keeping the archive current. The National Institute of Standards and Technology might play a role in working with the two boards, to ensure that the archive follows standards, making it an ongoing project that should become self-sustaining.

Work on the applications of the collections should of course go hand in hand with the creation of the digital collections. These should include both ways of using the material in education and using it for technology transfer, as well as for retrieval. Educational applications of documents, and projects that merge documents, teaching, and student research are particularly important to get maximum value from our efforts.

Current Related Research

The basic library materials needed for an electronic library database can be collected today. The creation of large databases of text is technically feasible. In addition, recent advances in the technology of document handling are enabling construction of substantially more powerful systems for information access. Among the major areas in which progress has been made and continues to be made, and in which exploitation in applications offers great promise, are:

- Text analysis. Technology now exists to do simple natural language analysis and to try conceptual rather than purely word-based representation of knowledge, fully automatically. Local context analysis, detection of phrases, and assignment of concept knowledge is all becoming possible. These areas, after many years of effort, are now starting to be practical and must be pursued, since they could make it easier for users of an electronic archive to describe their information need, and could help increase the thoroughness or "recall" of searching.

- Theoretical modeling. We can now begin to predict the performance of retrieval systems, rather than just measuring them after the fact. Studies in user behavior are teaching us how to adapt computer systems to particular user tasks. Modeling should be tried to improve the use of systems and accelerate the design of new systems. Systems could then handle more users, more effectively.

- Technology. Disks and display screens are both getting larger, faster and cheaper. Audio and video systems are going to be common. Transportable systems will be coming into use, using wireless networks. Most important is the ability of parallel computers to implement search algorithms that laboratory tests show improve effectiveness of searching by over 30%, but were previously thought too expensive, e.g., the use of relevance feedback on the Connection Machine.

- Intelligent systems. On the basis of user studies, knowledge representation, and text analysis, it is possible to design rules for improving searches for users. In this way, we can extend the range of people who find computer information systems easy and fruitful to use, so that end-users rather than intermediaries do most of the searches.

- Indexing and structuring information. With large collections of information we can exploit the structure and context of individual works and their inter-relationships. Citation and subject searching have been shown to be much more effective when used together. Hypertext and hypermedia facilitate the associative access recommended by Vannevar Bush and bring together multimedia information resources for easy manipulation with pointing devices and multisensory presentation. Knowledge-based techniques for indexing permit more controlled and more powerful ways for manual organization of information resources.

- Graphical displays. Two-dimensional information representations are now being developed to give better overviews of a collection of documents than can be presented with a traditional shelf-list. Knowledge viewers for both documents and conceptual dictionaries are possible. Work in this area holds great promise for decreasing the difficulties many users find in dealing with retrieval systems, just as scientific visualization systems can help scientists work with their data and models.

Complementary Research

In order to make the best use of progress to-date in research, most of which has been limited because of practical considerations dealing with technology and funding, the collection building initiative should be extended. In addition to building the collections, we need research in their exploitation.

Recommendation:

The NSF should fund an expanded program of research into information utilization. This should include projects in

- multimedia storage and retrieval, including audio, video and image processing and compression techniques;

- combined text and data storage and retrieval, tied into the scientific databases initiatives;

- knowledge representation, including frame-based and other representations of information conveyed in documents, and including indexing, classification and retrieval as possible with higher-level knowledge representation that support logical or probabilistic inferencing;

- user studies, involving the nature of retrieval requests, user models, and other topics related to what people of all types, and with a myriad different sorts of needs, want from retrieval systems;

- human interface design, including hypertext, graphical and audio display methodologies both to deal with non-textual material and also to assist in text retrieval;

- evaluation techniques, including both traditional relevance assessments and research on new ways of measuring utility and dealing with uncertainty in system performance;

- distributed systems, including information sharing and the administrative and economic complexity of shared systems;

- coordination systems, in which people work together on information tasks;

- uncertainty and "fuzzy" systems, including systems where retrieval, document description and relevance may be defined in probabilistic terms;

- "agents" or "knowbots", which execute user tasks autonomously in a non-interactive environment;

- organization and retrieval (clustering, classifying), i.e., the grouping or linking of documents, rather than their retrieval from queries;

- natural language text analysis, including lexical, syntactic, semantic and pragmatic considerations, but focusing on how natural language processing can improve text retrieval;

- theoretical foundations of information retrieval, such as models of collection structure, retrieval techniques, and methods for predicting retrieval performance.

All of these research topics should be in collaboration with and in the context of the data collection projects. Similarly, the data collection work should be done so as to facilitate research and evaluation on the collections. Observational research and controlled experimentation on users are both necessary and encouraged.

The goals of this research include developing: algorithms and interfaces to serve directly the broad class of users in our nation, that make them want to explore the electronic library; databases that support integrated keyword, concept, natural language, citation, and visualization approaches to search and browsing; methods to adapt to a spectrum of requests, from those requesting a few of the very best passages to others calling for a ranked list of all related materials; techniques to prepare multimedia educational materials in a cost-effective fashion by drawing upon data in the archive; mechanisms to stimulate action by users, through providing current awareness or notification of new useful data or explanations; capabilities to clearly show the interconnections and inter-relationships between raw data, primary research, commentary, and secondary sources; approaches to solve the cost and space problems of current libraries and at the same time to replace the confusing and ineffective world of scientific publishing; and procedures to scientifically compare systems that vary in terms of functions, efficiency, and effectiveness of delivering the right results.

NSF Role

The National Science Foundation is the obvious organization to launch this initiative. It coordinates scientific knowledge creation, scientific information transmission, and the HPCC and thus the NREN. It is also a major funding organization for basic scientific research; and can coordinate with those other parts of the government that apply scientific knowledge in specific disciplines, such as the Departments of Energy and Defense. The NSF is also the organization which takes the leading role in setting scientific goals for the United States, even when such projects are funded by other routes.

We estimate that a reasonable scale for this initiative would be to devote $10M per year for each of 5 years, or $50M total; to be divided approximately equally between the creation of the library and the research on its exploitation. We hope that it will be possible to make this a cooperative initiative strongly supported by all of the NREN organizers.

The total initiative would be able to fund about 10 data creation projects, each costing approximately $2.5M, which should suffice to build the basic library; and $25M in research to use the electronic library for education, technology transfer, and other purposes. Thus, there could be approximately 4 major research centers receiving $1M for each of four years plus 12 additional smaller 3 year research projects at $250K per year, which would be able to exploit the data resources.

Our funding scale is based on the assumption that an electronic library would cover perhaps 25 key areas, each demanding a mixture of books and journals involving 2 million pages. Scanning the pages for one of these areas might cost $300,000 and with other tasks (OCR, staff time, editing typesetting tapes, mark up, storage, and so on) the total effort would require perhaps $1M. By distributing the work of archive creation, this initiative would benefit from competition at the proposal and project stages, with the net effect of having many groups knowledgeable regarding handling of electronic texts. We suggest that these areas be grouped into perhaps ten data conversion tasks, to avoid excessive start-up costs for small conversion projects.

The result of these ten electronic library construction (data conversion) tasks would be approximately 50 million pages of on-line information. This is not a huge library; it is only about 10% of that planned for the French national library; Dialog adds more than this much material each year. In addition to data creation, we suggest that for each data file several exploitation and research projects should be performed, to get the maximum value from the data; each of these projects should be at the scale of at least two full-time investigators.

With the cooperation of NIST, and the involvement of the boards and researchers in the various projects, as well as the professional associations, publishers, authors, and others who have been involved through this initiative, the resulting electronic library would have such an impact on standards of practice that it should become a self-perpetuating institution. New authors would prepare their works for inclusion, and publishing habits would be reshaped based on the existence of the library. In a decade, it would become the envy of the world, and permeate the worlds of science, engineering, and technology.

Implications for the Nation

Building an electronic science library will have impact on education, industry, research and the general state of science in the U.S. It will permit people throughout the U.S. to immediately answer scientific questions and keep abreast of new technology more rapidly and with less effort. It will permit U.S. scientists, whether in industry or academia, to deliver their results immediately to educational or commercial users. It will stimulate development of computer-based training systems, informal information circulation, and other new forms of information transfer. And it will provide a base on which the U.S. can maintain a leading position in the international information industry.

Educationally, we should try to make a reference library suitable for any undergraduate in the major scientific and engineering disciplines at any U.S. educational institution. This library should include textbooks, handbooks, monographs below the graduate level, and major journals in the areas. It should store several major lecture courses, in audio and in some cases video form; and provide search systems to retrieve based on full-text searching. We should also encourage access to this material in advanced secondary schools to accelerate the education of advanced students. We should fund research, in cooperation with the Department of Education, on the best ways to use this material in conjunction with conventional or non-conventional courses to improve student education. For example, teachers could use the library to rapidly select the best sources on each subject they cover, and use hypertext methods to synthesize a new integrated whole, rather than having to pick a single text.

Our goal should be to provide the equivalent of the best library at every school, combining the advantages of the resources at the biggest institutions with the environment and diversity of the many colleges and universities throughout the U.S.

In the information industry, we should become leaders in developing the new basic technology that will be most commonly used. We also should develop leadership in delivery of this information over fast networks, and in the use of audio and image material to supplement plain text. We should build on the NREN to establish an information industry that serves the needs of U.S. industry and education and does not become dependent on outside data sources. The structure of the industry may change, but it should be possible, in new technology as well as old, to have a leading and self-sufficient industry in the U.S.

In terms of national competitiveness, we should see to it that all locations of technological industry are connected to the NREN and thus to the national electronic library. Each industry should be able to get current scientific and technical information; and should, of course, also benefit from the better education given to its employees. What is most important is that we routinely apply the best technical information to each problem. One possibility is to harness methods of computer assisted problem solving and visualization, employing them for educational purposes in the large. Electronic communication and electronic reference works should provide the best way to deliver up-to-date scientific solutions to real problems, especially with the young scientists of American trained in these techniques during their university years.

Research in information delivery will be facilitated by such a data library and network. Once people become accustomed to getting answers in seconds rather than hours or days, once they become accustomed to doing their own searches, they will get their information from the best sources rather than from those easiest to find. Researchers will use a library and network to experiment with new techniques related to data, audio, image and text handling and retrieval.

We must stress the urgency of building such a system now. If we start soon, we can effectively provide a standard platform and format for electronic information archives. This will encourage other publishers to join in and make all national resources available in a compatible way. If we do nothing, various small incompatible efforts will be started, as in the early days of railroads when, for example, Ohio had its own track gauge. This is likely to hold up the development of the entire area. Furthermore, if foreign countries take an early lead in this area, it may not make sense for any U.S. databases to try to compete, and we will be dependent on other nations for our technical information.

The overall economic impact of an electronic library would be enormous. It is not possible to state precisely what would happen, any more than one could have predicted thirty years ago the present rise of companies like Mead Data Central or Dialog. However, imagine:

- that a million college students studying science, and representing perhaps \$10,000 of tuition each, improve their learning by only one percent; that would save \$100M per year.

- that a million engineers in industry, each representing costs of \$100,000 per year, are able to work 1% more effectively - saving only two and a half days a year of information-seeking work; that would save \$1B per year.

- that ten million workers in manufacturing, producing \$100,000 per year each, are able to produce 1% more by employing more advanced technology; this would save \$10B per year.

The United States must develop such a basic technological library; the implications of not doing so imply an inability to effectively compete in the information and knowledge areas.

In summary, we recommend that the NSF establish a national initiative to build a basic scientific and technological library for availability on the NREN. We also recommend that the NSF provide for research in the exploitation of this library. This would support U.S. industry, education, and research by providing rapid access to current information throughout our country.

Chapter 1, Section B: Workshop Report

Workshop Report

Edited by

Jitender S. Deogun, Edward Fox, and Vijay V. Raghavan

Executive Summary

In recent years, the Division of Information, Robotics and Intelligent systems (IRIS) at the National Science Foundation (NSF) has sponsored a number of workshops to identify future directions for research in a number of disciplines within Computer and Information Sciences. The Workshop on Future Directions in Text Analysis, Retrieval and Understanding was supported by the Database & Expert Systems Program of the NSF in the same spirit and was held on October 10-11, 1991 in Chicago just before the 14th International Conference on Research and Development in Information Retrieval. The focus of the Chicago workshop was on documents and text, user models and interfaces, and intelligent document management systems. The impact of these research activities, in a 5-10 year period, will be the development of document management systems and tools to: serve a broad class of applications and user types; support advanced approaches to searching and browsing; facilitate the generation of multimedia educational products; solve the cost and space problems of conventional libraries; and scientifically compare document management systems that vary in terms of functions, efficiency, and effectiveness.

Introduction

On October 10-11, 1991, just prior to the 14th International Conference on Research and Development in Information Retrieval, 30 researchers representing academia, industry and government met in Chicago for an NSF sponsored workshop dealing with the broad areas of information retrieval, document analysis, text understanding, and intelligent document management systems. This workshop was unique in its focus on information as opposed to data, on supporting humans' fundamental needs to organize document databases and find documents in terms of their information content, and on the use of language processing and knowledge representation techniques to help in these processes. This report summarizes the discussion, highlights the research issues identified, and makes recommendations regarding the NSF's future involvement in the area of Information Resources and Document Management.

The remainder of the workshop report (i.e., Part I) is divided into two sections and four appendices. In Section 2 we describe the main research areas that are relevant to national needs and thus are recommended as candidates for sponsored research. In Section 3 specific recommendations and concluding remarks are presented. Appendix A.1 provides an overview of related NSF sponsored workshops and Appendix A.2 describes the relationships of this report's recommendations to the HPCC effort and the NREN. Appendix A.3 summarizes the accomplishments of the last two decades in Information Retrieval and related disciplines and Appendix A.4 identifies emerging technologies that are relevant to the information industry.

Further Background

In 1945, Vannevar Bush, President Roosevelt's science advisor, who played a key role in the development of the National Science Foundation, wrote of the need to develop and use tools to manage mankind's accumulated knowledge. [7] Although Bush's 'memex' idea didn't turn out to be of much practical value, his vision inspired many researchers in the areas of information retrieval, information science, and hypertext. As a result, in part, of his insight, the first text retrieval systems were developed in the 1950s.

The NSF was given a broad directive in its enabling act of 1950 "to foster the interchange of scientific information among scientists." The Office of Science Information Service (OSIS) at the NSF was responsible for providing research funding toward improved methods and techniques for handling scientific information, including mechanized systems, through the Documentation Research Program. The activities of the program was expanded considerably around 1956 under the directorship of Mrs. Helen Brownson. The National Defense Education Act of 1958 directed the foundation to undertake projects of a fundamental or general nature that may produce new insights, knowledge, or techniques applicable to information systems and services. [6] The projects supported at this time were featured in the semiannual reports on *Current Research and Development in Scientific Documentation.* [30]

During those early days, a number of conferences were sponsored by the NSF to chart a plan for research activities in the field of scientific documentation. National publicity prior to and following a conference on *Coordination of scientific and technical information relating to federal R&D activities* (Cleveland, February 3-4, 1958) attracted the attention of Senator Hubert H. Humphrey, who then arranged for the subcommittee of Senate Committee on Government Operations that he chaired to hold hearings on this topic. [44] An NSF sponsored conference, held in Washington, D.C. (October 2-3, 1964), was intended to encourage research on *Evaluation of Document Searching Systems and Procedures.* Active research programs at the Chemical Abstracts Service and the American Society of Physics were supported in large part by the NSF. OSIS supported research activities in information retrieval not only in the U.S., but on a worldwide basis. For example, the funds for the famous Aslib-Cranfield project was provided by the NSF to Aslib in England and, in turn, Cleverdon received the funds to execute the project. Agencies such as the Ford Foundation as well as the American Society of Metals initiated and supported many studies of library problems. [26] There is a general consensus among those doing research on scientific documentation at that time that "the 1960s were really a good time for IR research." In the 1970s, the significance of the area was further emphasized at the NSF by the establishment of, for example, the Division for Information Science and Technology. In the current structure at the NSF, IR and related areas come under the Database and Expert System Program (Director: Dr. M. Zemankova), which is located within the IRIS Division of the Directorate for Computer & Information Science & Engineering.

That research led to the growth of a vast industry, where the current systems support searching for relevant items in large collections, and where the items may correspond to library catalog entries, bibliographic records, citations with abstracts and/or keywords, or full-text documents. In addition to searching, users can browse the documents or indexes. Some classes of retrieval systems, such as those for Selective Dissemination of Information (SDI), route new documents to those users who have filed a profile that defines their information needs. These alert systems send notifications or documents to interested parties, upon the arrival of news or messages. Yet another class of retrieval systems is based on employing the, already mature, Database Management System (DBMS) technology. In these cases, a two step approach may be appropriate, where a *fact extraction system* can fill a database automatically from textual descriptions (such as from a patient record in a hospital) and then a database of these extracted facts is organized for retrieval via a DBMS. Any of the above mentioned classes of retrieval systems may be used either directly by users, or in connection with embedded applications. Important domains of applicability include: legal, industrial and academic research, libraries, consumer services, news, publishing, financial, marketing, and military logistics and intelligence.

Clearly, the support of the NSF, industries and professional societies, from the 1960s through the 1980s, provided the impetus for the development of a multi-billion dollar information industry, which was founded on early research findings. More recently, as the result of keen competition for limited research dollars as well as changes in the responsibility of the NSF vis-a-vis Federal Science & Technology Information activities, it appears that the support for basic research in this area is not at the levels that it previously was. This workshop identifies new research directions and proposes that the NSF supports a bold new research initiative in the area of managing information resources and document databases and pave the way for a National Electronic Library of the 21st Century.

Impact of the Proposed Research

In keeping with these challenges and opportunities that the information industry will face in the 1990s and beyond, we propose a broad program of research be initiated that will, in the next 5 to 10 years, lead to the development of:

- query languages and interfaces to directly serve the broad class of users in our nation, thus making them want to explore various information resources including a National Electronic Library;

- methods to adapt Information Resources and Document Management Systems (IRDMS) to a spectrum of requests, from those wanting a few of the most relevant passages to others calling for a ranked list of all related text materials;

- alert systems that stimulate action by users, through providing current awareness or notification of new and useful text or extracted facts that have just become available;

- IRDMSs that support integrated keyword, concept, natural language, citation, and visualization approaches to search and browsing;

- capabilities to clearly show the interconnections and inter-relationships between raw data, primary research, commentary, and secondary sources, as well as other types of metadata;

- electronic approaches to solve the cost and space problems of current libraries, by bringing together heterogeneous sources of information, and for streamlining the confusing and ineffective world of scientific publishing;

- procedures to scientifically compare IRDMSs that vary in terms of functions, efficiency, and effectiveness; and

- techniques to prepare multimedia educational materials in a cost-effective fashion by drawing upon information in document databases.

Research Agenda

Given the importance of the need for having access to information in our research, education, technology transfer, and other endeavors, it is crucial that a coherent plan emerge for NSF funding of basic research in this area. We call for a significant increase in the level of support for basic research efforts, believing it essential if we are to succeed in the goals of the High Performance Computing and Communications (HPCC) initiative, remain competitive in the world marketplace and improve our educational system to where it becomes more effective and affordable.

To ensure that sponsored research is relevant to national needs, we call for the new funding to be in the context of a Nationwide Electronic Library in Science, Engineering and Technology. In particular, we recommend research in the following areas.

Retrieval Theories and Models

As in any scientific discipline, new theories and models emerge to accommodate new observations and approaches. Thus, formal theories and models are required that address the following key issues and related questions:

- IR is fundamentally an interactive process, where a user's information need or the knowledge state prompts some type of dialog. What are suitable ways to model the processes of browsing and query formulation?

- The process is iterative, where the computer and human provide each other feedback of various types. What methods of learning should be employed in order to improve system performance relative to future interactions?

- Uncertainty is a fundamental issue, which must relate the degree of clarity of an information need, the expressive power of the (formal) query language, weights used in representations of queries and/or documents, ambiguity and relevance. Fuzzy set theory, probability theory, and other approaches, which facilitate the handling of possibly conflicting evidence, have been applied and each seem to have some validity. How can one choose one of these alternatives for a specific problem, or how can they be improved or combined with each other?

- A number of types of entities are involved, including documents, terms, and concepts. In certain applications there are others, such as geographic entities, graphical illustrations, historical dates, and time-lines. How can systems cope with such a variety of entities?

- Documents are in reality two dimensional objects, although in many systems they may be handled as a stream of characters. Thus, in decisions made about relevance, an understanding must be developed of the role of structure and context of document components, both physical (e.g., a component's coordinate position in the bit map image of a page) and logical (chapters, sections, section headings, etc.). For example, how can documents that have sections and subsections at various levels be marked in order to make explicit the semantic connotations of the position of headings?

- Techniques for matching and determining similarity between texts or portions of texts are of particular interest. An understanding of the distinctions between retrieval of sets and ranking of results should be further developed. Given the long standing experience of the retrieval community with such matching problems, it is of interest to extend these to "emerging" applications. For example, how can the key concepts used for text matching apply to important problems of genome mapping and chemical structure matching? What is required to assess the ability of text matching methods in "Functioning as intellectual amplifiers to detect similarities and differences in molecules whose size and complexity are too vast for the unaided human mind ..." [Grand Challenges, p. 40]?

- IR models should be developed to predict the relevance, relative to some user need, of newly acquired documents and the behavior of users, depending on their educational background, level of experience, etc. In particular, they should lead to accurate predictions in new domains. IR system models may involve concepts from object-oriented databases or file structures (e.g., inverted files, signature files), and must be applicable for efficient real-time implementations and when scaling up to large environments. Performance issues relate, and depend on data and file structures, as well as searching and clustering algorithms. How can current models be integrated and generalized to carry-over, from information retrieval to more general problems of document management?

Structures of Documents and Text

What is meant by a document, and which view (form) of document we adopt, are at the core of this field. How can we formalize these structural aspects?

- Instead of paper or sound waves or light, a digital representation of documents and their components is used. What are the effects of this conversion?

- Given documents that have multiple media types, with links and with timing and sequencing and synchronization, how can they be indexed? What attributes or features should be considered?

- Documents include data, maps, tables, images, chemical symbols, equations, especially when part of knowledge repositories, in general, and scientific databases, in particular. How do these relate, logically as well as physically? How can we improve the "development of scientific databases; and distribution of integrated system building kits and toolsets."?[Grand Challenges, p. 21]

- Document structure may be arbitrarily complex, though regularities are captured with SGML, HyTime, and other markup schemes. How can structure be used to advantage? How can passage retrieval be defined and improved in quality? How can automatic document analysis methods identify logical structure, and use that to help improve subject indexing?

- Issues relating to the relationships, groupings, or links that may be provided between the types and instances of the entities are critical. For example, a user may express a preference among retrieved documents, a thesaurus indicates various types of semantic relationships between concepts, concept networks frequently include domain specific relation types, a classification system will impose committee defined groupings, indexing adds associations between documents and the indexing vocabulary, citation data imposes a variety of link types based on scholarly studies, statistical dependencies may be estimated to quantify the degree of association, and clustering will impose some grouping or organization based on feature sets. In hypertext and hypermedia, unlike the case in text and business databases, the meaning of links is not unique. What kinds of theories can aid with the modeling of links in hyperbases, where not only multiple media, but also multiple link types, are involved? Further, what are the trade-offs between links specified in a database schema versus those whose existence is only known at the time of navigation?

User Models and Interfaces

Since IRDMSs are a type of interactive system, what results can we apply from the area of human-computer interaction to improve these systems and their performance? Key issues in this regard include:

- How do we relate the user's world (of people, places, events, and things) to the worlds of information objects, in particular: terms, concepts, classification systems, and documents?

- How can we design and understand various types of systems, where some operations occur autonomously (e.g., automatic indexing of documents), and others are fundamentally interactive (e.g., browsing)? How do those parts of systems relate?

- What are the benefits of graphical displays? How far can we push fast browsing of documents (including multimedia) by using advanced presentation methods? To what extent can graphical interfaces compensate for difficulties that seem inherent in one-dimensional query formulation approaches?

- How can we perform task analysis to most efficiently gather data needed for the system design? How can we really specify, design, and develop user-centered interfaces and systems? How do we deal with individualized handling of user needs, communication of those needs over time, and the complexity of information seeking behaviors? What approaches are needed to model problem solving that takes place over long periods of time including a sequence of sessions, where multiple related or un-related problems may be confronted in an interleaved fashion? How do we integrate users' access of reference works, passages, and other document components?

- What types of user models can be built and maintained with sufficient reliability to facilitate system tailoring or adaptation? ([9]: pp 75-79) What are the relative importances of facilities for browsing and exploring vs. searching, recognition of words vs. recalling them, and moving about in logical hierarchies? How can our understanding of users be enhanced through an analysis of the use of help and navigation assistance aids?

- What extensions and generalizations must be made when multiple humans and multiple computers all work together, for cooperative problem solving and information interchange? How can we better support: 'knowledge transfer exchange programs at national laboratories, centers, universities, and industry; and software dissemination through national databases and libraries."? [Grand Challenges, p. 21]

- Given these interactive systems, what are the best evaluation methods? What aspects and measures of the process should be recorded?

Intelligent Document Management Systems

Since human intermediaries often provide effective access to information and experts may provide service par excellence, won't approaches for building intelligent systems lead to better information access? Key aspects of this problem follow:

- What basic features should we identify about documents? What statistical knowledge is most important and why? Can we integrate statistical knowledge, the results of natural language processing, and knowledge of a domain and its terminology?

- Depending on whether our goal is finding facts, extracting important features, or summarizing content, what methods are most effective? How can the results be best used for common tasks within document management systems?

- How useful is document understanding to support information retrieval requirements? How accurate, complete and deep should parsing be? What type and level of discourse analysis is of value? Does in-depth text analysis help most if applied to queries, to documents, or to both? How can these be done in a robust way for heterogeneous document collections?

- What have we learned from natural language processing, in English as well as multilingual environments, to help understand documents? How does morphology relate to truncation or syntax relate to phrase identification? What characteristics of corpora are important in determining statistical and linguistic insights? What can be learned from machine-readable dictionaries?

- What media and knowledge representation methods are best suited to various types of retrieval needs? What are the roles of data encoding, frames, semantic networks, rules, neural nets, etc.? How do these support inference and concept-based retrieval? What methods for combining evidence are best suited to each of these contexts?

- Given the vast quantities of documents, how can they be best used to help with automatic knowledge acquisition?

- How can we fully develop the idea of agents or knowbots, so they accurately represent human interests and needs? What knowledge about users, in terms of their information seeking behavior and plans, are needed? What kinds of computer-to-computer protocols are needed to ensure proper, cooperative functioning of agents?

Conclusion and Recommendations

To tie together all of the above research problems, to relate them to important needs, to allow study of very large systems, and to solve many of the needs of our society, we recommend implementation of the plan proposed in the white paper submitted to the NSF for a Nationwide Electronic Library in Science, Engineering and Technology (see Part II). In the following we summarize the key benefits of such a library as well as the proposed initiative to fund basic research in the area of information resources and document database management.

- Researchers will have ready access to at least one very large information system, with at least 100 gigabytes of text, and many gigabytes of multimedia data, to study usability and performance issues, and other topics related to scale.

- Research progress will be speeded up, since researchers will have ready access to repositories (or test beds) for software, text resources, interfaces, and retrieval engines. In particular, re-use of existing techniques and knowledge will be easier.

- With large scale information resources and systems, the true potential of recent technological breakthroughs, such as special-purpose hardware architectures and storage systems, can be studied. As computers move to gigaops and then teraops, scalable parallel designs can be tested and techniques to apply them can be investigated.

- Distributed systems, at the national, state, local, and personal system levels, are being built. Many important issues arising in this context should be considered, regarding effectiveness, mix of local and remote storage options, interoperability, sharing and cooperation, applications to education, and use of agents.

- Human resource and behavioral issues regarding learning should be studied, as the "information engineer" profession emerges and as "knowledge workers" increase in number, in the context of libraries, publishing, research, education, and technology transfer.

References

[1] S.K. Bhatia, J.S. Deogun and V.V. Raghavan, Automatic Rule-Base Generation for User-oriented Information Retrieval, *Proc. of the Fifth International Symposium on Methodologies for Intelligent Systems*, Knoxville, TN., pp 118-125, Oct. 1990.

[2] P. Bollmann, V. V. Raghavan, G.-S. Jung, L.-C. Shu, On Probabilistic Notions of Precision as a Function of Recall, *IP&M*, v 28, no 3, pp 291-315, 1992.

[3] C. Borgman, D. Case, C. Meadow, The Design and Evaluation of A Front-End User Interface for Energy Researchers, *JASIS*, v 40, no 2, pp 99-109, Mar. 1989.

[4] Y. Breitbart, H. Garcia-Molina, W. Litwin, N. Roussopoulos M. Rusinkiewicz, G. Thompson and G. Wiederhold (Eds.), Multidatabases and Semantic Interoperability, Technical Report No. 188-91, Dept. of Computer Science, University of Kentucky, Lexington, KY.

[5] M.W. Bright, A.R. Hurson, S.H. Pakzad, A taxonomy and current issues in multidatabase systems, *IEEE Computer*, v 25, pp 50-60, 1992.

[6] H. L. Brownson, Research on Handling Scientific Information, *Science*, v 132, no 3444, pp 1922- 1931, 30 December 1960.

[7] V. Bush, As We May Think, *Atlantic Monthly*, v 176, pp 101-108, July 1945.

[8] W.B. Croft, Clustering large files of documents using the single-link method, *JASIS*, v 28, pp 341-344, 1977.

[9] J.S. Deogun, ed., Working Proceedings, *Workshop on Future Directions in Text Analysis, Retrieval and Understanding*, Chicago, IL., Oct. 1991, University of Nebraska-Lincoln Press, 1991, 160 pages, (forms Part III of this report).

[10] D.E. Egan, J.R. Remde, T.K.Landauer, C.C. Lochbaum, L.M. Gomez, Acquiring information in books and SuperBooks, *Machine Mediated Learning*, v 3, pp 259-277, 1990.

[11] M. Fischer, J. Fedder, The electronic information industry and forecast: in Europe, North America, and Japan, *On-line Information 90*, pp 515-525, 1990.

[12] R. Fidel, Towards Expert Systems for The Selection of Search Keys, *JASIS*, v 37, no 1, pp 37-44, 1986.

[13] E.A. Fox, G. Nunn, W. Lee, Coefficients for Combining Concept Classes in a Collection, *Proc. 11th Int'l Conf. on Research and Development in Information Retrieval*, Grenoble, France, pp 291-307, June 1988.

[14] E.A. Fox, Advances in Interactive Digital Multimedia Systems, *IEEE Computer*, v 24, no 10, pp 9-21, Oct. 1991.

[15] J.C. French, A.K. Jones and J.L. Pfaltz (Eds.), Scientific Database Management, Technical Report 90-21, Dept. of Computer Science, University of Virginia, Charlottesville, VA.

[16] J.C. French, A.K. Jones and J. L. Pfaltz (Eds.) Scientific Database Management -Panel Reports, Technical Report 90-22, Dept. of Computer Science, University of Virginia, Charlottesville, VA.

[17] G.W. Furnas, T.K. Landauer, L.M. Gomez, S.T. Dumais, The vocabulary problem in human-system communication, *CACM*, v 30, no 11, pp 964-971, 1987.

[18] W. Gardner, The Electronic Archive: Scientific Publishing for the 1990s, *Psychological Science*, v 1, n 6, pp 333-341, Nov. 1990.

[19] R.H. Giering, This is data central, Mead Data Central Inc., Dayton, Ohio, 1972, Report DTN-72-2.

[20] P. Hayes, S. Weinstein, CONSTRUE/TIS: a system for content-based indexing of a database of news stories, *Second Annual Conference on Innovative Applications of Artificial Intelligence*, 1990.

[21] IBM, Stairs VS -- A tool for the end user, IBM Scientific and Cross Industry Center, Uithoorn, Netherlands, 1978.

[22] C. Johnson, DARPA Eyes smart text sifters, *Electronic Engineering Times*, pp 35-39, February 1992.

[23] M. Lesk, The CORE Electronic Chemistry Library, *Proc. of the 1991 ACM SIGIR Conference*, Chicago, IL., Oct. 1991, ACM Press, pp 93-112, 1991.

[24] M. Liberman, Text on Tap: the ACL Data Collection Initiative, *Proc. of a DARPA Workshop on Speech and Natural Language*, Oct. 1989, Morgan Kaufmann, San Mateo, CA., pp 173-188, 1989 (ISBN 1-55860-112-0).

[25] R.S. Marcus, J.F. Reintjes, A Translating Computer Interface for End-User Operation of Heterogeneous Retrieval Systems; Part I: Design, Part II: Evaluations, *JASIS*, v 32, no 4, pp 286-303, pp 304-317, July 1981.

[26] B.P. McCune, R.M. Tong, J.S. Dean, D.G. Shapiro, RUBRIC: A system for rule-based information retrieval, *IEEE Trans. Software Engineering*, v SE-11, no 9, pp 939-945, 1985.

[27] E.J. McKinin, M.E. Sievert, E.D. Johnson, J.A. Mitchell, The Medline/full-text research project, *JASIS*, v 42, no 4, pp 297-307, 1991.

[28] I.A. McLeod, R.G. Crawford, Document retrieval as a database application, *Information Technology, Research and Development*, v 2, no 1, pp 43-60, 1983.

[29] J. Nielsen, *Hypertext & Hypermedia*, Academic Press, Inc., 1990, Boston, MA.

[30] NSF, *Current Research and Development in Scientific Documentation*, Rept. No. 6, Washington, D.C., 1960.

[31] A. Okerson, K. Stubbs, The Library "Doomsday Machine", *Publishers Weekly*, pp 36-37, February 8, 1991.

[32] A. Prozes, The Electronic Information Age, *Business Quarterly*, v 55, pp 80-84, 1990.

[33] S.E. Robertson, K. Sparck-Jones, Relevance weighting of search terms, *JASIS*, v 27, pp 129-146, 1976.

[34] G. Salton, *The SMART Retrieval System -- Experiments in Automatic Document Processing*, Prentice Hall Inc., Englewood Cliffs, NJ, 1971.

[35] G. Salton, A blueprint for automatic indexing, *ACM SIGIR Forum*, v 16, pp 22-38, Fall 1981.

[36] G. Salton, Developments in Automatic Information Retrieval, *Science*, v 253, pp 974-980, August 1991.

[37] G. Salton, E. Fox, H. Wu, Extended Boolean Information Retrieval, *CACM*, v 26, pp 1022-1036, 1983.

[38] T. Saracevic, P. Kantor, A. Chamis, D. Trivison. A Study of Information Seeking and Retrieving. I. Background and Methodology, *JASIS*, v 39, no 3, pp 161-176, May 1988.

[39] H.-J. Schek, P. Pistor, Data structures for an integrated database management and information retrieval system, *Proc. VLDB Conference*, Mexico City, pp 197-207, 1982.

[40] A. Silberschatz, M. Stonebraker and J.D. Ullman (Eds.), Database Systems: Achievements and Opportunities, Technical Report TR-90-22, Dept. of Computer Sciences, The University of Texas at Austin, Austin, TX.

[41] A.F. Smeaton, Information Retrieval and Natural Language Processing, in: *Prospects for Intelligent Retrieval, Informatics 10* (Ed. K.P. Jones), pp 1-14, 1990, Aslib, London.

[42] C. Stanfill, R. Thau, Information retrieval on the Connection Machine, *IP&M*, v 27, no 4, pp 285-310, 1991.

[43] H. Turtle, W.B. Croft, Evaluation of an Inference Network-Based Retrieval Model, *ACM Trans. on Inf. Systems*, v 9, no 3, pp 187-222, July 1991.

[44] U.S. Congress Senate, *Documentation, Indexing and Retrieval of Scientific Information*, Senate Committee on Government Operations, Senate Doc. 113: 1960.

[45] C.J. Van Rijsbergen, Foundations of evaluation, *Journal of Documentation*, v 30, pp 365-373, 1974.

[46] E.M. Voorhees, *The Effectiveness and Efficiency of Agglomerative Hierarchic Clustering in Document Retrieval*, Ph. D. Dissertation, Cornell University, Ithaca, NY, 1986.

[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?

CHAPTER 3

December 1992 Workshop

Chapter 3, Section A: List of Participants

Workshop on Digital Libraries
Xerox Palo Alto Research Center
December 9-10, 1992

Stanley Besen, Charles River Associates
Christine Borgman, UCLA
Esther Dyson, EDventure Holdings Inc.
Monica Ertel, Apple Library of Tomorrow
Georgia Finnigan, The Information Store, Inc.
Linda Gass, Adobe Systems Inc.
Tony Hall, Dialog Information Services
Pam Jajko, Syntex USA Corporate Library
Michael Lesk, Bellcore
David Levy, Xerox Palo Alto Research Center

Mark Liberman, University of Pennsylvania
Richard Lucier, University of California at San Francisco
Jean Mayhew, United Technologies Research Genter
Peter Menell, Stanford School of Law (1992-93)
James Michalko, The Research Libraries Group, Inc.
Eugenie Prime, HP Labs Research Library
Frank Urbanowski, MIT Press
Gio Wiederhold, DARPA/SISTO
Terry Winograd, Interval Research (1992-93), Stanford University

Xerox/PARC Participants

William Crocca
Margaret Graham
Giuliana Lavendel
Larry Masinter
Geoffrey Nunberg

NSF Participants

Su-shing Chen
Jane Caviness
Y.T. Chien
Larry Rosenberg
Maria Zemankova

Chapter 3, Section B: Agenda

Workshop on Digital Libraries — Xerox Palo Alto Research Center

Wednesday, December 9, 1992

8:30 Arrival (Main reception area-upstairs)

9:00 - 11:00 Introductory Remarks:

Mark Weiser (PARC welcome and overview)
YT Chien & David Levy (Background, Goals, Agenda)
Mike Lesk (The competitiveness issue)
Jane Caviness (NREN & Gigabit Testbeds)
Gio Wiederhold (A perspective from DARPA)

11:00 - 12:30 Around the Table - All participants (Vision statements)
12:30 - 2:00 Lunch
2:00 - 4:30 Group Breakout Session 1
4:30 - 5:30 Report Back from groups & General Discussion

Thursday, December 10, 1992

8:30 - 11:00 Group Breakout session 2
11:00 - 12:30 Report back & general discussion
12:30 - 3:00 Lunch and Summary
3:00 - 5:00 Demos (selected PARC projects)

Chapter 3, Section C: Lesk Summary

The Digital Library: What is it? Why should it be here?

Michael Lesk

The second in a series of NSF workshops on a digital science library took place in Palo Alto, Dec. 9 and 10, 1992. As time goes on, this project is becoming ever more vague; not only did we not agree during the workshop on what should be in the library, but many attendees did not even think this was important. We also reached no resolution on the problems of economic compensation for publishers and authors (not to mention intellectual property lawyers), but we did notice that there is enough non-protected material available to possibly start building a useful collection without addressing that issue. Nor did we settle why such a library would benefit the U. S. more than other countries if it were built here. However, we did see considerable agreement that a digital library would be a Good Thing.

Mark Weiser welcomed us to Xerox PARC with a description of Xerox work in information storage and access, social aspects of electronic information handling, and related areas such as linguistics, ubiquitous computing, and scanning/printing hardware. Xerox has, for example, done a project at Cornell University in which a thousand 19th century books were scanned and reprinted.

Y. T. Chien then explained the structure of his part of NSF and asked for input from the community about the importance of digital libraries and their relation to other technologies. Other government speakers were Jane Caviness who talked about the NREN, and Gio Wiederhold who described some of the information-related work in DARPA.

I described the risk of foreign countries, such as the French using Minitel and their proposals for the Bibliotheque Francais, dominating the next generation of information systems. I also described the original proposal ($50M over five years, half to convert a basic undergraduate science library of perhaps 100,000 books, and half to fund exploitation of this resource). Finally in the introductory speakers, Stan Besen of Charles River explained the increased risks that digital technology poses to intellectual property rights by facilitating piracy, private copying, and the creation of derivative works.

As each person gave their personal statement, we heard a sequence of people emphasizing the importance of social concerns, of bringing information to people rather than just piling it up, and of intellectual property issues. What was noteable was the lack of any fear that technology was unable to do the job; our questions are entirely to deal with context and politics. An interesting question raised was that this library, as envisaged, was in some ways an alternative to bookstores as well as to libraries, but we really did not follow up this point.

We then divided into several groups for the next day. These groups were intended to address more specific issues, with the thought of sketching out a possible report urging that the United States build a digital library, to see if it could be made convincing and if there was any agreement on its content.

One group was supposed to define the library. It came back with a statement that a digital library is a "distributed technology environment which dramatically reduces barriers to the creation, dissemination, manipulation, storage, integration and reuse of information by individuals and groups." They suggested the national initiative should contain some specific testbed projects, but gave no guidance on what these should be. In other words, they not only failed to define the collection, they didn't really even describe the system that would hold it.

The next group was asked what technologies were needed to build such a library: it rounded up the usual suspects, recognizing the particular importance among them of the NREN, OCR and conversion tools, software for filtering, retrieving, summarizing and so on, and user interface research. Security/privacy concerns were also considered important, as were some standardization issues.

The last group on the first day pointed to the conflict between intellectual property rights and free access, as well as the "political economy" problems (e.g., university tenure procedures today are intertwined with publication issues).

On the second day, we then re-formed into a new set of groups to deal with the most important question: why should the U. S. have a digital library? We hoped to sort the answers into three categories: research applications, business applications, and educational applications. I was disappointed by many of the answers, which suffered from circularity (e.g., a well educated work force is good for business), from a confusion of means and ends (e.g., a digital library provides access to multiple formats of information), and from a lack of quantification. Among the possible useful arguments that came out of this process:

Research:

1. One-third of a researcher's time is spent looking for information, and searching is more efficient in electronic systems;

2. Electronic libraries would let researchers get at existing data reserves which they have trouble finding now, and speed up their work;

3. Rapid identification of related work would reduce duplication;

4. The entry barriers to starting new research projects would be lowered.

Education:

5. An electronic library would help self-education and support distance learning.

6. The United States is falling behind in science education and must do something (although the countries ahead of us don't have digital libraries either).

7. Electronics can help libraries deal with increasing volumes of information.

Business:

8. Readily available information levels the playing field as between small and large business, and business in small and large towns.

9. Superconductivity and xerography were both found by random browsing, and we need more of that.

10. Better access to regulatory requirements, product literature, standards and the like would help business.

Social:

11. Rapid access to commercial resources in time of national emergency. For example, during the Gulf War it was necessary to identify quickly the manufacturers of the transistors that went into IFF devices.

The most convincing arguments, to my mind, were:

a) Biotechnology is a very productive area and makes enormous use of on-line information – if we could duplicate that productivity in other areas we would be way ahead.

b) Allegedly, studies in Blacksburg, Va. and in Pennsylvania show a positive correlation of information availability and economic growth.

c) Don Swanson (Univ. of Chicago) has demonstrated the ability of purely information investigation for finding medically promising ideas.

But all these benefits were a bit fuzzy, and we have few quantitative measurements as backup for these ideas.

The workshop then closed with final statements by Y. T. Chien and David Levy, and some follow-up discussion.

There was no really good answer to the question: why should the U. S. do this? Why will it benefit our country specifically? We all agree it would be best if such libraries were built internationally, but we also think that the NREN gives the U.S. an advantage in distributing this information within our country, and that we would not be better off letting other countries build such a system. If so, we will find ourselves paying for access to information, much of which was generated within the U.S.

What should be done at the next workshop? Important topics for the discussion at that workshop, I think, should include;

a) Specific consideration of possible compensation schemes for publishers: whether the concurrent-use site licenses now common in the software business, the unlimited use licenses favored by librarians, an adaptation or an equivalent of the Copyright Clearance Center, greater use of page charges, or something completely different (e.g., pledge breaks on the Internet).

b) Specific consideration of the goals list, and a resulting suggestion for content for at least one or two specific projects, including both form and subject matter. We need to know what kinds of material are to be in the library, both in terms of form and content.

c) Specific discussion of why this should be a United States funded activity; would the benefits of having such a library accrue primarily to us if we built it? Our advantages include the NREN and the creation of much of the research, plus our wide number of small businesses.

d) Coordination with other groups having similar discussions. For example, on the same days as our meeting there was another one in Washington with similar goals, also run by NSF, but that one focussed on the social sciences. As part of this, we should continue our efforts to include the publisher and academic library communities in our discussions.

e) And, on the side, perhaps NSF should require that each citation in a grant proposal should include the country of publication and the price of the publication to libraries, and reviewers should be instructed to ignore all of the proposers' publications which either appeared in foreign journals or in journals which charge libraries more than $1000/yr.

On balance, to get a report that would convince Congress to fund something,

a) we need to say what it is,

b) we need to explain why it is good for the United States,

c) we need to show that its economic benefits will exceed its costs, and

d) we need to do all this in simple words, with illustrative graphics, convincing testimonials, and respectable evidence as backup.

In particular, anyone reading this who can provide citations to studies showing that research, education, or business benefit from more information and from more electronic information in particular, is asked to communicate with the author.

Chapter 3, Section D: Statements of Participants

Questions and Assignments for Workshop Participants
by Y.T. Chien

During the workshop, we plan to look at digital libraries from three perspectives, from the perspectives of the users, the technology, and the industry. Below we've compiled a list of questions that can be asked from each of these perspectives (as well as some questions that don't naturally fall into any of the categories).

To get you thinking about the issues, and to help us organize the workshop's breakout sessions, we'd like to ask you to choose one of these topics (or more, if you wish) and prepare answers to some or all the questions associated with it. We're only looking for a one-page statement, but you should feel free to write more.

Questions:

1. The users view

a. In a few words or sentences, what is your vision of a digital library, say 10 to 15 years from now?

b. Given that vision, what are the grand challenge applications of a digital library?

(A Grand Challenge application can be roughly defined, in this context, as an application of digital libraries, which when fully implemented may have a significant impact on the scientific or economic well-being of our society.)

c. Who might use electronic libraries? What information and services do they need? What kind of support, training, and pricing policy is needed to foster more effective use?

d. What is the use of data from foreign sources? How is such use related to international competitiveness?

2. The technology view

a. What research is needed to realize a large-scale digital library? For example, research in computer science, linguistics, sociology/anthropology (work practice studies), business/economics, philosophy/history, etc.

b. Beyond high speed networks, what other infrastructure elements are needed to provide the foundation for digital libraries?

c. What component technologies are essential to the development of a digital library?

d. How do we characterize the functional capabilities of a digital library environment?

3. The Business View

a. What products or services might industry provide for use with or in digital libraries? What types of products or services might be important for your company to consider developing in the future?

b. What services provided by digital libraries would be particularly useful to industry? How would they enhance competitiveness, time-to-market, etc.?

c. What are the non-technical barriers to the realization of digital libraries in the foreseeable future, e.g., copyrights, standards, and other social factors?

d. What kinds of partnerships (e.g., research, training, experimentation) must be developed between industry, academia, and government to realize digital libraries on a national scale?

4. Miscellaneous

a. How is a national digital libraries effort related to other major initiatives, e.g., high performance computing and communications (HPCC), manufacturing, bio-technology, etc.?

b. What should NSF do with respect to development of a technical and social infra-structure linking researchers with each other and with consumers of research?

Christine Borgman, UCLA

NSF WORKSHOP ON DIGITAL LIBRARIES
Xerox PARC, 8-9 December, 1992

Response to "Questions and Assignments for Workshop Participants"

Of these 4 categories (Users, technology, business views, miscellaneous), my interests and contributions are primarily in the Users's View category. The discussion document I drafted in response to the July 1992 NSF Workshop on a National Electronic Library Initiative is relevant to this perspective, and thus is attached below. The paper also comments on other areas, particularly 2b (technology infrastructure elements), 2d (functional capabilities), 3b (business products or services), 4b (misc - technical and social infrastructure). Also, in relation to 4b (relation between the digital libraries effort and others such as the HPCC), Appendix A.2 (Relation to HPCC) of the July 18, 1992 Workshop report (referenced below) draws specific links among the HPCC and digital libraries efforts and the Grand Challenges. This also is a useful discussion document for the December meeting.

The Information Infrastructure Technology Act (also known as Gore II; distributed to December workshop participants) provides a useful definition and scope statement for "digital libraries" in Section 7, including 7 points for research and development. The role of NSF and other agencies is presented explicitly in that document. The scope appears to include all the major areas listed in my document and others previously generated in this workshop series. The Gore II document seems a valuable starting point for setting a specific funding and development strategy.

While this meeting should not be bound by these prior discussion documents, they are thoughtful pieces and we might make efficient use of our limited time by using them as a point of departure.

William T. Crocca, Xerox

1a. The digital library will be 1/3 equal partners along with publishers and retailers in the distribution and management of intellectual assets. Its role will be navigation, archiving and republishing as well as the traditional roles of access and storage. Distinctions will be made between content and delivery media, removing most artifacts from conservation labs and leaving only those artifacts which are intrinsically precious. Library economics will change. Since major institution holdings have economic value, they will have opportunity to create substantial revenue streams. Also, the shelving costs could come down, depending on collection management strategies.

1b. An important, near term "grand challenge" is indexing and navigation. These are unsolved problems which, when solved, should enable a dramatic opening up of the information base intrinsic to library collections. (I was originally going to use the phrase "knowledge base", but information has to be found, read and internalized before it becomes knowledge . . .)

1c. Potentially, anyone with either a TV set or a telephone. Furthermore, one would expect simple, low-cost printing systems (FAX, photo-CD, etc.) to be ubiquitous, thus relocating hardcopy to a different point in the creation - distribution - consumption chain.

1d. This is a large and complicated question even ignoring the competitiveness issue. However, my initial reaction is that the "competitiveness" issue can easily become divisive and should be proscribed from the libraries' role. Let information ownership and national practices deal with this one. The substantive issue seems to be around language and cultural context. This will effect navigation, understanding and presentation.

2a. Navigation and indexing, linguistics and translation, and indeed everything you mention. The question is what of these is special to the library? I'm not sure any of it is -- just the way it all comes together.

2b. High density, high reliability, VERY long term storage. This implies hardware, media and format standards. It also implies internal and mutual consistency across collections w/r/t format and indexing standards. We also need some sort of copyright/royalty/ownership collective in order to allow information to be marketed and used with at least as much facility as the technology provides. Finally, we need cross use of the various electronic common carriers (telephone, TV, etc.) to enable the kind of connectivity this vision will require.

2c. Compression/decompression chips at HD (High Definition) video rates with high resolution and LONG scan lines for large documents. Cheap, high quality display devices (both transient, e.g., CRT and hardcopy e.g., paper). A cultural mandate for real information disclosure, else arcane security systems will become necessary.

2d. . . . not sure I really understand this question

3a. Certainly the hardware, the software and the common carrier services. In addition, one can easily imagine other services such as production capturing, alternative delivery media manufacturing and other production operations a library might not want to manage directly -- especially those, like capture, which are one-time transformations.

3b. Document delivery and research. Identification of others with similar interests (this could also prove to be a liability). Need to brainstorm this one.

3c. The relationships between publishers, sellers and users. Revised skill sets and the necessary storage. New kinds of user terminal capability that is cheap (probably this will happen naturally), TV/cable/phone/etc. channels appropriately organized. The methods of use: the European approach to use it and then pay (in a standard way) for your use is more tractable than ours for the kind of future I am envisioning.

3d. I'd say similar in principle to the one between Xerox and Cornell.

4a. It offers a low-cost, high-performance information distribution infrastructure which is now only partially provided by specialty publishers.

4b. The government information resources could be organized into the first national electronic library. As such, it could experiment with the legal issues which are bound to arise with this kind of effort. The international questions could be explored by partnering with the Library of Congress efforts to provide information to former iron-curtain country government officials. This would be an interesting testing ground for the technologies and a strong signal to our private and quasi-public institutions of expectations for their future participation.

Esther Dyson, EDventure Holdings Inc.

A few answers, and more questions

What is the digital library? That term smacks of "filmed play," "horseless carriage," and the like. The digital library will be less like a library than we think, and more like itself. It won't have a physical location, nor will it really be a logical place. It will sit behind or around everything (electronic) we do. You will be able to use it while you're writing a letter, composing a presentation or analyzing a sudden sales drop in your Hungarian subsidiary; a query tool/knowledge navigator/agent savant will be just one more item on your menu (pick-an-icon or voice input) of choices.

In fact, we will perceive the digital library as an access method to information distributed all around the world. For content, it will contain elements of today's libraries -- books, printed periodicals and reference materials -- as well as videos and other more exotic media. But it will also include realtime access to designated experts (for a fee, at designated times, etc.) and other information sources such as government and private statistical and text bases (with appropriate privacy and property protections).

The main challenge of creating the electronic library is not to build it per se, but to ensure that people make their information (1) digital and (2) reachable. In the business community, that will require an overlay of allocation, accounting and payment mechanisms which should be mostly unobtrusive (although you may well want people to know how much things cost so that they don't spend recklessly).

The most interesting challenge will be to (build technology to) automatically cross-reference all this information, "understand" it well enough to find relevant items and avoid redundancy, and provide useful interfaces to make the content more intelligible. Some examples include hypertext maps such as gIBIS-CM/1, "reading" tools such as those forthcoming from (yes) Oracle, and overall layout and presentation tools.

Users will filter for each other, and good filterers will probably become professionals (like editors or active newsgroup managers today). Who will set the agenda? Today, editors/publishers and advertisers do; in the future, the reader will have more power. Should we leave filtering to imperfect software tools (as opposed to imperfect people)? If information is available all the time everywhere, what's news? What happened yesterday, last hour, last minute? Anything you didn't know yet?

All this of course assumes the requisite infrastructure:

- high-speed bandwidth,
- funding (preferably from competing private sources working hard to make their modules of the library better in a never-ending cycle of improvements),
- excellent software tools,
- ubiquitous information-delivery devices (PCs, personal digital assistants, video screens, information kiosks, etc.),
- appropriate financial underpinnings to manage use and compensation of owners/producers of information,
- appropriate legal and technical safeguards to guarantee authenticity and integrity of information and to protect privacy where appropriate. (Do you want people to know what you read or wanted to know? What are the rules governing anonymous comments? Responsibility for libel or incorrect information?)

Linda Gass, Adobe Systems, Inc.

1. The User's view:

a. In a few words or sentences, what is your vision of a digital library, say 10-15 years from now?

The digital library in the year 2005 will be available directly to end users from their personal computer (or whatever the device may be called) and will come with its own "digital librarian" - an intelligent agent that can assist you in finding whatever information you are trying to track down. You will be able to do unlimited searching - not only by author, subject, title but searches will include the content of the work. Searches will be intelligently filtered, and content will be visually rich and multi-media based with hypermedia links. Users will be able to get multiple views of the information: graphs, photos, etc., in addition to textual information. Users will also be able to access information from a variety of sources through a unified user interface and the details of the infrastructure will be hidden from the user.

b. Given that vision, what are the grand challenge applications of a digital library?

Scientific: A digital library will make it easier for scientists to work outside of academia and pursue hypotheses that are not in vogue and thus not funded or approved for research. This could lead to important medical and scientific breakthroughs.

Economic: A digital library will facilitate the work of the telecommuter, allowing people to work at what they please and live where they choose. This can help alleviate regional economic recessions. The library will also encourage small business startups as people discover that they can easily acquire the data they need to project the success of their business ideas. It will provide larger businesses with better competitive information, leading to less duplication and more innovation.

Social and Political: Impact beyond the scientific and economic should also be considered. For instance, access to digital information through the Freedom of Information Act can help insure that our environments are safe and our political choices are informed. The library can be an important safeguard of democracy.

c. Who might use electronic libraries? What information and services do they need? What kind of support, training, and pricing policy is needed to foster more effective use? Any current library user would use an electronic library: schools, universities, businesses. The key is to draw in a larger user base by making the library more accessible and more appealing than our existing libraries. The information and services needed are the digital librarian mentioned above, excellent searching capabilities, good response time. Support and training will present interesting problems as we will have to change the way people work. Even though computers are widespread in our work environment, they are still not widely used by most businesses in America today. Support will consist of some automated support tools and then there will also be an expert librarian who can assist users in using the system to its full effectiveness. Pricing policy is a tricky issue. Today's public libraries are free and we will exclude a significant portion of our population by charging for access to a digital library. Perhaps some sort of sliding scale is appropriate where users pay per view. The price per view can vary depending upon the amount of information, how current it is, etc.

d. What is the use of data from foreign sources? How is such use related to international competitiveness? This set of questions is actually rhetorical: We ARE in a global marketplace and access to foreign information can only enhance international competitiveness. A key issue is that of translation; much of the foreign information is published in the native language of the country and is of less utility to the average American user.

2. The technology view

a. What research is needed to realize a large-scale digital library? For example, research in computer science, linguistics, engineering sociology/anthropology (work practice studies), business/economics, philosophy/history, etc. The topic of effectively searching information needs research. This spans linguistics and computer science. The end result of a search could be overwhelming for a user unless intelligent relevance ranking techniques are used. Timeliness of search and retrieval on massive amounts of information is also important and we have a great deal to learn about how one can search multi-media (voice, images and video). Another area of research involves the sociological point of view: changing people's work habits. We have had 500 years of people using books and paper and now we are asking them to work in a new way: using electronic information; good user interface design will be critical in encouraging people to change the way they work. Another interesting area of research involves what happens to serendipity. If searching becomes too effective, one loses the accidental discoveries that are made by browsing. How can we preserve this feature? Moving towards electronic information also changes the business/economic structure for information. We are accustomed to paying for "widgets" - books, newspapers, etc. How will our economic structure evolve to account for electronic information? How will our legal system (copyright, piracy, privacy, etc.) evolve to handle the new issues that come with this electronic representation?

b. Beyond high speed networks, what other infrastructure elements are needed to provide the foundation for digital libraries?

- Standards for data interchange
- Visual fidelity of information
- Mass storage
- Security and accounting

c. What component technologies are essential to the development of a digital library?

- Natural language processing

- Filtering
- Better mass storage than CD-ROM

d. How do we characterize the functional capabilities of a digital library environment?
(Answered in question a. under the User's view)

3. The Business View

a. What products or services might industry provide for use with or in digital libraries? What types of products or services might be important for your company to consider developing in the future? The industry at large could provide things such as personalized newspapers, intelligent agents for processing information. Other potential products/services might be "seals of approval" by experts so that people could have some way to assess the quality and authenticity of information. Industry can provide much of the library information (the major publishers, etc.) in electronic form. Industry needs to provide tools to convert information to electronic form. The kinds of products Adobe considers important to develop are technologies that allow people to communicate electronically, especially as it relates to documents. We are working on a truly portable document format and platform independent viewing technology known as the Adobe Acrobat family of products. We are also working on technology that allows people to transform their existing base of documents, either in PostScript or paper form, into the portable document format (Acrobat Distiller and OCR technology).

b. What services provided by digital libraries would be particularly useful to industry? How would they enhance competitiveness, time-to-market, etc.? Timeliness of information and the ability to effectively search information such as new research findings, trends analysis, and general economic information will enhance competitiveness by providing industry with the tools they need to understand the market and the economic climate better. I cannot comment how an electronic library could benefit the general business as I'm only familiar with a software business but issues we need to be concerned with when creating a product span technological development, legal, marketing and sales activities. Being able to search through research papers on algorithm development may save development time. Being able to search patent and trademark databases can save legal time. Marketing and sales could be assisted by being able to search for reports/articles on customers who may have bought similar products.

c. What are the non-technical barriers to the realization of digital libraries in the foreseeable future, e.g., copyrights, standards, and other social factors? Yes, copyrights, standards, social factors such as changing work habits. Also cost.

What kinds of partnerships (e.g., research, training, experimentation) must be developed between industry, academia and government to realize digital libraries on a national scale? Anyone have any good ideas here? The best I could come up with is:

Although the Internet was sponsored by the DoD and developed by academia, I view it as one of the great successes in research. Can we use this model to put together a digital library? Can we prototype something and get it installed at various test sites to turn it into what it really needs to become?

4. Miscellaneous

a. How is a national digital libraries effort related to other major initiatives, e.g., high performance computing and communications (HPCC), manufacturing, biotechnology, etc?

Described above in section on Grand Challenge Applications.

b. What should NSF do wrt development of a technical and social infrastructure linking researchers with each other and with consumers of research?

Do it like the Internet. Start building something and try it out. One could start small by picking one area of research and implement the vision that we come up with. Depending upon the nature of the research, it may be interesting to make it available at the high school level to help put some energy back into our disintegrating educational system.

Vicky Reich, Stanford University

The ideal digital library will incorporate all the positive aspects of print-based libraries and will overcome all the barriers of the print form. Specifically:

Print libraries make it easy to:

1. find some information with little or no instruction
2. self-instruct through signs, brochures, etc.
3. locate new information serendipitously
4. browse
5. get help from a person
6. tailor the environment to specific communities and needs, so that different libraries provide different levels of information and access to different clienteles - but each different type of library has overlapping organizational constructs so the user has some familiarity as s/he moves from library to library.
7. provide a space for social interaction (play, learning, research)
8. offer inexpensive or free access at point of use for most resources
9. retain the principal and practice of fair use
10. provide organization and access to resources
11. be a cultural, scientific and technical public archive

Digital libraries will offer these advantages over print libraries:

1. materials are always available (not lost, stolen, defaced)
2. easy access to all formats, overcomes elitism of multimedia and includes all new media
3. use of resources is not physically limited
4. materials do not deteriorate with time or use/ replacement formatting is cheap, easy and keeps pace with hardware developments
5. materials are hyperlinked, and what is now "background" information (i.e., supporting research data, computer programs) is brought to the fore
6. "cataloging" and "indexing" provide rich and dynamic access - the problems of authority control (insuring that all the works of the same author are retrieved with one search command) have been overcome.

Jean Mayhew, United Technologies Research Center

DIGITAL LIBRARY

1. Users View

The digital library will be a concept from the viewpoint of the consumer or user to describe various forms of electronic information access. The electronic reader tucked in the pocket or briefcase, the self-contained unit in the home or office, and the system accessable halfway around the world will all fall into this concept.

Challenges exist in the evolution of the information culture and how it is regulated. The development of the culture will depend not just on the introduction of technology but also on a much deeper understanding of how the consumer uses information. Current laws and regulations do not reflect where the consumer needs to be in the use of information and are becoming the pacing item in the acceptance of technology into this culture.

Digital libraries will be used by the general populace, often without even realizing it. The newspaper delivered to the home will be an offshoot of the library. The scope of the information available will be limited only by the development of the culture. Services will be valued which sift through the digital library and quickly match user needs.

Training will be non-existent for products with the broadest distribution. Increasingly training will be built into the more sophisticated technologies.

Pricing will be dependent on the perceived value at the level of access. Retrieval of simple directory information will result in minimal cost. Simultaneous machine translation of foreign language documents may be relatively expensive.

Data from foreign sources is critical for entities competing in a global economy. Unless national security is an issue, multinational corporations require freedom to share information across borders.

Geoffrey Nunberg, Xerox PARC

The Digital Library: A Xerox PARC Perspective

Whatever form the digital library finally takes, its success will be measured by its ability to support current practices, media, and forms, while making provision for the new ones that are likely to emerge. And it must serve current constituencies (libraries, researchers, students) while opening itself to new users and communities, whose needs may not have been addressed by traditional libraries. From this simple principle -- the need to accommodate the old and the new -- some more specific principles follow:

1. First, the design of the digital library should be based on a rich understanding of current practice in specific domains -- libraries, research institutions, schools, hospitals, as the case may be. Some of this understanding will come from questionnaires, interviews, and experimental investigations of information-retrieval tasks. But this phase of the design process should also include detailed ethnographic observation of the production and use of documents. The object of this kind of investigation is not to build specific models of work-flow into the design, but rather to understand how the use of technology is accommodated to the work process. And these studies shouldn't be restricted to the narrow use of the technology to accomplish particular goals, but should also look at how people instruct each other in the use of documents and technology, and at how particular practices -- (e.g., looking up a reference) are embedded in the larger context of work activities (e.g., writing a paper, treating a patient). In the end, that is, we want to know how access to information affects the overall operation of organizations and the productivity of work practice. In this connection, it is also important that any digital library project be integrated with network facilities that allow users to exchange information both about the use of the system and about other sources of relevant information.

2. The design should be sufficiently flexible to accommodate the new practices and new document types that are likely to emerge over the coming years, partly as a result of changes in the technology of information management, and partly as a result of independent changes in the nature of work and communication. In addition, we should recognize that the digital library will reach users and communities who currently are not aggressive users of library services, and whose interests and needs may not be identical to those of current users. For these reasons, it would be a serious error to predicate the design of the project on a single model of the work process or on the practices of current users alone, or to presuppose a limited repertory of document types and genres.

3. The design should support all types of document use and all document media. In particular, long-term experience suggests that paper documents will continue to play an important role in all aspects of communicative practice, and that all components of the system should be designed with a view to permitting users to move back and forth between paper and electronic representations as their interests require. In particular, this means that it is important to plan for the creation and distribution of hardcopy documents. In addition, it is unlikely that a digital library will wholly replace traditional print documents, or that new institutions will wholly replace the role of traditional libraries and scientific publishers. Thus the digital library will be most effective if it is considered as one of a battery of forms and institutions that serve scientific communication, rather than as an omnibus substitute for them.

4. The digital library should support a range of search and browsing procedures. To this end, we require not just greatly improved content-based search tools (e.g., intelligent information access, automatic summarizers, translation tools), but also search tools and interfaces that support browsing of document databases and individual documents in ways that are comparably effective as the informal procedures that users now employ in dealing with paper documents.

Gio Wiederhold, DARPA and ACM Publication Board

I am proposing a rapid turn-around experiment to address one specific aspect of Electronic Publication: Referred, Archival Material of a quality that can be used for objectives as academic promotions, recordation of definite hypotheses proofs, etc.

Much of the pressure to publish and increase the number and size of journals is related to publication pressure. The information needs of the public might be as well served by publication on conference proceedings, etc. and requesting full reports in report form from the author's home institution. If we do not address this aspect of publication pressure we fail even with the best technology.

My specific plan is to gain experience with electronic journals by a rapid prototyping approach using predominantly existing resources, in various phases, each justified by success and analysis of the preceding phase:

The first phase establishes electronic dissemination of referred backlogged papers within ACM:

1. Only include completely referred and accepted, but publication backlogged from TODS, likely TOPLAS, and other journals interested in cooperating.
2. Papers to be removed when published in paper form in journals, to remove any nervousness due to competition with the paper world.
3. Papers so available could be listed in CACM, Member Net, the appropriate journal, and an Internet/Usenet listing.
4. To access a paper a single-use authorization code is to be obtained via some number as 900-ACM-TODS, with the charges collected by the telephone company to become income for ACM.
5. Charges are to be set equal to the copyright fee now published.
6. The single use code enables Internet FTP retrieval of the requested paper from an ACM controlled archival site.

7. Papers so distributed will be headed as 'Printed for the exclusive use of John Smith', to somewhat inhibit excessive re-distribution, i.e., electronic xeroxing.
8. Monitoring of usage and feedback preparatory to a possible phase 2.

A second phase would let us print the abstracts in the journal as soon as possible, with a reference to the electronic access mechanism.

A third phase could make lengthy appendices, corollary proofs, tables etc. available only in electronic form, but now persistently available.

A fourth phase could avoid publication of full papers, when the expected audience is limited. This would require cooperation of editors and authors.

In all phases high-quality refereeing must continue.

Terry Winograd, Interval Research and Stanford University

2. The Technology view

The term "digital library" is both suggestive and misleading. Misleading because it evokes the image of the traditional library - a storehouse of objects to be examined or checked out by "readers". As pointed out in some of the background papers, the digital medium will inevitably do away with this notion, which is grounded in the physical contingencies of print media. I will take the long term view here, without addressing the obvious questions about the intermediate waypoints and transformations that will get us from here to there.

We can look at an overall chain of functions in the traditional system as:

author publisher librarian reader

It is the intent of the author to produce materials that will be of interest to some readers, and of the readers to do with those materials what they will. These endpoints are the fixed part of the chain. The middle involves functions that have historically combined physical, commercial and informational roles:

PUBLISHER

Physical: Produce printed, bound objects

Commercial: Market the objects, advertise, collect money, pay royalties,...

Informational: Decide what should be published and what should be rejected. Provide editorial advice and assistance to authors.

LIBRARIAN

Physical: Maintain a physical collection, lay it out in an organized way, make it available to readers

Commercial: Use resources from some (academic or governmental) entity to purchase items and maintain the environmennt.

Informational: Decide what should be stocked, catalog it, provide organization and individualized information services.

When we move to the digital world, the physical functions become unnecessary. It is trivially easy for works to be stored by authors and provided on demand to readers, with no interventions of the kinds required by print. The commercial functions are still necessary if authors are to be remunerated for their work (a fact that is too easily overlooked if we take the model of current internet usage, where most of the proferred material is public domain or provided gratis). The exact form of this in the digital world has yet to be worked out. What is the equivalent of a library paying for one copy of a book (at an inflated price over the consumer version) and then letting many people read it? A site license? Is software a useful model?

The informational roles are still useful: quality control, editorial assistance and authoritative declaration from a respected institution that a work is "publishable"; and selection, cataloging, display (in a suitable medium) and information services by "librarians." However these roles need not be centralized as they are now, since the economies of scale suitable to print collections don't apply (for example, it would be outrageous in cost to have a separate librarian for each subspecialty area of books in a physical library, but not at all so for someone who performed that service nationwide on the net).

So what is the result? There won't be publishers as "manufacturers" of works, or librarians as maintainers of storehouses. But there will be many people doing the informational work they now do, in a more distributed way. New social/economic structures will evolve to suit this change. In particular, it will only work if there are consensus conceptual models that are used in the interchange among these various parties. This will no longer be enforced by the uniform training and professional unity of the publishing or library professions, but rather will be by emergent standards, in the sense that the computer graphic interface with its mice, menus, windows, scrollbars, etc. is an emergent standard that has provided a design language in which interfaces are built.

A. What research is needed?

In this light, a core piece of the research needed is the development of a consistent and manipulable "virtuality" -- the set of objects, relations, properties and views that constitute the world of "works" (which will integrate audio, video, and other media as well as the images and print that make up traditional libraries and most current on-line collections). Physical paper-based metaphors will be a useful starting point, but no more so than the physical desktop can serve as a basis for the computer "desktop" with elements such as scrollbars and icons, which have no direct analog. As with interfaces, it is unlikely that a committee setting out to define the "right guidelines" at an early stage in the evolution of the field will succeed. But people developing multiple versions and trying them out in the new information spaces will gradually converge on usable (not necessarily optimal) designs. By doing this in a thoughtful way from the outset, we can avoid some of the local optima which are blind alleys.

B. What infrastucture elements are needed?

Here again, I will focus on the interfaces. Not the screen and its resolution and portability, but on the set of manipulable "objects" that make up the world of information that the people in various roles can produce, manage and use. Possibilities range from simple keyboard/monitor combinations to imaginative uses of 3-dimensional images, sound, animations, etc. Also, in order to augment traditional methods of conscious cataloging (categories, call numbers, key words, ...), there will be a variety of derived indexing techniques, using statistical properties, heuristic analysis of document structure, massive fast pattern matching, recording of usage patterns, etc.

C. What component technologies are essential?

- High capacity storage with reliable, invisible management of copies, archiving, etc.

- High bandwidth communications nets

- High resolution visual displays with appropriate portability characteristics and networking

D. How do we characterize the functional capabilities of a digital library environment:

A library is a set of functions, not a set of information objects.
- individual search
- browsing, including news grazing, digests, magazine racks, new releases displays, etc.
- current flow of time-critical materials (e.g., newspapers, periodicals)
- forum, meeting place
- chat, etc.
- bboards
- expertise / skilled assistance, e.g., what a good reference librarian does

A library is not for words alone.
- Not just text (or graphics)- the electronic library will include all kinds of media, including video, databases, software, active servers,...

A library is not a physical place or piece of software.
- It is a collection of interfaces, including rooms, workstations, hand-held devices,...

A library's collection is not bounded by the walls of an organization that establishes it.
- It includes the entire network of resources that is available from within.

A library's users are not bounded by the organization.
- It is a place to put things for distribution as well as for use by your customers. It includes partners, people to whom you want to distribute things, the net community...

Librarians and patrons aren't different
- shared functions
- cooperative - users generate items and indices for each other
- find on-line resources
- catalog shared physical resources
- comments, annotations, recommendations

Chapter 3, Section E: Education Working Group Summary

NSF Digital Library Workshop
XEROX PARC
December 1992

Outline Summary from Working Group on Impact of National Digital Library (NDL) on Education

Dan Atkins, Chair

1. Approach
1.1. Try to flesh out a matrix along the following two dimensions
1.2. Functions, services of a digital library
1.3. Constituency in the education pipeline
1.3.1. pre-K
1.3.2. K-12
1.3.3. undergraduate and 2 year schools
1.3.4. graduate
1.3.5. continuing (life-long learning)
2. But first here is some consensus about the vision of a digital library
2.1. As we are using the term, a digital library represents a new, expanded domain for the library. The digital library supports the creation, editing, review, and use of information by individuals and groups. It encompasses both static (images of pages) and dynamic digital documents (software objects, integrated media, compound documents, etc.)
2.2. The inclusion of dynamic documents offers significant opportunities for education and enhances the marketing of the DL vision/initiative to Congress, et. al.
2.3. But we need to be careful in the phasing of provision of static and dynamic documents.
2.4. Need inclusion of static documents to achieve useful coverage of a subject domain.
2.5. Need dynamic documents to explore new applications and to accommodate

library services on new integrated media (multimedia) objects that are being created by academia and industry.

2.6. Suggest a balanced approach which would include moving some paper based collections to on-line images as well as some dynamic documents into a digital library environment.

2.7. Need an architecture which will accommodate both (static as special case of dynamic).

2.8. The digital library is not a single source. It is the aggregation of multiple, heterogeneous resources. In a virtual way, it creates one library out of many. There was some disagreement about whether or not transition from one library to another in the virtual single library should be seamless. Some felt that the context of being in a different type of library would be useful to the user.

2.9. The DL includes a communication (collaboration) function.

2.10. Warning: Need to be sure that we know what is really needed in education.- Focus on augmenting the teacher, not replacing the teacher.

2.11. There are and will continue to be important low-tech interactive media, e.g., crayons and paper.

2.12. Paradox: as digital libraries propagate and become ubiquitous, they recede out of sight and will become invisible, everywhere and nowhere just as dozens of electric motors are buried in modern automobiles but the automobile is not thought of as an instance of electric motors.

3. Functions and services of digital libraries relevant to education

3.1 New media and services for K-12 teachers and learners

3.1.1. Tailored, customized instructional material available on demand.

3.1.2. Augment the teacher with presentation media. "Teacher is the driver."

3.1.3. Provide interactive media directly to the students.

3.1.4. Real-time access to experts and to scarce facilities in a controlled learning environment.

3.1.5. Simulation of experiences ("learning by doing", "project-based learning", "CAD for kids"). Exploratory learning. Learning through "what if" and making mistakes. What are the limits of simulation? Do we need to smell the chemicals?

3.1.6. Customization to different rates and modes of learning

3.1.7. Note: Perhaps we should just document the extensive work under way in academia and industry on the creation and use of multimedia material and computers in general to enhance learning. We could motivate the National Digital Library as a distribution infrastructure for this intrinsically digital media. The NDL would be a repository for resources that can be pulled out on demand. It would facilitate standards and sharing and could serve a clearinghouse function.

3.1.8. Consider placing the first emphasis on the provision of digital versions of reference service, for example, encyclopedias. Focus on a few great products rather than a larger number of weaker ones. Be aware of the traditions of local control of textbook selection and try to avoid getting into that initially. An NDL will encounter tension between global access and local control.

3.1.9. Perhaps the NDL could make it easier to organize workshops, bring together experts, for example, to create new curriculum. Could also help capture and disseminate the results.

3.1.10. Some people expressed the opinion that on-line access was more justifiable for research than it was for education.

3.1.11. Some see the goal of education to be to produce "expert learners." Can the NDL help do this through, for example, supporting exploration, tutorials, experience? Law Schools are set up, not to expose students to every case, but rather to teach them how to become expert learners who can find what they need for a specific case. West Law and Mead's Lexus/Nexus are the digital libraries for law schools and are heavily used. Note that these companies heavily subsidize use of their products in law schools in an effort to create future paying customers.

3.2. New forms of information of contemporary importance to the general public and general literacy and welfare of citizens.

3.2.1. Example: Databases of environmental data which can be visualized and used through geographical information systems (GIS). Frank Caizonetti, Associate Dean for Research in College of Arts & Sciences at West Virginia University mentioned a statewide GIS system now being installed in West Virginia. He described increasing public demand for such services and also turf battles about who would provide them.

3.2.2. Example: Expert systems related to health awareness and self-diagnosis. There is a strong and growing consumer health movement in evidence in public libraries now.

3.2.3. Traditional types of library (public, school, academic/research, special) will be blended. Librarians and patrons at any of these will have access to holdings well beyond the local collection. At United Technologies (per Jean Mayhew) the corporate library collection is becoming more diverse to serve the personal needs of employees, for example, material on health, adoption.

3.2.4. Dramatic expansion of the holdings of public libraries and school media centers. Access to the NDL could make the difference between some and none information access in many rural communities. We did not, however, answer the question of where dollars would be available for on-line access if they were not available for print material.

3.2.5. The public library's role as a place and community center should continue. It however becomes an access point to a vast set of collections other than its local holdings. Also an access point to new services, e.g., a field trip through virtual reality.

3.2.6. Public libraries are an important part of a citizens' right (and obligation) to know in a democracy. Must be careful not to jeopardize this right as more and more information becomes available in digital form. Need to fund reinvention of the public library system for the 21st century just as Andrew Carnegie funded inventing the current system a hundred years ago.

3.2.7. NDL could provide lifetime information access, especially in rural communities.

3.2.8. NDL enables more frequent update of reference material.

3.3. Ideas on educational impact at the University level.

3.3.1. Many of the ideas under K-12 (above) apply here as well.

3.3.2. Humanists need access to special collections which frequently require travel. Could the NDL reduce the need for travel? Could it also promote more collaborative work in the humanities? Collaborative work is not as common as, for example in science, but there is increasing discussion in the humanities community that more should be promoted. This is the theme of a major humanities conference at the University of Michigan in April 93. Note too that Bill Wulf, former NSF/CISE Associate Director, is now involved in a "computing infrastructure for the humanities" initiative at UVA.

3.3.3. Could electronic publishing through a NDL help reduce the print publications back log problem?

3.3.4. Could it help bring intellectual control back within the academy?

3.3.5. In making the case for NDL application to higher education, we should probably distinguish between research and pedagogy. Graduate education will in general benefit from initiatives which support research productivity. The more difficult challenge might be to relate it to undergraduate education.

3.3.6. NDL could provide the library resources for "distance learning." Many distance learning environments now make little or no provisions for use of the host school's library. This is one reason that being a student at a distance may not be the same as being on campus.

3.3.7. NDL and collaboratory capabilities help accelerate linkages between research and education. The Arizona Collaboratory Project (Schatz, et al.) is illustrating this phenomena.

3.4. The NDL might be sold in part to help "people at risk."

3.4.1. But they must have access to the technology for the products of the NDL to have impact.

3.4.2. NDL could help level the access playing field. It might help leverage local investments through new forms of sharing.

3.5. An NDL might be sold (or not be sold depending on current Washington policy) as creating entrepreneurial business opportunities in educational markets.

3.6. An NDL might be sold in part to help support national competitiveness in math, sci-tech, and team-based creative activities.

3.6.1. But note that foreign countries that are doing better than the U.S. in K-12 math education are not doing it with information technology.

3.6.2. Perhaps with technology, both the number of students who are successful as well as the level of their success could go up, i.e., an enhancement in both depth and breadth.

3.6.3. Can the NDL be positioned as a critical support for education in the new information-centered industrial era?

3.7. Better interfaces to information (applicable to all of the above).

3.7.1. There is some evidence that interfaces to digital information can be built which are more productive for humans in some information seeking/use tasks than the traditional print interface. Some of the work by Mike Lesk, et al.. has shown this. Other examples are the work of Robinson and Card at PARC.

3.7.2. Can electronic publishers come up with better learning interfaces than printed pages?

3.7.3. On the negative side, there will be learning curves for new access technology and the inexperienced users might be less productive with the technology "assist" than without it.

3.7.4. Need to think about the NDL technology as helping not so much to find information, but rather to help better manage and use the fire hose of information that is directed at most of us. Use metaphors such as "harvesting" information that we need or want as individuals and as member of groups/organizations.

4. Possible major themes or threads

4.1. Most of our population has no support for learning in the home. The NDL could be a chance to link the home with the school learning environment.

4.2. Perhaps the NDL could help re-create the value system that we had when the community took responsibility for the education of its youth.

4.3. The NDL could enhance the possibility for the community and family to actively participate in education of the community.

5. Other comments and observations

5.1. There appears to be a growing willingness (necessity based on self-interest)

for industry to make investments in education
5.1.1. Adoption programs
5.2. It might be useful to look at the issue of NDL to support education along the following three dimensions:
5.2.1. New media
5.2.2. New practice
5.2.3. New types of communities
5.3. We need more participation from the electronic publishing community.

Chapter 3, Section F: Borgman Notes

NSF Workshop on Digital Libraries
Xerox Palo Alto Research Center (PARC)
December 9-10, 1992

Notes by Christine Borgman

These notes cover the general sessions, with summaries from small group presentations. They are fairly comprehensive but not complete. All omissions and interpretations are the responsibility of the notetaker. Interpretative comments are marked with initials (CB).

December 9, 1992 Wednesday

9:00 Welcome by ***David Levy***

9:15 ***Mark Weiser***, Welcome to PARC

Brief summary of work at PARC related to digital libraries.

He is head of computer science lab, former professor and entrepreneur.

His interests are in social science support for computer science, i.e., a different kind of research:

reverse technology transfer -- from industry to university deploy and use discoveries

PARC: 20 yrs of pioneering basic research

individual change, innovation continues

Reinforce Xerox commitment to long range research

labs ranging from physics to anthropology, all working together

Digital libraries as a nexus for the future of research

Four dimensions: technical + social + educational + political

technical: database, filtering, distribution

social: collaboration, help, reference, editing, communities, psychology, linguistics
education, teaching, browsing

political: copyrighting, intellectual property, international data flow, competitiveness

domains of research at PARC:
forms
conversion
content

Form domain
Cornell -- CLASS and bookstore
highspeed portable print service
work practice
scanning and printing hardware

Conversion domain
OCR
understanding
image processing
System 33 -- conversion among multiple forms

Content domain
information access
information visualizer
tapestry
liveboard
ubicomp -- ubiquitous computing

information access architecture
text database, various retrieval techniques

ubiquitous computing:
situated science: philosophy, psychology, anthropology
poverty of existing computer systems: heavy, GUI, intrusive, VR, agents, multimedia
technology trends: processing, display

goals of ubicomp:
invisible technology, integration of virtual and physical worlds, through desk, rooms, buildings, life

enhanced ability to act

using digital information should be as pleasant, refreshing, enjoyable as using a community library today.

YT Chien, NSF: Background on National Digital Libraries Initiative

Overview of IRIS supported research

HPCC -- support for educ and technology

Jane Caviness (next speaker) is in charge of NREN implementation for NSF

While digital libraries are part of NREN scope, no specific funds have been allocated for it. We need a next wave initiative to take advantage of the NREN infrastructure to develop digital libraries.

Gore II bill -- addresses specific work to be done. Digital libraries is one of most prominent activities in that bill. (CB note for record -- Gore II not passed; expect it to be resubmitted in some form in the 1992-93 Congress. It will have support in the

White House.)

The July Digital Libraries Workshop was for academic researchers. This one is for wider participation, especially participants from industry. Consider user and business point of view.

Goals of workshop:
Bring together people with expertise to help design blueprint for digital libraries. Get the communities involved. Design programmatic activities that have the endorsement and support of the community.

Digital libraries of the future

This term is used to denote a number of related concepts:

a knowledge network -- initially based on the NREN, which is expected to be in place in mid or late 1990s.

an enabling technology -- for the capture, store, distribution, connection, and access of a variety of information sources, materials, and tools: text, image, sound database, software, and scientific instruments

new information services -- packaged in unconventional ways beyond traditional library media and environments

public, personal, and ubiquitous

Technical issues that arise in developing electronic libraries from Computing the Future NRC report, 7/92, distributed to participants of this meeting.

database
information retrieval
pattern recognition
human computer interaction
data structures
algorithm engineering
data compression
distributed systems
parallel computation

linguistics
reliability
file system and management
networking
storage architectures

digital libraries
users view
technology
business view

Jane Caviness -- NSF

One of the reasons for being here is the need for such activities as the digital libraries initiative. NREN and digital libraries are complementary activities.

Focus of development in her group at NSF:
development of gigabit net with DARPA -- research
infrastructure -- NSFnet as interim NREN

Status report on NREN/NSFnet:

NSFnet -- backbone of 45 Megabits, 45 nodes(?), 6-700 major universities plus 1000 research labs

connections to 50 countries
started at 56 kbits, linking super computers

The network already is far beyond computational use -- need these many other applications. Even the rudimentary use of library catalogs on the net today have brought in multiple disciplines and groups.

155 Mbits next stage, next contract
already having a billion packet day on the net!

Building gigabit network testbeds; 5 testbeds:

AURORA, BLANCA, CASA, NECTAR, VISTANET
3 year research project

co-funding by NSF and DARPA
participants contribute from industry, university and national labs, super computer centers

different focus in each testbed

heavy industrial and academic participation -- she showed useful map of testbeds

showed applications list. Includes multimedia digital library applications

Moving to higher speeds very rapidly.

Knowbot project with CNRI is part of this effort.
We are trying to encourage library participation, but we are not funded to support it in our initiative.

We would like to see the network live up to its potential.

Gio Wiederhold -- DARPA perspective
I3 -- intelligent integration of information : an architecture for using the global information network

Objective of DARPA for information networks:

get useful information to
a user
in the appropriate form
amount and level of detail
at the right time

We want to exploit the many data and computing sources available and emerging.

data, data everywhere, and nothing helps you think. The ancient planner.

databases
experience and knowledge bases
simulations

They are working heavily in simulation technology.

Banking is still running at 55 baud!! on large pipelines, for reliability. Established standard around the world. (CB: Is this an example of setting standards too soon?)

Emerging technology advances needed: showed detailed list.

List includes
information retrieval,
description,
standards,
knowledge sharing standards,
data abstractions,
formal models.

They don't want to build one right system today that will be a great weight later.

Example of elaborate military inventory control system. Integrated system for 150 databases and applications around the world -- cannot change any part, because cannot know the effect on the rest.

Client server model: want intelligence in local system, not mainframe core -- too hard to adapt.

standards and scaling.
huge range of standards needed at different levels. data, telecomm, etc.

DARPA is not trying to set standards and improve them; rather to support testing and validation.

ex. KGML -- knowledge query and manipulation language

superset of SQL -- highly structured. (CB: based on example given, this interface is not for novice users).

new section documentation
syntax in BNF
semantics
semantic model
message ontology

Applications that should change:

electronic libraries with journal articles which permit replication of experiments. Where wet experiments are needed, interface to the experiments' equipment must be provided (doing this in genome data)

a design and utilization record for devices which can be effectively summarized.

an automated individual medical record, machine processable, relevant history of patient complaint, at an appropriate level for the care provider (physician or nurse)

materials database which can be plugged into design environment
current resources sparse
text which permits simulation of all examples
ditto for assignments given to students

Michael Lesk -- Bellcore

We have a shared vision -- people can sit at screens, call things up for use.

The vision works -- people can do it. CORE experiments at Cornell -- students with electronic access do much better than on equivalent paper tools.

People really do use the library resources on the net.
computer manuals are on-line or CDROM -- rarely get on paper now.
get software via ftp sources.

Right now the US dominates in these areas (computing and information services delivery). We have huge economies of scale -- we have control so far.
People in Europe go to US to get information and software. We'd like to keep it that way.

We don't dominate information services such as Minitel -- France has taken the lead.
France is close to ISDN, UK putting it in next year. Low penetration here in US.

Image areas -- Europe is taking off very quickly. Scanning books, European consortium for medical literature.
These need to be large scale efforts to happen.

Videotext -- failed in UK as customer supported technology.

UUCP -- Lesk is one of the inventors. This service/network has really taken off.

Who are the largest publishers in the world? -- 2 Japanese printing companies -- $6 billion companies. Printing everything from books to circuit boards and wallpaper. Europeans next tier of very large publishers.

Who cares if it's done in the US or not? WE do -- important strategic business. Good growth here, but much bigger in Europe and Japan

It wouldn't do any good to put US data on-line because others could read it too. Response of one reviewer of Lesks proposal (CB: This is the isolationist side of the competitiveness argument; discussed again on Dec 10)

We are still largest producer of scientific information.

The real issue is scientific education -- if were going to continue to be the best educators, we need to maintain control of scientific information.

Others will package our data and sell it back to us -- this is happening already.

what matters is:
putting the infrastructure in place
focusing on specific applications
economics and politics -- publishing and tenure.

summarized earlier proposal to Wash DC Digital Libraries workshop (July 92)
Get basic undergraduate science information on-line first -- get enough for everybody to use.
suggested $50 million over 5 years
1/2 for infrastructure, 1/2 for related research

get 5 or 10 areas, cluster of publisher, user, technology community

coordinating committees to keep things on track.

Proposal was presented to NSF advisory board in Dec 91 by ***Michael McGill***, where it was received favorably. This is why were doing more workshops.

The information business will have a few really big players. We really need to develop those businesses before we lose it to the Europeans.

So far we have the basic infrastructure, networks, information in place. We need ways to get the data available for lots of people to use. This is strong motivation for building a national digital/electronic library.

Stan Besen - Charles River Assoc.
Intellectual property issues.

Focus of talk: how does moving from print to electronic affect the infrastructure of the publishing industry?

Where would parts of the present industry be redistributed?

He is not focusing on the law of intellectual property. Rather when one goes to electronic form, can rights be appropriated? Can some or all of the rights be taken across the boundaries?

Scenarios:
Large scale interlibrary loan -- make copies available to everyone else. Only one copy of a journal would be sold.

Pirated copies of tapes, videos, are sold overseas.

Right now you find the guy and sue him -- threat is dis-incentive to violate intellectual property laws. Detection is hard to avoid.

In electronic environment it is harder to detect who is doing copying and distribution; easy to borrow copies of tapes, software, and hard to detect or trace.

It is hard to determine economic impact on the rights holder.
Part of the issue is how good is the copy as a replication -- it may be quite good, in the electronic case; i.e., a good substitute for the original. This is the problem for rights holders. What is the original in the electronic world?

The price of the original has to go down to the price of the copy, or it will be difficult to sell the product; would only get one sale. Result is a very messy economic model. The private copying problem is getting worse.

Derivative works -- e.g., novel and movie, foreign language translations; character becomes doll, toy. Rights holders make money from their work in a variety of ways. All revenue streams belong to originators in present scheme.

Court ruling: Someone purchased a book, sliced it up, and framed pictures to resell. The publisher sued over control of image and won.

When someone takes and manipulates another person's work electronically it will be much harder to detect the original source. Thus it is much harder to plan the revenue stream of derivative works.

How are firms going to compete? It is not clear. How to control intellectual property? They will try to do vertical integration and not put out lots of raw materials that might be used for other purposes.

We should have less restrictive rights, which lower the cost of this industry, and make more information available to people. Could have effect on the supply of works. Set of adaptations need cannot yet be anticipated.

Designation of breakout groups: ***Levy***

Day 1 Groups:

1. What is a digital library? Leader: ***Borgman***
Determine functions, dimensions. What are we trying to do?

2. What technology requirements must be met for a digital library? Leader: ***Lesk***
Topics: infrastructure, research

3. What are the economic, social, and political issues in the digital library? Leader: ***Besen***

Day 2 Groups:

Grand challenge applications from the point of view of

1. business
2. education
3. research

Charge to the groups:

1. Select one person to represent the group. Come back with a 15 minute presentation. Address these questions:

2. How is the digital library going to happen? What are the infrastructure issues? How are we going to build it, what has to be in it?

3. How much? Intellectual property questions. Industry structure. Political economy questions.

Group reports:

Group 1: What is a digital library? Leader: ***Borgman***

[Editor's note: see charts at end of this section.]

Summary from flip charts:

For the national initiative to develop digital libraries, the definition of a digital library is a service that enables users to access the information they need from any place at any time

an environment for creation, dissemination, and use of information

an archive for preservation over time -- stable versions of documents

that integrates data in multiple formats

that will select, acquire, manage, and organize information

a set of user friendly tools to locate, retrieve, and utilization the information resources

a distributed technology environment that dramatically reduces barriers...

Question from ***Michael Lesk***: what's in, what's out? (Definition is too general to be usefull)

CB response: This is intentionally a very inclusive definition. It would start with science and technology and move outward, but does not have a specific content as a stated scope. The breadth is deliberate -- YT Chien of NSF wants to be able to sell it to other federal agencies in multiple disciplines (medicine, engineering, etc.) and get their support and contribution to development. It is a service, not a place or a product.

Group 2: What technology requirements must be met for a digital library? Leader: ***Lesk***

[Editor's note: see charts at end of this section.]

Issues: collection/distribution/reception

collection: OCR/scanning. conversion. processable representations. develop collection. shared reference resources.

distribution: NREN, ubiquitous, distributed retrieval, natural language processing, tools to support indexing, summarization, classification; filtering; multimedia

computational requirements
security, privacy, authentication, protection, accounting
standards with extensibility

Technology for organization of database not included here; assumed it will be included in group 1 scope.

Group 3. What are the economic, social, and political issues in the digital library? Leader: ***Besen***

[Editor's note: see charts in the next session.]

This group issued no shoulds

Forecasting future of digital libraries is difficult.

The diversity of publications exacerbates problem. More research is needed.

Appropriability vs. free access: new mechanisms for appropriability in a digital environment

What will be the significance of gatekeeper function? We need to reward gatekeepers (CB: Besen is referring to publishers as gatekeepers, which is a much different definition than in scholarly communication research.)

How will the gatekeeper function change in a digital environment?

We expect changes in the number, identity, roles of participants in publishing.

Some interaction between intellectual property law and industry structure is expected.

What will be the political economy of introducing digital libraries?

How are digital publications different from other types of intellectual property that also have appropriability problems?

What is the role of government?
research and development
support for infrastructure
support for producers
support for users

Home delivery of information -- how can it piggyback on extant infrastructure?

December 10, 1992 Thursday

Day 2. Breakout into 3 groups; meet 2 hours and report back.

Reports of 3 groups:

What are the benefits of a national digital library to:

1. Research community (leader: ***Christine Borgman***)
2. Education community (leader: ***Daniel Atkins***)

3. Business community (leader: ***Michael Lesk***)

Reports:

Group 1. Research community. (***Borgman***)

[Editor's note: see charts in the next session.]

Group 2. Education community. (***Atkins***)

[Editor's note: see charts in the next session.]

Very detailed presentation with many flip charts; this is incomplete.

3 dimensions:
functions of digital libraries
constituencies K-12, college, research, continual life learning
subject domain

An infrastructure that supported new forms of dynamic media might be strongest case argument.

Communication processes are a very important part of the National Digital Library.

We must be careful not to suggest that high technology will supplant low technology -- still need paper and crayons.

Experiential learning still important -- chemistry labs and smelling -- olfactory senses.

There is the potential for customization to different rates and modes of learning.

Consider value of pulling things out that can be shared -- clearinghouse dimensions.

School libraries have minimal resources, as do public libraries -- a digital library could lead to expansion of virtual holdings.

The traditional role of a public library; the right to know, right to acquire information is a fundamental social right (in the US, anyway). We are reinventing the public library for the next century.

There is a growing demand for environmental, geographic information systems.

In the humanities, people spend much time to travel to resources. Some of this travel could be overcome.

Potential for personalization, customization of information resources.

Issue of selling our intellectual property offshore and buying it back -- can we recover some of this? (CB - discussion ensued later, part of isolationist issues.)

Graduate education and research are tightly coupled.

People at risk in not having adequate information access, particularly in K-12 (K-9?) level.

There is a tension between local control (textbooks, etc.) and remote access to materials.

Bootstrapping effect. Existence of digital library may make it easier to create materials for classroom use.

If we want to produce expert learners, we need to expose them at an early age to the resources that will be available to them later in life.

Consider a law school model of creating expert learners. Lexis and Westlaw can be seen as digital libraries. They provide the systems to law schools -- could we do the same for schools?

Teleconferencing technology is popular for education, but one of the major problems is that remote students do not have access to library materials; this might help compensate.

We need new types of interfaces -- better, more efficient access to resources.

Also consider health care as a national issue, what is relation to digital libraries?

Library useful for team based activities, competitiveness.

Students in low resource countries such as China are ahead of our students in science, which may make it difficult to make the argument that better information resources will improve our education system.

Could digital library increase the linkages between home and school?

Group 3. Business community. (***Lesk***)

[Editor's note: see charts in the next session.]

Three main issues for business community:

1. user impacts
2. producer impacts
3. social impacts

Assumption: the national digital library is good for business if it leads to a well educated workforce and higher productivity.

User benefits

Saving redundant work; processable and refinable information can reduce overhead.

Equal information access contributes to a level playing field
small/large/global business
businesses in small towns/large cities
isolated technology areas

Ways to improve productivity
superconductivity, xerography were discovered by random browsing
faster design making, shorter product cycles

industrial equivalent --as National Library of Medicine is to health

Example of identifying suppliers when technology needed quickly for Desert Storm. Who made IFF transistors? Identified a source, placed a contract in 2 days and got friend or foe detectors in 10 days.

Standardization, interagency agreements -- if we could get information out easily and quickly, it would increase use and dissemination (technology transfer).

We need rapid access to commercial resources in national emergency -- hurricane, need water purification information, etc.

We need access to grey literature -- government documents, etc. need better access and coverage.

a couple of months in the lab will save you a couple of hours in the library.

Other issues for business community:

advertizing, alerting, problem identification

technology transfer, both to and from universities and research

enable libraries to cope with greater volumes of materials

patent information

Some scenarios:

Blacksburg/Penn studies -- sharing correlation of information availability and economic growth

Swanson studies on medical searching

Productivity of biotechnology and their use of information -- new field, everything on-line.

software vendors -- for agents programs which display, search, filtering, abstracting

communications: services, telecommunications, etc.
equipment -- modems, portable devices

information vendors for other information sharing platforms
federal regulations,
materials information
product information
zoning building codes
new or substitute data

hardware: printers, displays
internal information system development

publishers alternative outlet for publication.

General discussion (after group reports)

social benefits of the national digital library:

small business opportunities as intermediaries, vendors, etc.
enable work at home
media changes for disabled
support democracy -- transfer Ben Franklin's argument for the post office to the NREN today
promote joint ventures -- anything that makes it easier for people to find each other will encourage joint ventures
business will locate where information is
easy to get
easy to distribute

Stated as nasty question: why won't benefits just accrue to foreign countries? (isolationist argument). Audience response:

1. Most of our information will be in English,which means more benefits to us.

2. Others already building national digital libraries and we need to trade with them.

3. If we don't trade, they are taking out information and selling it back to us.

4. Our economic survival is based on infrastructure and workforce -- we need it first to have the competitive advantage.

5. Same as general issues of international trade -- we need to have something to trade!

Other discussion point -- encouraging our staff to be multidisciplinary -- need multi disciplinary access

Closing session

YT Chien, NSF. Summary of what we've accomplished

Thanks to everyone, especially ***David Levy*** for making workshop happen.

Thanks also to Larry Masinter, Guiliana Lavendel, Geoff Nunberg.

NSF now has enough information that staff can assemble.

Our goal is to invent the future and what better place to do it than here at PARC, where Alan Kay started?!

What is the future of the national digital library?

Are we in the information business or the library business?

What economic model do we employ?
Free to the end user?
Bookstore?
Information kiosk?

How do you draw lines between who has free access and who does not?

One of the great future issues will be the models of who pays and how much and in what levels, etc.

Internal contradictions
free public and academic libraries
corporate libraries are willing to pay for Dialog, Lexis, etc.

Information delivery services are getting popular among academics.

People will pay for freelance services that are not called libraries.

We must not fail to resolve conflicts among these competing information sources.

NSF will distribute a resource book to all participants of the workshop.

December 9, 1992 - What is a Digital Library?

NSF workshop on National Digital Libraries
Breakout session on What is a national digital library?

Leader: Christine Borgman
Notetaker (these notes) Terry Winograd
Overheads for presentation are a subset of these notes

[Editor's Note: Formatting information has been partially lost for this breakout session report.]

1 Official notes on "breakout" session
2 What is a doable image of an electronic library that can be sold to Congress?
3 Gore bill says:
3.1 development of advanced storage
3.2 thousands of users
3.3 instantaneous
3.4 conversion of print to electronic
3.5 data base software for filtering and summarizing
3.6 standards for electronic data
3.7 training of users and librarians
3.8 simplify utilization
3.9 visualization technology for large volumes
4 Prototypes
4.1 digital libraries of scientific data
5 To be seductive to Congress as something worth doing it has to be a radical departure from what we have now
6 Break boundaries of traditional view of library
6.1 information is continuous
6.2 everything from classroom to homework to scholarship
7 Vision statement should be broader than the experiment
7.1 Talk about what it could be and have an experiment that fits in with today's structures and technologies
7.2 vision could be threatening to existing parties (e.g., publishers)
7.3 don't couch the vision as an attempt to put them out of business
8 A distributed technology environment which dramatically reduces barriers to the creation, dissemination, manipulation, storage, integration and reuse of information by individuals and groups
9 Any time, from any place, having access to the world's knowledge

9.1 reduce constraints of time and distance
9.2 all media
10 across communities
10.1 potential to bring connections from researchers to teachers to schools...
10.2 community of scholars can communicate their works and enter into dis course about it
10.3 collaborative group use is a key advantage of digital technology
10.4 Must be desirable because of what it does
10.5 leverage through digital replicability
11 A national initiative to develop digital libraries
12 Would include one or more projects, one or more libraries
12.1 includes research and specific experiments
12.2 continuum between different styles of information
12.3 bridging between roles of consumption and production
13 Testbed on a specific topic, e.g.,
13.1 Horizontal: Undergraduate science education (earlier proposal)
14 Vertical: e.g., chemistry from kindergarten to Nobel Laureate
15 allows synergy between different levels / teaching and research
15.1 as in the best institutions
15.2 Information for decision makers (e.g., Congress)
16 multiple levels of capacity
16.1 something completely portable
16.2 higher capacity available with some local concentration (e.g., large view, high resolution)
16.3 servers providing high-cost capacities
17 Capacity for archiving / preserving over time
17.1 long term preservation over technology and economics change
18 repository for guaranteed referentiality
18.1 stable versions of documents
18.2 Generally, widely and easily usable set of tools and capabilities to locate, retrieve, and utilize the information resources available.
19 Electronic vs. paper: electronic is better when materials are
19.1 highly volatile, e.g., air schedules
19.2 manipulation desired
19.3 scanning for names or words or phrases in lengthy texts
19.4 light use of remote materials
19.5 rapid communication in dispersed group
20 References
20.1 Redesigning Libary Services: A Manifesto, Michael Buckland
20.2 Rogers and Hurt, in Chronicle of Higher Ed on moving libraries to

community of scholars
21 Silicon Dreams, Bob Lucky
22 Digital information is fundamentally different
22.1 copiable, processable, rapid transmission,....
23 Different kinds of finding activities
24 Navigation in the "info space"
24.1 defuzzing and finding
24.2 negotiating the query
25 Access
25.1 How do I get it - obtaining access to the material, fact, or object
26 Serendipity
27 Browsing
27.1 encountering display
27.2 table of contents, distribution lists
27.3 general collections of abstracts
27.4 environmental scanning
27.5 network surfing
27.6 within structure of a conscious selection process
27.7 possibility of positive surprise
27.8 current awareness browsing
28 Reuse
28.1 Archival access: getting it over again
29 Putting it into something else
29.1 idea vs. expression
30 Participants
30.1 Chris Borgman
30.2 Jean Mayhew
30.3 Eugenie Prime
30.4 Vicky Reich
30.5 Tony Hall
30.6 Terry Winograd
30.7 Daniel Atkins
30.8 Maria Zemankova

31 What is it?
S = structure
B = behavior / function
31.1 S audience - education
31.2 S materials - science
31.3 S platform - NREN
32 S/B interface

32.1 agents
33 B economic model
33.1 B fee for service vs. tax/subsidy
33.2 S/B where is the librarian?
33.3 B structure for authoring / information generation
33.4 B scholarly communication process
34 B anticipate user needs / active vs. passive
34.1 B "don't tell me what I want to know, tell me what I should know"
34.2 S functions

Copies of Flip Charts

LEADER: CHRISTINE BORGMAN

BENEFITS #1: RESEARCH

1. LIBERMAN
2. BESEN
3. ZEMANKOVA
4. WINOGARD
5. MASINTER
6. LUCIER

LEADER: MICHAEL LESK

BENEFITS #3: BUSINESS

1. JAJKO
2. DYSON
3. ERTEL
4. PRIME
5. MAYHEW
6. GASS
7. WIEDERHOLD
8. LAVENDEL

LEADER: DAN ATKINS

BENEFITS #3: EDUCATION

1. CHEN
2. CAVINESS
3. ROSENBERG
4. MENELL
5. NUMBERG
6. REICH
7. CROCCA
8. URBANOWSKI

COLLECTION - OCR/SCANNING - CONVERSION
DISTRIBUTION - NREN - NII (NATIONAL INFO. INFRASTRUCTURE)
RECEPTION - USER INTERFACE (GRAPHIC)
UBIQUITOUS REPRESENTATION
ACCESS TO TRAINING
PROCESSABLE REPRESENTATION
PRINTING DISTRIBUTED RETRIEVAL
INTEGRATION DEVELOP COLLECTIONS
NL CONFERENCE
SHARED REFERENCE RESOURCES
ENTERTAINMENT
TOOLS TO SUPPORT ACCESSIBILITY
INDEXING
SUMMARIZATION DOMAINS
USER CLASSIFICATION, BACKGROUND
FILTERING
MULTIMEDIA RETRIEVAL
COMPUTATIONAL REQUIREMENTS
SECURITY/PRIVACY/AUTHENTICATION/PROTECTION/ACCOUNTING
STANDARDS WITH EXTENSIBILITY

LEADER: STAN BESEN
GROUP #3 ECONOMIC, SOCIAL, POLITICAL ISSUES

1. DYSON
2. LUCIER
3. MENELL
4. URBANOWSKI
5. LIBERMAN
6. GASS
7. ERTEL
8. JAJKO
9. ROSENBURG
10. NUNBERG

LEADER: MICHAEL LESK

GROUP #2 TECHNICAL REQUIREMENTS FOR A DIGITAL LIBRARY?

1. WIEDERHOLD
2. CAVINESS
3. CHEN
4. FINNIGAN
5. MASINTER
6. GIULIANA

LEADER: CHRISTINE BORGMAN

BREAK OUT GROUP #1 WHAT IS A DIGITAL LIBRARY?

1. BORGMAN
2. MAYHEW
3. PRIME
4. ZEMANKOVA
5. WINOGRAD
6. ATKINS
7. REICH

WHY

- LEVERAGE TECHNOLOGY TO IMPROVE RESEARCH & EDUCATION

- MULTIDISCIPLINARY SYNERGY/COORDINATION (THAT WON'T HAPPEN OTHERWISE WITHIN NSF DISCIPLINES)

- CREATE/ENABLE NEW INDUSTRIES

- DRIVER APPLICATION FOR NREN

- WIDER ACCESS/AVAILABILITY

- MORE COST EFFECTIVE

WHAT DOES THE INITIATIVE HAVE TO CONTAIN FOR YOU TO BUY INTO IT

I'M ALREADY IN!

RESEARCH INITIATIVES IN OPEN, DISTRIBUTED, EXTENSIBLE HYPER-BASE
SYSTEMS CLEARLY ADVANTAGEOUS FOR THE NATION'S EDUCATION (K-12 THROUGH POST DOC)

OPEN MINIMUM BARRIERS TO ENTRY

WHAT IS NEEDED TO ACCOMPLISH THIS

CRITICAL MASS OF RELEVANT INFORMATION
- ADD VALUE BY GOING BEYOND ELECTRONIC FORM OF EXISTING ITEMS
- CREATE NEW INTERACTIVE INFORMATION BASE

INFORMATION PEOPLE WANT AND NEED TO DO EXISTING WORD.

- $50m
- COOPERATION OF DIVERSE RESEARCH GROUPS
- COOPERATION OF CONTENT PROVIDERS
- INSTITUTIONAL BUY-IN (UNIVERSITIES)

- NEED TO DEFINE ROLE FOR TRADTIONAL LIBRARIES & PUBLISHERS
- ACCOMODATE LEGAL (COPYRIGHT) + BILLING CONCERNS
- CONSORTIUM BUILDING
- EVALUATION
- CONFERENCES / WORKSHOPS

WHY WILL IT BE GOOD FOR THE U.S.

PROVIDE FOR CROSS-FERTILIZATION OF IDEAS
- - APPLE PIE: EDUCATION, RESEARCH,...
- - SO WE WON'T BE A 3RD WORLD COUNTRY BY 2001
- - BECAUSE U.S. EDUCATION COULD STAND A LIFT!

MOON SHOT
MAINTAIN LEAD IN SCIENCE/TECHNOLOGY/ENGINEERING PUBLISH-ING

WHAT SHOULD BE IN THE NAT'L INITIATIVE

5 PARTS

(1) DISTRIBUTED/EXTENSIBLE/HETEROGENOUS/DATA + METADATA
DEFINE AN OPEN TECHNICAL/ARCHITECTURAL/STANDARDS FRAMEWORK TO CONTAIN THE MODULAR COMPONENTS OF AN ELECTRONIC LIBRARY
INFO SERVERS, REFINERS/BROKERS & CLIENTS - THAT FORMS AN OPEN PLAYING FIELD.

(2) FUND DEVELOPMENT OF SPECIFIC INSTANCES OF EITHER RESEARCH OF PRODUCTION LEVEL INSTANCES OF SERVERS, CLIENTS, ETC. THAT FILL SPECIFIC NEEDS OR INVESTIGATE/ LEVERAGE RESEARCH ACTIVITIES - BOTH NOW AND LONG TERM

(3) FUND LONGER - TERM RESEARCH ON SCALE UP, NAVIGATION, USER STUDIES,...
(4) DEVELOP PEER REVIEW MECHANISMS FOR VARYING COMPO-NENTS
- EVALUATION

(5) PROVIDE BACKBONE NETWORK INFRASTRUCURE THAT CAN SCALE TO MILLIONS OF USERS AND TERABYTES OF DATA IN PRODUCTION QUALITY, SECURE ENVIRONMENT.

HOW SHOULD IT BE COORDINATED NOW

TECHNICAL, ADMINISTRATIVE

USE E-LIB AS COORDINATION TOOL

PANEL AT NSF FOR NEW INITIATIVE IN IRIS

(1) MUST TAKE A VERY MULTIDISCIPLINARY APPROACH, AVOID PAROCHIAL INTERESTS OF NARROW RESEARCH COMMUNITIES

(2) SPONSOR CONFERENCES

(3) HAVE REPS OF: FEDERAL AGENCIES, ASSOCIATIONS

INVOLVE END USERS PROFESSIONAL SOCIETIES

NIST

RESEARCHERS

NEL Workshop Report

Christine Borgman
September 4, 1992

Response to workshop outcomes, as presented by Edward Fox on 21 July 1992.

Background:

A workshop was held in Chicago, 10-11 October, 1991, on Future Directions in Text Analysis, Retrieval and Understanding, in which I participated. Part II of the Final Draft Report (18 July 1992) of that workshop is a proposal for "A National

Electronic Science, Engineering, and Technology Library" (Lesk, Fox, & McGill). The proposal formed the basis of discussion for the 20-21 July 1992 workshop at NSF on developing a multi-agency initiative for a "National Electronic Library." Participants in the latter workshop were asked to respond to a number of issues raised in this report and in two days of discussion. This short paper is my response, organized by the suggested categories.

The above-mentioned reports do not contain an explicit, succinct definition of the term "National Electronic Library." For the purposes of discussion here, I am using the following definition, adapted from the workshop summary discussion:

The proposed National Electronic Library is

A service

An architecture that includes
—network requirements and expectations
—formal and de facto standards where they exist
—requirements where standards do not exist
—open framework for system component development

A set of information resources, databases of text, numbers, graphics, sound, video, etc.

A set of tools and capabilities to locate, retrieve, and utilize the information resources available

Users of the National Electronic Library include:
—students
—teachers/professors
—researchers/scholars
—librarians
—authors
—publishers
—information providers
—practitioners

Contributors of information resources to the National Electronic Library include:

—publishers
—universities
—professional societies
—libraries
—authors
—editors
—compilers

Responses to specific workshop questions

What does the National Electronic Library look like to different categories of users?

The National Electronic Library (NEL) initiative as outlined in the report of 7/92 would begin by building a set of information resources aimed at undergraduate level science, engineering, and technology, and would support related research in information technology, standards, and information behavior. The NEL would exist on the NREN. Research and development would continue to expand, refine, and support the services and system. It is assumed that most users of the NEL also would be users of other information network services such as those provided on the Internet: electronic mail, file transfer, and remote login. The following sketches are the images the network might offer to each of these groups after the first 5 years of development.

Undergraduates
Undergraduate students in science, engineering, and technology at U.S. colleges and universities would access databases of bibliographic, text, numerical, and graphic material supporting their coursework via their campus networks. (The NEL should be available to those outside the U.S. as well, but we discuss domestic users here as our primary interest for the NEL initiative.)

Specific implementations would vary depending on resources of the institution and individual student from dialup access in the academic library to campus workstations scattered through classrooms, laboratories, libraries, offices, and dormitories, connecting resources on the campus backbone and the NREN. Students with personal computers would be able to dialup or otherwise connect into the network from home or away.

At the low end, undergraduate applications of the NEL would replicate current library use; e.g., searching indexing and abstracting databases to obtain information for a term paper, then seeking the full publication in print form from the library shelf or interlibrary loan. For all but the most resource-rich libraries, NEL access would be a considerable service expansion by providing access to databases not held locally, or providing new databases created for the NEL. The broad access to a common network would equalize information resources between richer and poorer institutions and could greatly increase access for the part-time or commuting student, by providing easy access from home or non-university workplace. As the educational marketplace shifts to the older, part-time student, distributed access to information becomes increasingly important to the educational process.

Local computing resources will grow along with the addition of NEL databases, allowing the NEL to be used for higher end applications, such as manipulating numerical and graphical databases of scientific information and for conducting empirical analyses of data collected by others, such as geophysical data, genome data, or historical trends in social or political data.

While the NEL, as described in the proposal document, would be of greatest initial value to undergraduate students majoring in science, engineering, and technology, it would also be of immediate value to other undergraduates in supporting their basic liberal arts courses in these areas. Conversely, science cannot be taught with scientific data alone. Science must be applied to application domains, such as the environment, which require information drawn from the range of human knowledge. As the scope of databases expand, so will the scope of undergraduate students served, directly and indirectly.

K-12 Students

An NEL initially targeted at undergraduates will have immediate value for the high school student. Advanced high school students begin studying undergraduate material, sometimes placing out of lower division courses.

High schools have far less computing infrastructure in place to take advantage of an NEL than do colleges and universities, and thus adoption will be slower. As the databases expand beyond primary science materials and interfaces are improved, usage will expand downward into lower grades. Early uses of the NEL for K-12 are more likely to be teacher-driven, using the resources for classroom demonstrations and in-class explorations, than student driven, as students may have insufficient access to computers.

In the long term, the NEL offers great potential for enhancing education in the full K-12 range. Our studies of the Science Library Catalog (discussed further below) show that elementary school children (aged 8-12) are able to search for science information in a browsing-oriented system that is tailored to their level of cognitive development, reading level, and interests. They like to use computers and they are interested in looking for information. Wide adoption at the elementary school will require more computing resources in the schools, NEL databases designed for this age group, and rich but simple browsing mechanisms for general science and other information.

Graduate students

Graduate students may make the greatest use of the NEL, even if the system is aimed initially at undergraduates. Graduate work is much more oriented toward individual research and information seeking than is undergraduate work. Graduate students are more likely to have their own personal computers and more apt to invest their own resources in information seeking than are undergraduates. They are older, often more affluent than undergraduates, and have the incentive to pursue individual research so they can complete their degrees in a timely manner. Undergraduates are more oriented toward pursuing specific assignments using specifically-assigned materials in a fixed curriculum of courses.

Graduate students are rewarded for seeking and identifying new information and new sources not only in support of their own coursework and independent research, but in support of their professors' research. Graduate research assistants often have the responsibility of literature searching and locating information resources (textual or otherwise) in support of research projects. Thus they are even more likely than scholars, in many cases, to take advantage of the exploratory capabilities of the NEL and NREN.

Researchers

Researchers can be expected to use the bibliographic and textual databases selected for undergraduates for their own research, to the extent that they represent the basic sources of their discipline. Non-textual databases of numerical and graphic data may be of considerable value to researchers, both for reference purposes (e.g., engineering handbook data) and for empirical research (e.g., geophysical data). Given that research and teaching are two sides of a coin, university researchers will be seeking databases that are valuable for their own research and that can be used to teach undergraduates and graduate students. Better yet, if the information resources support projects that can be assigned to students that will advance the researchers' own projects, they will be even more readily adopted. As the NEL grows to include more primary scientific data, it will be increasingly attractive to this group.

University researchers' access to the NEL will be dependent upon university resources. Researchers are more likely than students to have ready access, through computers in their offices, laboratories, and homes. Non-university-affiliated researchers (government laboratories, corporate researchers, consultants, etc.) will have needs similar to those of university researchers, though without the complementary support of teaching.

An important aspect of the current Internet for researchers that was not addressed in the workshop report is the ability to mount locally-produced databases for access by others. We can mount files of data that we want to share with others, including research data, preprint files, backfiles of electronic newsletters, bibliographies, and various other material. Mounting our own databases allows easy dissemination of research materials, with minimal overhead of copying and mailing papers and disks. Provision of research data should increase the dissemination of research results and thus technology transfer as well, spreading resources of NSF and other funding agencies even further.

The workshop report did address the need for intelligent agents to locate network resources, such as the Archie and gopher utilities already extant. As we continue to mount large, formally-produced databases along with the informal locally-produced databases, researchers and other users will find the line between material formats increasingly fuzzy. The mixing of resources also will put pressure on standards initiatives and enforcement on the network, with extensive need for open standards and searching mechanisms that can handle a range of formats.

Practitioners

Practitioners is taken to mean people who do not fall into the other categories. Depending on what they are practicing, their needs may be comparable to students or researchers. The differences may be in infrastructure. Students, faculty, and non-university researchers typically have access to an infrastructure that assists in training, documentation, selection of hardware and software, alerts to new resources, etc. Independent practitioners may rely more heavily on sources internal to the NEL and NREN for such information. A new market may develop for providers of information about the NEL for the non-affiliated practitioner.

Publishers

Publishers are a large group with unclear boundaries. Traditional publishers of trade materials and scholarly books and journals will see the NEL as a large library to whom they may sell their wares. They may use it as practitioners as well. Attention to standards will be an increasingly important responsibility of publishers, who will be supplying computer-readable data directly. Even with the extensive use of computer-readable text in publishing now, some publishers are using the computer-readable data only as an interim step to typesetting, with disks later destroyed. While this approach means printing quickly and attractively, it complicates publishing the text electronically later and avoids some of the need to adhere to standards.

Currently, some publishers are contracting with large universities to mount their databases on local systems. These systems may be accessible over the networks, but with access restrictions to certified users on a database-by-database basis. For example, one can sign onto MELVYL anonymously, without fee, to search the bibliographic database of University of California holdings. To switch to a MELVYL database supplied by a commercial publisher, such as the MEDLINE or Inspec files, one is prompted for a password that is available only to UC affiliates. This barrier limits the access to these databases to members of institutions that have paid directly for this resource, while access to other resources is paid indirectly, in terms of each institution offering up some of its own files and expecting others to do the same. The solutions to more open access are not clear; policy studies are needed to address the access and revenue tradeoffs such that information is affordable while providing sufficient revenue to keep publishers in the marketplace.

The definition of "publishing" is stretched further when researchers begin mounting their own databases, essentially publishing them outside of the usual formal mechanisms. When information bases are in the public domain, it appears to be more cost-effective to mount them and let others search them than to deal with requests individually. I have no data to support this claim, except for the observation that the number of mounted files appears to be increasing exponentially, and those mounting them must consider it cost effective or would not otherwise do so.

Information Providers

Information Providers appears to be a general term to include publishers, who make available the work of authors, editors, and compilers; and the authors, editors, and compilers who make direct contributions to the NEL. All will view the NEL as a place to contribute their work for others to read and use. The degree to which they view the NEL as a competing entity to other formats or as a new opportunity will depend on their current distribution and payment mechanisms, if any. The NEL will pose the greatest challenge to those with established distribution mechanisms, who must determine how this new channel complements prior channels.

At the same time, the low overhead of contributing information in machine-readable form compared to traditional publishing should create a whole new class of information providers who see this as a grand opportunity to be heard. As multimedia applications increase and the capabilities of the NEL increase, so will the number and types of information providers. The usual problems of copyright, payment, and access tradeoffs apply.

Authors (and editors and compilers)
Authors are assumed to be academic authors for the most part, as the initial NEL will be supplying information aimed at the undergraduate. University faculty have competing interests in electronic publishing between being authors and readers. As authors, we want control over our writing and payment for its use. As readers, we want ready access to the writings of others at minimal or no cost! Academic authors routinely assign copyright to publishers in return for getting their work published, and get royalties for books, though rarely for journal materials, complicating the roles further.

Development of the NEL will contribute to the dismantling of the formal scholarly communication infrastructure of authors, publishers, and libraries. Various schemes have been proposed to bypass publishers and make materials directly available over the information networks, such that authors claim royalties directly. Similarly, publishers have various schemes to bypass libraries and gain more control over dissemination. Technical and policy issues abound, which are discussed further in Borgman (1992).

Current libraries
Libraries already have expanded their services well beyond their walls, having used commercial information retrieval systems since their development in the early 70s (and the batch systems prior to that time), and more recently using the existing Internet services. Academic libraries would be the most directly and immediately involved with the NEL, given the emphasis on undergraduate education. They would use the NEL to expand their own services, providing access to the databases on the NREN by training their user community, providing hardware and software access, and incorporating references to NEL resources in their own catalogs. They would also consider NEL resources in their own collection development, shifting duplicate resources to other materials and services, extending already strapped budgets.

Corporate and special libraries would be immediate users of the NEL, just as they are heavy users of other electronic sources. Special libraries often are staffed by one full or part-time professional serving dozens of practitioners, whether corporate researchers, managers, or attorneys. They have minimal print collections and may do the majority of their information searching by accessing retrieval systems. When they need the full text of materials for which only bibliographic data is available on-line, they get it through reprint services and by sending messengers to area university and public libraries. Large special libraries, such as those of international firms, may be well-staffed and have extensive local collections, but they also rely heavily on information retrieval services, company-developed databases, and electronic networks of their own corporate libraries. They will see the NEL as an extension of other on-line retrieval services and will be more than willing to pay reasonable fees for usage.

Public and school libraries currently make the least use of automated retrieval services. They may not yet have automated their own internal operations and lack the computing resources to make external networks available to their patrons. The exception are the large central and regional public library collections, which provide information services that individual branches cannot supply. Many of these services are fee-based; businesses without their own libraries often rely on these services or those of university libraries. They also may hire information brokers to find information in support of larger projects.

Public and school libraries have the potential to be the critical point of access to the NEL, as the databases expand and the service broadens to more general access. Public libraries can provide the training and access infrastructure for the general public. School libraries can provide the training and access infrastructure for the schools, especially as school libraries and classrooms become networked. Given the current funding level for public education, such networking probably is much more than 5 years off, however.

Educators

Educators are by no means last in consideration, but are an umbrella group, as are information providers. These include K-12 teachers, college and university faculty, and practitioner-teachers, training in environments other than full-term courses. Teaching applications run the full gamut of users of the NEL, including searching for information for one's own edification, designing curricula, demonstrating scientific topics by manipulating scientific data on the NEL in the classroom in real time, or making teaching assignments.

Early science database applications in education include not only resources intended for student use, but resources for teacher use (Science CD-ROM Conference, 1992). These databases include lesson plans for a variety of science topics and laboratory experiments that generate inventory lists of instructional materials.

Learning science requires not only the subject content of science, but how to find out more information about the topic. The NEL can be extremely useful in this respect, allowing students to search for new information beyond what exists in their classroom and texts, and allowing the teacher to locate new information in response to student questions. The current educational structure, with its minimal library support and lack of teacher training in information seeking, does not support well the teacher in assisting students in continuing their explorations. The NEL could fill a critical gap here.

Educator uses will vary depending on resources available and on technical sophistication. With one workstation and a display unit, a teacher can search for information for class preparation and can conduct classroom demonstrations. With multiple workstations, students can search on their own. With multiple networked stations, students can conduct searches and projects jointly with other students in the school, the district, and elsewhere, and librarians and teachers can work closely in supporting classroom instruction.

Problems with the National Electronic Library:

The National Electronic Library is a powerful concept and one whose time has come. The potential payoff in terms of shared resources, expanded library and educational resources for the individual institution, and technology transfer is vast. The problems will lay in coordination, implementation, and policy.

One problem, though a relatively minor one in my mind, is the inappropriate use of NEL resources. Students, faculty, and other users of the NEL will use the databases available on the NEL, often adapting their questions to the information rather than vice versa. When we mounted the New York Times Information Bank at the University of Pittsburgh in 1973 for free access, all undergraduate term papers seemed to become NYT topics. When Medline was mounted on MELVYL for open access in the UC system, the university community made massive use of the database, far beyond the needs of the health sciences alone. This follows from the principle of least effort -- people use the information available, whether or not it is the "right" or most appropriate information. NEL services will compete with traditional library services, particularly on a self-service basis. The alternative to using the wrong information often is getting no information at all. If people are searching and are finding useful material, they will continue to search when it is appropriate, and will eventually become more sophisticated users.

We must provide the service, police it in a cost-effective manner, and accept some level of misuse. We cannot fail to provide an information service due to misuse, anymore than we can fail to provide city streets because they will be used by prostitutes and drunk drivers. Rules of good management apply here as elsewhere.

Effect of delay on the National Electronic Library:

The time for the NEL is NOW! We have the necessary building blocks in place for the NEL: the outline of the NREN, with the Internet and the NSF backbone; large numbers of valuable databases that are underutilized due to lack of access; basic information technology standards for telecommunications interconnection and file sharing; a rapidly expanding base of personal computers in the hands of institutions and users; basic understanding of computing networks by the tens of millions of electronic mail users. The user community on the Internet is expanding exponentially and they are becoming increasingly sophisticated in the use of electronic mail, file transfer, and remote login services. Hardware costs are plummeting, which will continue to expand the base of workstations with access to the information networks.

If we do not take advantage of the infrastructure we have built, others will. We are faced with increasing competition in the information technology and information services marketplace by other countries, as outlined in the workshop report. We can maintain our (shrinking) competitive lead by developing the NEL, which in turn will set the information technology standards in communications, information retrieval, and interface design, that others will follow. We can mount the databases and design a reasonable cost and access structure before others control so much of the information that we are priced out of the market. Or we can wait, let others set the standards and control the information, so we are priced out of the marketplace, and lose control over the technology so we buy it from others rather than they from us.

The sooner we implement the NEL, the sooner we can take advantage of shared information resources and supply needed information to our classrooms and researchers. The longer we wait, the more we will lose in both quality of education and international competitiveness.

Approaches to the National Electronic Library:

The approach to building the NEL outlined in the workshop report begins by constructing a basic core of databases aimed at the undergraduate, while funding supporting research on the NEL infrastructure. This seems a reasonable approach, as the size of the undergraduate population is known and manageable, and is in great need of expanded access to science, engineering, and technology information. Databases aimed at the undergraduate population should be of immediate value to the advanced high school population as well as to the graduate student and research population, thus providing a fairly broad base of support.

The proposed approach also recognizes the need for research in a variety of areas, from specific standards to interface design. It needs also to recognize the need for policy research to address the complex issues of the shifting scholarly communication infrastructure, including copyright and ownership issues.

Beginning with a few databases and then expanding is useful, though reliance on a few large projects carries risks of large failure and the added costs of coordination. An approach favored by at least one workshop discussion group favored more, smaller projects in database building and implementation, in concert with the other research efforts.

The NEL initiative should include a prioritized funding plan for expanding the services and resources beyond purely science, engineering, and technology to a broader base of knowledge that supports the nation's educational and research and development infrastructure, from K-12 through graduate work.

Projects related to the National Electronic Library; Success stories, results from evaluations:

1. Science Library Catalog Project, UCLA and CalTech.
Abstract from Borgman, Gallagher, and Walter, 1992:

Computers are being introduced into libraries and schools, often without an adequate understanding of people's abilities to use them effectively. This is especially the case in elementary school libraries, where little is known about how children search for information in either manual or automated environments. Studies of cognitive development suggest that the skills required by standard, keyword command-driven on-line catalogs are beyond the abilities of elementary school children, and that a hierarchical, browsing-oriented hypertext system would be a better match.

We report results of two information-seeking experiments in two elementary schools. The first experiment compared children's search success and time on the Science Library Catalog, a hierarchical browsing system, in two schools with their own library databases. The second experiment compared a later version of the Science Library Catalog to a keyword system. We found that children aged 10 through 12 were able to use both systems and that younger children were more successful searching the browsing system than the keyword system. Younger children aged 9 through 11 were successful on the Science Library Catalog in both experiments at both schools, and when compared to the keyword system, preferred the Science Library Catalog. Children achieved comparable overall success rates and search times between the two systems, but the keyword system was more sensitive to differences in age, question set, and computer experience. We conclude that a hierarchical, browsing-oriented on-line catalog interface is more suited to the skills, abilities, and interests of children in the concrete-operational (age 7-11) stage of development. Children entering the formal-operational stage (age 11-), with adequate computer experience, perform equally well on a keyword system and prefer it, at least for simple search tasks.

2. Conferences related to the National Electronic Library

Creating CD-ROMs for Science Education Conference. April 1-3, 1992, Colorado Springs. Organized by the Center for the Improvement of Physics Instruction, University of Nebraska, with funding from the National Science Foundation, Educational and Human Resources Directorate, Award MDR-9253138, Gerhard Salinger, Program Officer. Proceedings available from University of Nebraska.

This conference brought together about 45 people representing 8 projects developing databases for science education, K-12 and undergraduate, plus representatives of professional organizations and industry.

For further information, contact, Robert G. Fuller, Center for the Improvement of Physics Instruction, University of Nebraska, 1110 Ferguson Hall, Lincoln, NE 68588-0109. Phone 402/472-2790 or 800/332-0265(ask for Toolkit); fax 402/472-6234. For subscription to mail server, contact Norman Chonacky: nchonacky@james.bowdoin.edu

Invitational Symposium on Knowledge Management. October 27-29, 1991, Coolfont Conference Center, Berkeley Springs, WV. Sponsored by the Council on Library Resources, Johns Hopkins University, and the University of California, San

Francisco. Synopsis available from Johns Hopkins.

This conference brought together about 70 people from universities, funding agencies, and related organizations to discuss the changing infrastructure of scholarly communication in the age of electronics. The major discussion groups were Scholarly and Scientific Communication, Intellectual Property, Institutional Infrastructure, Education, and New Technologies for Knowledge Management.

For further information, contact Nina W. Matheson, William H. Welch Medical Library, School of Medicine, 1900 East Monument St, Baltimore, MD 21205. Phone 410/955-3411; fax 410/955-0985.

User Interfaces for On-line Public Access Catalogs. November 1, 1991, Library of Congress, Washington DC. Sponsored by the Library of Congress.

This conference brought together about 40 people from LC, other National Libraries, and researchers in information retrieval and interface design, most of whom were in Washington for the Annual Meeting of the American Society for Information Science. Organizers of the conference include Edward Fox, Ben Shneiderman, Gary Marchionini, Ray Larson, and Dean Wilder. A larger conference is planned for 1993, with invited papers.

For further information contact Ray Larson, Graduate School of Information and Library Studies, University of California, Berkeley, Berkeley, CA 94720. phone 510/524-2275; ray@sherlock.berkeley.edu. A brief synopsis of the 1991 meeting is available from Trudi Bellardo, School of Library and Information Science, Catholic University, Washington, DC 20064. phone 202/319-5085; fax 202/319-5574. bellardo@cua

CIKM-92, First International Conference on Information and Knowledge Management. November 8-11, 1992, Baltimore, MD. Sponsored by ISMM and the University of Maryland Baltimore County in cooperation with AAAI, IEEE, ACM (SIGART and SIGIR), Bellcore, UMBC.

For more information, send mail to cikm@cs.umbc.edu

Testimonials related to the National Electronic Library:

Support for the initiative will be needed from all corners of the information world. This list of organizations is anecdotal, from my personal knowledge of the field. It would be useful for the developers of the initiative to do a thorough search of the directories of professional organizations and foundations to find other groups as well. Addresses and contact names for these organizations and others can be located in standard bibliographic sources. I've added some comments on the organizations where useful.

- American Library Association (50K plus members, active lobbying office in Washington, DC, major player, very important source of support for initiative.)
- Library and Information Technology Association (a subgroup of ALA, large, represents those involved in library automation and networks)
- National Commission on Libraries and Information Services (federally funded, organizes White House Conferences on Libraries and Information Services; held hearings on NREN and libraries in July 1992).
- Computer Professionals for Social Responsibility (interested in workplace issues, responsible use of technology, privacy issues. Testified before NCLIS on NREN issues. Washington office headed by Marc Rotenberg.)
- Association for Computing Machinery (especially special interest group on Information Retrieval)
- Coalition for Networked Information (Washington, DC. Working on standards for information delivery on Internet, cooperative issues. New but very active group.)
- Association for Research Libraries (Washington, DC. Umbrella for all major university and other research libraries in the U.S.)
- American Society for Information Science (Silver Spring, MD. Scholarly and professional society for broad range of theoretical and technical issues in information and delivery. Mix of librarians, information retrieval researchers, government and private industry.)
- Council on Library Resources (Washington, DC. Primary private funding agency for library-related research in the U.S.).
- Information Industry Association (Represents small and large electronic publishers)
- Corporation for National Research Initiatives (Vinton Cerf, Robert Kahn, et al. Active in developing network infrastructure.)
- Society for Scholarly Publishing (Represents university and private scholarly presses)

- Society for the Social Studies of Science (Scholarly society, studies scholarly communication infrastructure)
- Special Libraries Association
- Medical Library Association
- Art Libraries of North America (ARLIS)

(other library associations representing special groups)

- Museum Computing Network (building databases of museum information and networking; often modeled on work of libraries)
- Professional societies of major scientific disciplines who would be served (American Chemical Society, Physics, Biology, Geology, etc.)
- American Educational Research Association (interest in classroom technologies)
- Professional education societies (teachers, university educators, etc.)

Figures, tables related to the National Electronic Library:

I don't have much to contribute here. Will keep my ear to the ground and send snippets as I see them. One useful figure would be the rate of Internet growth. I've seen figures of 15-25% per month of users, though these numbers change quickly.

Bibliography related to the National Electronic Library:

Reference list from Borgman (in press).

Bailey, C.W. (1989). Public access computer systems: The next generation of library automation systems. Information Technology and Libraries, 8(2), 178-185.

Barlow, J.P. (1992). Electronic frontier: The great work. Communications of the ACM, 35(1), 25-28.

Barron, B. (1992). Internet on-line public access catalogs. Internet Society News, 1(1), 24-25.

Bates, M.J. (1992). Scholarly Communication. Manuscript, Graduate School of Library and Information Science, University of California, Los Angeles.

Bates, M.J.; Wilde, D.; & Siegfried, S. (1992). The Getty on-line searching project report no. 1: Categories of information sought by humanities scholars. Manuscript submitted for publication.

Borgman, C.L. (in press). Round in circles: The scholar as author and end-user in the electronic environment. In Hazel Woodward & Stella Pillig, (eds.) The International Serials Industry: Current and Future Issues for Publishers and Scholars, LIbrarians, and Suppliers. Aldershot, Hampsire, Eng: Gower Pub Co. (publication February 1993).

Borgman, C.L. (1990). Editor's introduction. In Borgman, C. L. (ed.), Scholarly Communication and Bibliometrics, Newbury Park, CA: Sage, pp.10-27.

Borgman, C. L. (1989). Bibliometrics and scholarly communication: Editor's introduction. Communication Research, 16(5), 583-599.

Borgman, C. L. (1986a). Why are on-line catalogs hard to use? Lessons learned from information retrieval studies. Journal of the American Society for Information Science, 37(6), 387- 400.

Borgman, C. L. (1986b). The user's mental model of an information retrieval system: An experiment on a prototype on-line catalog. International Journal of Man-Machine Studies, 24(1), 47-64.

Borgman, C.L.; Case, D.O.; Ingebretsen, D. (1985). University Faculty Use of Computerized Databases: An Assessment of Needs and Resources. On-line Review, 9(4), 307-332.

Borgman, C.L., Gallagher, A.L., & Walter, V.A. (1992). The design and evaluation of an information retrieval system based on children's abilities: The Science Library Catalog. Manuscript submitted for publication.

Borgman, C. L., Gallagher, A. L., Walter, V.A., & Rosenberg, J.B.

(1991). The Science Library Catalog Project: Comparison of children's searching behavior in a direct manipulation and a keyword search system. In J-M. Griffiths (ed.), Proceedings of the 54th American Society for Information Science Annual Meeting. October 27-31, Washington, DC. Medford, NJ: Learned Information, pp. 162-169.

Carrigan, D.P. (1991). Publish or perish: The troubled state of scholarly communication. Scholarly Publishing, 22(3), 131-142.

Carrigan, D.P. (1990). The political economy of scholarly communication and the American system of higher education. Journal of Academic Librarianship, 15(6): 332-337.

Case, D.O. (1991). Conceptual organization and retrieval of text by historians: The role of memory and metaphor. Journal of the American Society for Information Science, 42(9): 657-668.

Case, D.O. (1986). Collection and organization of written information by social scientists and humanists: A review and exploratory study. Journal of Information Science, 12, 97-104.

Case, D.O; Borgman, C.L.; & Meadow, C.T. (1986). End-User Information-Seeking in the Energy Field: Implications for End-User Access to DOE RECON Databases. Information Processing & Management, 22(4), 299-308.

Computer-Readable Databases. 8th ed. (1992). K.Y. Marcaddio (ed.), M.E. Williams (founding ed.). Detroit: Gale Research.

Crane, D. (1972). Invisible colleges: Diffusion of knowledge in scientific communities. Chicago: University of Chicago Press.

Crane, D. (1971). Information needs and uses. In C.A. Cuadra and A.W. Luke (eds.), Annual Review of Information Science and Technology, 6. Chicago: Encyclopedia Britannica, pp. 3-39.

Directory of On-line Databases, 13(1). (1992). Detroit: Cuadra Associates, Gale Research.

Directory of Portable Databases, 3(2). (1991). Detroit: Cuadra

Associates, Gale Research.

Duda, K. (1992). Real world Kerberos: Authentication and privacy on a physically insecure network. Intertek, 3(3), 22-25.

Fayen, E.G. (1989). Loading local machine-readable data files: Issues, problems, and answers. Information Technology and Libraries, 8(2), 132-137.

Garvey, W. D.; & Griffith, B. C. (1964). Scientific information exchange in psychology. Science, 146, 1655-1659.

Glicksman, M. (1990). Changing patterns of scholarly communication: Implications for libraries. Library Acquisitions: Practice & Theory, 14, 341-346.

Hayes, R.M. (1992). The promise and the peril of electronic information delivery in research libraries. Speech to OCLC Directors' Conference, 9 March 1992.

Hayes, R.M. (1990). Desktop publishing: Problems of control. Scholarly Publishing, 21(2), 117-123.

Hurych, J. (1986). After Bath: Scientists, social scientists, and humanists in the context of on-line searching. The Journal of Academic Librarianship, 12(3), 158-165.

Kent, S. (1992). Security initiatives in the Internet. Internet Society News, 1(1), 34.

Landweber, L. (1992). International connectivity. Internet Society News, 1(1), 3-5.

Latour, B., & Woolgar, S. (1979). Laboratory Life: The social construction of scientific facts. Beverly Hills, CA: Sage.

Lievrouw, L.A. (1990). Reconciling structure and process in the study of scholarly communication. In Borgman, C. L. (ed.), Scholarly Communication and Bibliometrics, Newbury Park, CA: Sage, pp.59-69.

Line, M.B. (1973). Information needs of the social sciences. INSPEL, International Journal of Special Libraries, 8, 29-39.

Lynch, C.A. (in press). Networked information: A revolution in progress. Proceedings of the Illinois Clinic on Library Applications of Data Processing, 1991. Urbana, IL: University of Illinois.

Lynch, C.A., & Preston, C.M. (1990). Internet access to information resources. In M.E. Williams (ed.), Annual Review of Information Science and Technology, 25. Amsterdam: Elsevier, pp 263-312.

Lyons, P. (1992). Rules of the road: Network law. Internet Society News, 1(1), 41.

McCain, K.W. (1990). Mapping authors in intellectual space: Population genetics in the 1980s. In Borgman, C. L. (ed.), Scholarly Communication and Bibliometrics. Newbury Park, CA: Sage, pp.194-216.

Meadows, A.J. (1974). Communication in Science. London: Butterworth's.

Meadows, A.J. (1989). Communicating Research - Past, Present, and Future. Inaugural lecture, professor of library and information studies, Department of Library and Information Studies, Loughborough University of Technology, Loughborough, Leicestershire, U.K.

Menzel, H. (1967). Planning the consequences of unplanned action in scientific communication. In A. DeReuck and J. Knight, Communication in Science: Documentation and Automation. London: Churchill. pp 57-77.

Merton, R.K. (1957). Priorities in scientific discovery. American Sociological Review, 22(4), 635-659.

Morrison, P. (1979). Since Bath: A review of published information studies in social and behavioral sciences, 1974 through 1978. Behavioral and Social Sciences LIbrarian, 1, 5-22.

Nelson, T.H. (1987). Literary Machines. Edition 87.1. Self published; no place of publication given. Also published by The Distributors, South Bend, IN.

Okerson, A. (1990). Scholarly publishing in the NREN. ARL, 151, 1-4.

Osburn, C.B. (1989). Structuring of the scholarly communication system. College & Research Libraries, 50(3), 277-286.

Paisley, W.J. (1965). The flow of (behavioral) science information: A review of the research literature. Stanford, CA: Institute for Communication Research, Stanford University.

Pierce, S.J. (1990). Disciplinary work and interdisciplinary areas: Sociology and bibliometrics. In Borgman, C. L. (ed.), Scholarly Communication and Bibliometrics. Newbury Park, CA: Sage, pp. 46-58.

Price, D.J.deS. (1965). Networks of scientific papers. Science, 149, 510-515.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? Journal of Documentation, 25, 348-349.

Quarterman, J.S. (1990). The Matrix. Bedford, MA: Digital Press.

Rawlins, G.J.E. (1991). The new publishing: Technology's impact on the publishing industry over the next decade. Technical report. Department of Computer Science, Indiana University, Bloomington, IN 47405. Available via ftp (file transfer protocol) from iuvax.cs.indiana.edu: /usr/ftp/pub/techreports/tr340.ps.z

Rogers, S.J.; & Hurt, C.S. (1989). How scholarly communication should work in the 21st century. Chronicle of Higher Education, 36(7), A56. Reprinted in College & Research Libraries (1990), 51(1), 5-8.

Silverman, R.A. (1989). Desktop publishing: Its impact on the academic community. Scholarly Publishing, 21(1), 57-63.

Skelton, B. (1973). Scientists and social scientists as information users: A comparison of results of science user studies with the investigation into information requirements of the social sciences. Journal of Librarianship, 5, 138-155.

Small, H.G. (1978). Cited documents as concept symbols. Social Studies of Science, 8, 327-340.

Smith, L.C. (1981). Citation analysis. Library Trends, 30, 83-106.

Sproull, L., & Kiesler, S. (1991). Connections: New Ways of Working in the Networked Organization. Cambridge, MA: MIT Press.

Stone, S. (1982). Humanities scholars: Information needs and uses. Journal of Documentation, 38(4), 292-313.

Uncapher, W. (1991). Trouble in cyberspace. The Humanist, 51(5), 5-14.

White, H.D. (1990). Author co-citation analysis: Overview and defense. In Borgman, C. L. (ed.), Scholarly Communication and Bibliometrics. Newbury Park, CA: Sage, pp. 84-106.

CHAPTER 4 *Notable Events*

Conferences

CIKM-92, First International Conference on Information and Knowledge Management

November 8-11, 1992, Baltimore, MD. Sponsored by ISMM and the University of Maryland Baltimore County in cooperation with AAAI, IEEE, ACM (SIGART and SIGIR), Bellcore, UMBC.

For more information, send mail to cikm@cs.umbc.edu

Computer Professionals for Social Responsibility Annual Meeting

October 16 - 17, 1993
University of Washington, South Campus Center
Seattle, Washington, USA

Envisioning the Future:

A National Forum on the National Information Infrastructure and Community Access

Co-sponsored by the American Society for Information Science Pacific Northwest Chapter (ASIS-PNC)

The National Information Infrastructure (NII) is the proposed medium or "electronic highway" that will connect schools, libraries, government agencies, and citizens using state-of-the-art computer and communications technology. Media, computer, and telecommunications companies are also acutely interested in the NII as it dramatically increases their opportunity for delivery of communication, services, and entertainment. CPSR is convening this meeting for both computer professionals and the informed public to discuss the broad range of choices facing society regarding the NII.

Saturday, October 16th

- 8:00 - 9:00 Registration/Coffee & Tea
- 9:00 - 9:15 Welcome to the CPSR Annual Meeting - Aki Namioka
- 9:15 - 10:15 Keynote Address - Bruce McConnell, Office of Management and Budget

"Shaping National Information Infrastructure Policy"

Bruce McConnell, Chief of Information Policy at the Office of Information and Regulatory Affairs in the Office of Management and Budget (OMB), will present his views on the major NII issues now facing the administration. He has been with OMB since 1985 and became head of Information Policy in 1992. He now chairs the inter-agency task force responsible for developing federal information policy for the Information Infrastructure Task Force.

- 10:15 - 10:45 Break

- 10:45 - 12:15 Panel Discussion - Moderated by Eric Roberts

"Public Access to Internetworks"

Public access to the Internet (and other major networks) is a critical issue in any discussion about an "electronic highway". Panelists representing a wide variety of perspectives, including representatives from the Pacific Northwest, will present their views.

Panelists:
Phil Bereano, College of Engineering, University of Washington
Eric Hood, NorthWestNet
Kenneth Kay, Computer Science Policy Project
Jim Taylor, Coordinator Automated Services, Seattle Public Library

- 12:15 - 1:45 Lunch Break
- 1:45 - 3:00 Panel Discussion - Moderated by Andrew Gordon

"Municipal Information Infrastructure"

City and other government agencies are exploring possibilities for developing municipal networks. In this panel a city official as well as a representative from the state regulatory agency and a representative of commercial interests will offer their insights and interests.

Panelists:
Joe Hommel - Washington Utilities and Transportation Commission A representative from the Department of Administrative Services for the City of Seattle

- 3:00 - 4:30 Panel Discussion - Moderated by Doug Schuler

"Networking in the Community"

Community networks exist and are being developed all over the U.S. Panelists from various community networks will present their perspectives on the state of community networking now and in the future.

Panelists:
Tom Grundner, National Public Telecomputing Network

Parker Lindner, New Media Matters
Evelyn Pine, CPSR/Berkeley member and former Executive Director of the Community Memory Project
Roy Sahali, CLAMDYP (Computing Literacy and Access Making a Difference for Youth Projects)

- 4:30 - 4:45 Break
- 4:45 - 6:15 Panel Discussion - Moderated by Marc Rotenberg

"Computers and Democracy - What's the Connection?"

How might computers facilitate the democratic process? How might they pervert it? "Electronic Democracy" is currently in the public eye and aspects of it will probably be tested soon. What can be done to promote wise uses of computers in this critical area?

Panelists:
Jeff Chester, Center for Media Education
Jamie Love, Taxpayers Assets Project
Leah Lievrouw, Department of Telecommunication and Film, University of Alabama

- 6:15 - 6:30 Closing Remarks - Jeff Johnson
- 7:00 - 7:30 No Host Bar at Banquet Site
- 7:30 CPSR Banquet - Fundraiser - Waterfront Activities Center (Vegetarian food will be available)

+ Presentation of the 1993 Norbert Wiener Award to The Institute for Global Communications (IGC)
Presenter: Eric Roberts

The CPSR Board of Directors has chosen to award the 1993 Wiener Award to IGC in recognition of the work the organization has done to use network technology to empower previously disenfranchised individuals and groups working for progressive change. Geoff Sears, IGC's Executive Director, will be present to accept the award.

+ Banquet Address and Interactive Event - Kit Galloway, Electronic Cafe International

Kit Galloway of Electronic Cafe International in Santa Monica, California will present his innovative approach to electronic communication using phone lines, slow-scan television, and other technology. Using videotapes and a live demonstration with CPSR chapters in Los Angeles and other locations, Kit will discus how the Electronic Cafe concept has been used in a variety of settings. Electronic Cafe International has staged global events with poets, children, and communities in France, Nicaragua, Japan, as well as a variety of American cities.

Be sure to attend the banquet and participate in this provocative encounter with multimedia community networks !!

Sunday, October 17th

The second day of the conference will feature a more interactive format including several workshops, a discussion of CPSR's NII vision, and a special - free and open to the public - panel discussion on "Privacy Implications of Health Care Reform" Sunday evening.

- 8:30 - 9:30 Coffee & Tea
- 9:30 - 11:30 Workshop Sessions

Gender Relations in Networking - Judi Clark
Information Policy: A Framework for Action - Karen Sy
Computer Professionals and Social Action - Jeff Johnson

- 11:30 - 1:00 Lunch Break
- 1:00 - 2:30 CPSR Discussion
- 2:30 - 3:00 Break
- 3:00 - 4:30 CPSR NII Vision Document Discussion - Moderated by Todd Newman
- 4:30 - 5:00 Closing Remarks - Future CPSR NII Program - Marc Rotenberg
- 7:30 - 9:30 Special Panel Discussion

"Health Care Reform and the National Information Infrastructure: Opportunities and Risks"

This special event is free and open to the public.
Kane 210, University of Washington

ABOUT CPSR

Computer Professionals for Social Responsibility stands alone as the only national, non-partisan, public-interest organization dedicated to understanding and directing the impact of computers on society. Decisions regarding the use of this technology have far-reaching consequences that necessarily reflect the basic values and priorities of the people who govern their use.

Founded in 1981, CPSR has 2000 members from all over the world and 22 chapters across the country. Each of our members is an important participant in the dialogue that is helping to shape the future use of computers in the United States. Our National Advisory Board includes one Nobel laureate and three winners of the Turing Award, the highest honor in computer science.

We believe that as the influence of computers continues to permeate every aspect of our society, it is important that professionals become active participants in formulating the policy that governs computer use and access. CPSR welcomes any and all who share our convictions.

Creating CD-ROMs for Science Education Conference

April 1-3, 1992, Colorado Springs. Organized by the Center for the Improvement of Physics Instruction, University of Nebraska, with funding from the National Science Foundation, Educational and Human Resources Directorate, Award MDR-9253138, Gerhard Salinger, Program Officer. Proceedings available from University of Nebraska.

This conference brought together about 45 people representing 8 projects developing databases for science education, K-12 and undergraduate, plus representatives or professional organizations and industry.

For further information, contact, Robert G. Fuller, Center for the Improvement of Physics Instruction, University of Nebraska, 1110 Ferguson Hall, Lincoln, NE 68588-0109. Phone 402/472-2790 or 800/332-0265 (ask for Toolkit); fax 402/472-6234. For subscription to mail server, contact Norman Chonacky: nchonacky@-james.bowdoin.edu

Document Delivery & Internet Conferences

Conferences & Exibits focusing on the issues, utilization and implications that these services will have in the library environment.

Wyndham Franklin Plaza, Philadelphia, Pennsylvania December 7-8, 1992

Sponsored by: Meckler - Managing Information Technology

Meckler, a leading provider of information technology publications and conferences for librarians is pleased to present two conferences of significant technology-based services that will be greatly affecting libraries and patrons in the coming years.

Document Delivery: While commercial document delivery services are not new, what has changed is the high level of integration of computers, telecommunications, networks and fax machines for the identification and delivery of information. This conference will discuss: - Trends and Issues - Relative Costs of Interlibrary Loan Delivery System - Access vs. Ownership - Case Studies of Commercial Services - Case Studies of Library Cooperatives - Copyright Issues - Transboarder Issues and Programs - Subject Specialty Issues - OCLC's and RLG'S Document Delivery

Internet: The Internet is currently used by nearly 3 million people. Internet is a collective term for the many backbone, regional, and site data networks that it comprises - more than 5000 networks in 33 countries. From on-line catalog access to news services, remote control of CD players, and browsing through archives, the Internet has become part of everyday library operations. This conference will discuss: - Virtual Library Projects - Connectivity Options - Emerging Interlink Standards - Bibliographic Management of Internet - Gateways to European Databases - US Gateways - GPO WINDO: Gateway to Government - Special, K-12 and Public Library Connectivity

The Electronic Library: Administrative Issues for Organization & Access Institute

IRLIST Digest ISSN 1064-6965
September 20, 1993
Volume X, Number 36 Issue 180

I.A.3.

Fr: Arnold Hirshon
<AHIRSHON@WSU.BITNET>

Re: ALCTS Institute on the Electronic Library: Administrative Issues

THE ELECTRONIC LIBRARY: ADMINISTRATIVE ISSUES FOR ORGANIZATION & ACCESS INSTITUTE Sponsored by The Association for Library Collections & Technical Services, A Division of the American Library Association Hyatt Regency Cincinnati, OH Oct. 16-17, 1993 (Program will occur immediately prior to the EDUCOM meeting)

FOR COMPLETE INFORMATION CONTACT: ALCTS/ALA, 50 E. Huron St., Chicago, IL 60611. 800-545-2433, extension 5032.

OBJECTIVES: In what promises to be a watershed event in defining the future of electronic scholarly information systems, ALCTS is sponsoring this program to provide senior-level decision makers with information which they need to articulate a vision and formulate strategies for their institutions as they move from trends and concepts (like Gophers and WAIS) to the concrete reality of electronic libraries. Administrators need information about * design principles * governance issues * staffing and service concerns * potential outcome of electronic libraries. Geared specifically to the needs of senior level decision makers (such as directors and associate directors) of libraries and computing facilities, this institute will focus on administrative concerns and will be generally non-technical in nature. Participants will receive information necessary to * articulate the vision * formulate the strategies for the local institutional planning process.

METHODS: Presentation of papers and case studies with live demonstrations will be among the educational techniques used. In addition, breaks will provide an opportunity to speak with vendors who are in the forefront.

AUDIENCE: The institute, to be held immediately prior to the 1993 EDUCOM meeting in Cincinnati, has been designed for upper-level administrators (deans, directors, associate directors) from libraries and computer centers in academic and other institutions who are interested in providing electronic information to their users.

FEES: Fees for the institute are $225 for ALCTS members; $265 for ALA members; $305 for non-ALA members. Registration fees include instructional materials, refreshment breaks, and continental breakfast and full lunch on Sunday.

FACULTY: - - Arnold Hirshon, University Librarian, Wright State Univ. - - Jill Ellsworth, Assistant Professor, Southwest Texas State Univ. - - Michael Buckland, Professor, Univ. of California at Berkeley SLIS - - Brian L. Hawkins, Vice President for Academic Planning & Administration, Brown University - - Jeffrey Trzeciak, Head, Automation Services, Wright State Univ. - - Beth Shapiro, University Librarian, Rice Univ. - - Kevin Long, Director of Educational and Research Computing, Rice Univ. - - Peter Graham, Associate University Librarian for Technical and Networked Information Services, Rutgers Univ. - - Malcolm Getz, Associate Provost for Information Services and Technology, Vanderbilt Univ. - - David Penniman, President, Council on Library Resources - - Others to be announced

PROGRAM OUTLINE

SATURDAY, OCT. 16

- 12:00-1:00 p.m. Registration
- 1:00-1:30 p.m. Welcome and Overview of the Institute: From Virtual Libraries to Working Reality: The Challenge to "Just Do It!" Arnold Hirshon, Wright State Univ.
- 1:30-2:30 p.m. Keynote Address: The Electronic Library: If You Build It, Will They Come? Jill Ellsworth, Southwest Texas State Univ.
- 2:30-3:15 p.m. Vendor Demonstrations and Break
- 3:15-4:45 p.m. Electronic Information: What Is It and How Do We Organize It?: Putting It Together: The Principles of Information Access - Michael Buckland, Univ. of California at Berkeley SLIS Case Study #1: Example of an Early Implementation of an Electronic Library System - Jeffrey Trzeciak, Wright State Univ.; Case Study # 2: Example of an Advanced Implementation Presenter to be Announced
- 5:00-6:00 p.m. Reception and Cash Bar
-

SUNDAY, OCT. 17

- 8:00-9:00 a.m. Continental Breakfast
- 9:00-10:30 a.m. Internal Organizational Issues: Re-engineering the Organization for the Virtual Library: Staffing and Service Issues - Beth Shapiro and Kevin Long, Rice Univ. Moving Beyond Traditional Electronic Services: Collection, Access and Text Management Issues - Peter Graham, Rutgers Univ.
- 10:30-11:15 p.m. Demonstrations and Break
- 11:15 a.m-12:15 p.m. The Economics of Information: The Price of Doing Business in the Virtual Environment - Malcolm Getz, Vanderbilt Univ.

- 12:15-2:00 p.m. Lunch
- 2:00-3:15 p.m. External Organizational Issues: The Role of the Library in the Development of the Electronic Library and The Role of Campus Computing Services in the Development of the Electronic Library - Brian L. Hawkins, Vice President for Academic Planning & Administration, Brown University - Additional Speaker to be Announced
- 3:15-4:00 p.m. Approaching the Future: National Trends and Local Challenge: The New Electronic World Order: Implications for Local Development - David Penniman, Council on Library Resources

GopherCon '93: Internet Gopher Workshop and Internet Gopher Conference

April 12-13, 1993
Minneapolis, Minnesota

Prentiss Riddle
riddle@rice.edu

GopherCon '93 consisted of two events: an invitation-only workshop for about 75 participants focusing on technical details mostly related to the Gopher+ development effort, and a larger conference of some 250 participants covering a broader range of topics including CWIS and library applications of Internet Gopher. Both events were jointly sponsored by the University of Minnesota and by CICnet.

Day 1: the Workshop

GOPHER HISTORY: After greetings from our hosts, things got rolling with reports on the past, present and future of Gopher from the UMinn GopherTeam. Farhad Anklesaria began by recounting the history of Gopher development: UMinn's first attempt to design a CWIS resulted in a classic design-by-committee monstrosity, to which the Gopher Team responded by proposing a very simple protocol which they believed they could actually implement. The proposal was not well received by the

CWIS committee, but after an attempt to out-source computing services at UMinn failed, its opponents retreated and the Gopher Team was permitted to proceed. (This was enlightening not only as an example of the effect of political history on technology, but also because it may explain something about the Gopher Team's mistrust of design by committee, about which more later.) The talk also illustrated how much ground Gopher has covered: from in-house demo to a very popular technology and an RFC (#1436) on the IETF standards track in just two short years.

GOPHER PRESENT AND FUTURE: Mark McCahill gave an overview of the features being developed in Gopher+, the upwardly-compatible extension to the protocol (also known as "a gopher with a bag on the side to hold more"):

-- Item attributes: meta-information about Gopher objects.

-- Multiple views of objects.

-- ASK blocks: a method of passing information from the user to the server. While originally conceived as a forms fillout mechanism, ASK blocks are now expected to provide much more, including a way to make structured queries against a database.

-- "AdmitOne", a home-brewed authentication mechanism based on single-use tickets.

McCahill also mentioned some future directions for Gopher development:

-- Support for URNs and URLs ("Unified Resource Names" and "Unified Resource Locators"), a pair of standards in the IETF process for identifying Internet resources.

-- A Gopher->Z39.50 gateway.

-- The idea of combining ASK blocks with an authentication mechanism to create read-write Gopherspace. (While I find the prospect intriguing, I wonder about the wisdom of jerry-rigging it this way rather than designing the necessary protocol extensions to support it explicitly.)

-- Support for Whois++ (an IETF effort intended as a simpler alternative to X.500).

The other members of the Gopher development team reported on their work:

-- Paul Lindner, author of the Unix server and curses client, continues to enhance the base Gopher versions while working on Gopher+. The latest version of the base

Gopher client now offers a download menu for dialup users; a better CSO interface; faster I/O; and snazzy dialogue boxes. The base Gopher server now uses a config file to eliminate compile-time defaults; user-configurable filename extensions; better numbering support ("Numb=" is fixed); and improved IP-based security. The Gopher+ server and client are coming along: they already implement forms via ASK blocks (the resulting information is processed by being handed off to perl scripts); alternate views and languages; extended information about items; and AdmitOne authentication. Upcoming enhancements for the Gopher+ client include specification of preferred data types and languages to take advantages of alternate views and a pretty, integrated pager, while Paul's plans for the Gopher+ server include automated mirroring of data and the integration of a relational database.

-- Bob Alberti reported on fixes to the Gopher-FTP gateway.

-- Dan Torrey spoke of his work on the PC client: he had programmed himself into a corner with the previous version, so he threw it out and started over. The result is PC Gopher III, which is multi-threaded to allow for menu accessibility even in the middle of a retrieve and appears much faster to the user. Future plans include a new memory management scheme to allow for Gopher+ and files over 360K; support for other TCP/IP stacks; internal support for 3270 sessions; and improved "helper" applications (image viewers, etc.).

-- Dave Johnson reported on his Mac client work. He has made various speed and user interface enhancements to the base Gopher client, and also has written a Gopher+ client which supports item attributes, alternate views, ASK blocks and AdmitOne authentication.

VERONICA: Fred Barrie and Steve Foster of the University of Nevada at Reno led a session on Veronica. Veronica is making great strides in the direction of reliability and responsiveness by dividing the work: at the time of the conference there were two Veronica servers up (at UNR and CNIDR), with three more expected soon. Recent improvements to Veronica software include faster server code, boolean support (AND and OR, but not NOT or parens), and better data harvesting and cleaning.

Some Veronica statistics: UNR has been receiving 150,000 queries per month; it takes 30 hours to prowl "known Gopherspace", meaning the 400+ servers in the UMN master list and their children (amounting to 1200 servers all told); Veronica detects over 10 million Gopher items, which it reduces after eliminating redundancies to about 1.5 million items; Veronica's database now takes up 350 Mb for raw data and 40 Mb for the index; up to 10% of the items in Gopherspace turn over in 1-2 weeks.

There is a lot to be gained by making Veronica use Gopher+ attributes, such as abstracts, keywords, or a controlled-vocabulary classification scheme. Not only will Gopher+ allow for better searching, but it should help with one of the most common user requests, namely additional context surrounding the items returned from a Veronica search.

Problems still to be solved: scaling (although this may not be a problem -- Barrie and Foster think that 12 servers should be sufficient through the end of the year); how to let people control their own data harvesting (ideas: a cache at the top of the Gopher tree, perhaps in a ".veronica" file, and/or coordination through regional networks); lack of permission to publish information gathered by Veronica ("we haven't been sued yet"); "Where am I?" questions for users; topic-oriented Veronica servers; meta-searches (like a "monster mindex").

GOPHER+ DIRECTIONS: This session began with a "lively exchange of views" between the Gopher development team and members of the Gopher community who felt that Gopher would benefit by paying more attention to standards which either already exist or are in parallel development. To overdramatize the split, one could call it an argument between "design by committee" and "design over a couple of weekends". Among the points of the critics: Gopher should try to be more expandable and more easily integrated with other protocols (we should avoid the need for gateways if possible); Gopher should support MIME types, which is distinct from supporting *a* MIME type; Gopher has confused the encoding type (e.g., compressed) with the content type (e.g., text), while MIME keeps these distinct. The standards which were most often discussed as being of potential benefit to Gopher were MIME, WWW, WAIS (and Z39.50), URLs and URNs.

Paul Lindner gave a short but provocative talk on a proposed "geographic" view type, "directory/GVIEW", which would present a menu as an image with "hot spots". His original idea was for "geographic" menus like maps of bus routes or Gopher sites, but it could also be used to create Gopher menus which would resemble VNS pages or WWW hypertext documents. (I am struck by the tendency of these technologies to start resembling one another.)

ACROBAT DEMO: Gary Cassamiti (?) of Adobe gave a talk on a very interesting forthcoming product called Acrobat which is intended to move Adobe from being a printer company to being a document language company. Evidently Adobe was surprised to see PostScript become a means of transmitting documents from place to place, ridden as it is with printer dependencies. Acrobat is Adobe's answer: a program which converts between PostScript and a PostScript-like document language called PDF ("Portable Document Format"). It is Adobe's bid to come up with a standard so ubiquitous that they are calling it "electronic paper". In addition to increased portability, PDF has several features desirable in a document exchange format: it retains the structure of the document, allows direct access to any page in the document, allows for embedded special fonts and replaces missing fonts on the fly, includes efficient compression techniques, and is extensible to include multimedia and programmatic objects in the future. PDF will be an open format and Adobe plans to help develop "save as" modules for common document preparation tools. Adobe plans to offer Windows and Mac viewers for PDF this summer and DOS and Unix viewers soon thereafter. If it catches on this would be a good candidate for a Gopher data type and could solve many document incompatibility problems.

ATTRIBUTE REGISTRY: This was dominated by another exchange between the Gopher development team and the proponents of compatibility with other standards. The main proposal was that rather than Gopher developing its own attribute registry, Gopher should use the IANA ("Internet Assigned Numbers Authority") registration process, which according to some only takes 3 to 4 weeks. There was also further discussion of MIME. Result: the UMinn GopherTeam agreed to look harder at MIME.

SOFTWARE LICENSING: The much-awaited confrontation between Shih-Pau Yen of UMinn and the Gopher community at large was heated, as expected. Mr. Yen presented a draft of a three-tiered licensing scheme: separate categories of licenses for educational institutions, small businesses and large businesses. Mr. Yen pointed out that all UMinn software released prior to GopherCon would be considered freeware and not restricted by the license, hence the license covered effectively only future Gopher+ development; furthermore, the licenses restricted only Gopher servers, not clients. Under the proposal, educational institutions would be exempt from licensing fees and in fact need not even sign the license. Rates for businesses would vary according to the size of the business and the use to which it put its servers (whether it charged for information served out); rates would go as high for large businesses as the greater of $5000 per server or 7.5% of fees charged for Gopher services.

Among the criticisms of Mr. Yen's proposal: By charging fees exclusively for servers, he would penalize organizations which wish to put information into Gopherspace to the benefit of us all; it might make more sense to charge for clients. The license agreements defined fees per year and made no guarantees that fees won't go up in subsequent years, raising fears of price hikes once the market was hooked on Gopher. The language of the license agreements defined the size of a business by the gross incomes of all entities owning a share of or partially owned by the business, which would mean that one wealthy investor owning 1% of an otherwise small business could kick it up into the large business bracket. Finally, there were the familiar complaints that the UMinn code wasn't written entirely by UMinn, that it has or will soon have freeware competitors, and that it doesn't come with the sort of support structure one might expect from commercial software. Mr. Yen seemed determined to proceed with licensing fees of some kind, and it was unclear to me how much he would modify his proposal in the light of the comments at GopherCon.

Day 2: the Conference

The second day began with a recap of the Gopher update from day 1.

REPORTS FROM MONDAY'S BREAKOUT SESSIONS:

Using Other Technologies -- MIME, Kerberos, Z39.50, etc (Jane Smith, CNIDR): Repeated the plea that the Gopher developers take a harder look at other standards and strive for interoperability.

Cataloging E-journals and Resources (Billy Barron, UTDallas): Spoke about the gains to be made from improved Veronica, including a controlled vocabulary for searching and use of Veronica data for research on the most common terms currently used in Gopherspace titles and Veronica searches. Experiments in the past few months suggest that Gopher classification efforts based on schemes like the Library of Congress or Dewey Decimal system are less successful than ad hoc "used bookstore" schemes.

Security, Authentication and Signatures (Tamara Iversen, Northwestern): Session focused on a hole in the design of the AdmitOne authentication scheme discovered by Lee Brintle of UIowa; suggested a modular authentication scheme selectable by each Gopher site.

Gopher Front-Ends to Databases (Paul Lindner): A number of efforts have taken place to integrate both relational and flat-file databases with Gopher. Two basic methods are common: a "shadow" system, in which reports are generated from the database and moved into Gopherspace, and on-line access to an interactive query engine. Paul plans to work toward integrated Gopher support of one flat-file db (DBase is the most popular) and one generic SQL db.

GOPHER+ ATTRIBUTE REGISTRY: The Gopher development team agreed to consider using the IANA registration process in the future. For now, new attribute types are to be registered by sending mail to the gopher-news list and reaching some sort of consensus in the ensuing discussion.

NEW GOPHER USES AND TOOLS: unless otherwise specified, any tools mentioned here are said to be available in the archives on boombox.micro.umn.edu.

Jughead (Rhett "Jonzy" Jones, UUtah): "Jonzy's Universal Gopher Hierarchy Excavation And Display", this is like a local veronica -- it allows for search by title of items in a local Gopher hierarchy. It supports boolean searches (AND, OR and NOT). A Gopher tree of some 45000 items takes about 30 minutes to walk and 10 minutes to index.

Procmail (Marie-Christine Mahe, Yale): a mail-to-whatever gateway which Marie-Christine has used as a way to put things into Gopherspace.

WELL Gopher (Paul Holbrook of CICnet): an interesting collection of original text resources from a well-known Internet watering hole. Highlights include works by Bruce Sterling and Stewart Brand. Gopher to nkosi.well.sf.ca.us port 70.

Ac (Eric Morgan, NCSU Libraries): an interactive tool for the systematic acquisition of electronic journals.

OS/2 Gopher client and server (David Singer, IBM Almaden Research Center): the client is already on boombox, the server is expected soon.

Gmail/gmailcal (Prentiss Riddle, Rice University): a mail-to-gopherspace gateway with an accompanying events calendar manager.

Grabdata (Mahesh Subramaniam, Notre Dame): a tool to gather data from users' AFS directories, perform various validation checks, and move it into Gopherspace.

Relational database of course schedules (Ed Moy, UCal): A gateway between Gopher and a relational database containing a course schedule. Works by selecting a set of courses and successively limiting it by such categories as day, department, starting time, ending time, room, instructor, etc.

MOOGopher (Erik Ostrom, Gustavus Adolphus College, and Larry Masinter, Xerox PARC): An object-oriented MUD (multi-user dungeon/dimension) with an integrated Gopher client. Objects in the MUD can contain Gopher items, and users can move about Gopherspace analogously to exploring a MUD. To see this in action, telnet to Jay's House MOO at theory.cs.mankato.msus.edu port 1709.

[Editors' Note: We were unable to connect to the MOO]

Overview of MSU tools and applications (Rich Wiggins and Dennis Boone): Lots of nifty things at MSU, including sound archives (historic voices and 1992 Presidential debates), tools to generate a Gopher "road map" and analyze Gopher logs, an "on-line photo gallery" of campus sites, and a "guest book" for recording user comments.

The coveted stuffed gopher was won by Erik Ostrom and Larry Masinter for MOOGopher.

SERVER ADMINISTRATION: The Server Administration breakout session was in some ways a continuation of the tool-off, but it got beyond tools to cover Gopher administration and design philosophy as well. Joel Cooper spoke more about how Notre Dame copies data from individuals' AFS directories into Gopherspace. The procedure includes much error-checking and requires files to include a template containing information about author, expiration, etc.

Dennis Boone spoke about the design rules for the MSU CWIS, which may bear repeating here:

-- Ample "About" files.
-- No "hunt the wumpus" games -- hierarchy is preferable to a maze.
-- Lots of navigational tools (e.g., a title indexer)
-- The root directory should include help and be understandable
-- Directories should not exceed a screenful in length
-- Identity of data providers should be attached to all documents
-- Reliability: 24 by 7 support for Gopher services
-- Lowest common denominator: assume screens <= 72 columns wide and <= 16 rows high
-- No dead ends, no sparse directories
-- Local-use information is most important
-- Break these rules when necessary

Paul Linder spoke about the trials and tribulations of getting a CWIS off the ground at UMinn. Although he and the rest of the Gopher development team are celebrities in Gopherspace, he's "just another joe" at his home institution. Some interesting tidbits: the principal servers at UMinn are a mirrored pair of Macintosh IIci's running AUX; they carry out some 50000 transactions a day. UMinn is interested in ways to provide redundant servers, including the use of DNS route shuffling. The UMinn method of CWIS design emphasizes decentralization using small, cheap hardware, typically Mac servers. They see decentralization as a proper response both to the fact that "machines go down" and the need to put power in the hands of the people with the data.

I spoke about our experiences at Rice University with RiceInfo and some of the formal steps we've taken (the "RiceInfo Management Proposal", CWIS committees, Data Maintainer's Agreements, etc.). I also noted that, like MSU, we see reliability issues as encouraging reliably maintained central services over a potential chaos of less reliably maintained small servers.

ORGANIZING SUBJECT-MATTER BASED GOPHERS: A lively breakout session which made me wish I could have attended the other library- and resource-oriented sessions earlier in the conference.

Dan Jacobson of Johns Hopkins spoke about his biology Gopher server; he believes in the need for control to be in the hands of specialists in a field, not librarians. He also spoke of the need for context: guess what happens when a biologist does a veronica search on the word "mouse"? :-)

Mike McGervin of Stanford spoke of his work building a Gopher server of medical informatics sources based on ten years' experience with the SUMEX-AIM ftp archives.

Librarian Nancy John of the University of Illinois mentioned an Educom paper by Howard Strauss entitled "Anarchy in the CWIS". She spoke of the need to go beyond old methods developed for print resources (including the LC and Dewey classification schemes). One promising direction is the use of Veronica with Gopher+ attributes (Academic Discipline, Audience, Format (newspaper or novel), etc.) to provide needed context for searches.

(I piped up from the audience with a plea for feedback mechanisms from users and a community organized around each subject area, rather than depending on Gophermeisters discovering resources in isolation. I'd like to see a set of discipline-specific newsgroups or mailing lists devoted to the discovery of resources. I also talked a bit about "linkmerge", my ad-hoc scheme to merge subject-tree directories from other sites. Much to my surprise, people seem to like it, even the librarians who I expected to be horrified.)

CLOSING: The conference ended with a general question and answer session.

Highlights: Someone suggested a common conference at which Gopher, WAIS, WWW, etc. developers could meet (a good candidate to host such a conference would be someplace like CNIDR). Paul Lindner agreed to put ".cap" support back into the Unix Gopher server as an #ifdef option. The Gopher development team talked again about farming out documentation to outside volunteers.

My conclusions? There's lots happening in Gopherspace. It was very exciting to see all these talented people carrying out all of these great ideas. Gopher's explosive growth is continuing with no end in sight. At the same time, I was disappointed that Gopher+ has not matured more than it has -- there are still nagging details about the protocol, the feasibility of small clients, and the licensing of UMinn-developed software which could make Gopher+ less successful at fulfilling its goals than basic Gopher has been. I had hoped that these issues would be put behind us at the conference, but it is now clear that we will continue to struggle with them for some time.

- -- Prentiss Riddle ("aprendiz de todo, maestro de nada") riddle@rice.edu
- -- Unix Systems Programmer, Office of Networking and Computing Systems
- -- Rice University, POB 1892, Houston, TX 77251 / Mudd 208 / 713-285-5327
- -- Opinions expressed are not necessarily those of my employer.

InfoFair '92 - Mastering the Internet Maze

Thursday, April 2, 1992

University of Utah
Spencer S. Eccles Health Sciences Library
Salt Lake City, Utah

Sponsored by The University of Utah Health Sciences Center

INFOFAIR is an annual event at the University of Utah Health Sciences Center with the goal of providing up-to-date information on computer applications, resources and servcies as well as a glimpse into the future of computers and computing in the health sciences.

The theme of INFOFAIR '92 is: Mastering the Internet Maze. The morning presentations will focus on the Internet: what is the Internet, how to access the Internet, what resources and services are currently available. Sharing sessions and hands-on Internet sessions will be offered throughout the afternoon for new and experienced Internet users. Several workshops and demonstrations are also planned for the afternoon. These afternoon presentations offer a chance to learn more about new HyperCard features, Windows 3.0, Grateful Med and Loansome Doc. The exibits present a unique opportunity to try out new hardware and software, to become aquainted with computer services offered on and off campus, and to become familiar with computer applications developed at the University of Utah.

The Internet is a collection of computer networds linked through bridges, gateways and routers connecting univerities, government agencies and commercial companites. Examples of networks are NSFNET, WestNet, NYSERnet, etc. Today in the United States and around the world, there are over 500,000 computers connected to the Internet.

The Internet is an effective tool for academic and administrative users. Internet users can discuss common problems, write joint papers, keep up with current developments in their field, and access specialized databases, retrieve information from on-line archives, exchange files with colleagues, login to supercomputer centers and lookup information in on-line library catalogs.

The staff at the Eccles Health Sciences Library regularly connects to the Library of Congress, the New York Public Library and the University of California (to name a few prominent on-line catalogs). A dynamic environment, Internet access to special resources is a unique benefit to faculty, staff, students, residents and researchers here at the University.

From IAIMS Newsletter, April 1992, Vol 7, Num 7, University of Utah Health Sciences Center, published by the Spencer S. Eccles Heath Sciences Library

Information Exchange Telecommunications as Part of the National Information Infrastructure

You are cordially invited to attend the IEEE-USA Information Exchange on Telecommunications as Part of the National Information Infrastructure.

The United States needs to develop a world-class national information infrastructure. Fundamental to achieving this goal is establishing a national telecommunications policy that fosters effective application of evolving technologies, incentives for investment, and responsiveness to users. Many groups and associations are initiating projects to represent the views of their members on such an infrastructure.

Because IEEE's wide-ranging membership touches almost all entities addressing this issue, the IEEE Committee on Communications & Information Policy will convene an Information Exchange May 4-5, 1993, Rosslyn Westpark Hotel, Arlington, Virginia.

The purpose of this exchange is to allow the groups and associations developing a vision of U.S. information infrastructure to share views and better understand their respective issues. We envision a dialogue among those already active in this arena and other key individuals who are directly concerned with telecommunications. This will be a neutral ground where different interests can target particular issues important to them and this country. Panelists of national stature will cover the range from technology through policy. We expect many officials of the Executive and Legislative branches to participate as well.

The ultimate goal of this effort is to bring together the elements for a national policy to guide the nature and continuing development of a superior infrastructure. This Information Exchange is a first step in the process.

Attendance is limited. The conference fee of $150.00 and the completed registration form are due by April 15. A block of rooms has been reserved in the name of "IEEE" at a discounted rate of $87.00. Hotel reservations must be made directly with the hotel at [PHONE (703) 527-4814; FAX (703) 522-8864] or before April 3, 1993.

For registration and housing information, and a copy of the advance program, please contact:

Deborah Rudolph Manager, Technology Policy Council IEEE-USA 1828 L Street NW, Suite 1202 Washington, DC 20036 PHONE: (202) 785-0017 FAX: (202) 785-0835

Invitational Symposium on Knowledge Management

October 27-29, 1991, Coolfont Conference Center, Berkeley Springs, WV. Sponsored by the Council on Library Resources, Johns Hopkins University, and the University of California, San Francisco. Synopsis available from Johns Hopkins.

This conference brought together about 70 people from universities, funding agencies, and related organizations to discuss the changing infrastructure of scholarly communication in the age of electronics. The major discussion groups were Scholarly and Scientific Communication, Intellectual Property, Institutional Infrastructure, Education, and New Technologies for Knowledge Management.

For further information, contact Nina W. Matheson, William H. Welch Medical Library, School of Medicine, 1900 East Monument St, Baltimore, MD 21205. Phone 410/955-3411; fax 410/955-0985.

National Net '93

April 14-16, 1993

Loews L'enfant Plaza Hotel Washington, D.C.

Sponsoring Organizations: American Library Association, Association of Research Libraries, CAUSE, Coalition for Networked Information, Coalition for School Networking, Computing Research Association, Corporation for Research and Educational Networking, EDUCOM, Networking and Telecommunications Task Force, FARNET, Internet Society, National Association of State Universities and Land-Grant Colleges.

Corporate Sponsors: AT&T, IBM, Novell, Xerox

Topics: The NREN and the National Information Infrastructure - Competing Visions?; National Computing Environment & Grand Challenges; Citizen Access At The Statehouse; The 'E' in NREN; Internet Architecture for 100 Million Users; International Internet Report; Froma Network of Networks to a Community of Communities; K-12 Session; Images and Documents; Productive Partnering; Multimedia on the Internet; Reaching the Classroom; Sex on the Internet: Right or Wrong?; Listening of the Users; Internet Video: Meltdown or the Next E-Mail?; Easy Access Infrastructure and the 103rd Congress

New York Technology Summit - The Conference and Exhibition for Networking, Telecommunications, and Computing

Global Access Tools for Knowledge Workers

Jabob Javits Center, December 8-11, 1992 New York City, NY

The 1st Annual Conference and Exhibition on Telecommunications, Networking and Computing

As the world emerges from the age of information management and data communications and enters the next stage in the technology chain where the disciplines of communications, computing and telecommunications are wholly merged, an entirely new approach to management and problem solving will become necessary. These changes are revolutionary. Data communcations is being joined by high speed digital imaging; e-mail, voice and video mail are merging into a powerful mix that might be called multimail; video teleconferencing is making way for person-to-person desktop video computers and corporate morning meetings via LANS and WANS and mobile communications is evolving into wireless communications including facsimile, cellular phones, computer notepads and portables where total personal interactivity is becoming the standard.

The corporations which adapt to these technological innovations first will gain vital leadership positions in their fields and the resulting growth of these companies will quickly force all business to assess and evaluate their technology infrastructures or fear being lost in the world of the corporate have-nots. NYTS, the New York Technology Summit, will bring together the thousands of industry professionals - the executives, the information networking & telecommunication managers, technologists, developers, marketing directors and engineers from what used to be three distinct industries, Computer, Telecommunications & Networking, into one high technology gathering.

ACM SIGIR '93 16th International Conference on Research and Development in Information Retrieval

Pittsburgh, PA, USA
June 27-July 1, 1993

In co-operation with ACM:
AICA-GLIR (Italy) BCS-IRSG (UK)
DD (Denmark) GI (Germany)
INRIA (France)

PRELIMINARY PROGRAM AND REGISTRATION INFORMATION

The annual ACM SIGIR conference is the primary international meeting for the presentation of research and development in information retrieval (IR). The conference will be valuable to those interested in the theory of information retrieval as well as those responsible for system design, testing and evaluation. Topics include association methods, query processing, data and file structures, DBMS integration, structured text, inference networks, interface issues, and natural language processing.

This year the conference will be held in Pittsburgh, PA, USA, within easy reach of major east coast centers. The city centre conference site overlooks the three rivers, providing access to parks, restaurants and shopping, cultural amenities and sporting events.

Pre-conference tutorials will cover both beginning and advanced topics. The main program consists of contributed papers as well as panel sessions and demonstrations. There will be two post-conference research workshops, User Interfaces for On-line Public Access Catalogs, and Information Access and the Networks.

Conference highlights include:

Sunday * Tutorials
* Reception
Monday * Keynote address: Bruce Buchanan

	* Text REtrieval Conference (TREC) overview
Tuesday	* Panel session: NLP for Information Management
	* Cruise and Banquet on Pittsburgh's Three Rivers
Wednesday	* Demonstrations
	* Reception
Thursday	* Post-conference research workshops: -User Interfaces for On-line Public Access Catalogs -Information Access and the Networks (see details at very end of this message)
	* Tour of Frank Lloyd Wright's Fallingwater
	* Tour and Open House at Carnegie Mellon University

Full conference registration includes technical sessions, proceedings, conference banquet, lunch at the SIGIR annual meeting, and two receptions.

Conference Organization:

Chair: Edie Rasmussen, University of Pittsburgh

Program Chairs: Robert Korfhage, University of Pittsburgh
Peter Willett, University of Sheffield

Tutorial Chairs: Donna Harman, National Institute of Standards and Technology
Joan Morrissey, University of Windsor

Program Committee: Ijsbrand Jan Aalbersberg - Philips Research Laboratories
Maristella Agosti - Universita' di Padova
Abraham Bookstein - University of Chicago
Christine Borgman - UCLA
Giorgio Brajnik - Universita' di Udine
Chris Buckley - Cornell University
Yves Chiaramella - LGI-IMAG
Kenneth Church - Bell Laboratories
Richard Fowler - University of Texas - Pan American
Ophir Frieder - George Mason University
Norbert Fuhr - Universitaet Dortmund
Michael Gordon - University of Michigan
Peter Ingwersen - Royal School of Librarianship, Copenhagen
Kalervo Jarvelin - University of Tampere
Marcia D. Kerchner - The MITRE Corporation
Kui Lam Kwok - Queens College, CUNY

D. Lucarella - CRA-ENEL
David Lewis - AT&T Bell Laboratories
Sung H. Myaeng - Syracuse University
Michael Nelson - University of Western Ontario
Kai A. Olsen - Molde College, Norway
A. Stephen Pollitt - University of Huddersfield
Lisa Rau - GE Research and Development Center
Stephen E. Robertson - The City University, London
Ernst Schuegraf - St. Francis Xavier University, Nova Scotia
Alan F. Smeaton - Dublin City University
Linda C. Smith - University of Illinois
Padmini Srinivasan - University of Iowa
Craig Stanfill - Thinking Machines Corporation
Keith van Rijsbergen - Glasgow University
Jean-Luc Vidick - Telinfo, Brussels

SIGIR '93 PROGRAM

Tutorials
Sunday, June 27, 1993

AM: Introduction to Information Retrieval
Keith van Rijsbergen, University of Glasgow
Developing Information Retrieval Applications using Object Database Technology
David J. Harper, University of Glasgow
Trends in Multimedia Development
Alan Griffiths, University of Sheffield

PM: Natural Language Processing for Information Retrieval
David Lewis and Elizabeth Liddy, AT&T Bell Labs. & Syracuse University
Information Retrieval and Databases
Norbert Fuhr, University of Dortmund, Germany
Non-Textual Compression in Full-Text IR Systems
Shmuel T. Klein, Bar-Ilan University, Israel

Technical Sessions

Monday, June 28, 1993

Session 1: Keynote Address, From Data to Information: Where is the Knowledge?
Bruce Buchanan, University of Pittsburgh

Session 2: Inference Networks
David Haines & W. Bruce Croft, Relevance Feedback and Inference Networks
P.D. Bruza & L.C. van der Gaag, Efficient Context-Sensitive Plausible Inference for Information Disclosure
Kostas Tzeras & Stephan Hartmann, Automatic Indexing Based on Bayesian Inference Networks

Session 3: Donna Harman, Overview of the First Text REtrieval Conference (TREC)

Session 4: Full Text Analysis
Gerard Salton, J. Allan, & Chris Buckley, Approaches to Passage Retrieval in Full Text Information Systems
Marti A. Hearst & Christian Plaunt, Subtopic Structuring for Full-Length Document Access
Chris D. Paice & Paul A. Jones, The Identification of Important Concepts in Highly Structured Technical Papers

Session 5: Compression & Signature Files
A. Bookstein, S.T. Klein, and Timo Raita, Is Huffman Coding Dead?
Gordon Linoff & Craig Stanfill, Compression of Indexes with Full Positional Information in Very Large Text Databases
Deniz Aktug & Fazli Can, Analysis of Multiterm Queries in a Dynamic Signature File Organization

Tuesday, June 29, 1993

Session 6: Association Methods
S.K.M. Wong & Y.J. Cai, Computation of Term Associations by a Neural Network
Rodrigo A. Botafogo, Cluster Analysis for Hypertext Systems
Douglass R. Cutting, David R. Karger, & Jan O. Pedersen, Constant Interaction-Time Scatter-Gather Browsing of Very Large Document Collections

Session 7: Query Expansion
Peter G. Anick & Rex A. Flynn, Integrating a Dynamic Lexicon with

a Dynamic Full-Text Retrieval System
Efthimis N. Efthimiadis, A user-centered evaluation of ranking algorithms for interactive query expansion.
Yonggang Qiu & Hans-Peter Frei, Concept Based Query Expansion

Session 8: Linguistic Analysis
Ellen M. Voorhees, Using WordNet to Disambiguate Word Senses for Text Retrieval
Julian Kupiec, MURAX: A Robust Linguistic Approach for Question Answering Using an On-Line Encyclopedia
Robert Krovetz, Viewing Morphology as an Inference Process

Session 9: Structured Text
Michael Fuller, Eric Mackie, Ron Sacks-David, & Ross Wilkinson, Coherent Answers for a Large Structured Document Collection
Pekka Kilpelinen & Heikki Mannila, Retrieval from hierarchical texts by partial patterns

Session 10: Panel on Natural Language Processing for Information Management
David Evans, Lessons from the CLARIT project
Jaime G. Carbonell, Lessons from TIPSTER/ SHOGUN/JANUS
Sergei Nirenburg, Lessons from PANGLOSS

Session 11: Processing for Japanese Text
Yasushi Ogawa, Ayako Bessho, & Masako Hirose, Simple Word Strings as Compound Keywords: An Indexing and Ranking Method for Japanese Texts
Hideo Fujii & W. Bruce Croft, A Comparison of Indexing Techniques for Japanese Text Retrieval

Wednesday, June 30, 1993

Session 12: Interface Issues
Edward A. Fox, Robert K. France, Eskinder Sahle, Amjad Daoud, & Ben E. Cline, Development of a Modern OPAC: From REVTOLC to MARIAN
Daniel E. Rose, Richard Mander, Tim Oren, Dulce B. Poncelen, Gitta Salomon, & Yin Yin Wong, Content Awareness in a File System Interface: Implementing the 'Pile' Metaphor for Organizing Information
Gert Schmeltz Pedersen, A Browser for Bibliographic Information

Retrieval, Based on an Application of Lattice Theory

Session 13: Mathematical Models
Yiming Yang & Christopher G. Chute, An Application of Least Squares Fit Mapping to Text Information Retrieval
Joon Ho Lee, Won Yong Kim, Myoung Ho Kim, & Yoon Joon Lee, On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework
Carlo Meghini, Fabrizio Sebastiani, Umberto Straccia & Costantion Thanos, A Model of Information Retrieval based on a Terminological Logic

Session 14: DBMS/IR integration
Norbert Fuhr, A Probabilistic Relational Model for the Integration of IR and Databases
Peter Schuble, SPIDER: A Multiuser Information Retrieval System for Semistructured and Dynamic Data

Session 15: Query Processing and Evaluation
David Hull, Evaluation and Relevance Feedback
Nicholas J. Belkin, C. Cool, W. Bruce Croft, & James P. Callan, Effect of Multiple Query Representations on Information Retrieval System Performance
James P. Callan & W. Bruce Croft, Evaluation of Query Processing Strategies Using the TIPSTER Collection

Demonstrations (concurrent with sessions)
Coordinator: Philip J. Smith

Thursday, July 1, 1993

Post-conference research workshops:

User Interfaces for On-line Public Access Catalogs

Information Access and the Networks
(see more information on this at the very end of this entire message)

Tours:
Tour and Open House at Carnegie Mellon University
Frank Lloyd Wright's Fallingwater

FOR FURTHER PROGRAM AND REGISTRATION INFORMATION CONTACT:

SIGIR '93 Conference Secretariat
School of Library and Information Science
University of Pittsburgh
Pittsburgh, PA 15260, USA

tel: (412) 624-9459; fax: (412) 648-7001
email: sigir@lis.pitt.edu

PRE-CONFERENCE TUTORIALS
Sunday, June 27, 1993

SIGIR tutorials provide an opportunity to learn the basics of IR or to learn about a new or specialized area from experts in the field. This year, six half-day tutorials are available prior to the main program, held in parallel sessions during the morning and afternoon. Separate registration is required for tutorials.

Morning Tutorials: 8:30 a.m. - 12:30 p.m.

Introduction to Information Retrieval

Keith van Rijsbergen, University of Glasgow

This tutorial will enable participants to reach an understanding of the science and engineering underlying information retrieval research and development. It is aimed at anyone who wants to:
(a) do research in IR but has little or no basic knowledge in the subject;
(b) examine the state of research and development in IR for commercial purposes;
(c) teach IR at advanced undergraduate or postgraduate level but has no prior knowledge;
(d) think about IR but has rarely thought about it before and does not know where to start.
The tutorial will answer the following questions:
(a) What is IR?
(b) What are sensible models for IR?
(c) How to measure things in IR.
(d) What impact has IR research had on the existing technology? The tutorial will

include a demonstration of a current system. The reward for participants: a feel for the excitement of state-of-the-art in IR!

Keith van Rijsbergen was born in Holland in 1943. He was educated in Holland, Indonesia, Namibia and Australia. He took a degree in mathematics at the University of Western Australia. As a graduate he spent two years tutoring in mathematics while studying computer science. In 1972 he completed a Ph.D. in computer science at Cambridge University. After almost three years of lecturing in information retrieval and artificial intelligence at Monash University he returned to the Cambridge Computer Laboratory to hold a Royal Society Information Research Fellowship. In 1980 he was appointed to the chair of computer science at University College Dublin; from there he moved in 1986 to the Glasgow University where he is now, and indeed is head of department.

Since about 1969 his research has been devoted to information retrieval, covering both theoretical and experimental aspects. He has specified several theoretical models for IR and seen some of them from the specification and prototype stage through to production. His current research is concerned with the design of appropriate logics to model the flow of information. He is also involved in several Esprit projects concentrating on the engineering issues associated with the building of IR systems. He is a fellow of the IEE and a member of the BCS. In 1993 he was appointed Editor-in-Chief of The Computer Journal. He has also recently been appointed a member of the Advisory Committee for GMD in Germany.

Developing Information Retrieval Applications using Object Database Technology

David J. Harper, University of Glasgow

The field of information retrieval (IR) has yielded a range of techniques for efficiently and effectively storing and retrieving information items based on their text content. The utility of many application systems could be greatly enhanced by employing these techniques. Examples of such applications include office information systems, hypertext systems, and advanced information systems for government and industry. In this tutorial we describe how modern IR techniques, namely automatic indexing and best-match retrieval, can be efficiently implemented and conveniently delivered to application developers using object-oriented design and object database technology. We examine some sample applications and thereby demonstrate that the resultant IR framework, which takes the form of a class library, is both useable and extensible by application developers. We argue that our approach has a number of advantages over traditional mechanisms of delivery such as standalone packages and conventional programming language libraries. The tutorial is aimed at software designers and engineers who want to employ modern IR techniques within application systems, academics and researchers who are interested in implementing, experimenting with or prototyping IR systems, and managers wanting to assess the benefits of implementing IR applications using object database technology. Basic knowledge of IR and database concepts is assumed. A brief introduction to object-oriented concepts and object database technology will be provided.

David J. Harper is a lecturer in the Department of Computing Science, University of Glasgow, Scotland. He is a principal investigator in Glasgow on the EEC-funded ESPRIT project Comandos, which is concerned with the construction and management of distributed open systems, and he leads a major research stream on object data management services. His other research interests include information retrieval systems, object-oriented design and programming, formal specification of database systems, and office information systems. He received his B.Sc.(Hons) in Computer Science from Monash University, Australia, in 1973 and his Ph.D. in Computing Science from the University of Cambridge, England, in 1980.

Trends in Multimedia Development
Alan Griffiths, University of Sheffield

New products and applications for multimedia are being announced almost daily and there is a considerable hype over what is still a commercially insecure field with ill-defined standards and user groups. This tutorial will provide a stimulating overview of multimedia and will discuss a wide range of examples including computer-assisted-learning (CAL), image banks, electronic books and virtual reality. The currently available multimedia products and services will be examined to look for trends in development and these will be extrapolated into the future to propose products and services which will take advantage of the hardware, software and intellectual combinations to create imaginative learning and information environments. This enthusiastic view will be tempered by discussing the very real, and unresolved, problems in designing large multimedia systems. The aims of the tutorial are:

(a) To provide an overview of current and future trends in multimedia;
(b) To provide thought-provoking possibilities for information products incorporating multimedia;
(c) To stimulate discussion on how information retrieval techniques can and are being implemented into multimedia products.

The attendees for this tutorial will be people interested in the stimulating prospects that multimedia offers, but not necessarily specialists in the field. Multimedia will not provide a solution to information retrieval but it can provide front-end systems which use the algorithms and techniques outlined elsewhere in the conference.

Alan Griffiths is an independent consultant working in the areas of database and multimedia design, and is also a part-time lecturer in multimedia at the Department of Information Studies, University of Sheffield, England. Over the past ten years he has completed consultancies for a diverse group of clients ranging from heavy industry to public sector organizations. These clients have included Birmingham City Council, the Council for British Archaeology, Hewlett Packard, various parts of the British government, Basil Blackwells, Bridon Ropes and the Health & Safety Executive. His hazardous chemical software, produced for the Health & Safety Executive, is marketed by HMSO and used world-wide. His presentations on multimedia have been given to academic groups, software houses, companies and the public in England, France, Italy, Canada and Mexico.

Afternoon Tutorials: 1:30 - 5:30 p.m.

Natural Language Processing for Information Retrieval

David Lewis and Elizabeth Liddy, AT&T Bell Laboratories and Syracuse University

Our subject is the application of natural language processing (NLP) methods to information retrieval (IR) systems. We will discuss the characteristics of human language that make IR difficult, and will provide an extensive treatment of both NLP methods and their non-NLP alternatives for addressing these characteristics. We will also discuss the newer area of using IR components to aid NLP systems. This tutorial should be of interest both to researchers interested in past work and current research directions, and to application developers and managers who need to choose practical strategies for immediate implementation. Linguistic examples will be drawn primarily from English and Japanese, with a scattering from other languages. A bibliography on NLP and IR will be distributed.

David D. Lewis (lewis@research.att.com) is a Member of Technical Staff at AT&T Bell Laboratories in Murray Hill, NJ, where his research areas are IR, NLP, machine learning, and their intersections. Lewis has implemented a variety of experimental and operational IR and NLP systems, as well as being an organizer for several large scale evaluations of such software.

Elizabeth D. Liddy (liddy@suvm.syr.edu) is an assistant professor in the School of Information Studies at Syracuse University, where her research areas are discourse linguistics, semantic disambiguation, and the use of natural language processing in information systems. She is a co-principal investigator on the DR-LINK project funded under the DARPA Tipster initiative on text retrieval and filtering.

Information Retrieval and Databases

Norbert Fuhr, University of Dortmund, Germany

Besides basic database functions such as concurrency, recovery, security and integrity, future IR systems will also need data modelling concepts for coping with structured documents and powerful query languages in order to ask for any kind of object in the database (for example authors, affiliations, journals). On the other hand, database systems are required to handle textual attributes and to provide appropriate text retrieval facilities. This tutorial will present specific database concepts which are important for both types of systems.

After a short introduction to basic database concepts, three major lines of research will be considered: modelling data structure (relational and NF2 models); modelling structure and behavior (object-oriented databases); and modelling uncertainty (Fuzzy, Bayesian and Dempster-Shafer approaches). For each of these approaches, examples for their application to IR problems will be presented.

The target audience is those who have basic knowledge in IR and now want to learn about the database aspects of the field. The course would be most useful for people designing or building IR or database systems.

Norbert Fuhr is professor in the computer science department of the University of Dortmund, Germany. He holds a diploma and doctorate in Computer Science from the Technical University of Darmstadt. He is well-known for his theoretical and experimental work on probabilistic IR models. Recently, his research interests have focused on the application of IR methods to factual databases and the integration of IR and database systems.

Non-Textual Compression in Full-Text IR Systems

Shmuel T. Klein, Bar-Ilan University, Israel

Data compression has received increased attention lately, resulting in a large number of recent publications in this area. The tutorial presents an overview of these recent developments. We shall first classify the various types of files that appear in a full-text retrieval system. The focus will be on the auxiliary structured files, such as dictionaries, concordances and various forms of bitmaps, (which may significantly enhance the retrieval algorithms) and will present state of the art approaches for compressing them. The tutorial is intended for researchers, designers and users of large textual IR systems. By the end of the tutorial, the participant should be able, for a given file, to choose the appropriate compression technique and design his/her own encoding and decoding routines.

Shmuel T. Klein received his Ph.D. from the Weizmann Institute of Science in Israel. He then spent three years at the University of Chicago and is now at Bar-Ilan University in Israel. Dr. Klein has worked on two of the world's largest natural language full-text information retrieval systems: the Responsa Project at Bar-Ilan University (in Hebrew),and the Tresor de la Langue Francaise at the University of Chicago (in French), and has published several papers on various aspects of compression in IR systems.

User Interfaces for On-line Public Access Catalogs

November 1, 1991, Library of Congress, Washington DC. Sponsored by the Library of Congress.

This conference brought together about 40 people from LC, other National Libraries, and researchers in information retrieval and interface design, most of whom were in Washington for the Annual Meeting of the American Society for Information Science. Organizers of the conference include Edward Fox, Ben Shneiderman, Gary Marchionini, Ray Larson, and Dean Wilder. A larger conference is planned for 1993, with invited papers.

For further information contact Ray Larson, Graduate School of Information and Library Studies, University of California, Berkeley, Berkeley, CA 94720. phone 510/524-2275; ray@sherlock.berkeley.edu. A brief synopsis of the 1991 meeting is available from Trudi Bellardo, School of Library and Information Science, Catholic University, Washington, DC 20064. phone 202/319-5085; fax 202/319-5574. bellardo@cua

WAIS Inc. Announcements and Technology Showcase

I am pleased to announce that WAIS Inc. will hold an Announcement and Technology Showcase on Thursday, April 29, in Menlo Park, California. This is a very important event for WAIS Inc., as we are announcing our company and product line. This event will also showcase the many uses of WAIS technology. You are cordially invited to join us.

We will announce several products and relationships that will substantially advance network publishing on the data superhighway. Bill Dunn and I, along with several of our technology partners, will speak about how WAIS and the Internet are being used for network publishing.

After the announcement, we'll have a Technology Showcase, where you'll see how WAIS is being used in business, government, education and publishing. Organizations like Lockheed, U.S. Geological Survey and TRADE will show how they're using WAIS. Ensemble will show how WAIS can be used with up-to-the-minute information from The Wall Street Journal and Dow Jones. And MCC will demonstrate their clients for Macintosh, Windows and Unix. In addition, we're working to confirm several other important participants.

Finally, each of the participants in the Technology Showcase will give a 15-minute presentation on their use of WAIS, providing more detailed information than they may be able to give at the Showcase.

This will be a truly exciting event. I hope you can join us on April 29. Details and directions follow. If you plan to attend, please RSVP to Barbara Lincoln Brooks at barbara@wais.com.

Brewster Kahle
President
WAIS Inc.

Workshops

HCI Workshop

An Agenda for Human-Computer Interaction: Science and Engineering Serving Human Needs

Report of an Invitational Workshop Sponsored
by
The National Science Foundation

Edited by:
Gary Marchionini and John Sibert

Workshop Participants

Christine Borgman, University of California at Los Angeles
Deborah Boehm-Davis, George Mason University
Stuart Card, Xerox PARC
John Carroll, IBM Watson Research Laboratory
James Foley, Georgia Institute of Technology
Rex Hartson, Virginia Polytechnic Institute and State University
John Hestenes, National Science Foundation
Scott Hudson, University of Arizona
Robert Jacob, Naval Research Laboratory
Bonnie John, Carnegie Mellon University
Tom Landauer, Bellcore
John Leggett, Texas A & M University
Clayton Lewis, University of Colorado at Boulder
Gary Marchionini, University of Maryland at College Park
Jim Miller, Hewlett-Packard Laboratories
Brad Myers, Carnegie Mellon University
Dan Olsen, Brigham Young University
Judith Olson, University of Michigan
Randy Pausch, University of Virginia
Laurence C. Rosenberg, National Science Foundation
Ben Shneiderman, University of Maryland at College Park
John Sibert, George Washington University

Table of Contents

Executive Summary

Human-computer interaction (HCI) research is concerned with the design of interfaces that allow easy and efficient use of computer systems. This report is the result of a workshop held to define the state of the art and to identify HCI research directions. The workshop was held on March 4 and 5, 1991 at George Washington University. The participants considered four areas of HCI research: theory and models; input/output devices; tools and architectures; and computer-supported cooperative work (CSCW). This report contains information about each of these areas.

HCI models are categorized as scientific or engineering, and as prescriptive or descriptive. Recommendations are made for extending and integrating scientific models and for developing new engineering models.

Input and output devices are described and their potential application in advanced interfaces discussed. Strategies for increasing user to computer bandwidth (input devices) are stressed by highlighting the use of advanced devices such as data gloves, eye trackers, and position sensors in parallel. Trends toward direct manipulation and non-command interaction styles are also noted.

Tools to aid interface designers include toolkits, user interface management systems, and user interface design/development environments. The report recommends continuing existing lines of research to develop tools that support the design process through programming by example and by automatically generating code for I/O management. Additionally, the report recommends that the overall design process become more formal by systematic application of task analysis, usability specification, rapid prototyping, and formative evaluation techniques.

HCI support for computer supported collaborative work is considered mainly from structured message and interface sharing perspectives. They include the development of shared interfaces that allow messages to be classified according to standard categories or by explicit links and development of interfaces for multiple communication channels.

Finally, there are discussions of infrastructure support for HCI. These focus on the requirements for special equipment and expertise and on difficulties associated with interdisciplinary research. Resource sharing strategies are recommended to minimize some of these problems.

1. Introduction

The National Science Foundation sponsored a workshop about research on Human-Computer Interaction (HCI). It convened many leading HCI researchers for two days of structured discussions at The George Washington University Washington, DC on March 4-5, 1991. The workshop was organized in four areas of concentration: HCI models; input/output devices and interaction styles; tools and architectures for HCI; and relationship to computer supported collaborative work (CSCW).

1.1. Motivation for HCI Research

Human-computer interaction (HCI) research is concerned with the design of interfaces that allow easy and efficient use of computer systems. Computers are used for a variety of work tasks and recreation. Users vary in skill, education and frequency of computer use. This ubiquitous use of computers leads HCI researchers to seek general models of HCI, to develop a variety of devices and techniques for facilitating computer use, and to develop new software tools to improve the design and implementation of interfaces.

HCI research is important to business and science. For example, the Computer Systems Policy Project (a group of computing industry CEOs) has identified human interfaces as one of the 16 technologies critical to the U.S. computer industry; most computer manufacturers (e.g., Apple, DEC, IBM, HP) have HCI researchers and laboratories; Apple, IBM, Sun, and other companies have defined interface specifications for software developers who create packages for their platforms; and both the Japanese and European communities are devoting extensive effort to develop HCI research projects. Also, the development of significant markets outside the U.S. is beginning to spur research and development in graphic and multilingual interfaces.

Interfaces that allow scientists in all fields to take direct advantage of high performance computing will allow them to focus on their scientific problems rather than on programming or mastering technological details. Accurate and efficient interfaces are essential in complex, life-critical situations such as nuclear power plants, air traffic control, medical and military applications. Improved interfaces also offer economic benefits in business and government by facilitating training and improving productivity. Research in HCI can also act as a catalyst in other fields by providing new contexts for the systematic study of human performance, group behavior, and information management.

2. HCI Models

Scientific disciplines employ models to study, test, and explain phenomena. This section reviews some of the models developed in HCI. The purpose is to reveal underlying factors at work in the interaction of people with computers. This knowledge should help in the HCI design and engineering process.

There are two principal roles for models in HCI: The first is to help reveal mechanisms behind the phenomena or behavior exhibited by some systems. Roughly, this could be called science. In this case there is relatively low sensitivity to the costs of producing and using the model. The second purpose is as part of a technique for building artifacts. This could be called engineering or perhaps design. In this case there is relatively high sensitivity to the costs of producing and using models, and usability is important.

HCI models can be further divided into descriptive and predictive models. The intent of descriptive models is to provide systematic information on some phenomenon. For example: what are the set of actions taken in learning to use a system? The intent of predictive models is to enable us to use information about the mechanisms at work to ask 'what if' questions. For example, if the commands of a system were to be designed in a certain way, how long would it take a user to do a certain task?

2.1. The State of the Art

The broad set of topics in HCI for which models could be developed is roughly depicted in Figure 1. Basically, these reflect (I) the use to which computers are put and the context of that use, (II) characteristics of the human that constrain interfaces, (III) the design elements of the system, and (IV) the process of design itself. In fact, models will often address themselves to combinations of these topics (e.g., the effect of command structure on learnability). Using this categorization of HCI topics, Figure 2 lists some of the principle models concerned with the direct human-machine interaction process. Models of the social and task context of computers are mainly covered in the CSCW section. Models of the design process are beyond the scope of the review, but are reflected in the tools and architectures section. The figure attempts to group models according to scientific versus engineering models and predictive versus descriptive models. It is to be understood that any such classification is only approximate since some models may serve more than one role or concern multiple topics. Also, the figure is not intended to be exhaustive, but concentrates on the major groupings of models that have appeared multiple times in the HCI literature, e.g., the ACM Conference on Human Factors in Computing

Systems and the IFIP Interact Conference.

2.1.1. Scientific Models

2.1.1.1. State Transition Models

The most sustained development of models in HCI has centered around describing and predicting the characteristics of interactive dialogue. One of the earliest of this was based on state transition models. Later versions of such models have been used to formalize the semantics and interactive processes of computer science, and formal models specifically aimed at HCI have come from this tradition. For example, Squeak is a notation for formally describing mouse-based interfaces and the PIE model has been used to analyze the properties of prototype systems at York University, where much of the work on formal models of user interfaces is centered. Researchers have described model-based systems for specifying dialogues and others model systems based on the PAC architecture, that structures the software system into a hierarchy of interacting agents. The formalization of interaction semantics is intended to allow properties of the system to be identified, proved, and guaranteed. Some of the state-transition models are intended as the basis of implementation systems.

2.1.1.2. Grammar-Based Models

The state-transition models look at interaction from an internal system point of view. Another set of models attempt to use grammars to describe the dialogue structure as seen by the user. BNF has been used to analyze the consistency and learnability of a dialogue. Multiparty grammars have been used to separate human and machine actions. The dialogue structure has been separated into four levels, syntactic, semantic, lexical, and pragmatic, and a grammar (the Command Language Grammar, CLG) proposed for describing these levels. Task-action mappings from the ideas in CLG have been used to partition the task space implied by a command language and CLG was then augmented by External Task Internal Task (ETIT), a proposal that separately described the external and internal mental task spaces and their mappings. These developments led to a Task Action Grammar (TAG) that differed from the previous grammars by replacing the

Figure 1. HCI Topics
[figure omitted]

atoms of the grammar with feature lists, as in an attribute or case grammar. This allowed descriptions to partially match and gave a way of talking about family resemblances between different parts of a dialogue or different dialogues and hence to give some technical meaning to the notion of consistency. Later developments have sought to strengthen the formal base of TAG by shifting to Jackendorff semantics (ETAG) or by adding new display concepts of the language (DTAG).

2.1.1.3. GOMS and Related Cognitive Models of Skill

A third set of models attempt to model dialogue structure in terms of the cognitive processing of the user. Goals, Operators, Methods, Selection rules (GOMS) models the sequence of user actions in tasks of routine cognitive skill. In these cases the environment for action is stable and well-learned by the human. An example is much routine text-editing. By assigning mean times (or time distributions) to the operators, method duration can be predicted. The model has been used to study method tradeoffs and method times for on-line text-editing, spreadsheets, and other systems. GOMS as originally stated did not handle learning. To model some forms of learning within the same framework a production-system form of the model called Cognitive Complexity Theory (CCT) was developed. CCT has been used to predict learning transfer between systems as well as dialogue times. A form of the model (NGOMSL) has been developed and used as a practical tool. GOMS also did not model errors. Recently, research has extended the GOMS formulation to look at errors caused by the overload of human working memory for spreadsheet programs and SQL database queries.

2.1.1.4. Unified Models of Cognition

Another set of cognitive models attempt to model human cognition more generally. The Model Human Processor summarized part of the psychological literature into a parameterized, engineering style theory capable of computing expected properties of human information processing in simple situations. A combined GOMS/Model Human Processor analysis has been used to compute stimulus-response compatibilities, and hence response times, command name recall times, and times for speech input and output. Soar is partially a descendant of the Model Human Processor and is based on methods that allow the automatic generation of subgoals and automatic learning. It has been used to model a number of cognitive tasks. GOMS analysis is beginning to be integrated with it. ACT* is based on the separation of declarative and procedural memory. Declarative memory is modeled as a semantic network and procedural memory is modeled as a set of productions. This theory has been used to give an account similar to that from the GOMS/CCT analysis of learning transfer from one system to another.

2.1.1.5. Fitts' Law Model of Input Devices

Another sustained line of HCI model development is around the use of Fitts' Law to model pointing device use. Fitts' Law says that movement time for the hand depends on the ratio of the distance to be moved and the size of the target to which the movement is directed. Movement time = K log2 [2D /W] where D is the distance moved, W is the width of the target, and K is a constant, called the Fitts Index of Difficulty. It has been found that the speed of movement of the mouse and other pointing devices is predicted by this relation. Fitts' Law has been shown to fit many sorts of manual movements and an extensive literature of Fitts' Law and its descendants have developed that includes several hundred articles.

2.1.2. Descriptive Models

In addition to models developed directly within the HCI context, work from a number of other disciplines has contributed models, especially descriptive models of phenomena of relevance to HCI. Perceptual psychology has contributed analyses of color and many other perceptual phenomena. Descriptive models of the basic human-machine interaction cycle and slips made during interaction have been developed by cognitive psychologists. Psychology and education have contributed descriptive models of user individual differences--for example, psychometric analysis of novice users and analysis of UNIX knowledge. Information science has contributed models of user interactions with information retrieval systems. On the machine side, there are a number of models that attempt to analyze and group input devices, including those that classify devices by the number of dimensions and the property sensed, by the primitive tasks they perform, and by force, position, and type of motion.

2.1.3. Engineering Techniques

Models have been the basis of a number of engineering techniques, although most design and engineering is presently done on a cut and try basis. Although intended as scientific models to understand the structure of human-machine interaction, GOMS and related models have actually been applied to a number of design problems. GOMS has been used in the design of a Xerox text editing system, to the analysis of learning an oscilloscope design at Tektronix, to the analysis and implementation of scenarios for helicopter cockpit design at NASA, and to the evaluation of a telephone company operator workstation. A simplified version of GOMS called the Keystroke-Level Model was developed to allow back-of-the envelope style calculations of method times and an executable modeling language Task Modeling Language (TML) intended for developers at IBM based on CCT and GOMS has also been developed. An interesting attempt to provide a modeling-based design technique based on this class of models--mainly, Soar, but also ACT*, TAG, and Interacting Cognitive Systems--are what are called Programmable User Models (PUMS). PUMS is a constrained cognitive architecture in terms of which a developer programs a model of his task. PUMS has been used to predict time and usability and the difficulty of learning task-action mappings. Another interesting modeling-based design technique is the Cognitive Walkthrough technique in which the CE+ model of exploratory learning is used to generate a set of evaluation questions.

Fitts' Law has been used for a number of engineering purposes. First, it has been used to devise rapid engineering tests for input devices. This is a prime case where empirical testing work is made both faster and sounder because of some theoretical basis. Second, Fitts' Law has allowed some inferences about the design space itself and is partially responsible for the commercial introduction of the mouse. Some of the formal modeling work discussed under state-transition models should also be discussed in connection with engineering models. This work is oriented toward a certain methodology of systems engineering that advocates vigorous analysis as a part of design.

2.1.4. Models of Task and Context

Models of tasks have traditionally depended on task analytic methods to identify possible states that may occur as the user moves from initial conditions to goal conditions. In such problem spaces, tasks are modeled as Markovian processes or production systems and those models are then applied during the design process to develop specific dialogue procedures for state transitions. Specific task-analytic models have been developed for processes such as text editing, information retrieval, and programming. These approaches have been successful in guiding initial design iterations, but do not take into account the total context of problem situations. An alternative approach is to consider situated actions--organizational settings for HCI. Such a holistic approach is appealing because it unites user, machine, and problem situation within a larger social context. Cognitive artifact theory links machines and problem situation to users' cognitive processing of the overall environment. Artifacts critically determine how tasks can be accomplished and what it feels like to do them. Thus, artifact analysis links up system models (artifact features), user models (psychological consequences) and task analysis (the boundary conditions, or context of the relation). For example, a system that incorporates animated demos to support learning, embodies the claim that animated demos (artifact feature) provide models for learning-by-doing user activity (psychological consequence) in a learning situation in which the learner may be wondering what sorts of actions or activities are appropriate (task context). Discussion of social aspects of context are included in the CSCW section.

2.2. Research Directions

Theoretical work on HCI has produced state-transition oriented formal models, grammar techniques for describing dialogue, information processing oriented cognitive models, Fitts' Law analysis of input device operation, models of situated problems, and a number of others. It has been easiest to get models of small, relatively context-free operations (e.g., mouse movement), and most difficult to get models of extensive user knowledge. Presently, much more effort has gone into scientific models than to models oriented towards aiding the design process. Scientific models of HCI phenomena are important for insight and for codifying the results of the field. Experience suggests, however, that explicit attention to model-based techniques usable in design is also necessary if models are to be used directly in engineering.

An issue of debate in the HCI community is the amount of effort that should be given to pursuit of a unified model of HCI and to development of new models or extensions of existing models. A unified model of HCI must be robust enough to explain performance by wide ranges of users with varying physical, cognitive, and affective characteristics; to predict optimal interface characteristics for given user-problem situations; and to specify what knowledge of an interface is necessary for a user to successfully execute complex tasks. Development of unified models that explain all aspects of HCI phenomena is desirable, but may be premature given the current state of the science (although some work has begun on the integration of scientific models, such as GOMS/CCT, Soar, and TAG).

Development of specific models, for example, perceptual/cognitive models of display devices, are important for interface design but are not well-defined in current models. One approach is to develop proposals for the relationship between modeling the small pieces of interaction most current HCI models achieve and the larger context of the problem situation.

The workshop participants believe that design decisions based on engineering models for particular users of specific systems are needed. For example, it would be useful to have models that inform designers about certain aspects of I/O devices or interface styles such as: optimal design for specific users or tasks; interface features most critical to support in a user interface management system; and human physical or cognitive characteristics that can be accommodated with devices or styles. Two specific examples are engineering advances in the characterization of design spaces (e.g., similar to studies of local area network options and corresponding perfo mance) and in compositional analysis for 2D and 3D displays.

3. I/O devices and interaction styles

Research on human-computer interaction arises in the context of two powerful information processors (human and computer) attempting to communicate with each other via a narrow-bandwidth, highly constrained interface. Research in this area attempts to increase the useful bandwidth across that interface and make best use of what there is of it by designing faster modes of communication, by reducing the amount of information that must be transmitted, and by providing better human adaptability (both physical and cognitive) to a human/machine dialogue.

Technological development in the HCI field has tended to concentrate on the computer-to-user direction (output devices in computer science terms) resulting in much higher bandwidth from the computer to the user than from user to the computer (input). Research seeking to improve the user to computer side of the dialogue by finding ways to obtain data from the user conveniently and rapidly may have a large payoff. In this regard, it should be noted that input is the source of the single biggest success story for theoretically-based research in HCI--the mouse.

3.1. State of the art

The invention of the mouse, a cheap and reliable pointing device, has stimulated much work on graphical user interfaces (GUI). GUIs represent the current state of the art in human computer interfaces. Interfaces incorporating more advanced technology, such as data gloves, position sensors, eye trackers, etc., are in their infancy and availability is primarily limited to research laboratories. Moreover, the development of these interfaces has typically been triggered by the availability of the technology rather than by a recognized need for them. We wish to move from a "gadget-driven" paradigm to a "human-driven" one. The study of human responses in a particular user-communication setting should lead researchers to say, "we need a device to transduce this particular human action to computer input," and that kind of information should guide the deployment of efforts in input device research. The workshop participants believe research in input and output devices should start with studies of the characteristics of human communication channels and skills and proceed from there to the development of devices that communicate effectively to and from those channels.

The hardware necessary for this high-bandwidth communication is characterized by many degrees of freedom, probabilistic input, and high input data rates. Software should support real time response, input data reduction, parsing of gestures, and multiple simultaneous I/O paths. Embryonic hardware with all of these characteristics exists, but much further research is needed. The following section gives some specific examples.

3.2. Input Devices

- Three-dimensional pointing and manipulation: A magnetic tracker (e.g., Polhemus or Ascension) can provide the three-dimensional analogue of the mouse or data tablet. Other technologies, such as ultrasonic ranging, can also be used in this way. All of them are still limited in one way or another--in precision, stability, proximity to CRTs, or speed. The ideal solution may lie in sensors that observe the user, without requiring him/her to hold or wear anything.

- General gesture input: The three-dimensional mouse can be taken a step further. Rather than simply designating a location in three-space, it can allow a user to make natural, continuous gestures in space. Progress in this area requires not only a non-encumbering three-dimensional tracking technology but also a way to recognize human gestures, which occur dynamically with poor precision and repeatability.

- Simultaneous two-hand input: Aside from touch typing, computer input scenarios typically assume the use of only one hand at a time, though people are quite good at manipulating both hands, and even feet, in coordinated or different tasks. This area requires hardware that can be replicated (certain technologies permit tracking only one object at a time and are thus unsuitable) as well as research into interaction techniques and styles that make appropriate and advantageous use of both hands.

- Stereo display in conjunction with three-dimensional manipulation: A natural extension of three-dimensional tracking and gesturing is to allow the direct manipulation of virtual, displayed three-dimensional objects. This suggests stereoscopic display, but any perceivable display of depth might be adequate. The key is linking the three-dimensional input to the three-dimensional output in a faithful and convincing way.

- Virtual input devices: The notion of three-dimensional object manipulation can be extended to manipulating virtual tools, so that the user first obtains a tool (by three-dimensional direct manipulation) and then applies it to a three-dimensional virtual object. A virtual tool, can of course, metamorphose as needed for the job at hand and otherwise improve upon the properties of non-virtual tools.

- Speech: Speech input is a long standing area of research, and while progress is being made, it is slower than originally predicted, and further work remains in this field. Although the goal of continuous speech recognition remains elusive, unnatural, isolated-word speech is often appropriate and even natural for communicating with a computer, rather than another human. There is not a good understanding of how to use speech in an interface. There is not a comprehensive, principled framework for coordinating speech with other modes, such as manual input.

- Eye input: A promising area of investigation is eye movement-based input. This can be an unusually rapid and natural means of communication as well as a good example of the passive-sensing quality of a non-command interaction style. Eye-tracking technology is progressing slowly. It is still not adequate for use in practical applications for non-disabled users. An important consideration for eye movement input research is to make wise and effective use of eye movements, ideally in a non-command- based style. Eye movements, like other passive inputs (e.g., gesture, conversational speech) are often non-intentional or not conscious, so they must be interpreted carefully to avoid annoying the user with unwanted responses.

- Device taxonomy: Even without adding new devices to the range of available choices, there is considerable confusion among current position and gesture-sensing devices, touch tablets, joysticks, and mice. For example, sophisticated forms of joysticks are available for controlling six degrees of freedom, both isometric, isotonic, and in combinations. Software for using such input devices must typically be modified for each device. It might be possible to start with a taxonomy of these devices and design user interfaces with respect to the taxonomy and the user's task rather than specifically for each individual device.

The following two areas of research generated some concern in the workshop about their implications for privacy, and other user rights.

- Passive monitoring of user attitude: A user-computer dialogue could be improved if the computer knew the answer to such simple questions as, Is the user still sitting in her chair? Is he/she facing toward the computer? Is he/she using the telephone? Is another person present in the room?

- Further physiological measurements: In addition to three-dimensional position tracking and eye tracking, a variety of other physiological characteristics of the user might be monitored and the information obtained used to modify the user computer dialogue. Pupil diameter and galvanic skin response are examples of measurements that are relatively easy and comfortable to make, although their accurate instantaneous interpretation is an open question. A more difficult measure is an electro-encephalogram, although progress has been made in identifying specific evoked potential signals in real time.

3.3. Output Devices

- Graphics display packaging: The quest for a large, flat, high- resolution display screen continues, with limited success. Head-mounted displays with wide-angle optics can also provide some of the same benefits.

- Autostereoscopic displays: There are a variety of methods currently available for displaying stereo images, but most are too specialized for conventional day-to-day user-computer interaction. Graphic displays were once used only for specialized purposes. Now, they are used for routine data file manipulation. A high-quality, readily-convincing, easily fusible, high-resolution autostereoscopic (no glasses) display would make three-dimensional images similarly accessible to user interface designers.

- Touchable three-dimensional displays: A further improvement in the realism of a three-dimensional display would be to permit the user to reach out and feel the displayed object. Force and resistance feedback have particular power when combined with a convincing stereoscopic display.

- Techniques for representing complex abstract data: "Scientific visualization" issues involve both inventing new visualization techniques and understanding which techniques work best for which abstractions.

- Directional audio: It is possible to create the effect of a sound that comes from a specific location in three-space. Improvements in robustness, convenience, and user-independence are also possible.

- Force and tactile feedback: Improved force and tactile feedback may make computer-generated objects more realistic and allow the use of everyday motor skills with them.

3.4. Interaction Styles

Interaction styles include the display and selection mechanisms supported by an interface. Primary styles include: commands, form fill-ins, menus, and direct manipulation. GUIs are primarily based on direct manipulation styles that require direct selection of visual objects with immediate feedback. Touch panel and non-continuous speech interfaces are also based on direct manipulation styles. The development of advanced technology (e.g., data gloves, position sensors, etc.) has led to non-command interaction styles that support parallel human-computer exchanges along multiple channels.

3.5. Summary

Research in new input/output devices for interactive systems can lead to two kinds of benefits. The first is progress in our scientific understanding of human-computer communication: it can improve our understanding of the modes by which people can communicate with machines and lead to understanding of the characteristics of the devices, interaction techniques, and dialogue styles that allow us to exploit these modes. The second is in application of these results: it can lead to faster and more natural communication with computer systems, enable better quality and efficiency in the operation of such systems, and improve the working conditions of their users by providing them with richer and more natural means of communication.

4. Software Tools and Architecture

The tools available to the user interface architect include: pre-packaged interface objects (normally known as widgets), toolkits and interface builders which help the architect to assemble a collection of widgets into a coherent interface, and user interface management systems (UIMSs) which provide a run-time support environment for the user interface. Researchers are developing user interface design/development environments (UIDEs) which provide a more or less integrated set of tools, many of which are knowledge based. Examples of such tools include:

- tools to aid the designer in selecting widgets;
- tools for developing screen layouts and graphic presentations;
- tools to provide automatically generated help messages; and
- tools to support interface layout by demonstration.

Examples of current architectures include those provided by operating systems, window systems, object-oriented programming environments and UIMSs. One of the more interesting extensions to these traditional environments is the use of constraint systems to maintain relationships among interface objects and between interface and application objects.

4.1. Research Issues

Research issues for user interface tools will be discussed in three parts:

- (1) The interface of the future;
- (2) The interface development environment; and
- (3) The Interface Development Process.

4.1.1 The interface of the future

The form and content of computing tools can be under the direct control of their users. The end user "grabs" tools from the environment, combines them, and modifies their behavior in order to accomplish specific application tasks. These changes are not made through traditional programming activities, but through interactive techniques, much as users today build spreadsheet applications. Such malleable interfaces have two distinct advantages over the interfaces of today. First, they allow construction of the interface by the person who knows the application best--the user. Second, they support an optimal division of labor between users and programmers in co-development of systems where programming remains necessary.

The interface incorporates intelligent and unobtrusive processes, called "agents" and "critics". Agents off-load some of the tedious and repetitive work and allow the user to concentrate on the more imaginative and conceptual task level. Critics observe the user's behavior and can suggest more efficient uses of the interface.

The interface is context sensitive. This means it is both aware of its current state and also has knowledge of the current session history. It also has knowledge of both the application and the user. This context sensitivity allows the interface to provide reasoned and efficient guidance and help in learning and using the system.

4.1.2 The interface development environment

In this section tools for the designer and developer of user interfaces are discussed. There are a variety of ways in which these tools can help the developer.

4.1.2.1. Automatic generation of interfaces

Current research on user interface software is showing promising results in the automatic generation of user interfaces. This research explores a variety of possibilities such that, on the one hand, the interface is generated entirely from the source code of the application, and on the other hand, the interface is generated entirely from a demonstration of what it should look and act like without any specification of the code.

Some existing systems can generate a set of dialogue boxes and menus from a list of the commands supplied and data needed by the application program (e.g. Mickey). Others require a semantic specification of the interface in a special purpose specification language. Some sort of rule based system is then typically used to generate screen layouts, menu contents, etc., based on the code or specification.

An ideal tool would allow the designer to quickly draw a sketch of what the interface should look like and then demonstrate its behavior. The tool would learn from the demonstrations and generate an appropriate interface automatically.

Existing tools allow creation of some parts of the interface by drawing. Many tools (UIMX, NeXT interface builder and a host of others) allow designers to drag predefined widgets (such as menus, buttons, and text boxes) into place to create dialogue boxes and windows. These tools can generate program stubs for which application routines can be written. Other systems allow a limited form of actual drawing to create widgets (Peridot) and some graphical objects (Lapidary).

These tools are still very limited in scope. For example, no tool will allow text editing operations to be demonstrated or how commands can change graphic objects. Research on these tools can lead to expanding the range of interfaces that can be generated automatically and to learning how to improve their quality. The tools also need to be improved to make them easier to use and to facilitate construction by demonstration.

4.1.2.2. Easy modification of the interface design

Iterative design is important. But, current tools make it difficult to modify a design once it has been created. So, a research goal is to make it possible to easily and directly adjust the design in response to analysis and user testing of the interface or to extend the system. With many of the tools it is difficult to know what to modify to achieve a change in the interface since the tool has generated much of it automatically. Even with the more conventional tools, there are often complex interdependencies among various parts of the interface so that to change the behavior of one widget may require reprogramming a number of call-back procedures.

4.1.2.3. Tools for early parts of the design process

Most tools are basically code generators which are useful after a design has been specified. Tools are also needed for task analysis and to incorporate models into the design process. These tools can be connected to code generating tools analogous to those which currently exist so that the interface can be generated from the original requirements analyses, thus greatly facilitating iterative design.

4.1.2.4. Metrics

There is no generally accepted way of measuring the quality and usefulness of design tools. For example, Is the Macintosh toolbox easier to use than the Motif-toolkit? Some possible metrics to incorporate into user interface tool evaluation include: the time it takes to develop an interface; run time efficiency of the generated interface; and learning time for the tool. Measures of the quality of the resulting interface are also important.

4.1.2.5. Services beyond the interface design

Useful additions to developer's toolkits include items that help provide services to the end user that are frequently ancillary to the interface design itself. Some examples include: Help messages that are generated automatically; System documentation that is generated along with the interface and automatically modified when the interface is changed; Support for undo and redo capabilities; Evaluation tools to aid in analysis of designs and to measure effectiveness of the interface when it is used.

4.1.3. User Interface Development Processes

Increased usability in software systems is influenced by the user interface development process. A number of techniques and methods have been developed to support iterative design. These include: task analysis in early stages of design; design representation techniques (behavioral representations for the interface and constructional representations for the interface software); psychological design rationale for analyzing psychological claims embedded in interface designs; usability specifications; rapid prototyping; and formative evaluation (a means for testing as part of the development process).

Formal methods for a variety of steps in the design process are under development, including: new ways to get a deeper understanding of the user's tasks into the interface design, bringing task analysis fully into the interface development process, usability specification, rapid prototyping, and formative evaluation.

Specific suggested approaches to this research include field studies of real development projects, essential in defining these activities and the associated developers' roles. These studies can be used to determine, through observation and deduction, an operational model of what real interface developers presently do, as well as what they should be doing.

5. Computer Supported Cooperative Work (CSCW)

There is increasing recognition that the complexity of scientific inquiry requires collaborative efforts among scientists. Technology such as electronic mail, group writing tools, and large-screen displays play important roles in facilitating communication and data exchange among remote and local groups. Such support for communication and exchange are as vital as the computational and representation support computing provides for the conduct of science, and there are concerted efforts to coordinate remote use of scientific instruments and group communication and data exchange.

Collaboration in decision making and policy execution in government and industry has also given rise to varied CSCW efforts such as virtual meetings. CSCW tools for large-scale engineering projects are emerging. Tools to facilitate group learning have found application in K-12 as well as colleges and universities.

HCI is especially relevant to CSCW because interfaces are critical to the effectiveness, efficiency, and acceptance of technological solutions to the problems of collaboration. Conversely, CSCW has presented new challenges to HCI research because the multiplicity of users and machines involved complicates the design by demanding coordination as well as implementation of interfaces--adding new layers of bureaucracy to the overall system. Thus, it is no accident that there is a large overlap among the HCI and CSCW communities and HCI researchers have contributed to the growing collection of software known as "groupware". In a sense, a single user collaborates with a machine through the interface and CSCW extends the interface to multiple users and machines.

5.1. HCI Issues for CSCW

Two dimensions that are critical to CSCW are time and space--whether the work is being done synchronously or asynchronously, and whether the collaborators are working in the same location or remotely. These conditions affect the degree to which interfaces must be shared and temporally constrained. Shared interfaces offer less noisy communication channels and easier data exchange, but must be more robust in terms of computing platforms and individual user differences. Shared interfaces for local groups can take advantage of wider single communication channels or multiple media since they can be augmented by the high fidelity of humans working within direct sensory proximity. Interfaces that facilitate asynchronous communication can take advantage of time delays to add computational value to messages or data by prioritizing, reformatting, or filtering mechanisms, but lack the immediate feedback synchronous interfaces allow.

A central principle in developing shared interfaces for CSCW is the notion of structured messages--interface strategies for standardizing and controlling information in CSCW. System interfaces provide users with facilities to label information according to structural categories or to form hypermedia links among other information structures. These categories are then used to improve dissemination and retrieval by the group. Implementations use various menu designs, form fill-ins, or direct manipulation techniques but the critical design decisions must be based on analysis and understanding of the group work tasks. Structured messages play key interface roles in special-purpose systems such as the Virtual Notebook for medical researchers and gIBIS for software designers and in generic systems such as the Object Lens. Mapping complex tasks to message structures is difficult in its own right, but the success of such interfaces depends on multiple users assigning appropriate labels to the information they create or use. Throughout this document it is suggested that interfaces must provide intelligent assistance to users and much research remains to be done on this problem.

The social nature of activities supported by CSCW places requirements on interface designers to provide very high bandwidth for this social communication. The common approach to this challenge is to create interfaces that employ multiple channels for supporting CSCW. Research platforms such as Video walls (e.g., Colab), virtual hallways (e.g., Cruiser), and various electronic meeting rooms (e.g., Capture Lab, etc.) illustrate just how complex and parallel human interactions are, and illustrate the need for interfaces that allow collaborators to both project and perceive emotive as well as cognitive information. Such interfaces necessarily involve parallel input and output devices and new paradigms for polylog management. In this regard, fidelity in communication channels that support CSCW has been improved by interfaces such as drawing surfaces, telepointers, and corporate memory storage and retrieval mechanisms.

Interfaces for CSCW have also illustrated a number of additional issues that motivate basic research in coordination theory and complicate continued work in HCI itself. Such issues relate to ergonomics, privacy, new standards of behavior, standards of representation, intellectual property rights, and distributed cognition.

5.2. Research Directions

Progress in developing interfaces to support CSCW is complicated by the need to share not only the content that scientists or decision makers are interested in but also the interfaces themselves. As science increasingly depends on high-performance computing it will concomitantly require communication tools that allow sharing and discussion in real time, three-dimensional motion images, highly structured data sets and maps, real-time sketches, and voice or text information. Given the bandwidth of future networks, interfaces to create, retrieve, send, display, edit, discuss, and store such information must be designed and tested. Interfaces are needed: that allow users to work in parallel without being overloaded or confused; for managing multiple inputs and outputs and facilitating human turn taking exchanges; that seamlessly integrate multiple media; that provide easily manipulatable and varied levels of private and public broadcasts; and that take advantage of human communication channels such as gestures, eye movements, and voice intonation.

6. Infrastructure for HCI Research and Development

All scientific fields require a variety of support systems that facilitate research. These include: instruments and tools for conducting research; professional societies; journals; curriculums rooted in the principles and techniques of the field; and mechanisms for attracting new scientists to the discipline. HCI infrastructure development is complicated by the fact that the best research in HCI is highly interdisciplinary. It does not fall directly within the current disciplinary structure of Computer Science or Psychology, but right at the boundary between the two. Institutional pressures tend to lead researchers toward the centers of their disciplines, not the edges. For example, it is surely easier to achieve tenure in a department by doing work that is close to the center of its discipline.

Base-level support for HCI research includes workstations and some funds for software, not differing from other areas of Computer Science. But many projects require significant investment in equipment beyond this base. Examples of technologies requiring special equipment that will play an increasing role in HCI research include virtual reality, teleoperations, digital photography, and hypermedia. In addition, empirical user studies often require special facilities for collecting data from groups of participants simultaneously, so that multiple stations and video recording are needed.

Acquiring staff support to operate, maintain, and adapt equipment and software is expensive and requires a commitment of continuity of employment. Universities fund staff support positions for major externally-funded projects, but not usually for small ones, such as have been typical in HCI.

Many scholarly journals serve the HCI field and new journals have appeared in the past few years, including a new abstracting publication. Likewise, professional societies such as SIGCHI hold well-attended conferences that include a mix of theory and practice.

The ACM SIGCHI curriculum recommendations call for single course offerings at the undergraduate and graduate levels. Introducing ideas and methods from HCI into undergraduate introductory courses is another idea. Yet another possibility is joint curriculum development with other areas of CS, in order to help integrate HCI into the overall field. For example, HCI material can be used in database and software engineering courses. Useful material exists in videos produced by SIGGRAPH and SIGCHI, but is not well packaged for instructional use. Finally, no widely accepted textbook has emerged.

6.1. Leveraging Infrastructure Resources

There are opportunities for resource sharing between HCI projects. Staffing needs could be reduced to some extent, and overall research productivity increased, by pooling some kinds of work across projects. For example, research in I/O devices requires that prototype devices be designed and built; one laboratory could do these jobs and make multiples of the devices available to other groups. As another example, researchers doing cognitive modelling could share models more effectively if resources were available to produce distribution versions of software and provide some level of consulting and support.

"Virtual center" project structures could provide useful support for HCI research.

Such support could come from the establishment of large research centers,such as NSF Engineering Research Centers and Science and Technology Centers. Hub projects within virtual centers could provide infrastructure support. Virtual centers could include one or more hub projects funded at a higher level than other projects. Such projects would undertake to perform hardware and software development as required for the center as a whole, to distribute devices and software. The larger funding for hub projects could be used to leverage university staff support at the hub sites.

Smaller, more flexible, funding arrangements would also be important for resource sharing. One such model of sharing is exemplified by the community of researchers who have developed the Soar cognitive model. They have developed a kind of "virtual center" in which a high degree of sharing of work, information, and ideas has been achieved without any collective funding arrangement. Soar researchers raise funds separately for their own projects but then participate in well-organized community-wide workshops and share software development.

7. Summary

This report summarizes the state of the art and identifies research directions in HCI.

A variety of scientific and development models of users, systems, and tasks have been developed and extended. These models serve as the basis for a cumulative science of HCI. Progress toward integrating some of the scientific models has begun and payoffs can be expected from the development of design and engineering models or taxonomies based on scientific models.

Many devices for input (e.g., gesture, speech, gaze, etc.) and output (e.g., autostereoscopic displays, 3D displays, etc.) have been developed. Progress has also been made in visualization techniques and the development of alternative modes of interactions such as directional audio and force and tactile feedback. What remains problematic is mapping particular devices to tasks and users and the integration of devices and non-command interaction styles. Strategies for natural integration of parallel I/O devices are also required.

Toolkits and user interface management systems have been developed to support interface designers. New research aims to discover ways to support end user programming and to develop higher levels of software/interface abstraction. It is hoped this will lead to broader reusability of interface code. Progress is also being made toward improving the overall development process (e.g., iterative refinement, usability specifications, evaluation).

HCI research has supported CSCW through the development of interfaces for structured messages, development of real-time interfaces, and the definition of shared standards and local customizable agents. CSCW and HCI also share theoretical models for social context vis-a-vis human-computer interactions.

Recommendations for improving the infrastructure to support HCI research and education include ways to share resources and expertise in order to: develop laboratories with special equipment and staff; foster cooperation and support from professional organizations; and develop curriculums and programs for HCI education.

Bibliography

The following sources give additional details of research and development discussed in this report. Citations are organized by the four main topics areas of HCI research around which the workshop was organized. The bibliography is not meant to be exhaustive. For a comprehensive bibliography of HCI literature, see The HCI Bibliography Project directed by Gary Perlman. This bibliography is available on-line. To access the bibliography, send electronic mail to hcibib@cis.ohio-state.edu. To register as a user, the first line of the message should be Register: followed by your electronic mail address and additional lines with your name and address. The receive documentation or an index, the first line of the message should be send: readme or send: index.

HCI Models

Anderson (1983). The Architecture of Cognition. Cambridge, Massachusetts: Harvard University Press.

Anderson (1987). Skill acquisition: Compilation of weak-method solutions. Psychological Review 94, 192-211.

Atwood, M. E., Gray, W. D., & John, B. E. (in preparation). Project Ernestine: From basic research to application and back again, and again, and again. NYNEX Science and Technology Center, White Plains, New York.

Barnard, P. (1987). Cognitive resources and the learning of human computer interaction. In J. M. Carroll (Ed.), Interfacing Thought: Cognitive Aspects of Human-Computer Interaction, 112-158. Cambridge, Massachusetts: MIT Press.

Bennett, J., Lorch, D., Kieras, D. E., & Polson, P. G. (1987). Developing a user interface technology for use in industry. Proceedings of INTERACT 87, Second IFIP Conference on Human-Computer Interaction, 21-26. Amsterdam: Elsevier.

Berry, G. (1984). The ESTEREL synchronous programming language and its mathematical semantics. Proceedings of the NSF/SERC Workshop on Concurrency, CMU, 1984.

Boff, K. R., Kaufman, L., & Thomas, J. P. (1986). Handbook of Perception and Human Performance. New York: Wiley.

Bovair, S., Kieras, D. E., & Polson, P. G. (1990). The acquisition and performance of text-editing skill: A cognitive complexity analysis. Human-Computer Interaction 5, 1-48.

Buxton, W. (1983). Lexical and pragmatic considerations of input structures. Computer Graphics 17(1), 31-37.

Card, S. K. (1989). Theory-driven design research. In McMillan, G. R., Beevis, D., Salas, E., Strub, M. H., Sutton, R. & Van Breda, L. Applications of Human Performance Models to System Design. New York: Plenum Press.

Card, S. K. (1990). Modeling scenarios for action. In Elkind, J. I. , Card, S. K., Hochberg, J., & Huey, B. M. (Eds.), Human Performance Models for Computer-Aided Engineering., Boston: Academic Press, Inc., 1990.

Card, S. K., English, W. K. & Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT. Ergonomics 21, 601-613.

Card, S. K., Moran, T. P., & Newell (1980a). Computer text-editing. An information-processing analysis of a routine cognitive skill. Cognitive Psychology 12, 32-74.

Card, S. K., Moran, T. P., & Newell (1980b). The Keystroke-Level Model for user performance time with interactive systems. Communications of the ACM

23, 396-410.

Card, S. K., Moran, T. P., and Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, N. J.: Lawrence Erlbaum Publishing.

Cardlli, L. & Pike, R. (1985). Squeak: A language for communicating with mice. Computer Graphics 19(3), 199-204.

Carroll, J. M. (1989). Evaluation, description and invention: Paradigms for Human-Computer Interaction. In Yovits, M. C. (Ed.), Advances in Computers 29, 47-77.

Corker, K., Davis, L., Papazian, B., & Pew, R. (1986). Development of an advanced task analysis methodology and demonstration for Army-NASA Aircrew/Aircraft Integration.

Coutez, J. (1987). PAC, an object oriented model for dialgoue design. In H. J. Bullinger & B. Shackel (Eds.), Proceedings of INTERACT 87, Amsterdam: North-Holland, 431-436.

Doane, S. M., Pellagrino, J. W., & Klatzky, R. L. (1990). Expertise in a computer operating system: Conceptualization and performance. Human-Computer Interaction 5(2&3), 267-304.

Drury, C. (1975). Application of Fitts' law to foot-pedal design. Human Factors 17, 368-373.

Durrett, H. J. (1987). Color and the Computer. Boston: Academic Press, Inc.

Egan, D. E., Bowers, C., & Gomez (1982). Acquiring text editing skills: Learner characteristics that predict success in using a text-editor tutorial. Proceedings of the Conference on Human Factors in Computing Systems, 337-3.

Elkind, J. I., Card, S. K., Hochberg, J., & Huey, B. M. (1990). Human Performance Models for Computer-Aided Enginnering. Boston: Academic Press, Inc.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. Journal of Experimental

Psychology 47, 381-391.

Foley, J. D. Wallace, V. L., & Chan, P. (1984). The human factors of computer graphics interaction techniques. IEEE Computer Graphics & Applications 4(11), 13-48.

Gasen, J.B. , Baecker, R., Card, S. K., Carey, T., Perlman, G., Strong, G., & Verplank, W. (1991). Designing Curricula in Human-Computer Interaction. Tutorial Notes, CHI 91, (New Orleans, LA, April 29, 1991).

Gomez, L. M., Egan, D. E., Wheeler, E. A., Sharma, D. K., & Gruchacz, A. M. (1983). How interface design determines who has difficulty learning to use a text editor. Proceedings of the CHI'83 Conference on Human Factors in Computing Systems, 176-181. New York: ACM Press.

Hill, Ralph D. (1987). Event-response systems: A technique for specifying multi-threaded dialogues. Proceedings of CHI+GI 1987 (Toronto, April 5-7). New York: ACM Press, pp. 241-248.

Hoare, C. A. R. (1978). Communicating sequential processes. Communications of the ACM 21(8), 666-678.

Howes, A. & Payne, S. J. (1991). Display-based competence: Towards user models for menu-driven interfaces. University Lancaster.

Howes, A. & Young, R. M. (1991). Predicting the learnability of task-action mappings. Proceedings of the CHI 91 ACM Conference on Human Factors in Software, 113-118. New York: ACM Press.

Kieras, D. E. (1989). Towards a practical GOMS model methodology for user interface design. In, Helander (ed.), Handbook of Human-Computer Interaction. Amsterdam: Elsevier.

Kieras, D. & Polson, P. G. (1985). An approach to the formal analysis of user complexity. International Journal of Man-Machine Studies 22, 365-394.

Lerch, F. J. (1988). Computerized Financial Planning: discovering Cognitive Difficulties in Model Building. Unpublished doctoral dissertation, University of Michigan, Ann Arbor, Michigan.

Lerch, F. J., Mantei, M. M., & Olson, J. R. (1989). Translating ideas into action: cognitive analysis of errors in spreadsheet formulas. Proceedings of the CHI T89 Conference on Human Factors in Computing Systems, 121-126. New York: ACM Press.

Lewis, C., Polson, P., Wharton, C., & Rieman, J. (1990). Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. Proceedings of the CHI '90 ACM Conference on Human Factors in Software, 235-241. New York: ACM Press.

Jagacinsky, R. J. & Monk, D. L. (1985). Fitts' law in two dimensions with hand and head movements. Journal of Motor Behavior 17, 77-95.

John, B. E., Rosenbloom, P. S., & Newell, A. (1985). A theory of stimulus-response compatibility applied to human-computer interaction. Proceedings of the CHI 85 Conference on Human Factors in Computing Systems, 213-219. New York: ACM Press.

John, B. E., Vera, A. H., & Newell, A. (1991). Towards real-time GOMS. Poster at CHI 91, ACM Conference on Human Factors in Computer Systems (New Orleans, April 28-May 2, 1991). New York: ACM Press.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. Artificial Intelligence 33, 1-64.

Lee, A. Y. Polson, P. G., & Bailey, W. A. (1989). Learning and Transfer of Measurement Tasks. Proceedings of the CHI 89 Conference on Human Factors in Computing Systems, 115-120. New York: ACM Press.

MacKenzie, I. S. (1991). Fitts' Law as a Performance Model in Human-Computer Interaction. Unpublished PhD thesis, Department of Education, University of Toronto. Toronto, Canada.

Mackinlay, J, Card, S. K., & Robertson, G. G. (1990). A semantic analysis of the design space of input devices. Human-Computer Interaction 5(2&3), 145-190.

Milner, R. (1980). A calculus of communicating systems. Lecture Notes in Computer Science, no.92. Springer-Verlag.

Milner, R. (1982). Four combinators for concurrency. ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing., Ottawa, Canada.

Moran, T. P. (1981). The Command Language Grammar: A representation for the user interface of interactive computer systems. International Journal of Man-Machine Studies 15, 3-50.

Moran, T. P. (1983). Getting into a system: External-internal task mapping analysis. Proc. CHI 83 Human Factors in Computing Systems. New York: ACM Press.

Newman, W. M. (1968). A system for interactive graphical programming. Spring Joint Computer Conference, AFIPS Press.

Norman, D. A. (1981). Norman, D. A. (1981). Categorization of action slips. Psychological Review 88, 1-15.

Norman, D. A. (1984). Four stages of user activities. In Shackel, B. (ed.). Proceedings of INTERACT '84, IFIP Conference on Human-Computer Interaction. Amsterdam: Elsevier Science Publishers, B. V. (North-Holland).

Olsen, D. R., Jr. (1990). Propositional production systems for dialog description. Proceedings of the CHI 90 Conference on Human Factors in Computing Systems, 57-63. New York: ACM Press.

Olson, J. R. & Olson, G. M. (1990). The growth of cognitive modeling in human-computer interaction since GOMS. Human-Computer Interaction 3(2&3), 221-263.

Payne, S. J. (1984). Task-Action Grammars. Human-Computer Shackel, B. (ed.), INTERACT 84, Proceedings of the IFIP Conference on Human-Computer Interaction, pp. 527-532. Amsterdam: Elsevier Science Publishers, B. V. (North Holland).

Payne, S. J. & Green, T. R. G. (1986). Task-Action Grammars: A model of mental representation of task languages. Human-Computer Interaction 2(2): 93-133.

Polson, P. G., & Kieras, D. E. (1985). A quantitative model of the learning and performance of text editing knowledge. Proceedings of the CHI '85 Conference on Human Factors in Computing Systems, 202-212. New York: ACM Press.

Polson, P. G., & Lewis, C.H. (1990). Theory-based design for easily learned interfaces. Human-Computer Interaction 5(2&3), 191-220.

Polson, P. G., Muncher, E., & Engelbeck, G. (1986). A text of a common elements theory of transfer. Proceedings of the CHI T86 Conference on Human Factors in Computing Systems, 78-834. New York: ACM Press.

Olson, J. R., & Nilsen, E. (1988). Analysis of the cognition involved in spreadsheet software interaction. Human-Computer Interaction 3, 309-350.

Reisner, P. (1977). Use of psychological experimentation as an aid to development of a query language. IEEE Transactions on Software Engineering SE-3(3), May 1977, 218-229.

Reisner, P. (1981). Formal grammar and human factors design of an interactive graphics system. IEEE Transactions on Software Engineering SE-7(2), March 1981, 229-240.

Reisner. P. (1984). Formal grammar as a tools for analyzing ease of use: Some fundamental concepts. In Thomas, J. C. & Schneider, M. L. (Eds.), Human Factors in Computing Systems, Norwood, New Jersey: Ablex Publishing Corp.

Schiele, F. & Green, T. HCI formalisms and cognitive psychology: the case of Task-Action Grammar. In Harrison, M. & Thmbleby, H. (1990). Formal Methods in Human-Computer Interaction, 9-62.

Singley, M. K., & Anderson, J. R. (1988). A keystroke analysis of learning and transfer in text editing. Human-Computer Interaction 3, 223-274.

Shneiderman, B. (1982). Multiparty grammars and related features for defining interaction systems. IEEE Transactions on Systems, Man and Cybernetics 12(2, March-April), 148-154.

Tauber, M. J. (1988). On mental models and the user interface. In Van der Veer, G. C., Green, T. R. G., Hoc, J.M., & Murray, D. M. (eds.), Working with Computers: Theory vs. Outcome. London: Academic Press, 89-119.

van den Bos, J., Plasmeijer, M. J., & Hartel, P. H. (1983). Input-output tools: A language facility for interactive and real-time systems. IEEE Transactions on Software Engineering, SE-9(3), 247-259.

Welford , A. T. (1968). Fundamentals of Skill. London: Methuen.

Young, R. M., Green, T. G. R., & Simon, T. (1989). Programmable user models for predictive evaluation of interface designs. In K. Bice & C. Lewis (Eds.), Proceedings of HCI '89 Human Factors in Computing Systems, 15-19. New York: ACM Press.

I/O Devices

Bleser, T.W. and Sibert, J. "Toto: A Tool for Selecting Interaction Techniques, (1990). Proc. ACM SIGGRAPH Symposium on user Interface Software and Technology, pp. 135-142, Snowbird, Utah.

Card, S., English, D. & Burr, B. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a CRT, Ergonomics, vol. 21, pp. 601-613.

Card, S., Mackinlay, J. & Robertson, G. (1990). The Design Space of Input Devices, Proc. ACM CHI'90 Human Factors in Computing Systems Conference, pp. 117-124.

Foley, J.D., van Dam, A., Feiner, S. & Hughes, J. (1990) Computer Graphics: Principles and Practice, Addison-Wesley, Reading, Mass.

Jacob, R.J.K. (1990). What You Look At is What You Get: Eye Movement-Based Interaction Techniques, Proc. ACM CHI'90 Human Factors in Computing Systems Conference, pp. 11-18.

Ledgard, H., Whiteside, J.A., Singer, A. & Seymour W. (1980). The Natural Language of Interactive Systems, Comm. of the ACM, vol. 23, pp. 556-563.

Lewis, G.W., Trejo, L.J., Nunez, P., Weinberg, H. & Naitoh, P. (1987). Evoked Neuromagnetic Fields: Implications for Indexing Performance, Biomagnetism 1987, Proc. of the Sixth International Conference on Biomagnetism, Tokyo.

Minsky, M., Ouh-young, M., Steele, O., Brooks,F.P. & Behensky, M. (1990). Feeling and Seeing: Issues in Force Display, Computer Graphics, vol. 24, no. 2, pp. 235-244.

Nielsen, J. (1990) Trip Report: CHI'90, SIGCHI Bulletin, vol. 22, no. 2, pp. 20-25.

Tufte, E.R. (1989). Visual Design of the User Interface, IBM Corporation, Armonk, N.Y., 1989.

Wenzel, E.M., Wightman, F.S. & Foster, S.H. (1988). A Virtual Display System for Conveying Three-dimensional Acoustic Information, Proc. 32nd Annual Meeting of the Human Factors Society, pp. 86-90, Anaheim, Calif.

Wickens, C., Kramer, A., Vanasse, L. & Donchin, E. (1983). Performance of Concurrent Tasks: A Psychophysiological Analysis of the Reciprocity of Information-Processing Resources, Science, vol. 221, pp. 1080-1082.

Tools and Architectures

Bennett, W. E., Boise, S. J., Gould, J. D., Green, S. L., & Wiecha, C. F. (1989). Transformations on a dialog tree: Rule-based mapping of content to style. Proceedings of the ACM SIGGRAPH Symposium on User Interface Software and Technology (Williamsburg, VA, November 1989), pp. 67-75.

Borning, A. (1979). Thinglab: A constraint-oriented simulation laboratory. Xerox Palo Alto Research Center Technical Report SSL-79-3, July, 1979.

Borning, A.,The Programming Language Aspects of ThingLab, a Constraint-Oriented Simulation Laboratory, ACM Transactions on Programming Languages and Systems, October 1981, pp. 353-387.

Borning, A., and Duisberg, R., Constraint-Based Tools for Building User Interfaces, ACM Transactions on Graphics, October 1986, pp. 345-374.

Card, S. K., Moran, T. P., & Newell, A. (1983) The psychology of human-computer interaction. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cardelli, L. (1988). Building User Interfaces by Direct Manipulation. Proceedings of the ACM SIGGRAPH Symposium on User Interface Software and Technology: UIST'88. (Banff, Alberta, Canada. Oct. 17-19, 1988). pp. 152-166.

Crowley, T., Milazzo, P., Baker, E., Forsdick, H., and Tomlinson, R. (1990). MMConf: An infrastructure for building shared multimedia applications.

Proceedings of CSCW 90 , (Los Angeles, October 7-10, 1990), pp. 329 - 342.

Fischer, G. & Girgensohn, A. (1990). End-user modifiability in design environments. Proceedings of CHI 90 (Seattle, April 1-5, 1990), pp. 183-191.

Foley, James. Transformations on a Formal Specification of User-Computer Interfaces, Computer Graphics. April, 1987, Vol 21, No, 2, pp. 109-112.

Freeman-Benson, B., Maloney, J., and Borning, A., An Incremental Constraint Solver, Communications of the ACM, January 1990, pp. 54-63.

Gillian, D. J., Holden, K., Adam, S., Rudisill, M., & Magee, L. How does Fitts' Law fit pointing and dragging? Proceedings of CHI90 (Seattle, April 1-5, 1990), pp. 227-234.

Hudson, S.E., Incremental Attribute Evaluation: A Flexible Algorithm for Lazy Update, (to appear) ACM Transactions on Programming Languages and Systems.

Hudson, S.E., An Enhanced Spreadsheet Model for User Interface Specification, University of Arizona Technical Report TR 90-33, October 1989.

John, B.E. (1990) Extensions of GOMS analyses to expert performance requiring perception of dynamic and auditory information. In Proceedings of CHI, 1990 (Seattle, Washington, April 30-May4, 1990) ACM, New York, 107-115.

Lai, K., Malone, T., and Yu, K.., Object Lens: A "Spreadsheet" for Cooperative Work, ACM Transactions on Office Information Systems, October 1988, pp. 332-354.

Leler, W., Constraint Programming Languages: Their Specification and Generation, Addison-Wesley, 1988.

MacLean, A., Carter, K., Lovstrand, L. and Moran, T., User-Tailorable Systems: Pressing the Issues with Buttons, Proceedings of CHI '90, Seattle, April 1990, pp. 175-182.

Maulsby, D.L., Kittlitz, K.A., and Witten, I.H., Metamouse: Specifying Graphical Procedures by Example, Proccedings of SIGGRAPH '89, Boston, July 1989, pp. 127-136.

Myers, B.A., Visual Programming, Programming by Example, and Program Visualization: A Taxonomy, Proceedings of CHI '86, Boston, April 1986, pp. 59-66.

Myers, B.A., Creating Dynamic Interaction Techniques by Demonstration. Proceedings of CHI '87, Toronto, pp. 271-278.

Myers, B.A. (1988). Creating User Interfaces by Demonstration. Boston: Academic Press

Myers, B.A., Vander Zanden, B. & Dannenberg, R.B. (1989). Creating Graphical Interactive Application Objects by Demonstration. Proceedings of the ACM SIGGRAPH Symposium on User InterfaceSoftware and Technology: UIST'89. Williamsburg, VA Nov. 13-15, 1989. pp. 95-104.

Nardi, B. & Miller, J.R. (1990). An ethnographic study of distributed problem solving in spreadsheet development. Proceedings of CSCW90, (Los Angeles, October 7-10, 1990), pp. 197-208.

Olsen, D.R. and Halverson, B.W., (1988). Interface Usage Measurements in a User Interface management System, Proceedings of ACM SIGGRAPH Symposium on User Interface Software, Oct 1988.pp. 102-108.

Olsen, D.R. (1989). A Programming Language Basis for User Interface Management. Proceedings of CHI '89,(Austin, TX, April 30-May 4, 1989), pp 171-176.

Olsen, D. R. (1990). Propositional production systems for dialog description. Proceedings of CHI 90 (Seattle, April 1-5, 1990), pp. 57-63.

Rosson, M.B., Carroll, J., M., and Bellamy, R., K., E., Smalltalk Scoffolding: A Case Study of Minimalist Instruction, Proceedings of CHI '90, Seattle, April 1990, pp. 423-429.

Szekely P. A., and Myers, B. A., A User Interface Toolkit Based on Graphical Objects and Constraints. Proceedings of the ACM Conference on Object-Oriented Programming, Systems Languages and Applications, Nov.

1988, pp. 36-45.

Wiecha, C. & Boies, S. (1990). Generating UserInterfaces: Principles and use of ITS Style Rules, Proceedings of the ACM SIGGRAPH Symposium on User Interface Software and Technology: UIST'90. Snowbird, Utah. Oct. 3-5, 1990. pp. 21-30.

CSCW

Conklin, J., & Begeman, M. (1988). gIBIS: A hypertext design tool for exploratory policy. In Proceedings of the 1988 Conference on Computer Supported Cooperative Work, (Portland, Oregon, Sept. 26-28, 1988). NY: ACM. p. 140-152.

Elwart-Keys, M., Halonen, D., Horton, M., Kass, R., & Scott, P. (1990). User interface requirements for face to face groupware. In Proceedings of CHI, 1990 (Seattle, Washington, April 1-5, 1990). NY: ACM. p.295-301.

Fish, R.S., Kraut, R.E. & Chalfonte, B.L. (1990). The VideoWindow system in information communication. In Proceedings of the Conference on Computer-Supported Collaborative Work, (Los Angeles, Oct. 7-10, 1990). p. 1-12.

Gorry, G.A., Burger, A.M., Chaney, R.J., Long, K.B. & Tausk, C.M. (1988). Computer support for biomedical work groups. In Proceedings of the 1988 Conference on Computer Supported Cooperative Work, (Portland, Oregon, Sept. 26-28, 1988). NY: ACM. p.39-51.

Grudin, J. (1991). CSCW: The convergence of two development contexts. In Proceedings of CHI, 1991 (New Orleans, April 27-May 2, 1991). NY: ACM. p. 91-98.

Hiltz, S.R. (1988). Collaborative learning in a virtual classroom. In Proceedings of the 1988 Conference on Computer Supported Cooperative Work, (Portland, Oregon, Sept. 26-28, 1988). NY: ACM. p.282-290.

Kraut, R., Egido, C., & Galegher, J. (1988). Patterns of contact and communication in scientific research collaboration. In Proceedings of the 1988 Conference on Computer Supported Cooperative Work, (Portland, Oregon, Sept. 26-28, 1988). NY: ACM. p. 1-12.

Malone, T.W. & Crowston, K. (1990). What is coordination theory and how can it help design cooperative work systems? In Proceedings of the Conference on Computer-Supported Collaborative Work, (Los Angeles, Oct. 7-10, 1990). p. 357-370.

Malone, T.W., Grant, K.R., Lai, K.Y., Rao, R. & Rosenblitt, D.R. (1987). Semistructured messages are surprisingly useful for computer-supported coordination. ACM Transactions of Office Information Systems, 5(2), 115-131.

Mantei, M. (1988). Capturing the Capture Lab concepts: A case study in the design of computer supported meeting environments. In Proceedings of the 1988 Conference on Computer Supported Cooperative Work, (Portland, Oregon, Sept. 26-28, 1988). NY: ACM. p.257-270.

Neuwirth, C.M., Kaufer, D.S., Chandhok, R. & Morris, J.H. (1990). Issues in the design of computer support for co-authoring and commenting. In Proceedings of the Conference on Computer-Supported Collaborative Work, (Los Angeles, Oct. 7-10, 1990). p.183-195.

Root, R.W. (1988). Design of a Multi-media vehicle for social browsing. In Proceedings of the 1988 Conference on Computer Supported Cooperative Work, (Portland, Oregon, Sept. 26-28, 1988). NY: ACM. p. 25-38.

Stefek, M., Foster, G., Bobrow, D., Kahn, K., Lanning, S. & Suchman, L. (1988). Beyond the chalkboard: Computer supported collaboration and problem solving in meetings. In I. Greif, (Ed.), Computer-supported cooperative work: A book of readings, p. 335-366. Morgan Kaufmann.

Workshop on Hyperbase Systems

May 1992

Executive summary

Advanced information management systems of the future will not resemble the systems of today. Ubiquitous access to high-speed computer networks will lead to massive data and process distribution and entirely new forms of collaborative work processes. New media and new data types will revolutionize computing environments and require petabyte storage capabilities. Even today we see specifications for new information management systems that call for terabyte storage of multimedia data and national network access. Future digital electronic libraries will allow scholarly access to the world's literature through local, national and global networks.

Requirements for advanced information management applications are overwhelming existing database theory and practice. Current architectures of database management systems will not scaleto terabytes. Existing data models and database technology provide inadequate support for these new environments, and, as a result, hyperbases (hypermedia databases) are emerging as an important new area of research. We are currently at a crucial point in the development of hyperbase systems. Fundamentally new models, architectures, data representations and algorithms are required. Although several researchers have designed new data models and architectures for hyperbases, the field has reached consensus on neither.

This report presents the results of a workshop held to define the state of the art and to identify research directions for the hyperbase community. The workshop brought together 24 participants from the areas of hypermedia, database, collaborative and information retrieval systems for an intensive two-day workshop. The workshop was held on October 15 and 16, 1992 in Washington, D.C. at the Vista hotel. This report identifies critical research issues facing hyperbase system researchers, current progress on the issues, potential methods of approach (including literature citations) and a general research agenda for the field.

The hyperbase system architectures discussed at the workshop can be characterized as being large-scale (tens of terabytes), open, distributed, heterogeneous and extensible with capabilities for computation. Typical application environments discussed were archival electronic libraries, large-scale collaborative systems and large engineering enterprises. The workshop concentrated on hyperbases and did not include hypermedia environments or user-interface issues except as they might impact the hyperbase. The participants considered the following six key topical areas: models and architectures; node, link, and structure management (data and metadata management); browsing/search and query; version control; concurrency control, transaction management, and notification control; impacts of application environments on hyperbases; and, core functionality of a hyperbase management system. This report contains the summary of the discussion of the issues in each of the six key topical areas.

The workshop report will be available in Spring, 1993 from the Hypermedia Research Laboratory at Texas A&M University.

Workshop on Hypermedia and Hypertext Standards 22nd-23rd April Amsterdam

A major international workshop to discuss hypermedia standards issues will take place 22nd-23rd April 1993 at the World Trade Centre, Amsterdam.

The workshop is sponsored by the Commission of the European Communities (DG XIII) as part of the IMPACT programme for the development of the European information market - Open Information Interchange initiative. OII seeks to improve awareness of information coding standards and promote their use in the electronic-services market.

Sessions will be devoted to structure standards (Hytime, MHEG, AVI, ODA-hyper extensions) and content standards (MPEG,JPEG,JBIG,GDID,IIF,etc.) as well as applications.

Further details and registration forms can be obtained from the organisers: Johan van Halm, c/o HYPWKS, PO Box 688, 3800 AR Amersfoort, Netherlands. fax +31 33 650945

Information Access and the Networks - A Research Workshop

(held after SIGIR'93)

July 1, 1993 --- Pittsburgh, PA

Information access involving networks is one of the most important growth areas in the broad field of information technology. User demand has caused a massive growth in the size of the Internet and the amount of information accessible on it. This has made it imperative to develop new technologies to provide an infrastructure for this new information space. This workshop will bring together researchers and developers to discuss network-based information services in both general and specific terms, drawing upon those attending ACM SIGIR '93 and the communities related to:

* Archie
* Digital Libraries

* Gopher
* HyperPage
* Knowbots
* NCSA Mosaic
* NNTP/Usenet
* Prospero
* Veronica
* WAIS
* WWW (WorldWideWeb)
* Z39.50

along with other related initiatives, protocols, systems, and services.

Attendees will include persons with an active interest in network-based information systems and relevant research.

Presentations will include design and implementation of client-server architectures, resource discovery, distributed file systems and services, development of integrated network tools, and innovative applications of more basic technologies, such as bulletin board systems and electronic mail.

The workshop format involves a plenary session from 8:30-11:30 a.m. with invited speakers, each of them addressing a major theme in the area of wide-area networking. After some discussion, the workshop will break up into groups, starting with a working lunch. The organization of these groups and assignment of participants to them will be based on a short interest statement provided by each participant.

The workshop will conclude with brief summary presentations by group leaders. Notes from the plenary talks and interest statements will be provided to attendees and made available for on-line access after the workshop.

In addition to discussion of the initiatives, protocols, systems, and services listed above, other topics of interest include:

- scaling - to handle more users, more databases or documents, bigger databases, larger data items (e.g., multimedia)
- efficiency - algorithms, data structures, simulations, experiments
- effectiveness - evaluation methods, studies, designs
- interfaces - platforms, user needs, development methods, evaluations, integration with other applications
- limitations and future plans for enhancement

- progress in developing standards or using existing ones
- technical details relating to protocols, implementations
- user studies, application surveys, innovative uses
- integration - plans, designs, requirements, implementations of systems to integrate information access over the networks
- novel applications using network-enabled information access technology.

If you are interested in attending, please prepare a 1-page statement of interest, covering your background, experience, and topics of interest. Priority, in case of limited space, will be given to SIGIR attendees and those who have submitted interest statements.

The Program Committee includes:
George Brett, CNIDR
Edward Fox, Chair (of workshop, SIGIR), Virginia Tech
Jose-Marie Griffiths, U. Tenn.
Brewster Kahle, WAIS Inc.
Clifford Lynch, U. California
Craig Stanfill, Thinking Machines
Craig Summerhill, Coalition for Networked Information
Chris Tomer, Co-chair and local arrangements coordinator

Interest statements should be submitted by May 15. They may accompany registration forms or be sent directly to:
Chris Tomer
School of Library and Information Science
University of Pittsburgh
Pittsburgh, PA 15260, USA
tel: (412) 624-9448; fax: (412) 648-7001
email: ianet@lis.pitt.edu

Further information is available from:
Ed Fox, (fox@fox.cs.vt.edu; foxea@vtcc1.bitnet) or
Chris Tomer (ianet@lis.pitt.edu)

Cost for this one-day session is $50,
which includes coffee breaks and lunch.

National Research and Education Network (NREN) Workshop

Monterey, California
September 16-18, 1992

Sponsored by

Computing Research Association
Educom
IEEE U.S. Activities Board

With assistance from the NSF.

Major Issues
(Taken from the Executive Summary of the Proceedings of the NREN Workshop)

The workshop provided time for lively discussion both in small groups and in plenary sessions. A number of conclusions about major issues were shared widely among the participants:

The High-Performance Computing Act and the current Administration's High Performance Computing and Communications program are important first steps toward the realization of a national network. The Congressional intent to "promote the more rapid development of an information infrastructure" should be met through wide participation in the emerging NREN system.

- The NREN should be more than a program supporting high performance computing. It has the potential to:

- facilitate development of the National Information Infrastructure

- provide tools for increasing the effectiveness of research, education and technology transfer at all levels, and

- demonstrate network applications and technologies which can aid in addressing critical social needs.

- Some members of Congress and others who have been involved with the prior development of research networks appear to assume that the NREN will be a federal network of dedicated physical links. In contrast, the workshop participants see the NREN as a national network program whose advanced transmission requirements will be met by the rapidly evolving commercial communications infrastructure of the country.

- More emphasis should be placed on making the NREN a truly national program which includes strategic partners in higher education, state government and industry.

Recommendations on Policy Questions

The specific Congressional questions serve to test the commonality of visions of the network, its services, its governance and funding. The workshop discussion of the questions reflected general agreement that:

- Historical models for the evolution of the NREN such as the highway system, telephone system, national electrical power grid and the agricultural extension system all contain elements which may be useful in developing the NREN. However, workshop participants believe that historical precedent must be balanced with empirical experience gained from actual network implementation.

The final shape of the NREN, and its balance of public and private activites, will be unique.

- Federal funds for the existing interim NREN (principally NSFNET) have been significantly leveraged by non-federal funds provided though a variety of means such as user fees, industry support, state and university investment. However, the total funding currently available falls far short of the amount needed to realize the goal stated in the Act of connecting research, education and libraries at all levels in every state.

- NREN governance currently includes Congressional oversight of an Administration program involving multiple agencies as well as coordinating and advisory committees. While this structure may work for the current federal program, it is not likely to be sufficient for the development of the NREN as a national program, since major investors and stakeholders currently participate at most in an advisory capacity.

- As the NREN and other large scale computer networks continue to expand and gain use for a wide range of activities, both institutional and individual, the Congress must be sensitive to possible needs for revision and strengthening of federal statues, regulations and policies covering security and intellectual property (i.e., copyright) protection.

- Charging for commercially provided services is technologically feasible and can be dealt with during NREN implementation.

NSF Workshop on Visual Information Management Systems

February 24-25, 1992

Redwood, CA

Excerpt from Workshop Report:

One of the most important technologies needed across many traditional areas as well as emerging new frontiers of computing, is the management of visual information. For example, most of the Grand Challenge applications, under the High Performance Computing and Communication (HPCC) initiative, require management of large volumes of non-alphanumeric information, computations, communication, and visualization of results. Considering the growing need and interest in the organization and retrieval of visual and other non-alphanumeric information, and in order to stimulate academic projects in this area, a workshop on Visual Information Management Systems (VIMS) was sponsored by the National Science Foundation. This workshop was held in Redwood, CA, on Feb 24-25, 1992. The goal of the workshop was to identify major research areas that should be addressed by researchers for VIMS that would be useful in scientific, industrial, medical, environmental, educational, entertainment, and other applications.

Selected Legislation in the 103rd Congress Related to Electronic Information Delivery

Infrastructure Development

H.R. 707 (Dingel)
Companion Measure S.335
Emerging Telecommunications Technologies Act of 1993

Amends the National Telecommunications and Information Administration Organization Act to require the Assistant Secretary of Commerce for Communications and Information and the Chairman of the Federal Communications Commission (FCC) to conduct, at least biannually, and to report to specified congressional committees, the FCC, and the Secretary of Commerce annual on, joint electromagnetic spectrum planning with respect to: (1) future spectrum requirements for public and private uses and the allocation actions to accomodate those uses; and (2) actions to promote the efficient use of the spectrum. Introduced February 2, 1993. Referred to the House Committee on Energy and Commerce, Subcommittee on Telecommunications and Finance. Passed House on March 2, 1993. Received in Senate and Referred to Senate Committee on Commerce, Science and Transportation on March 3, 1993.

These bills are incorporated in H.R.2264, the Omnibus Budget Reconciliation Act of 1993.

H.R. 1091 (Clinger)

A bill to establish the Commission on Information Technology and Paperwork Reduction. Introduced on February 24, 1993. Referred to House Committee on Government Operations, Subcommittee on Legislation and National Security. Establishes the Commission on Information Technology and Paperwork Reduction in order to minimize the information reporting burden imposed by the Federal Government, consistent with the information needs of the Government for policy purposes. Lists specific Commission functions, which include the study and review of former Commission on Paperwork recommendations for paperwork reduction.

Requires a final Commission report to Congress and the President and action by the Office of Management and Budget on Commission recommendations.

H.R. 1312 (Boucher)
Local Exchange Infrastructure Modernization Act of 1993

A bill to recognize the unique status of local exchange carriers in providing the public switched network infrastructure and to ensure the broad availability of advanced public switched network infrastructure.

Amends the Communications Act of 1934 to require the Federal Communications Commission (FCC) to exercise its authority to: (1) preserve and enhance universal telephone service at reasonable rates; (2) achieve universal availability of advanced network capabilities and information services; (3) assure a seamless nationwide distribution network through joint network planning, coordinations, and service arrangements between and among local exchange carriers (LECs); (4) maintain high standards of quality for advanced network services; and (5) assure adequate communications for the public health, safety, defence, educations, national security, and emergency preparedness.

Introduced on March 11, 1993. Referred to the House Committee on Energy and Commerce, Subcommittee on Telecommunications and Finance and to the House Committee on the Judiciary, Subcommittee on Economic and Commercial Law.

H.R. 1504 (Boucher)
Communications Competitiveness and Infrastructure Modernization Act of 1993

A bill to encourage the modernization of the Nation's telecommunications infrastructure, to promote competition in the cable television industry and to permit telephone companies to provide video progamming. Introduced on March 29, 1993. Referred to House Committee on Energy and Commerce, Subcommittee on Telecommunications and Finance.

Amends the Communications Act of 1934 to allow a common carrier to provide video programming directly to subscribers in its telephone service area through its own facilities or an affiliate. Authorizes the common carrier to provide channels of communications, pole line conduit space, or other rental arrangements to any entity which is directly or indirectly owned, operated, or controlled by it if such facilities or arangements are used for, or in connection with, the provision of video programming directly to subscribers in the telephone service area of the common carrier.

Prohibits a common carrier from providing video programming directly to subscribers in its telephone service area unless the programming is provided through a separate video programming affiliate. Requires business arrangements and transactions between a common carrier and its video programming affiliate to be pursuant to regulations prescribed by the Federal Communications Commission and to be without cost to the telephone service ratepayers of the carrier.

Requires any common carrier which provides video programming directly to subscribers through an affiliate in its telephone service area to establish a basic video dial tone platform.

Requires any such common carrier to make a maximum of 75 percent of the equipped capacity of its basic video dial tone platform available to unaffiliated video program providers. States that the carriage of local broadcast signals shall not constitute the provisions of affiliated video programming under this Act.

Sets forth prohibitions on: (1) cross-subsidization between telephone service and video programming by common carriers; and (2) common carrier buyouts of cable systems located in the carrier's telephone service area.

Requires the Commission to convene a Federal-State Joint Board to establish practices, classifications, and regulations necessary to ensure proper jurisdictional seperation and allocation of the costs of providing broadbound services, including affiliated video programming.

Makes provisions of the Act inapplicable to video programming provided in a rural area by common carrier that provides telephone exchange service in such area.

H.R. 1613 (Collins, C.)
Telecommunications Policy Coordination Act of 1993

A bill to improve coordination in the formulation of telecommunications policy within the executive branch. Introduced on April 1, 1993. Referred to House Committee on Energy and Commerce, Referred to Subcommittee on Telecommunications and Finance.

Establishes an Office of Telecommunications Policy (OTP) in the Executive Office of the President. Directs OTP to: (1) prepare national telecommunications policy options; (2) serve as the principal advisor to the President on telecommunications issues; (3) arbitrate telecommunications policy disputes among Federal agencies; (4) communicate the views of the agencies and the President concerning telecommunications policy to the Federal Communications Commission (FCC) and the Congress; and (5) monitor developments in telecommunications technology. Requires the Director of OTP to: (1) establish an Advisory Committee on Telecommunications Policy; and (2) report to the President and Congress annually on OTP activities and on emerging trends in telecommunications.

Requires the FCC to report to the President and the Congress on its reasons for taking any final action which is inconsistent with views received from OTP.

H.R. 1757 (Boucher)
High Performance Computing and High Speed Networking Applications Act of 1993
National Information Infrastructure Act of 1993

A bill to provide for a coordinated federal program to accelerate development and dissemination of applications of high performance computing and high-speed networking. Introduced April 21, 1993. Referred to House Committee on Science, Space, and Technology, Subcommittee on Science. Hearings held by Subcommittee on May 6 and 11, ordered and reported as amended on June 30, 1993.

Amends the High-Performance Computing Act of 1991 to direct the Federal Coordinating Council for Science, Engineering, and Technology to: (1) establish an interagency applications program to develop applications of computing and networking advances under the National High-Performance Computing Program; and (2) develop a Plan for Computing and Networking Applications which shall identify application program goals and priorities and set forth specific Federal agency responsibilities.

Requires the Plan to: (1) foster local network access programs and their connection with Internet; and (2) develop projects and technologies in the fields of education, health care, libraries, and government information access.

Provides for the establishment of a high performance computing and applications advisory committee.

H.R. 2264 (Sabo)
Companion Measure, S. 1134
Omnibus Budget Reconciliation Act of 1993
Licensing Improvement Act of 1993
***Emerging Telecommunications Technologies Act of 1993**

A bill to provide for reconciliation pursuant to section 7 of the concurrent resolution on the budget for fiscal year 1994. Incorporates H.R. 707 and S. 335. Introduced on May 25, 1993. House Committee on the Budget Reported an Original Measure, report No: 103-111. Passed House on May 27, 1993, Passed Senate in lieu of S. 1134 on June 25, 1993, Senate also requested a conference.

S. 4 (Hollings)
Companion Measure H.R. 820
Manufacturing Technology and Extension Act of 1993
Information Infrastructure and Technology Act of 1992
National Competitiveness Act of 1993
Wind Engineering Program Act of 1992

A bill to promote the industrial competitiveness and economic growth of the United States by strengthening and expanding the civilian technology programs of the Department of Commerce, amending the Stevenson-Wydler Technology Innovation Act of 1980 to enhance the development and nationwide deployment of manufacturing technologies, and authorizing appropriations for the Technology Administration of the Department of Commerce, including the National Institute of Standards and Technology. Introduced on January 21, 1993. Referred to Senate Committee on Commerce, Science and Transportation. Committee hearings held February 24, 1993 and March 25, 1993. Ordered and reported as amended on May 25, 1993.

TITLE III--CRITICAL TECHNOLOGIES, Sec. 313. Technical Amendments ...(3) The Office of Technology Monitoring and Competitive Assessment is authorized to (A) act as a focal point within the federal government for the collection and dissemination, including electronic dissemination, of information on foreign process and product technologies, including information collected under the Japanese Technical Literature Program; (B) coordinate the extensive foreign technology monitoring and assessment activities already under way in the federal government; (C) act as an electronic clearinghouse for this information or otherwise provide this function.

S. 335 (Innouye)
Companion Measure H.R. 707
Emerging Telecommunications Technologies Act of 1993

A bill to require the Secretary of Commerce to make additional frequencies available for commercial assignment in order to promote the development and use of new telecommunications technologies. Introduced on February 4, 1993. Referred to Senate Committee on Commerce, Science and Transportation, Subcommittee on Communications. Hearings held on March 17, 1993.

Directs the Assistant Secretary of Commerce for Communications and Information and the Chairman of the Federal Communications Commission (FCC) to conduct joint spectrum planning meetings. Directs the Secretary of Commerce to: (1) identify bands of frequencies that may be reallocated to commercial users; and (2) establish a related advisory committee. Directs the FCC to submit to the President and the Congress a plan for the distribution of the reallocated banks of frequencies under this Act.

Authorizes the President to reclaim reallocated bands of frequencies for reassignment to government stations.

This bill was also incorporated in H.R. 2264, the Omibus Budget Reconciliation Act of 1993.

S. 473 (Johnston)
Department of Energy National Competitiveness Technology Partnership Act of 1993

A bill to promote the industrial competitiveness and economic growth of the United States by strengthening the linkages between the laboratories of the Department of Energy and the private sector and by supporting the development and application of technologies critical to the economic, scientific and technological competitiveness of the United States. Introduced on March 2, 1993. Referred to Senate Committee on Energy and Natural Resources, Subcommittee on Energy Research and Development. Committee hearings held on March 18, 23 and 24, 1993. Reported to Senate (Amended) by Senate Committee on Energy and Natural Resources On June 24, 1993, report no.:103-69.

Amends the Department of Energy Organization Act to authorize the Secretary of Energy and the directors of departmental laboratories (laboratoires operated by or on behalf of the Department of Energy (DOE) or facilities that would be considered to be laboratories under the Stevenson-Wylder Technology Innovation Act of 1980) to enter into any partnership that will enhance the economic, scientific, or technological competitiveness of U.S. industry.

Directs the Secretary to develop a multi-year critical technology strategy for each critical technology listed in the National Critical Technologies Report. Authorizes the Secretary and the directors of departmental laboratories to enter into partnerships that build on the core competencies of the laboratories to conduct research, development, demonstration, or commercial application activities in areas listed in Report or in energy efficiency or supply, high-performance computing, the environment, human health, advanced manufacturing technologies, advanced materials, transportation, space, or quality technologies, or technologies listed in the annual defense critical technologies plan.

Amends the High-Performance Computing Act of 1991 to provide for cost-shared projects involving DOE or DOE laboratories and non-Federal entities to test and apply high-performance computing and high-speed networking technologies.

S. 570 (Grassley)
Local Exchange Infrastructure Modernization Act of 1993

A bill to recognize the unique status of local exchange carriers in providing the public switch network infrastructure and to ensure the broad availability of advanced public switched network infrastructure. Introduced on March 11, 1993. Referred to Senate Committee on Commerce, Science, and Transportation.

Amends the Communications Act of 1934 to require the Federal Communications Commission (FCC) to exercise its authority to: (1) preserve and enhance universal telephone service at reasonable rates; (2) achieve universal availability of advanced network capabilities and information services; (3) assure a seamless nationwide distribution network through joint network planning, coordination, and service arrangements between and among local exchange carriers (LECs); (4) maintain high standards of quality for advanced network services; and (5) assure adequate communication for the public health, safety, defense, education, national security, and emergency preparedness.

Requires the FCC to prescribe regulations that require: (1) joint coordinated network planning, design, and cooperative implementation among all LECs in the provision of public switched network infrastructure and services; (2) development of standards for interconnection between the LEC public switched network and others by appropriate standard-setting bodies; and (3) a LEC to share public switched network infrastructure and functionality with requesting LECs which serve a geographic area for which they lack economies of scale or scope for the particular required network functionality.

S. 1086 (Danforth)
Telecommunications Infrastructure Act of 1993

A bill to foster the further development of the Nation's telecommunications infrastructure through the enhancement of competition. Introduced on June 9, 1993. Referred to Senate Committee on Commerce, Science and Transportation.

Government Information

H.R. 629 (Owens, M.)
Improvement of Information Access Act of 1993

A bill to amend title 44, United States Code, to promote improved public dissemination of Government information. Introduced on January 26, 1993. Referred to House Committee on Government Operations, Subcommittee on Information, Justice, Transportation and Agriculture.

Requires agencies to: (1) disseminate information in diverse modes and through appropriate outlets that will permit and broaden public access to Government information; and (2) use depository libraries, national computer networks, and other distribution channels that improve and assure free or low-cost public access to Government information.

H.R. 1328 (Rose)
Companion Measure S.564--P.L. 103-40
Government Printing Office Electronic Information Access Enhancement Act of 1993

A bill to establish in the Government Printing Office a means of enhancing electronic public access to a wide range of Federal electronic information. Introduced on March 11, 1993. Referred to House Committee on House Administration. On April 1, 1993 reported to House by Committee, report no.:103-51.

Requires the Superintendent of Documents, under the direction of the Public Printer, to establish a means for providing the public with on-line access to electronic public information of the Federal Government.

Sets forth guidelines for determining fees for accessing such information. Permits depository libraries to access information through such means without charge.

Requires the Public Printer to report to the Congress on the savings resulting from such on-line public access to government information and on the status of the system providing such access.

S. 560 (Nunn)
Paperwork Reduction Act of 1993

A bill to further the goals of the Paperwork Reduction Act to have the Federal agencies become more responsible and publicity accountable for reducing the burden of Federal paperwork on the public, and for other purposes. Introduced on March 10, 1993. Referred to Senate Committee on Governmental Affairs.

TABLE OF CONTENTS
Title I: Authorization of Appropriations
Title II: Reducing the Burden of Federal Paperwork on the Public
Title III: Enhancing Federal Agency Responsibility and Accountability for Reducing the Burden of Federal Paperwork
Title IV: Enhancing Government Responsibility and Accountability for Reducing the Burden of Federal Paperwork
Title V: Enhancing Agency Responsibility for Sharing and Disseminating Public Information - Provides for Government-wide standards for sharing and disseminating public information.

Imposes certain responsibilities on Federal agencies for sharing and disseminating public information.

Abolishes the Federal Information Locator System established in the Office of

Information and Regulatory Affairs and replaces it with a system in each agency for providing access via electronic and other means to a comprehensive inventory of agency information dissemination products.

Provides for the use of electronic information collection and dissemination techniques to reduce the Federal paperwork burden.

S. 564--P.L. 103-40 (Ford)
Companion Measure, H.R. 1328
Government Printing Office Electronic Information Access Enhancement Act of 1993

A bill to establish in the Government Printing Office a means of enhancing electronic public access to a wide range of Federal electronic information. Introduced on March 11, 1993. Referred to Senate Committee on Committee on Rules and Administration. Passed Senate on March 22, 1993, passed House on May 25, 1993, signed into law on June 8, 1993.

Requires the Superintendent of Documents, under the direction of the Public Printer, to establish a means for providing the public with on-line access to electronic public information of the Federal Government.

Sets forth guidelines for determining fees for accessing such information. Permits depository libraries to access information through such means without charge.

Requires the Public Printer to report to the Congress on the savings resulting from such on-line public access to governmnet information and on the status of the system providing such access.

S. 681 (Glenn)
Paperwork Reduction Reauthorization Act of 1993
Regulatory Review Sunshine Act of 1993

A bill to amend chapter 35 of title 44, United States Code, relating to Government paperwork reduction, to modify the Federal regulatory review process. To ensure the greatest possible public benefit from information collected, maintained, used, disseminated, and retained by the federal government. Introduced on March 31, 1993. Referred to Senate Committee on Governmental Affairs.

Educational Applications

H.R. 856 (Owens, M.)

Educational Research, Development, and Dissemination Act

A bill to improve education in the United States by promoting excellence in research, development, and the dissemination of information. Introduced on February 4, 1993. Referred to House Committee on Education and Labor, Subcommittee on Select Education and Civil Rights. Subcommittee hearings held on May 27, 1993. Field hearings held in New York, New York on June 19, 1993.

TABLE OF CONTENTS:

Title IV: - Amends GEPA to establish within OERI an Office of Reform Assistance and Dissemination (Dissemination Office), through which the Secretary shall carry out a national education dissemination system for school improvement. Provides for Dissemination Office functions and duties, including: (1) identification, designation, and dissemination of exemplary and promising programs including certain training, technical, and financial assistance; (2) 16 Education Resources Information Clearinghouses; (3) dissemination through new technologies; (4) an electronic network for sources of materials and research about teaching and learning for improving nationwide education (SMARTLINE) to link various educational research and other entities; (5) an electronic networking and resource-sharing for school improvement program of grants to State education agencies; (6) a networked system of the ten regional educational laboratories; (7) an America 2000 Communities Special Assistance Program, with grants for Learning Grant Institutions and District Education Agents within eligible communities, development of a comprehensive America 2000 plan for assuring educational success for all students in the community, and implementation of a community-wide plan for educational improvement; (8) the Teacher Research Dissemination Network (regional partnerships for teacher change agents); and (9) the existing National Diffusion Network and its Developer-Demonstrator and State Facilitator projects.

H.R. 2268 (Brown)

A bill to facilitate the development of an integrated, nationwide telecommunications system dedicated to instruction by guaranteeing the acquisition of a communications satellite system used solely for communications among State and local instructional institutions and agencies and instructional resource providers. Introduced on May 26, 1993. Referred to House Committee on Education and Labor, Subcommittee on Elementary, Secondary and Vocational Education.

S. 264 (Bingaman)
Technology for the Classroom Act of 1993

A bill to establish a Classrooms for the Future Program. Directs the Secretary of Education to award competitive grants to eligible consortia to develop instructional programs and technology-based systems for complete courses or units of study for a specific subject and grade level, if these are commercially unavailable locally. Introduced on January 28, 1993. Referred to Senate Committee on Labor and Human Resources.

S. 1040 (Bingaman)
Technology for Education Act of 1993

A bill to support systemic improvement of education and the development of a technologically literate citizenry and internationally competitive work force by establishing a comprehensive system through which appropriate technology-enhanced curriculum, instruction, and administrative support resources and services, that support the National Education Goals and any national education standards that may be developed, are provided to schools throughout the United States. Introduced on May 27, 1993. Referred to Senate Committee on Labor and Human Resources.

Table Of Contents:
Title I: Leadership for Technology in Education
Title II: School Technology Support
Title III: Information Dissemination, Technology Training and Technical Assistance
Title IV: Educational Technology Product Development, Production, and

Distribution
Title V: Educational Technology Research, Development and Assessment

Library Applications

S.345 (Pell)
Library of Congress Fund Act of 1993

A bill to authorize the Library of Congress to provide certain information products and services.

Expresses the intent of the Congress that core Library of Congress services shall continue to be provided at no cost.

Defines: (1) "core library products and services" as domestic interlibrary loan and information products and services customarily provided by libraries to users at no charge; and (2) "specialized library products and services" as specified customized information products and services that exceed core services, that are not national library products and services, and that are designed for individuals or discrete groups of persons or entities.

Declares that this Act shall not modify Federal copyright law.

Introduced on February 4, 1993. Referred to Senate Committee on Committee on Rules and Administration. Committee Consideration and Mark-up Session held on May 20, 1993. Reported to Senate (amended) by Senate Committee on Rules and Administration on May 26, 1993, report no.:103-50. Questions were raised by Senators DeConcini and Feinstein which resulted in a meeting of all interested parties including House and Senate Judiciary Committee staff on June 4. Discussions focused upon amendments proposed by the information and publishing industries.

S. 626 (Kerrey)
Electronic Libraries Act of 1993

A bill to establish a system of State-based electronic libraries. Provides for a system of State-based electronic libraries which (1) allows delivery of or access to a vast array of interactive, multimedia educational programs, research and information data and services, and networking opportunities; (2) seeks to make such materials available to the public via public libraries, electronic databases and telecommunications systems such as the Internet and other networks. Authorizes the National Science Foundation, the Department of Education, the Department of Commerce, the Defense Advanced Research Projects Agency, and the Library of Congress to make multi-year grants to states to develop electronic libraries. Introduced on March 22, 1993. Referred to Senate Committee on Commerce, Science and Transportation.

Health Services

S. 1088 (Harkin)
Rural Telemedicine Development Act of 1993

A bill to amend the Public Health Service Act to provide grants for the development of rural telemedicine. Introduced on June 10, 1993. Referred to Senate Committee on Agriculture, Nutrition, and Forestry.

S. 1143 (Baucus)

A bill to improve the delivery of health care services in rural areas by creating an Assistant Secretary for Rural Health, to amend title XVIII of the Social Security Act to provide that medical assistance facilities be reimbursed based on reasonable cost, to establish a grant program for the use of interactive telecommunications systems, and to adjust the payments made for certain direct graduate medical education expenses. Introduced on June 22, 1993. Referred to Senate Committee on Finance.

Privacy and Intellectual Property

H.R. 12 (Hughes)

A bill to amend title 17, United States Code, with respect to infringement of copyright. Makes a television broadcast station an infringer of copyright and subject to civil remedies (including attorney's fees and litigation costs) if such station, without the express written consent of the copyright owner, authorizes the secondary transmission by a cable system or other multichannel video programming distributor of a copyrighted work broadcast by such station. Introduced on January 5, 1993. Referred to House Committee on the Judiciary, Subcommittee on Intellectual Property and Judicial Administration.

H.R. 135 (Collins, C.)
Individual Privacy Protection Act of 1993

A bill to amend the privacy provisions of title 5, United States Code, to improve the protection of individuals information and to reestablish a permanent Privacy Protection Commission as an independent entity in the Federal Government.

Establishes an Individual Privacy Protection Board to: (1) study the data banks, automated data processing programs, and information systems of public and private organizations to determine standards and procedures in force for the protection of personal information; ... and (5) comment on the implications for data protection of proposed Federal, State, or local statutes, regulations, or procedures. Provides penalties for violations of privacy rights.

Introduced on January 5, 1993. Referred to House Committee on Government Operations, Subcommittee on Information, Justice, Transportation and Agriculture.

H.R. 759 (Boucher)
Compulsory License Clarification Act of 1993

A bill to amend chapter 1 of title 17, United States Code, to include in the definition of a cable system a facility which makes secondary transmissions by microwave or certain other technologies. Introduced on February 3, 1993. Referred to House Committee on the Judiciary, Subcommittee on Intellectual Property and Judicial Administration. Subcommittee hearings held on March 17, 1993.

H.R. 897 (Hughes)
Companion Measure - S. 373
Copyright Reform Act of 1993

A bill to amend title 17, United States Code, to modify certain recordation and registration requirements, to establish copyright arbitration royalty panels to replace the Copyright Royalty Tribunal. Introduced on February 16, 1993. Referred to House Committee on the Judiciary, Subcommittee on Intellectual Property and Judicial Administration. Subcommittee hearings held on March 3 and 4, 1993.

H.R. 1103 (Hughes)

A bill to amend title 17, United States Code, with respect to secondary transmissions of superstations and network stations for private home viewing, and with respect to cable systems. Introduced on February 24, 1993. Referred to House Committee on the Judiciary, Subcommittee on Intellectual Property and Judicial Administration. Subcommittee hearings held on March 17, 1993.

H.R. 2576 (Hughes)
Digital Performance Right in Sound Recording Act of 1993

A bill to amend title 17, United States Code to provide an exclusive right to perform sound recordings publicly by means of digital transmissions. Introduced on July 1, 1993. Referred to House Committee on the Judiciary, Subcommittee on Intellectual Property and Judical Administration.

S. 23 (Hatch)

A bill to amend title 17, United States Code, to clarify news reporting monitoring as a fair use exception to the exclusive rights of a copyright owner. Introduced on January 21, 1993. Referred to Senate Committee on the Judiciary, Subcommittee on Patents, Copyrights and Trademarks.

News Items from "The Computists' Communique" 1992

- The U.S. prints 600M pages of computer printout each day, plus 234M photocopies, 76M letters, and 24M documents, or 45 sheets of paper per worker per day. Electronic document "imaging" is going to save a lot of trees. [Herbert A. Edelstein, Euclid Associates. Steve Kaplan, CC, 5/19.]

- ACM has contracted with Engineering Information Inc. (New York, NY) to fax you articles from 11 journals and 40 conference proceedings of the last 10 years. The document delivery service (EiDDS) also handles literature from

Elsevier, IEEE, Pergamon, John Wiley & Sons, and AAAI. For coverage and prices, contact Ruth Miller at (800) 221-1044 or (212) 705-7301, fax (212) 832-1857. [ACMemeberNet, 5/92.] You might also check your phone book or library for local document delivery services.

- McGraw-Hill Inc. (New York) is installing an Electrobook digital press that works like a laser printer -- at 350 feet/minute, printing both sides of a page. (Newspaper presses run at 2,000 feet/minute.) Initial use will be for McGraw-Hill's Primis line of custom textbooks. R.R. Donnelley is expected to install an Electrobook press for its Books on Demand Module in Harrisonburg, VA. [Patrick M. Reilly, WSJ, 4/14.]

- R.R. Donnelley (Chicago) and Compton's NewMedia, a San Diego- based Britannica company, are forming a partnership for CD ROM and digital publishing. Donnelley's Database Technology Services division (Willowbrook, IL) will handle all CD-ROM pre-production services for Compton's NewMedia, which has links with Merriam Webster, McGraw Hill, National Textbook Co., and others. One product is Compton's SmarTrieve Publishing Toolkit, which provides indexing and text search software for interactive books and reference sources to new electronic publishers. (708) 655-7705. [Business Wire, 3/11. agentsee.]

- The New Publishing: Technology's Impact on the Publishing Industry Over the Next Decade is a 62-page technical report by Gregory Rawlins. You can get a copy for $5 P&H if you write to ICDS, Indiana University, Lindley Hall, Bloomington, IN 47405. [Copler@iubacs.bitnet, PACS-L, 3/16.]

- Fortune magazine reports that 76% of subscribing managers use a computer at work, averaging 15 hours per week. A quarter of these are hooked into networks, mostly at large companies; many more (40%) are planning to add networked computers this year. More than half also use a computer at home, averaging 4.4 hours on work-related tasks. (Another 49% plan to buy desktop or portable PCs for home use next year.) Popular software programs are Lotus 1-2-3 and WordPerfect, followed by Harvard Graphics, dBase, Windows, and Excel. Only 20% of the offices have Macs available. [Robin Stacy, The Macon Telegraph, 3/11. agentsee.] The managers also plan to buy software, but I get the impression that they're not looking forward to the process. One of the nice things about Macs is that there's less to choose from -- and it's nearly all good to excellent. With a PC, you need a wizard around somewhere -- especially with Windows, which runs on top of DOS. Macs are easier, and you can use eight or ten programs with less trouble than learning DOS WordPerfect and Excel.

- The American Newspaper Publishers' Association seems to be losing its fight with the Bell companies due to fraternization. Member newspapers keep breaking ranks to use phone services in the delivery of their information services. This is partly a defense against the cable and microwave-based communication companies. [Dana Blankenhorn, Newsbytes, 3/19.]

- RELCOM Joint Stock Company has a network information department to spread science and technology news from the former Soviet Union. Publications include an RENEWS bulletin and a Computer Market monthly supplement to the POISK scientific newspaper. Contact Vladimir Shliemin (nev@renews.relcom.msk.su) to request RENEWS.

- ANS, the carrier of NSFnet backbone services, has reached a trial agreement with CIX, the Commercial Internet Exchange. Although ANS has not yet joined CIX, it agrees to share traffic in both directions. Physical implementation in 60 days will open a way for NEARNet, BARRnet, and other internet regionals to exchange commercial traffic with Alternet, PSI, and all other CIX members -- provided that certain ANS CO+RE agreements are signed (and paid for). [Mitchell Kapor (mkapor@eff.org), com- priv, 6/9.] NSF's Acceptable Use Policy can thus be bypassed at the backbone level, although local host policies would still apply.

- The Special Libraries Association is having its 83rd (!) annual conference this week. Master's-level library professionals and information analysts make about $45K. Most are computer-proficient, including word processing, spreadsheets, and databases. Apple has a library staff of 18, each with a high-end Mac and access to 600 databases, 750 magazines, and 8,000 books. Raychem's library handles 35 technical inquiries per day. Dialog Information Services (Menlo Park, CA) says there are over 4,500 commercial databases. (Search costs average about $100/hour.) [Gavin Power, SF Chronicle, 6/9.] (Sounds like a great market, although competition is keen and profit margins are slim.)

- There will be a shortage of MLS-level librarians well into the decade, especially in youth services, school librarianship, cataloging, and special librarianship (such as corporate libraries). People often enter librarianship as a second or third career. Retirements are up and enrollments are down. Salaries range from entry-level $20K to $100K for the director of a large library. [Jenny Upchurch, SJM, 6/14.] (Maybe AI techniques could help with cataloging.)

- Jack Kessler (kessler@ocf.berkeley.edu) says that the Tres Grande Bibliotheque (TGB) or Bibliotheque de France (BDF) project, with initial funding of US$1B , "easily ranks as the century's premier library project." One goal is to give the French people electronic access to hitherto secret treasures of the Bibliotheque Nationale. [PACS-L, 11/25/91.]

- The British Library has become the first overseas member of the Research Libraries Group, Inc. (Mountain View, CA). Harvard University is another new member. RLG priorities include preservation and on-line access to historic collections. [Jennifer Porro (bl.jmp@rlg.stanford.edu), LIBRES, 3/5.]

- The MARC standard for networked information is being coordinated by Sally H. McCallum (smcc@seq1.loc.gov), LM 639, Library of Congress, Washington, DC 20540. Two important papers are "Dictionary of Data Elements for On-line Information Resources" (MARBI 49, Spring '91) and "Providing Access to On-line Information Resources" (MARBI 54). Send a "get dp54 doc" to listserv@maine.maine.edu to retrieve the latter. [Barbara.Shipman @med@umich.edu, CNIDIR-L. David Mattison(dmattiso@cue.bc.ca), PACS-L, 4/23.]

- BRS has announced an AI-based Domain Independent Intelligent Information and Services Network Interface (DISNET) from IDE. The Unix or DOS interface negotiates an ASCII or GUI query, then uses BRS/SEARCH and RDBMS (Oracle, Ingress, etc.) software to access multiple local and remote databases. Videotex, email, and natural-language interfaces are under development. BRS Software Products Benelux, Gele Rijdersplein 16, 6811 AP Arnhem, the Netherlands; (+31) 85 579600. [MeckJournal, 3/92.]

- Charles W. Bailey, Jr., has examined "Intelligent Library Systems: AI Technology and Library Automation Systems" in Advances in Library Automation and Networking, 4(1991), pp. 1-23. He lists useful systems and reasons why there aren't more of them -- lack of AI training, tools, expertise, willingness to take risks -- then speculates on fruitful R&D areas. [Lis Rowlinson, Current Cites, 3/92.]

- Publishers, bookstores, and libraries are all intermediaries in getting information from authors to users. Publishers and bookstores make money; libraries do not. Perhaps libraries should cut out the middlemen, dealing with authors directly. [Sanjay Chadha (sanjay chadha@library.tmc.edu), PACS-L, 3/1.] Unfortunately, librarians wouldn't know how to make money even if authors contributed their works for free. It's more likely that publishers will bypass libraries by selling information services and CD ROMs.

- Terry Tobias (ttobia01@gumedlib.georgetown.edu) recommends two articles on the likelihood of on-line information retrieval displacing traditional journals in medicine: "BioSYNTHESis: Bridging the Information Gap," by Naomi C. Broering et al., Bull. Med. Libr. Assoc., 77(1), Jan. 1989, pp. 19-25; and "Medical Informatics: An Emerging Academic Discipline and Institutional Priority," by Robert A. Greenes, JAMA, Feb. 23, 1990, Vol 263 (8), pp. 1114-1120. [on-line, 3/20.]

- Several libraries report that they've saved money by canceling Chem/Bio Abstracts and subscribing to bibliographic search services like Dialog and STN. The arrangement does not endanger ACS accreditation. [John Piety (piety@jcvaxa) and Mark G.R. McManus (mmcmanus@s850.mwc.edu), PACS-L, 3/20.] Monthly costs aren't predictable, but they are predictably low.

- The ERIC indexing service is the first to abstract an electronic journal, New Horizons in Adult Education. The next issue of Serials review is introducing an Electronic Journal Forum and will discuss New Horizons and the problem of indexing e-journals. [Linda Langschied (langschied@zodiac.bitnet), PACS-L, 4/10.] The MLA Bibliography and Directory of Periodicals also cite e-journals. [Daniel Uchitelle (mlaod@cuvmb.bitnet), ibid.]

- The ERIC CIJE clearinghouses are poised to annotate and index electronic journal articles. Resources are limited, so e-journals must satisfy the usual criteria for substantiveness, contribution to knowledge, authority of author and source, timeliness, availability, etc. RIE archiving can help make copies available through microfiche collections or ERIC reproduction. [Nancy R. Preston (eric@suvm.acs.syr.edu), PACS-L, 4/20.]

- Project Gutenberg has the 1911 Roget's Thesaurus. Ed Vielmetti (emv@msen.com) has created an indexed version named roget-thesaurus on WAIS host wais.cic.net. [alt.wais, 12/11.] 1990 U.S. Census information is also available, along with the Oedipus Trilogy and Herland. Also corrected versions of the

King James' Bible and of Far From the Madding Crowd, and a new edition of Paradise Lost (plrabn10.txt). FTP at night from etext or etext92 on mrcnext.cso.uiuc.edu. Newsletter info is in directory etext/articles. Supporters of the project include Apple, Calera, TextPert, Caere, and Expervision. [Michael S. Hart (hart%vmd.cso.uiuc.edu), PACS-L, 2/20 and 4/20.]

- Meckler will be publishing a book (Fall, 1992) on text-management and bibliographic programs for DOS and the Mac. Systems include askSam, Pro-Cite Mac, WordCruncher, Inmagic, Papyrus, Library Master, EndNote, IZE, Notebook II, and Nota Bene. Contact Erwin Welsch (ewelsch@vms.macc.wisc.edu). [PACS-L, 10/1.]

- For a review of bibliographic software, see Health Libraries Review, 1991, V8, pp. 11-20. I believe the article is Diana Herman's "Downloading from Medline: A Comparison of Personal Database Software." It covers Papyrus, Bib/Search, DMS4CITE, Notebook II, Pro-Cite, Ref-11, Reference Manager, Refsys Super and the Ref-Filer. [Geoff Smith (ql23@liverpool.ac.uk), PACS-L, 10/2.] Another review of Papyrus, ProCite, and Reference Manager is in Appendix II of Sue Stigleman's "Bibliographic Formatting Software," published by the Institute for Academic Technology, UNC Chapel Hill (iat@unc.bitnet). [Jim Morgan (izie100@indyvax.bitnet), PACS-L, 10/2.]

- "Bibliography Formatting Software: A Buying Guide," by Sue Stigleman, can be found in Database 15(1), 2/92, pp. 15-27. The review includes Pro-Cite, Artfile, and Notebook. [Vivienne Roumani, Current Cites, 4/22.]

- Bibliographic programs let you format your citations using keywords (author, editor, etc.), then print them out in your choice of formats. Programs can import data from other sources, including downloads from Dialog, Medlars, and other on-line or CD-ROM services. (Conversion programs may cost extra.) Each citation can typically contain 32K characters or more. You can attach notes to any record, and can do Boolean searches on any field. The two Mac programs below work with most word processors, and can produce footnotes as well as bibliographies. Each supports only a single bibliographic database, with size limited to 32Mb or 100K records.

- EndNote Plus ($249/$149 street; Niles & Associates; (510) 655-6666) is good for quick and easy citation entry and formatting. It offers 14 reference types (book, journal, thesis, etc.). Common "rapid entry" abbreviations are permitted for journal fields. Importing can be done in Pro-Cite or Unix Refer/Bibix [Bibtex?] format. File conversions are via the EndLink program, which does not include several popular library formats. EndNote Plus will identify duplicate records but will not delete them automatically. A "quickfind" database indexing function is available, but doubles or trebles the size of your database. Document text can be in almost any format, including MacWrite II. You can copy citations text ([author, year record#]) from a database menu, making insertion very easy. The program will automatically reformat in-text citations and then insert your bibliography at the end of the document. You can have more than one document (or chapter) using the same database, with the bibliography being appended to the final document. You can also print out your entire database as a bibliography, but the capability is limited compared to Pro-Cite. [Aaron K. Poffenberger, BMUG Memo, 11/91.]

- Pro-Cite ($395/$195 street; Personal Bibliographic Software, Inc.; (313) 996-1590) is available in both Mac and DOS versions, and is good for advanced database work. It has 26 reference types, including dissertation, music, and movie. An "authority list" stores common abbreviations for use throughout the database. Only ASCII importing is supported. File conversions require a separate program for each on-line service or library format, and this can become quite expensive. Pro-Cite will delete or merge duplicate records, and has a global search-and-replace function. You can save commonly used search criteria in files. Document text must be in the original MacWrite format. Selecting citations from a database menu is a bit more complicated than in EndNote Plus, as a "punctuation file" must be defined. Browsing the database is more convenient in Pro-Cite, though. Pro-Cite produces a separate bibliography in MacWrite format. (If you don't like the [author, year record#] in-text citation format, you must edit it by hand.) It is particularly good at printing out a categorized bibliography of your entire database. [Ibid.]

- Frank Norman (f-norman@nimr.mrc.ac.uk) says that Papyrus is cheaper and easier to use than Procite, especially for import reformatting. It's a good package for individual scientists. Procite has more power as a general-purpose database manager. A new version of Procite will include italics, boldface, and subscripts. [PACS-L, 10/2.]

- Mark Zimmerman (science@oasys.dt.navy.mil) has some experimental information-retrieval source code that he's willing to share. qndxr.c builds inverted index files for large text files (at 50-80Mb/hour on Sun and NeXT workstations); brwsr.c is a command-line-driven browser program that lets you see word lists, keyword-in-context displays, and full text on demand, and does very simple proximity search. Todd Kaufmann of CMU has a GNU Emacs interface to the brwsr which makes it easier to use. Mark also has a Free Text 1.02 HyperCard/C interface to the indexed text files; it's 250Kb in binhex'd stuffit'd form. Mark recommends the PARA group (para-request@cs.cmu.edu) for discussion of hypertext, free-text IR, Emacs interfaces, etc. Liam Quin's lq-text Unix IR source code would be discussed there. [IRList, 4/25.]

- Senator Gore (D-TN) has entered a bill, The Information Infrastructure and Technology Act of 1992, that calls for NSF to connect K-12 schools to NSF-NET and to sponsor educational software and teacher training ($300M/5 years); NIST to develop networking technology for manufacturing ($250M); NIH, NLM, and NSF to develop applications for health care ($300M); and NSF, NASA, and DARPA to develop digital libraries and databases ($300M). OSTP would coordinate the activities under the HPCC initiative. Gore's office is (202) 224-4944. [David J. Farber (farber@central.cis.upenn.edu), com-priv, 7/1.] Text of Gore's bill is available via FTP as gorebill.1992.txt in /nren/iita.1992 on nic.merit.edu. You can also mail a "send gorebill.1992.txt" message to nis-info@nic.merit.edu. [Mark Davis-Craig (mad@merit.edu), com-priv, 7/6.]

- Something is happening. Windham Hill sells music through the Nautilus multimedia magazine-on-disc service. It also bypasses record stores by shipping encrypted discs to potential purchasers. Justice Records has an electronic record store on CompuServe. Cable companies are selling 30-channel "digital radio" and pay-per-listen programs. Syracuse University has hundreds of thousands of digitized recordings ready for downloading. [Peter Newcomb, Forbes, 7/20.] Newsletter and e-journal publishers are bypassing traditional publishers. Ross Perot was bypassing political parties and the media establishment. Shareware bypasses software publishers. Adobe sells pay-as-you-need-them fonts on CD ROMs. Bypassing middlemen is an age-old route to wealth. (The A&P grocery chain, for instance, started as a straight-off- the-ship tea retailer. McDonalds in Russia brings in its own beef. 7-11 in Japan is bypassing traditional distributors.) What's new is that computers are putting producers in direct contact with customers. That lowers costs, encourages small businesses, and strengthens the producer-consumer relationship. It's bound to be good for the economy and for our standard of living.

- The Oxford English Dictionary (2nd ed.) is now available on CD ROM, and can be searched with software from AND Software (Rotterdam). The last of its 616,500 entries is Zyxt, an obsolete word meaning "see." (Its new meaning of "the last word in the OED" isn't listed. :-) [BW, 8/3.] Oxford University Press sold about 1K copies of a 12-floppy set in 1987. This $875 ROM edition has 40% more data (including 2.4M illustrative quotations and 249K etymologies) and much better search software. The $2,750 hardback version fills 20 volumes, takes four feet of shelf space, and weighs 137 pounds. [Steve Lohr, NYT. SJM, 7/26.]

- Lance Leventhal sent me a review copy of Ellen Thro's "The Artificial Intelligence Dictionary" (Microtrend Books, Slawson Communications, (619) 744-2299). Its 1,300 definitions cover topics in logic, expert systems, neural networks, fuzzy systems, etc., and seem aimed at writers for the trade press. The Yale Shooting Problem is present (although cryptic); the DAI problem of The Dining Philosophers is not. ROV (remotely operated vehicle) is present, ALV (autonomous land vehicle) is not. The appendices seem likewise a bit sparse and idiosyncratic. Under abbreviations, DARPA is listed but MCC is not. I'm not sure why one would want lists of common operators in Lisp, Prolog, and OPS5, or multilingual equivalents of computer terms. Ellen does give a nice list of addresses -- but no phone numbers -- for 15 professional organizations, 6 magazines, 47 journals, 19 newsletters, 8 compendia, and 17 recurring conferences. (No wonder we can't keep up!) One appendix mentions 25 landmark expert systems; another gives affiliations -- but not addresses -- of 90 prominent researchers. This is an interesting browsing book for someone who half knows the material (e.g., students), but seems only half there as a reference book.

- The CD-ROM Directory (TFPL Publishing, London; UK#81) is a 870-page compilation CD-ROM and multimedia publishers, distributors, 2,212 titles, and related publications -- indexed by title, subject, country, computer type, and software. $150 from Omnigraphics, (313) 961-1340. Also available in CD ROM. [John McCormick, Newsbytes, 3/10.]

- Newsbytes is a computer-related news service carrying over 600 original write-ups each month. You can get two email deliveries or one PC/Mac floppy disk per week for $25/month. (DOS users get front-end software for menu and keyword access.) Newsbytes News Network, 822 Arkansas Street, San Francisco, CA 94107; (415) 550-7334. One month free if you cancel. Usenet and BBS distributions are also available. Material for the articles is often culled from other news wires, but tends to be pretty good.

- "Discovering RLIN" is a 52-page booklet introducing the Research Libraries Information Network database for reference librarians and scholars. RLIN materials include journals, rare books, manuscripts, letters, photographs, films, sound recordings, computer files, incunables, and realia. Distribution Services Center, The Research Libraries Group, Inc., 1200 Villa Street, Mountain View, CA 94041-1100. [Jennifer Porro (bl.jmp@rlg.bitnet), IRLIST, 6/15.] I presume that the booklet is free.

- "Information Resource Guides: Technology Sources and Bibliographies" is available free from the Institute for Academic Technology, P.O. Box 12017, Research Triangle Park, NC 27709-2017; (919) 560-5031, (919) 560-5047 Fax. You can also request it from pubs.iat@mhs.unc.edu or from Carolyn M. Kotlas (carolynk.iat@mhs.unc.edu), or FTP it from gandalf.iat.unc.edu. Topics include CD ROM and Laserdisc sources, computer lab design, copyright guidelines, computers and the humanities, software for higher education, distance education, internet resources, multimedia developers and associations, and wide-area information servers. [CCNEWS, 6/19.]

Publisher/Editor: Dr. Kenneth I. Laws, 4064 Sutherland Drive,
Palo Alto, CA 94303, USA. Phone: (415) 493-7390.
Internet: laws@ai.sri.com. (Courtesy of SRI International.)

News Items from "Electronic Frountier Foundation (EFF) Library"

EFFector On-line July 29, 1992 Issue 3.1 / Part 1
A Publication of the Electronic Frontier Foundation
ISSN 1062-9424

UPDATE ON THE EFF LIBRARY

The EFF Library was set up over a year ago when it became clear that, regardless of the digital revolution, we were being overwhelmed by a wave of hard copy. At that time, we had a backlog of around 1,000 documents, books and magazines concerning issues relevant to the Electronic Frontier. We engaged a professional librarian, Hae Young Wang, to bring order to chaos, and to provide us with a method that would enable us to file and retrieve material necessary to the work of the foundation.

Today, the EFF library in Cambridge houses over 2,300 items. The holdings cover journal articles, newspaper articles, conference proceedings, court documents, legislation, magazines, books, and brochures. The subject areas include such things as information infrastructure, computers and civil liberties, intellectual property and copyrights, and EFF archives.

The Library also maintains over 130 subscriptions to magazines and newsletters.

In addition, the EFF library maintains, classifies and indexes EFF's anonymous ftp archive files. These files, which are accessible to everyone with Internet access, have recently been re-organized into what we hope is a more user-friendly and informative manner. In the EFF ftp directory, you can find documents about the EFF, back issues of its on-line newsletter, notes on eff-issues, historical items, legal issues, current legislation, local chapters, and a host of other material germane to the Electronic Frontier.

While the ftp files are open to all, the EFF Library can now serve only the staff here and in Washington. We hope to be able to provide service to EFF members and the general public in the future, as funding and staffing allow.

In the meantime, we have recently acquired new scanning software which we hope will reduce the work involved in moving hard-copy information into digital form. With this in place we will be adding items to the anonymous ftp archive at an increased rate throughout the rest of the summer.

Recent additions to the EFF ftp files are:

The EFF Open Platform Proposal. This is the full text of the EFF's plan to create a national public network through the deployment of ISDN technology.
(pub/EFF/papers/open-platform-proposal)

Howard Rhinegold's "A Slice of Life in My Virtual Community".

This meditation on what it means to be on-line in 1992 was first serialized in EFFector On-line. (pub/EFF/papers/cyber/life-in-virtual-community)

Senator Al Gore's High-Tech Bill (S.2937) as introduced on July 1, 1992. This bill provides funding to both NSF and NASA to develop technology for "digital libraries", huge data bases that store text, imagery, video, and sound and are accessible over computer networks like NSFNET. The bill also funds development of prototype "digital libraries" around the country. This is the full-text of this bill along with the press release from Gore's office announcing the bill. (pub/EFF/legislation/gore-bill-1992)

An information packet on the GPO/WINDO legislation before congress as S.2813/H.R. 2772. This discusses the function of the proposed "gateway" for on-line public access to government databases. From the Taxpayers Assets Project. (pub/EFF/legislation/gpo-windo-info)

These files are also available through WAIS as eff-documents.src. WAIS clients are available for the Mac, PC, NeXT, X11, and GNU Emacs environments via anonymous ftp from think.com. A "guest" WAIS client is available by telnetting to quake.think.com and logging in as 'wais'.

To retrieve these files via email, send mail to archive-server@eff.org, containing (in the body of the message) the command

send eff <path from pub/EFF>

So to get the Gore bill, you would send

send eff legislation/gore-bill-1992

If you have any trouble obtaining these documents, send email to ftphelp@eff.org.

EFFector On-line 5.07
Summary of Rep. Boucher's Bill for NREN applications
Keywords: NREN,Boucher

Date: 30 Apr 93 20:13:12 GMT

In this issue:
Congressman Boucher Introduces NREN Applications Bill

"Future of Computing" Program in Palo Alto, CA
-==--==--==-<>-==--==--==-

Congressman Boucher Introduces NREN Applications Bill

--Offers greatly expanded vision of applications program
for widespread social benefit

by Andrew Blau
EFF Associate for Telecommunications Policy

On April 21, Congressman Rick Boucher (D-VA) introduced legislation to create computer and networking applications to serve the education, library, and health care communities, and to promote access to government information. The bill, H.R. 1757, significantly expands on similar provisions found in last year's "Information Infrastructure and Technology Act" (often referred to as "Gore II," then-Senator Gore's follow-up to his NREN bill, the High Performance Computing Act ("HPCA")), and the Senate bill to promote U.S. competitiveness, S. 4.

Boucher, who chairs the House Science Subcommittee which oversees the NSF, has held oversight hearings on the development of the NREN program at which EFF Chairman Mitch Kapor testified. Many of EFF's suggestions, and the suggestions of EFF's partners in the education, library, and health care sectors, have been included in this legislation.

Highlights include:
*a substantial broadening of the focus of NREN to accelerate progress toward "a universally accessible high-capacity and high speed data network for the nation";

*a significant commitment to public libraries, K-12 schools, and support for hardware purchases;

*the creation and inclusion of local 'civic networks' of local libraries, schools, and local and state government offices, which would be connected to the Internet;

*an emphasis on promoting access to government information; and

*a codification of the distinction between research and production networks.

This bill also shifts away from the manufacturing focus of the earlier bills; it has no provisions for manufacturing applications at all.

There are a handful of weak spots, most notably that the bill seems to emphasize broadband connections to the Internet, which EFF believes could drive up the costs of the connections program and reduce the number of beneficiaries; and the lack of any coordinating or responsible agency for the government information program, the network security program, the privacy program or the ease of use program.

EFF supports the approach outlined in this bill, and will be working to secure passage of it. We will also seek some minor modifications in order to improve the bill at the margins -- for example, to improve the access to information section in order to support putting federal information on-line and enabling innovative non-profit groups to make it available as demonstration projects, and to clarify that the broadband provisions are an option, not a mandate. Overall, however, EFF believes this is a substantive advance that merits widespread discussion and support.

EFF will make a copy of the full text of the bill in our ftp archives (ftp.eff.org).

Section-by-section review:
****

Sections 1 and 2 include the bill's title ("High Performance Computing and High Speed Networking Applications Act of 1993") and the Congressional findings that support the need for this legislation.

Sec. 3. Applications of the High Performance Computing Program. Contains the major provisions, which are proposed as an amendment to the original HPCA. Sections 301 through 305 cover administrative issues.

Sec. 301 establishes the applications program. The bill improves on S. 4 by specifying that the applications should be "designed to be accessible and usable by all persons in the United States"; adds the provision of government information to the program purposes, and mandates that the Plan to create applications must take into account the recommendations of the High Performance Computing Advisory Committee, which this bill also mandates will include representatives of the research, K-12, higher education, and library communities, consumer and public interest groups, network providers, and the computer, telecommunications and information industries.

Sec. 302 describes the Plan to implement the program. The Plan must: (a) be submitted within one year and revised at least once every two years; (b) include goals and priorities, specific responsibilities of agencies and departments to meet goals, recommend funding levels to departments; and (c) include progress reports, evaluations and recommendations.

Sec. 303 describes the role of the Federal Coordinating Committee for Science, Engineering, and Technology (FCCSET) for coordination among agencies and budget review.

Sec. 304 creates a new "Coordinator" position, which is to be chosen from the staff of the White House Office of Science and Technology Policy. The Coordinator is to monitor the agencies, report any discrepancies to the OSTP Director, assist in interagency coordination, and act as Congressional and public liaison.

Sec. 305 describes the annual reports that each agency is to submit to OMB and OMB's review and report to the President.

The major application areas:
****

Sec. 306 creates a program to foster network access. This is a new provision to create local networks of K-12 schools, libraries, state and local governments, etc. It includes support for buying hardware and connecting those local nets to the Internet; it also expands training to teachers, students, librarians, government personnel to use networks and the Internet. Note however, that the provisions specify broadband connections, which could slow down the program, increase the costs, and reduce the beneficiaries if institutions are not free to choose the most appropriate-sized connection for their needs. NSF is the lead agency. Over the next five years, it authorizes 20, 60, 70, 80 and 80 million dollars (i.e., $310 million). Sec. 307 calls for research into security and privacy of information, integrity of digital information, and ease of use for non specialists. This is also a new provision with no counterpart in S. 4. It authorizes 10, 30, 35, 38, and 38 million dollars over the next five years for these activities (i.e., $151 million). No lead agency is specified.

Sec. 308 outlines educational applications. H.R. 1757 broadens the range of educational applications compared to S. 4, and adds additional features to support the intent of this section. New provisions include: support for hardware and software purchases in order to demonstrate the educational value of the Internet; support for systems, software and networks for "informal education" including job training and life-long learning applications outside of school; a mandate to address the needs of rural and urban communities; a clearinghouse of K-12 network projects and available educational resources; and the creation of undergraduate level course materials for student teachers to familiarize them with the Internet and educational uses of computer and networking applications. Other elements are similar to or better specified versions of provisions found in S. 4 that call for projects to enable K-12 students and teachers to communicate with peers and university level students and teachers, and to gain access to educational materials and other computing resources. NSF is directed to be the lead agency, and the section authorizes 24, 70, 82, 94 and 94 million dollars over the next five years ($364 million) for education.

Sec. 309 outlines health care applications. This is a substantially expanded version of S. 4's health care section. The lead agency is shifted from the National Library of Medicine to the Department of Health and Human Services, which is to implement it through the NLM, the National Institutes of Health, and the Centers for Disease Control. H.R. 1757 also splits health care applications into three subsections. Besides clinical information systems, which repeats S. 4's six health care provisions for clinicians, H.R. 1757 adds two sections of entirely new provisions: health information to public, and health delivery systems and population data sets for epidemiology. The section authorizes 24, 70, 82, 94, and 94 million dollars ($364 million) over the next five years.

Applications for health information to the public include: consumer-oriented, interactive, multimedia materials for health promotion and distribution of such materials to public access points, such as community health and human service agencies, schools and public libraries; interactive, multimedia materials to assist patients in deciding among health care options; interfaces to allow non specialists ease of access and use; and the means to provide customized preventative and treatment information to non specialists.

Applications for health delivery systems and population data sets include: networks and software for communication among local public and private health and human service providers, e.g., health centers, clinics, entitlement offices, and school based clinics to enable social service providers to deliver coordinated services; access for health care providers to current clinic-based health promotion and disease prevention recommendations and two-way links with prevention specialists at state and local health departments; and database technologies to help clinicians diagnose, treat, and provide preventative information to patients and facilitate the gathering of systematic population data sets in order to measure treatments and national health trends.

Sec. 310 describes the applications programs for libraries. Most of this section describes the same digital library applications found in S. 4: terabit storage systems accessible by thousands of simultaneous users; high speed digitizing of printed and photographic materials; tools to search huge volumes of stored text, imagery, data and sound; encouragement of the development and adoption of standards; smart systems to categorize and organize information; training for librarians and database users; making networked databases easy to use; and visualization tools to help browse through large volumes of imagery. The subsection on the development of prototypes, however, is expanded in three significant ways. H.R. 1757 specifies that the prototypes should be testbeds for all the features noted above. Most importantly, H.R. 1757 specifies that the prototype libraries will be accessible to the public via the Internet. Lastly, H.R. 1757 requests an evaluation of the suitability and utility of distributing electronic information over the Internet, including an assessment of the barriers that hinder the use of the Internet for this purpose. H.R. 1757 also directs NASA to develop databases of remote-sensing images to be made available over computer networks. NSF is named as the lead agency, and 10, 30, 35, 44, and 44 million dollars ($163 million) is authorized over five years. For its part, NASA is authorized 6, 16, 20, 20, and 20 million dollars ($82 million) for the same period.

Sec. 311 calls for applications for government information. H.R. 1757 has a set of new provisions to promote public access to information generated by Federal, state and local governments. H.R. 1757 calls for projects that connect depository libraries and other sources of government information to the Internet to enable access to Federal, state and local government information, and access to "related resources" as well as linkages among libraries in order to enhance the use of that information. H.R. 1757 also calls for the creation of technologies to increase access to and effective use of government information in support of three goals: research and education; economic development; and an informed citizenry. Finally, the section mandates the creation of a Federal information locator to help the public find and retrieve government information. No agency is given coordinating or lead responsibilities, but the bill authorizes 8, 24, 26 30 and 30 million dollars over the next five years ($118 million).

Other provisions:
****

Section 4 changes the High Performance Computing Advisory Committee into a Computing *and Applications* Advisory Committee. It also adds representatives from K-12, consumer and public interest groups, and computer, telecommunications, and information industries. Among the Committee responsibilities is to assess whether the applications that are developed successfully address the needs of the targeted populations and to estimate the number of users served by the applications.

Section 5 rewrites Section 102 of the HPCA. Whereas HPCA proposed that portions of the NREN would reach gigabit transmission rates "to the extent technically feasible," this bill appears to assume gigabit networking and moves on to redefine test-bed networks separately. The Network Program now would have three parts: R&D to support gigabit transmission speeds; experimental test-bed networks to develop advanced networking technologies in the quest for gigabit networks and to support applications that exceed what commercial networks can handle; and a connections program to help researchers, educators and students obtain access to and use of the Internet.

H.R. 1757 adds a new section to the HPCA, 102(d), that would codify the distinction between experimental, "bleeding-edge" research networks and services available off-the-shelf from commercial service providers. The bill specifies that eighteen months after the bill is enacted, test bed networks are forbidden to provide services that could otherwise be provided satisfactorily over commercial networks.

Other sections include one that creates a new OSTP Associate Director to oversee Federal efforts to disseminate scientific and technical information, and a handful of miscellaneous provisions.

Program Announcement for Palo Alto, California from Ted Haynes of the Churchill Club

Terry Winograd and Jim Warren will speak on "The Future of Computing and Its Impact on Society", May 27, 1993, at the Hyatt Rickey's, Palo Alto, California; sponsored by the Churchill Club (415-321-9016). A reception and a light dinner begin at 6:00 PM with the program starting at 6:45 PM.

Terry Winograd is a Professor of Computer Science at Stanford and a founder of Computer Professionals for Social Responsibility. Jim Warren is a MicroTimes columnist, founder of Infoworld, and a founder of the Computers, Freedom and Privacy Conferences. They will be joined by Denny Brown, founder of Coherent Thought and President of Expert Support.

Will more powerful computers turn into twenty-first century servants or Big Brother? What are the implications for employment, economic growth, privacy, education, and the family? Come and find out!

The Churchill Club, founded in 1985, is a non-profit public affairs organization in Silicon Valley that provides a non-partisan forum on timely issues. Past speakers include Edward Teller, Bill Joy, Bill Clinton and Sandra Kurtzig. The club has 1100 members of which about 66% work in a "high tech" related company.

News Items from ALAWON

ALAWON
ALA Washington Office Newsline
An electronic publication of the
American Library Association Washington Office

Volume 2, Number 18
May 10, 1993

In this issue: (321 lines)
NREN APPLICATIONS BILL SUMMARIZED
HEARINGS HELD ON BOUCHER NETWORKING APPLICATIONS BILL
UPCOMING HEARINGS

NREN Applications Bill Summarized

HR 1757 WOULD ESTABLISH NREN PROGRAM

On April 21, Rep. Rick Boucher (D-VA) introduced HR 1757, the High Performance Computing and High Speed Networking Applications Act of 1993. The bill amends Section 102 of the High-Performance Computing Act of 1991, which defined the National Research and Education Network. Under HR 1757, the NREN Program would have three components: (1) research and development of gigabit networking software and hardware; (2) experimental test bed networks for developing and demonstrating gigabit networking technologies and providing connections requiring network performance not available from privately operated commercial networks; and (3) support for researchers, educators, and students to obtain access to and use of the Internet for communications with others in the research and education communities and to allow for access to high-performance computing systems, electronic information resources, other research facilities, and libraries.

Federal agencies involved must have a plan for providing financial assistance to educational institutions, public libraries and others to enable them to access the Internet. Test bed networks are not to be used "to provide services that could otherwise be provided satisfactorily using privately operated commercial networks" after 18 months from enactment of the bill. The Internet is defined as "the network of both Federal and non-Federal interoperable packet-switched data networks."

ACCESS, TRAINING, AND RESEARCH

HR 1757 would also amend the HPCA by adding a new title III for applications of computing and networking. This title would require the Director of the Office of Science and Technology Policy to establish the Applications Program, including a plan and responsibilities of the Federal Coordinating Council for Science, Engineering, and Technology, as well as designating a coordinator at OSTP for the Program.

A section on network access would authorize a Connections Program at the National Science Foundation for creation of local networks in communities. The networks would connect institutions of higher education, elementary and secondary schools, libraries, and state and local governments to each other and to the Internet. The section also calls for programs to train teachers, students, librarians and government personnel. "Training programs for librarians shall be designed to provide skills and training materials needed by librarians to instruct the public in the use of hardware and software for accessing and using computer networks and the Internet."

Reports are required within a year of enactment on the extent to which the education, library, and state and local government communities have access to the Internet, broadband connections to the Internet, the factors limiting access, an estimate of the cost of providing universal broadband access for those institutions, and recommendations for collaborative public-private sector programs to expand connectivity. Authorizations: FY94 \$20M, FY95 \$60M, FY96 \$70M, FY97 \$80M, FY98 \$80M.

A section on research in support of applications requires the plan to specify research in areas such as computer science and engineering, math, computer visualization, and human cognition needed to achieve the plan, plus research needs in network security and privacy, and ease of Internet use. Authorizations: FY94 $10M, FY95 $30M, FY96 $35M, FY97 $38M, FY98 $38M.

APPLICATIONS IN K-12, LIBRARIES, GOVT. INFORMATION

Applications for education would authorize NSF to develop and apply computing and networking technologies at all levels of education. K-12 is specifically identified for connections, acquisition of hardware and software, dissemination of information about use of computing and network technologies, development of courses and training, and development of education software. Authorizations: FY94 $24M, FY95 $70M, FY96 $94M, FY97 $94M, FY98 $94M.

Applications for health care would authorize the Department of Health and Human Services, through the National Library of Medicine and other HHS agencies, to focus on clinical information systems (including test bed networks for linking hospitals, medical libraries and other locations, as well as database technology for medical information and literature); health information for the public (including distribution to public access points such as public libraries); and health delivery systems and population data sets. Authorizations: FY94 $24M, FY95 $70M, FY96 $82M, FY97 $94M, FY98 $94M.

Applications for libraries would authorize NSF to support development of digital libraries and prototypes, including advanced data storage systems, advanced digitizing techniques, development of database software, development of standards, visualization methods, user-friendly tools, and training. Authorizations: FY94 $10M, FY95 $30M, FY96 $35, FY97 $44M, FY98 $44M. NASA would be authorized to develop databases of remote-sensing images. Authorizations: FY94 $6M, FY95 $16M, FY96 $20M, FY97 $20M, FY98 $20M.

Applications for government information would authorize unnamed agencies to support connecting depository libraries and other sources of government information to the Internet for access to federal, state, or local government information and related resources, and to support linkages with other libraries and institutions to enhance use of government information. Agencies could also support demonstration projects for improved public access to government information. This section includes establishment of a federal information locator system accessible over the Internet. Authorizations: FY94 $8M, FY95 $24M, FY96 $26M, FY97

$30M, FY98 $30M.

The bill also adds a high-performance computing and applications advisory committee with widely representative non-federal membership, including the elementary and secondary education, higher education, and library communities.

HR 1757 WOULD MAKE NREN A TRUE PROGRAM

The strength of HR 1757 is that it would implement the NREN Program as originally envisioned in the HPCA, adding substantive provisions for school and library connections and support for applications and training in areas useful to these institutions and the public. A welcome section on government information applications includes many of the provisions suggested by ALA. The science agencies to which funds would be authorized are those within the jurisdiction of the Science Subcommittee, but these agencies, particularly NSF, are to work in cooperation with other appropriate agencies.

The troublesome provisions relate to an artificial distinction between experimental test bed networks and production uses of networks, with a rigid 18-month timetable for such separation. These provisions reflect attempts to accommodate the telecommunications companies who recently produced a policy statement identifying their interest in network and national information infrastructure issues. However, the provisions would potentially affect the current structure of backbone and regional networks which are operated on top of private-sector communications facilities (which receive revenue for transport), but add services, training, and applications support which are not provided by the large communications companies.

Hearings Held on Boucher Networking Applications Bill

ADMINISTRATION VIEWS

Chairman Rick Boucher (D-VA) presided at hearings on his bill, the High Performance Computing and High Speed Networking Applications Act (HR 1757) on April 27 and May 6, 1993. The first hearing opened with John Gibbons, Director of the Office of Science and Technology Policy, providing the Administration's view. He discussed the Clinton Technology Initiative, of which the High Performance Computing and Communications Program is a critical part. The HPCC program has five components in the Administration's FY94 budget: (1) High

Performance Computing Systems, (2) National Research and Education Network, (3) Advanced Software Technology and Algorithms, (4) Basic Research and Human Resources, and (5) Information Infrastructure Technology and Applications.

Gibbons said HR 1757 is "important, forward-looking legislation and is largely consistent with the Administration's proposal for developing the nation's information infrastructure," including "very important provisions for improving the dissemination of Federal information." He felt roles for other federal agencies were needed, and he said the experimental/production distinction and the 18-month timetable could limit agency flexibility and complicate the difficult tasks of spurring private sector development and assisting the research and education communities.

INDUSTRY VIEWS

A panel of industry witnesses, Thomas Tauke of NYNEX, Robert Ewald of Cray Research (representing the Computer Systems Policy Project), W. B. Barker of BBN Communications, and Richard Rashid of Microsoft Corporation, were generally supportive of the bill. However, Tauke felt the restrictions in the bill in not allowing general ("production") use of federally subsidized test bed networks should extend to other areas of the bill as well. For instance, Tauke did not think the federal government should encourage development of local networks, nor did he see why schools and libraries would want to be connected to each other in local networks.

In the question and answer session, Boucher attempted to get general agreement from the panel on his vision of the government role which he saw as funding for new network technology research and development, support for applications such as those in the bill, creation of a test bed to develop technologies created through government R&D, stimulating setting of common standards and protocols, and helping institutions such as schools and libraries to get connected. The main disagreement was not over whether schools and libraries should be helped, but over how they should be helped. Tauke said local schools and libraries should be helped by providing government funding to them to purchase commercial services. He felt educational or other government supported networks might market their excess communications capacity. He admitted commercial providers do not always offer comparable services to Internet data communications at comparable prices, but said this was a regulatory problem that Congress should remedy. Barker's view was that government should support school and library connections by procuring them in the most economical manner, not by dictating one manner independent of costs.

APPLICATIONS AREAS - SCHOOLS, LIBRARIES

On May 6, the Science Subcommittee heard from several groups representing users of networks and applications. First to testify was Rep. Major Owens (D-NY), who stressed that barrier-free access to information was critical for all Americans, including people in urban and rural areas, people with disabilities, minorities, and the poor. He recommended that the bill "recognize public libraries as central to a free and comprehensive information delivery system available to everyone."

Don Detmer, Vice President for Health Services, University of Virginia, discussed the implications of high speed networking for health care. Connie Stout, Director of the Texas Educational Network, described the TENET system, a state-supported school network with 21,000 users. TENET was successful, she said, because it was simple to use and training was easily available. She also noted the specific difficulties in procuring equivalent services (connection to the network, training, etc.) from the commercial environment.

John Masten, Chief Operating Officer of the New York Public Library, commended the committee for including library applications, for the bill's focus on training, and for support of public libraries as network points of access. He emphasized that libraries of various types have a role in all program components of the bill. The NYPL experience in local and statewide cooperative efforts to increase access to electronic information, including a menu driven Internet gateway, involvement in creating and distributing a simple access kit for small libraries and schools, and close work with the New York State Internet Access Network (NYSERNET), made the library a believer in using local and regional network to reach the real needs of users.

Martin Massengale, President of the University of Nebraska, represented the National Association of State Universities and Land-Grant Colleges. Massengale made several points supported by ALA, EDUCOM, and others. He said he understood there were concerns that government-supported networking activities would interfere with commercial networks, and that the Internet was something separate from the public switched network. In actuality, he said, they all work together; "telephone lines connecting the campuses to the mid-level networks and the mid-level connections to the NSFnet are leased from regional, local telephone companies and interexchange carriers."

While strongly supporting most of HR 1757, Massengale expressed concern over the rigid timetable in Section 5 for change in the current structure for Internet support. Any transition from government support of networks to serve public missions should take into account performance based tests such as price, accessibility, breadth of service, competitive options, availability, ubiquity, and timing. Without such considerations, the bill's language may unnecessarily restrict federal and state government actions in the public interest.

INFORMATION INDUSTRY VIEWS

Cynthia Braddon, Vice President of McGraw-Hill, presented testimony for the Information Industry Association. She said the value of information content was missing from the bill. She recommended that development and implementation of a mechanism for compensating copyright proprietors be added to the list of digital library applications. IIA recommendations for government information applications included designating a lead agency, reflecting the essential role of the industry in enabling public access to information held by the government, and requiring determination of the competitive impact on other sources of information of any government agency initiatives.

Boucher asked the panel if the bill identified all the applications that needed to be supported. Stout replied that the education community hasn't had access, so it is hard for them to identify needed applications. Many schools do not have the hardware or the phone lines for electronic networking: an average elementary school has two phone lines; a junior high school has three, and a high school has six. Training is another need; 65% of TENET's budget is spent on training. Other witnesses also stressed the need for training.

A third hearing on HR 1757 is planned for Tuesday, May 11 (see below).

Upcoming Hearings

The following hearings were announced in the Congressional Record of May 7, 1993. Locations, dates, and times are subject to change.

HOUSE COMMITTEE ON APPROPRIATIONS, Subcommittee on Interior. Public witnesses for National Endowment for the Arts, National Endowment for the Humanities, and the Institute of Museum Services. Wednesday, May 12, 9:30 am and 1:30 pm, 2360 Rayburn.

HOUSE COMMITTEE ON SCIENCE, SPACE, AND TECHNOLOGY, Subcommittee on Science. Continuing hearings on HR 1757, the High Performance Computing and High Speed Networking Applications Act of 1993. Tuesday, May 11, 9:00 am, 2318 Rayburn.

JOINT COMMITTEE ON PRINTING. Hearing to review Congressional printing and other activities of the Government Printing Office. Tuesday, May 11, 9:30, SR-301.

About ALAWON:

ALAWON is an irregular publication of the American Library Association Washington Office, 110 Maryland Avenue, N.E., Washington, DC 20002-5675. Phone: 202-547-4440; Fax: 202-547-7363; Internet: alawash@alawash.org. Editor and List Owner: Fred King (fdk@alawash.org). All or part of ALAWON may be redistributed, with appropriate credits.

ALAWON is available free of charge and is available only in electronic form. To subscribe, send the message "subscribe ala-wo [your name]" to listserv@uicvm (Bitnet) or listserv@uicvm.uic.edu (internet). Back issues of ALAWON are available from the list server. To find out what's available, send the message "send ala-wo filelist" to the listserv. The ALA-WO filelist contains the list of files with the exact filename and filetype. To get a particular file, issue the command "send filename filetype" to the listserv.

Other News Items and Announcements

ALCTS Announces Publication of "After the Electronic Revolution, Will You Be the First To Go?

ALCTS announces the publication of AFTER THE ELECTRONIC REVOLUTION, WILL YOU BE THE FIRST TO GO?, the proceedings of the 1992 Association for Library Collections & Technical Services President's Program held on 29 June 1992 at the American Library Association Annual Conference in San Francisco, CA. The editor of the publication and 1991-1992 President of ALCTS is Arnold Hirshon, University Librarian at Wright State University in Dayton Ohio.

The idea for the program was inspired by the pioneering work of those who have spoken and written about paradigm shifts, such as Thomas Kuhn and Joel Barker. Kuhn spoke about paradigm shifts in the scientific community, and Barker applied those observations to the world of management. Barker instructs us that when a paradigm shifts, our past success will guarantee nothing -- "When the paradigm shift occurs, everyone goes back to zero." The electronic revolution currently occurring in libraries is such a paradigm shift.

In his introductory chapters, on "The Convergence of Publishing and Bibliographic Access," Hirshon provides the general perspective for the proceedings and sets forth a general premise that the information access is moving from an indirect two-stage process of searching-then-retrieving information to a direct process where the search of the index and the delivery of information is done virtually at the same time.

The keynote speaker, Theodor Holm Nelson, coined both the term and the concept of hypertext over 25 years ago. Today he is actively engaged in Project Xanadu, a developmental project that seeks to create a world-wide electronic publishing network. In Nelson's presentation "You the Guardians of Literature Still," he addresses the changes in electronic communication and observes that the system of literature of which librarians have been the guardians is undergoing a paradigm shift. Nelson asserts that we are moving from a system of providing whole documents to one where the user purchases only separate pieces, with the quoted material bought from the publisher at the moment of request and with automatic royalty payment to the author.

Peter S. Graham (Associate University Librarian for Technical and Networked Services at Rutgers, The State University of New Jersey) writes about "Intellectual Preservation in the Electronic Environment." Addressing the technical services perspective of libraries, Graham speaks particularly to the preservation of electronic information, and explores the implications to the profession when electronic texts remove the confidence that we have had in the past in the fixity of text in the print world.

Thomas Duncan (Faculty Assistant for the Museum Informatics Project and Associate Professor of Integrative Biology at the University of California at Berkeley) writes on the "Implementation of Electronic Information Systems in Universities and the Implications for Change in Scholarly Research." Writing from the perspective of an informed user, Duncan notes that university faculty must develop a new structure for using electronic information, and this will have a serious impact on future research, teaching, and public service in universities.

Susan K. Martin (University Librarian, Georgetown University) in "Librarians on a Tightrope: Getting from Here to There and Loosening Up in the Process," provides a public services administrator's perspective. Explicating the Council for Library Resources funded "Strategic Visions Statement," Martin notes that if librarianship is to assume responsibility for molding the information environment, we must take steps to reinforce and emphasize our leadership role, which is now minimal or nonexistent.

These proceedings challenge all librarians to be electronic revolutionaries. The 1992 ALCTS President's Program was the first of two President's Programs to address the effects of electronic publishing on libraries. The second program will occur at the 1993 ALA Annual Conference. This 1992 program included a distinguished panel of individuals whose role was not to provide answers, but rather to raise questions about where information access is going both outside of librarianship and within it. It is up the reader to decide whether after the electronic revolution, will you be the first to go?

To order AFTER THE ELECTRONIC REVOLUTION, WILL YOU BE THE FIRST TO GO?" (ISBN 0-8389-7650-6), please contact the ALA Order Department, 50 East Huron St., Chicago, IL 60611; telephone 800-545-2433, press 7.

Clinton OKs Computer Access To Congressional Record, Federal Register

WASHINGTON (AP) _ The Congressional Record and the Federal Register, major sources of information about Congress and federal agency rules, will be made available through computers under a bill signed Tuesday by President Clinton.

The measure directs the Government Printing Office to establish a system for electronic dissemination of the material. A statement by Clinton said the system "will complement, not supplant, commercial information services and federal agency information dissemination programs."

AP-DS-06-08-93 1907EDT

Counterpoint Publishing and The Internet Company, (Internet.COM) announce

Gopher and WAIS Access to the Federal Register via the Global Internet

March 1, 1993

Introduction

The full text of the U. S. Federal Register is now available on the Global Internet via Gopher and WAIS. Anyone with a direct connection to the Internet can now browse, search and retrieve the full text of any article printed in the Federal Register. Articles appear the same day that the GPO makes them available in electronic format.

Methods of Access

The Internet Federal Register is available to any Internet connected host. There are three access methods which require a "client program" to run on your computer. All client programs mentioned are available free of charge to any Internet connected site, via FTP from various archives. The final access method requires only that you run the standard "TELNET" client on your computer. All TCP/IP packages available for popular computers support TELNET.

Gopher Access

Gopher is a "user-friendly" menuing interface to information, developed by the University of Minnesota. To access a Gopher server on the Internet, you run a client program on an Internet connected host. Clients are available for VAX/VMS, VM/CMS, MS-DOS, and Macintosh computers.

Gopher also integrates WAIS within itself, so using your single Gopher client program you can conduct WAIS searches, as well.

WAIS Access

WAIS (Wide Area Information Server) is a power search and retrieval engine developed by Brewster Kahle, (then of The Thinking Machines Company). WAIS support "relevance feedback searching" which allows a user to specify their search parameters not only as keywords, but to refine the search to favor documents or articles which are "like" a particular document. To access WAIS, you run a client program on an Internet accessable host. Clients are available for most popular computers.

NNTP (Usenet News) Access

NNTP is the standard way that news articles are transferred over the Global Internet. The Federal Register/NNTP Service breaks the Federal Register out into +/- 18 separate newsgroups in categories like:

argiculture
commerce
defense
education
energy
environ
finance
foreign
govern
health
humanserv
legal
science
transport
misc

Federal Register/NNTP Service allows you to keep the entire text of the Federal Register on-line, browsable using any standard Usenet newsreader.

TELNET Access

If you are unable to use Gopher/WAIS locally on your computer, or do not wish to contract for the multiple user pricing, access to a Gopher/WAIS "TELNET" account is also provided. This account will allow you to use Gopher and WAIS without having to run the client locally on your computer. The only requirement is that your computer support the standard Internet "TELNET" terminal access program.

Pricing

Gopher/WAIS Server Access

Gopher/WAIS pricing is measured by the number of regular users on your site who will have access to the Federal Register. Prices are yearly. Access to this service can be limited by Internet host or domain address. Domain address limiting allows an entire Internet domain, (eg. *.purdue.edu) to access this service.

Limitations

There are no usage limitations on this information. You are free to use any and all information gathered from this service within your own organization. You are not allowed to serve this information to any site or individual not affiliated with your organization. Re-publication rights are available by arrangement.

To Order, either call Counterpoint Publishing at 1-800-998-4515, or send email to "fedreg@internet.com" with your name, affiliation, and telephone number. Someone will contact you as soon as possible.

'Library of the Future' Takes Shape at Columbia University Law Library

New York City, February 1, 1993..........Today, amidst the nation's third largest collection of legal materials, teeming with well over 700,000 volumes and a half-million sheets of microfilm, Columbia Law School publicly launched the nation's first 'virtual' library by announcing the installation of the world's leading parallel supercomputing system, the Connection Machine System, CM-2*, from Thinking Machines Corporation of Cambridge, Massachusetts. The 'virtual' library is part of Columbia's Project JANUS, a 5 year project aimed at making the University's vast resources more available to its students and researchers by allowing researchers to use electronic retrieval of complete text with unaltered format and graphics instead of less sophisticated text-only electronic sources or the traditional paper-based library system. Recent advances in the conversion and storage of full-text page images digitally are being coupled with a powerful new generation of search software on the CM-2 supercomputer developed by scientists at Thinking Machines.

"Columbia's system will include easy-to-learn, easy-to-use interfaces, more effective search methods, and the ability to make scattered text databases appear to be a single database. All these are important advances.

"Our goal is to create a library which is completely free-text searchable," said Willem Scholten, Columbia's Director of Computer Systems and Research.

"The library must allow users to 'browse' comfortably through the library environment, while also providing enhanced ability to locate a particular piece of information."

According to Scholten, free-text searches using natural language queries are more useful and accurate than keyword searches. A free-text system does not require users to have prior knowledge of the discipline they are searching and allows users to search the full text of documents for arbitrary combinations of words. Keyword systems require that indexers read all documents and assign each one keywords from a controlled vocabulary, a process is which is laborious, as well as imprecise and inconsistent.

With the JANUS system, for example, a researcher of free-trade agreements typing in "lowering US automobile production costs" would receive within seconds dozens of references. References would be listed in descending order based on how closely the text matches the typed in request. The system uses a complex weighing algorithm that focuses on uncommon words to direct the search.

A researcher could proceed to retrieve in seconds an optically exact text of a page or document, highlighting a paragraph that describes how the auto industry cut production costs, and request an additional search of texts matching that paragraph. The system would again in seconds list references in descending order based on how closely the text relates to the highlighted paragraph. The researcher could then download or print the documents -- or could continue on an even narrower search based on the new information.

Key to Project JANUS is that complete works with text, format, illustrations and graphics are digitally stored and electronically searchable. At Columbia, books and documents are scanned using an optical character recognition scanner both to produce a digital text file in which any word or phrase can be recognized by the computer and used by the search software and to produce a high definition optical image file which is an exact copy of the document seen by the user.

As the project progresses, Columbia will work with publishers to receive copyrighted materials directly in electronic form and to develop programs to track and verify the use of licensed materials electronically to comply with copyright laws. The project will also work towards establishing standard formats for scanned materials and making the system available nationally.

Thinking Machines Corporation is the world's largest manufacturer of highly parallel supercomputers. The CM-2, installed at the Columbia Law Library, is equipped with 32,000 processors and has 256 megabytes of main memory and 20 gigabytes (billions of bytes) of hard disk storage.

The Columbia Law School Library is the nation's third largest collection of legal materials, with over 700,000 volumes and over 500,000 microform items.

The collections cover the law in all its forms and for every country, and include the Toshiba Library for Japanese Legal Research. The library by 1996 expects to convert 10,000 to 12,000 volumes annually to computer storage, which is about the number of volumes it adds to its holdings annually.

Project JANUS is funded by Columbia Law School and Columbia University Libraries. The School seeks funding for the remaining development program.

*Connection Machine, CM-2 and Thinking Machines are
Trademarks of Thinking Machines Corporation.

Project JANUS - The Electronic Library Development Project
Columbia University School Of Law

National Science Foundation Notice of Briefing Meeting

The NSF will hold a Briefing Meeting for the NSF/ARPA/NASA "Research on Digital Libraries" Initiative (Announcement NSF-93-141). The meeting will take place on Dec. 6, 10 am -12 noon at the Auditorium of the National Academy of Sciences, 2100 C Street NW, Washington DC. This meeting will be open to all parties interested in responding to this Initiative. For further information, please contact gbarber@nsf.gov

OCLC Awards Three Research Grants

DUBLIN, Ohio, June 17, 1993--The OCLC office of research has awarded three Library and Information Science Research Grants (LISRG) to university researchers for 1993. "We are pleased to be able to support these projects," said Martin Dillon, former director of the office of research and current director of OCLC's library technical services division. "University-based research adds an important dimension to our research agenda, and these projects promise findings that should be of broad interest to the library and information science community."

The grant recipients and their projects are: Carolyn O. Frost, Ph.D., associate dean, School of Information and Library Studies, University of Michigan: "An Empirical Test of Gopher Searching Using Three Organization Schemes." Gopher is an example of a commonly used protocol for searching networked information. Although the use of networked information is proliferating at an astounding rate and is providing unprecedented access to information, the organization of this information has not kept up with its use. Likewise, there is need for a greater understanding of information searching in order to design better searching tools, organize networked information more effectively, and assist information providers in the mounting of networked sources. There has been little research on how people use Gopher, or its effectiveness as a searching tool. The proposed project will study Gopher users' information searching behavior, identify patterns and problems in the searching behavior or with the information retrieval, and recommend changes for improvement.

<text deleted>

The OCLC Library and Information Science Research Grant program awards grants of up to $10,000 to help foster quality research by faculty in schools of library and information science. Projects are generally completed within one year, and findings are published in the OCLC Annual Review of Research and in other scholarly communications. Application materials for 1994 will be available this November. For more information, contact the office of research. OCLC is a nonprofit computer library service and research organization whose computer network and products link more than 16,000 library in 47 countries and territories. (KS/NC)

FOR IMMEDIATE RELEASE. FOR MORE INFORMATION CALL:
Keith Shafer (614) 761-5049
Nita Dean (614) 761-5002

The On-line Book Initiative - Announcement

The On-line Book Initiative is being formed to make available freely redistributable collections of information. There exists huge collections of books, conference proceedings, reference material, catalogues, etc. which can be freely shared. Some of it is in machine-readable form, much of it isn't.

The purpose of the On-line Book Initiative is to create a publicly accessible repository for this information, a net-worker's library.

Information in the On-line Book Repository will be available for free redistribution. On-line access, magnetic media and other methods of distribution will involve reasonable charges for the services provided, not the information.

WHAT WE WISH TO ARCHIVE

All on-line materials (other than software collections) such as books, journals, catalogues, conference proceedings, magazines, manuals, maps, images, technical documentation, reference works, etc. The only software we are interested in is software specific to the viewing, manipulation, searching and maintenance of information in the repository.

Materials must be free of copyrights limiting redistribution by us or any individual or organization who receives them.

We also need pointers to collections of materials which may be available. For example, there are government collections of interesting data which are available at reasonable costs and do not limit further redistribution of copies obtained.

WHAT WE NEED FROM YOU

Beyond machine-readable material there are huge collections of printed material which could be redistributed if put on-line. We need people willing to organize informal projects to scan, type or otherwise get this material on-line for inclusion in the On-line Book Repository.

We need to get in touch with Library and Information Scientists interested in helping us create formats and structures for organizing the repository.

We need international participation to help ensure our efforts are useful to people everywhere.

We need people willing to participate in a Technical Advisory Board to help us guide our efforts.

We need involvement from academia, industry and governments to help us enrich this effort without bounds and make available a first-rate, freely available information utility.

We need involvement from publishers who have materials which can be included in the On-line Book Repository. Many books and reference works become unprofitable to publish by ordinary paper means. It's time to make these materials available!

We need involvement from the technical community to choose and implement multi-media software standards such as hypertext, mark-up languages, index and catalogue software, text retrieval, network access methods and more. Standards are critical to our efforts.

WHAT WE ARE OFFERING

WORLD.STD.COM is a public access UNIX system which will serve as the initial repository. WORLD.STD.COM is a large Solbourne file server (over 100MIPS, 4 CPUS, 256MB, 8GB.) We soon (2/2/93) plan to put OBI onto its own server machine which will be called OBI.STD.COM; that name will work now and will be correctly aliased on our end as needed.

Anyone can dial into the system and set up an account if they wish direct access (617-739-WRLD.) Accounts are charged and proceeds will be used to build the On-line Book Repository.

HOW TO GET INVOLVED

If you think you can help or want more information send electronic mail to: obi@world.std.com

Or call us at Software Tool & Die, 617-739-0202. Or drop by our office and chat if you're in the area: 1330 Beacon Street, Brookline, MA 02146.

POSTSCRIPT

This started as an informal discussion group which called themselves "The KiloMonkeys Project" ("Strong Typing For Weak Minds") who wanted to figure out how to get useful materials on-line and generally available. I have decided to make Software Tool & Die a home for this activity and formalize the project under the new name "The On-line Book Initiative."

CHAPTER 5

Directory of Interested Parties

[Editor's Note: If you notice errors or omissions here, please let us know so we can keep this up-to-date.]

People

Akeroyd, Richard G., Jr.

Connecticut State Library
State Librarian
231 Capitol Avenue
Hartford, Connecticut 06106-1537
Phone: 203-566-4301
Fax: 203-566-8940
rakeroyd@eagle.wesleyan.edu

Arms, William Y.

Carnegie Mellon University
Vice President for Academic Services
Warner Hall, 304
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213
Phone: 412-268-2122
Fax: 412-268-7886
wya@andrew.cmu.edu

Anderson, Greg

MIT Libraries
Massachusetts Institute of Technology
Associate Director for Systems and Planning
Room 14S-216
Cambridge, Massachusetts 02139
Phone: 617-253-5654
Fax: 617-253-8894
ganderso@athena.mit.edu

Andre, Pamela Q. J.

National Agricultural Library
Associate Director for Automation
Information Systems Division
10301 Baltimore Boulevard
Beltsville, Maryland 20705
Phone: 301-504-6813
Fax: 301-504-7473
pandre@asrr.arsusda.gov

Atkins, Dan

University of Michigan
Dean, Library and Information Science
Chair, Working Group on Inpact of National Digital Library (NDL) on Education

Atkinson, Ross W.

Cornell University
Assistant University Librarian for Collection Development
201 Olin Library
Ithaca, New York 14853-3689
Phone: 607-255-3393

Fax: 607-255-9346
rwa@cornellc.bitnet

Badger, Robert C.
Springer-Verlag New York
Manager, Electronic Media
175 Fifth Avenue
New York, New York 10010
Phone: 212-460-1622
Fax: 212-533-5781

Becker, Herbert S.
Library of Congress
Director, Information Technology Services
Madison Building
Information Technology Services LM G51
Washington, D.C. 20540-2300
Phone: 202-707-6207
Fax: 202-707-0955

Besen, Stanley
Charles River Associates
Phone: 202-393-5840
Fax: 202-393-3432

Billick, David J.
UMI
Vice President of Document Acquisition and Imaging
300 North Zeeb Road
Ann Arbor, Michigan 48106
Phone: 313-761-4700
Fax: 313-761-1204

Billington, James
Library of Congress
Librarian of Congress
Washington, DC 20540

Borgman, Christine
Graduate School of Library and Information Science
Professor

300 Circle Drive North
University of California at Los Angeles
Los Angeles, California 90024
Phone: 310-825-6164
Fax: 310-206-4460
IIN4CLB@MVS.OAC.UCLA.EDU

Braid, James Andrew
British Library
Head of Development, Document Supply Care
Boston Spa
Wetherby, West Yorkshire
ENGLAND LS23 7BQ
Phone: +937 546030
Fax: +937 546236
andrew.braid@london.bl.uk

Brett II, George H.
Clearinghouse for Networked Information Discovery and Retrieval (CNIDR)
Director
Center for Communications - MCNC
P.O. Box 12889
Research Triangle Park, North 27709-2889
Phone: 919-248-1886
Fax: 919-248-1405
GEORGE.BRETT@CNIDR.ORG

Burke, Jane
NOTIS Systems, Inc.
President
1007 Church Street, 2nd Floor
Evanston, Illinois 60201-3622
Phone: 708-866-0170
Fax: 708-866-0178

Butler, Meredith
State Univeristy of New York at Albany
Dean & Director of Libraries
University Libraries 123

Albany, New York 12222
Voice: 518-442-3568
Fax: 518-442-3567
mB801@uacsc1.albany.edu

Caviness, Jane S.

National Science Foundation
Deputy Divison Director
Division for Networking and Communications Research and Infrastructure
4201 Wilson Boulevard, Room 1175
Arlington, VA 22230
Phone: 202-306-1950x1117
jcavines@nsf.gov

Cerf, Vinton G.

Corporation of National Research Initiatives (CNRI)
1895 Preston White Drive, Suite 100
Reston, VA 22091
Phone: 703-620-8990
Fax: 703-620-0913
vcerf@nri.reston.va.us

Chen, Su-Shing

National Science Foundation
Intelligent Search Algorithms and Natural Language Processing
4201 Wilson Boulevard, Room 1115
Arlington, VA 22230
Phone: 703-306-1925x1178
schen@nsf.gov

Christian, Eliot

U.S. Geological Survey
Information Systems Divison
802 National Center
Reston, Virginia 22902
Phone: 703-648-7245
Fax: 703-648-7069
echristi@usgs.gov

Chien, Y. T.

National Science Foundation
IRIS (Information, Robotics, and Intelligent Systems) Division
Director
4201 Wilson Boulevard, Room 1115
Arlington, VA 22230
Phone: 703-306-1927x1173
ytchien@nsf.gov

Chesler, Steve

Apple Library of Tomorrow
Manager
10381 Bandley Dr. M58C
Cupertino, CA 95014
Phone: 408-974-3258
sac@apple.com

Crocca, William T.

Xerox/PARC
CLASS Project since 1989 with Cornell Libraries
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Conley, Laura M. H.

John Wiley & Sons, Inc.
Director of Research and Planning
605 Third Avenue
New York, New York 10158
Phone: 212-850-6219
Fax: 212-850-6154
lconley@jwiley.com

Crystal, Tom

DARPA

Deogun, Jitender S.

University of Nebraska - Lincoln
Department of Computer Science
Lincoln, NE 68588-0115

Phone: 402-472-5033
402-472-3200
Fax: 402-472-7767
deogun@csr.unl.edu

Dillon, Martin

OCLC, Inc.
Director, Office of Research
6565 Frantz Road
Dublin, Ohio 43017-0702
Phone: 614-764-4365
Fax: 614-764-2344
mjd@oclc.org

Dyson, Esther

EDventure Holding, Inc
Writer, Publisher EDITs Project

Eaton, John

Virginia Tech
Associate Dean - Graduate School
Sandy Hall
Blacksburg, Virginia 24061-0325
Phone: 703-231-5645
Fax: 703-231-3714
eaton@vtvm1.cc.vt.edu

Eisenburg, Michael

ERIC Clearinghouse on Information Resources
Director
030 Huntington Hall
Syracuse University
Syracuse, New York 13244
Phone: 315-443-3640
Fax: 315-443-5732
mike@suvm

Ellis, William

Library of Congress
well@seql.loc.gov

Engelbart, Douglas C.
Bootstrap Institute
Director
6505 Kaiser Drive
Fremont, California 94555
Phone: 510-713-3552
Fax: 510-793-2362
engelbart@bootstrap.sri.com

Ertel, Monica
Apple Library of Tomorrow
Library Manager

Faloutsos, Christos
University of Maryland
Department of Computer Science
A.V. Williams Building
College Park, MD 20742-3255

Fields, Martha
State University System of Florida
Director, Information Resource Management
325 West Gaines Street
Suite 1420
Tallahassee, Florida 32399-1950
Phone: 904-488-6030
Fax: 904-487-4568
maf@nwbr400.mail.ufl.edu

Finnigan, Georgia
The Information Store, Inc.

Fox, Edward A.
Virginia Tech
Department of Computer Science
638 McBryde Hall
Blacksburg, Virginia 24061-0106
Phone: 703-231-5113
Fax: 703-231-6075

fox@fox.cs.vt.edu

Garcia-Molina, Hector
Stanford University
Department of Computer Science
Stanford, CA 94305-2140
Phone: 415-723-0685
Fax: 415-725-7411
hector@cs.stanford.edu

Garrett, John
Corporation of National Research Initiatives (CNRI)
1895 Preston White Drive, Suite 100
Phone: 703-620-8990
Fax: 703-620-0913
jgarrett@nri.reston.va.us

Gass, Linda
Adobe Systems, Inc
1585 Charleston Road
P.O. Box 7900
Mountain View, CA 94039-7900
Phone: 415-961-4400
Fax: 415-961-3769

Graham, Margaret
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Graham, Peter S.
Rutgers University
Associate University Librarian for Technology & Networked Information Services
169 College Avenue
New Brunswick, New Jersey 08903
Phone: 908-932-7505
Fax: 908-932-7637
graham@zodiac.rutgers.edu

Grisham, Frank P.

Southeastern Library Network, Inc. (SOLINET)
Executive Director
1438 West Peachtree Street
Suite 200
Atlanta, Georgia, 30309-2955
Phone: 800-999-8558
Fax: 404-892-7879
sgrisham@mail.solinet.net

Grycz, Czeslaw Jan

University of California, Office of the President
Chair, Scholarship and Technology Study Project (CNI)
Kaiser Center, Eighth Floor
Oakland, California 94612-3550
Phone: 510-987-0561
Fax: 510-527-4512
cjg@dla.ucop.edu

Hankins, John L.

CICNet, Inc.
Assistant Director
2901 Hubbard, Pod G
Ann Arbor, Michigan 48105
Phone: 313-998-6102
Fax: 313-998-6105
hankins@cic.net

Hall, Tony

Dialog Information Services
Director of Planning

Hartzband, David

Digital Equipment Corporation (DEC)
hartzband@cals.enet.dec.com

Henry, Charles

Vassar College Libraries
Poughkeepsie, New York 12601
Phone: 914-437-5782
Fax: 914-437-5864
chhenry@vassar.edu

Heterick, Robert Jr.
EDUCOM
President
1112 16th Street, N.W., Suite #600
Washington, D.C.
Phone: 202-872-4200
Fax: 202-872-4318
heterick@educom.edu

Hoffman, Ellen S.
Merit Network, Inc.
Manager, Network Information Services
2901 Hubbard, Pod G
Ann Arbor, Michigan 48109-2016
Phone: 313-936-3000
Fax: 313-747-3185
ellen@merit.edu

Hopcroft, John
Cornell University
Department of Computer Science
4130 Upson Hall
Ithaca, NY 14853-7501

Hunter, Beverly
National Science Foundation
Former Directorate for Education and Human Resources
Former Program Director, Application of Advanced Technologies,
Division of Materials Development, Research and Information Science

Hunter, Karen
Elsevier Science Publishers BV
Vice President and Assistant to the Chairman
655 Avenue of the Americas
New York, New York, 10010
Phone: 212-633-3787
Fax: 212-633-3764

khunter@pucc.princeton.edu

Hunter, Larry

National Library of Medicine
Building 38A, MS-54
Bethesda, MD 20894
Phone: 301-496-9300
Fax: 301-496-0673
hunter@nlm.nih.gov

Jagadish, Hosagrahar

AT&T Bell Labs
Murray Hill, NJ
jag@research.att.com

Jajko, Pam

Syntex USA Corporate Library

Kahn, Robert

Corporation of National Research Initiatives (CNRI)
1895 Preston White Drive, Suite 100
Reston, VA 22091
Phone: 703-620-8990
Fax: 703-620-0913

Kantor, Paul

Rutgers University
(SCILS) School of Communication, Information and Library Studies
4 Huntington St.
New Brunswick, NJ 08903
kantor@zodiac.rutgers.edu

Lasher, Rebecca

Stanford University
TR Effort with Garcia-Molina
Head Librarian
Math/Computer Science Library
Stanford, CA 94305-2125
Phone: 415-723-0864
RLasher@forsythe.stanford.edu

cn.mcs@forsythe.stanford.edu

Lannom, Larry

Meridian Corp. (for DARPA)
Phone: 703-552-6410
lannom@darpa.mil

Lavendel, Giuliana

Xerox/PARC
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Leggett, John

Texas A&M University
Department of Computer Science
College Station, TX 77843-3112
Phone: 409-845-0298
Fax: 409-847-8578
leggett@bush.tamu.edu

Lenzini, Rebecca T.

CARL Systems, Inc.
President
3801 East Florida Avenue
Denver, Colorado 80222
Phone: 303-758-3030
Fax: 303-758-0606
rlenzini@carl.org

Lesk, Michael

Bellcore
Phone: 201-829-4070
Fax: 201-829-5981
lesk@bellcore.com
lesk@parody.bellcore.com

Levy, David

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

DLevy.parc@xerox.com

Liberman, Mark

University of Pennsylvania
Computer Science and Linguistics
Phone: 215-898-0141
Fax: 215-573-2091
myl@unagi.cis.upenn.edu

Liston, David

Council on Library Resources
1785 Massachusetts Ave., N.W.
Washington, D.C. 20036-2117

Lucier, Richard E.

University of California at San Francisco
UCSF Health Sciences
University Librarian and
Academic Vice Chancellor for Information Management
The Library
530 Parnassus Avenue
San Francisco, California 94143-0840
Phone: 415-476-8293
Fax: 415-476-4653
REL@LIB.UCSF.EDU

Lyman, Peter

University of Southern California
University Librarian and Dean of the University Libraries
3550 Trousdale Parkway, Room 100
Los Angeles, California 90089-0182
Phone: 213-740-7317
Fax: 213-749-1221
lyman@vm.usc.edu

Lynch, Clifford A.

University of California at Berkeley
Office of the President
Director, Library Automation
300 Lakeside Drive, 8th Floor
Oakland, California 94612-3350

Phone: 510-987-0522
Fax: 510-839-3573
calur@uccmvsa.ucop.edu
lynch@postgres.berkeley.edu

Lynn, M. Stuart

Cornell University
Vice President, Information Technologies
308 Day Hall
Ithaca, New York 14853-2801
Phone: 607-255-7445
Fax: 607-254-5333
msl1@cornell.edu

Malamud, Carl

Internet Talk Radio
Founder
302 W. W. Glendale Avenue
Alexandria, Virginia 22301
Fax: 703-548-1280
carl@malamud.com

Masinter, Larry

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304

Mayhew, Jean

United Technologies Research Center
Responsible for library network all over but seven countries in the world

McCray, Alexa T.

National Library of Medicine
8600 Rockville Pike
Besthesda, MD 20894
Phone: 301-496-9300
Fax: 301-496-0673
mccray@nlm.nih.gov

McGill, Michael J.

University of Michigan
Director, Network Systems
1075 Beal Avenue
3208 Computing Center
Ann Arbor, Michigan 48109-2112
Phone: 313-764-4262
Fax: 313-763-4050
Michael.McGill@um.cc.umich.edu

Menell, Peter

Stanford University
School of Law (1992-93)
Stanford, CA 94305
Phone: 415-723-0886
Fax: 415-723-0253

Michaiko, James P.

The Research Libraries Group, Inc
President
1200 Villa Street
Mountain View, California 94041-1100
Phone: 415-691-2243
Fax: 415-964-0943
bl.jpm@rlg.stanford.edu

Moline, Judi

National Institute of Standards Technology (NIST)
Computer Scientist
Building 225, Room B266
Gaithersburg, Maryland 20899
Phone: 301-975-3351
Fax: 301-590-0932
moline@mml.ncsl.nist.gov

Nelson, Mike

Executive Office of the Vice President
Washington, D.C. 20500
Phone: 202-395-6175
Hearings on Electronic Libraries

"Son of Gore" Legislation
mnelson@nsf.gof

Noreault, Terry

OCLC, Inc.
6565 Frantz Road
Dublin, Ohio 43017-0702
Phone: 614-764-4392
terry@rsch.oclc.org

Nunberg, Geoffrey

Xerox/PARC
Stanford University

Okerson, Ann Lillian

Association of Research Libraries
Director, Office of Scientific and Academic Publishing
21 Dupont Circle
Washington, DC 20036
Phone: 202-296-2296
Fax: 202-872-0084
ann@cni.org

Olson, Don

Eastman Kodak
Director of New Market Development
343 State Street
Rochester, New York 14650-0811
Phone: 716-781-1520
Fax: 716-724-0826
olson@pixel.kodak.com

Palmer, Roger

Getty Art History Information Program
Technical Manager
401 Wilshire Boulevard
Suite 1100
Santa Monica, California 90401-1455
Phone: 310-395-1025
Fax: 310-451-5570

Penniman, David W.
Council on Library Resources
President
1785 Massachusetts Ave, N.W.
Washington, D.C. 20036-2117
dpenn@cni.org

Perry, Susan Lane
Stanford University
Director, Departmental Systems Group
Stanford Data Center
Stanford, California 94305-4136
Phone: 415-725-8425
Fax: 415-723-2011
gd.slp@forsythe.stanford.edu

Peters, Paul Evan
Coalition for Networked Information (CNI)
Executive Director
21 Dupont Circle
Washington, D.C. 20036
Phone: 202-296-5098
Fax: 202-872-0884
paul@cni.org

Polly, Jean
Public Library Prospective
polly@lpl.org

Prime, Eugenie
Hewlett Packard Labs Research Library
Previous Publisher of printed indexes became a database on Dialog, BRS, and Datastar

Raghavan, V. V.
Center for Advanced Computer Studies
University of Southwestern Louisiana
P.O. Box 44330
Lafayette, LA 70504-4330
raghavan@cacs.usl.edu

Reich, Vicky
Libraries at Stanford University
Fax: 415-723-4775
Stanford University
Stanford, CA 94305-2125
vicky.reich@forsythe.stanford.edu

Rodgers, David L.
American Mathematical Society
Electronic Products and Services Director
Electronic Journal Project with A. Okerson
416 Fourth Street
Ann Arbor, MI 48103
Phone: 313-995-4123
Fax: 313-996-2916
dlr@math.ams.org

Rose, Dan
Apple
rose@apple.com

Rosenberg. Lawrence C.
National Science Foundation
Deputy Divison Director
Information, Robotics & Intelligent Systems Division
4201 Wilson Boulevard, Room 1115
Arlington, VA 22230
Phone: 703-306-1927x1174
lrosenbe@nsf.gov

Ryland, Jane Norman
CAUSE
President
4840 Pearl East Circle, Suite 302E
Boulder, Colorado 80301-2454
Phone: 303-449-4430
Fax: 303-440-0461
jryland@cause.colorado.edu

Sabelli, Nora H.
National Science Foundation

Program Director, Applications of Advanced Technologies
4201 Wilson Boulevard, Room 885
Arlington, VA 22230
Phone: 703-306-1651x5888
nsabelli@nsf.gov

Saltzer, Jerome H.
MIT
Department of Electrical Engineering and Computer Science
Professor
NE43-513
Cambridge, MA 02139
Phone: 617-253-6016
saltzer@mit.edu

Sato, Andrea
Columbia University School of Law
Janus Project Director
435 West 116th Street, Room 7W1
New York, New York, 10027-7297
Phone: 212-854-2635
Fax: 212-854-3295

Schnase, John L.
School of Medicine Library and
Biomedical Communications Center
Washington University School of Medicine
Director, Advanced Technology Group
660 South Euclid Avenue (Box 8132)
St. Louis, Missouri 63110
Phone: 314-362-3117
Fax: 314-362-0190
schnase@medicine.wustl.edu

Seiden, Peggy
Skidmore College
Head of Public Services
North Broadway
Saratoga Springs, New York 12866-1632
Phone: 518-584-7400 Ext. 2126
Fax: 518-584-3023

pseiden@skidmore.edu

Shaw, Ward

CARL Systems, Inc.
Chairman and Chief Executive Officer
3801 East Florida Avenue
Building D, Suite 300
Denver, Colorado 80210
Phone: 303-758-3030
Fax: 303-758-0606
wshaw@carl.org

Simpson, Donald

Center for Research Libraries
President
6050 South Kenwood Avenue
Chicago, Illinois 60637-2804
Phone: 312-955-4545
Fax: 312-955-4339
simpson@uhuru.uchicago.edu

Smarr, Larry

NCSA
University of Illinois Library School
Professor, Director of Supercomputing Applications Division
155 Comp. Appl. Building, MC 476
Champaign, IL 61820
Phone: 217-244-0079
smarr#ncsa.uiuc.edu

Soloway, Elliot

University of Michigan
Department of Electrical Engineering and Computer Science
Associate Professor
Phone: 313-936-1562
soloway@caen.engin.umich.edu

St. George, Art

University of New Mexico
CIRT
Executive Network Services Officer

Albuquerque, New Mexico 87131
Phone: 505-277-8423
Fax: 505-277-8101
stgeorge@bootes.unm.edu

Thaker, Al

American Math Society
Electronic Journal Project with A. Okerson

Tompkins, Philip

Estrella Mountain Community College Center
Director of Library Information Services
3000 Dysart Road
Litchfield Park, Arizona 85340
Phone: 602-935-8077
Fax: 602-035-8008
tompkins@names.maricopa.edu

Turner, Judith

Chronicle of Higher Education
Director, Chronicle Information Services
1255 23rd Street, NW
Washington, D.C. 20037
Phone: 202-466-1063
Fax: 202-296-2691
JUDITH@PAGE1.com

Urbanowski, Frank

MIT Press
Director
55 Hayward Street
Cambridge, MA 02142
Phone: 800-346-0343
Fax: 617-258-6779

Vassallo, Paul

National Institute of Standards and Technology
Director, Office of Information Services
Admin. E106
NIST

Gaithersburg, Maryland 20899
Phone: 301-975-2786
Fax: 301-869-8071
pvassa11@micf.nist.gov

von Wahlde, Barbara
State Univeristy of New York at Buffalo
Associate Vice President for University Libaries
Buffalo, New York 14260
Phone: 716-636-2967
Fax: 716-636-3844
unlbvw@ubvms.cc.buffalo.edu

Watts, Carol B.
National Environmental Satellite Data and Information Service
National Oceanic & Atmospheric Administration
Chief, Library and Information Services Division
6009 Executive Boulevard
Rockville, Maryland 20852
Phone: 301-443-8287
Fax: 301-443-0237
watts@nodc2.nodc.noaa.gov

Webster, Duane E.
Association of Research Libraries (ARL)
Executive Director
21 Dupont Circle
Washington, D.C. 20036
Phone: 202-296-2296
Fax: 202-872-0884
duane@cni.org

West, Richard Paul
University of California, Office of the President
Associate Vice President, Information Systems and Administrative Services
300 Lakeside Drive, 8th Floor
Oakland, California 94612-3350
Phone: 510-987-0405
Fax: 510-839-3573

west@rio.ucop.edu

Wiederhold, Gio

DARPA/SISTO
ACM Publication Board
Phone: 703-696-2249
Fax: 703-696-2202
gio@darpa.mil

Williams, James F. II

University of Colorado at Boulder
Dean of Libraries
Norlin Library N210C
Campus Box 184
Boulder, Colorado 80309-0184
Phone: 303-492-7511
Fax: 303-492-8775
williams_j@cubldr.colorado.edu

Williams, Cynthia

MCC (Microelectronics and Computer Technology Corporation)
Manager Communications
3500 West Balcones Center Drive
Austin, TX 78759-6509
Phone: 512-338-3512
Fax: 512-338-3892
williams@mcc.com

Winograd, Terry A.

Stanford University
Professor, Department of Computer Science
Bldg 460 Room 238
Stanford, CA 94305-2140
Phone: 415-723-2780
Fax: 415-725-7411
winograd@cs.stanford.edu
Interval Research (1992-93)
Phone: 415-354-0854
Fax: 415-354-0872
winograd@interval.com

Woelk, Darrell

NCR/MCC
3500 W. Balcones Center Drive
Austin, TX 78759
Phone: 512-338-3468
woelk@mcc.com

Zemankova, Maria

On Leave 1993-4 Mitre Corporation
National Science Foundation
Program Director Database and Expert Systems
4201 Wilson Boulevard
Arlington, VA 22230
mzemanko@note2.nsf.gov

Groups and Organizations

Testimonials related to the National Electronic Library

C. Borgman

Support for the initiative will be needed from all corners of the information world. This list of organizations is anecdotal, from my personal knowledge of the field. It would be useful for the developers of the initiative to do a thorough search of the directories of professional organizations and foundations to find other groups as well. Addresses and contact names for these organizations and others can be located in standard bibliographic sources.

- American Library Association (50K plus members, active lobbying office in Washington, DC, major player, very important source of support for initiative.)
- Library and Information Technology Association (a subgroup of ALA, large, represents those involved in library automation and networks)

- National Commission on Libraries and Information Services (federally funded, organizes White House Conferences on Libraries and Information Services; held hearings on NREN and libraries in July 1992).
- Computer Professionals for Social Responsibility (interested in workplace issues, responsible use of technology, privacy issues. Testified before NCLIS on NREN issues. Washington office headed by Marc Rotenberg.)
- Association for Computing Machinery (especially special interest group on Information Retrieval)
- Coalition for Networked Information (Washington, DC. Working on standards for information delivery on Internet, cooperative issues. New but very active group.)
- Association for Research Libraries (Washington, DC. Umbrella for all major university and other research libraries in the U.S.)
- American Society for Information Science (Silver Spring, MD. Scholarly and professional society for broad range of theoretical and technical issues in information and delivery. Mix of librarians, information retrieval researchers, government and private industry.)
- Council on Library Resources (Washington, DC. Primary private funding agency for library-related research in the U.S.).
- Information Industry Association (Represents small and large electronic publishers)
- Corporation for National Research Initiatives (Vinton Cerf, Robert Kahn, et al. Active in developing network infrastructure.)
- Society for Scholarly Publishing (Represents university and private scholarly presses)
- Society for the Social Studies of Science (Scholarly society, studies scholarly communication infrastructure)
- Special Libraries Association
- Medical Library Association
- Art Libraries of North America (ARLIS)

(other library associations representing special groups)

- Museum Computing Network (building databases of museum information and networking; often modeled on work of libraries)
- Professional societies of major scientific disciplines who would be served (American Chemical Society, Physics, Biology, Geology, etc.)

- American Educational Research Association (interest in classroom technologies)
- Professional education societies (teachers, university educators, etc.)

Internet Society

The Internet Society is a new international, professional membership organization to promote the use of the Internet for research and scholarly communication and collaboration.

The Society provides a forum for government, industry, educators and users to debate and recommend technical standards and procedures for the global Internet and private internets.

The Society seeks to advance open scholarship in all countries.

THE INTERNET SOCIETY ....
- began operations in January 1992
- is governed by an elected Board of Trustees
- seeks to encourage international scientific collaboration through networking

Joining now will ....
- entitle you to the quarterly member newsletter and on-line information recording important developments in the technical and operational evolution of the Internet
- help shape the international agenda for the Society
- support Internet Society objectives

Abstract

The purpose of this document is to provide a brief description of the Internet Society and its goals and objectives. It functions as a professional society to facilitate, support and promote the evolution and growth of the Internet as a global research communications infrastructure. The suggestions and recommendations of all parties interested in the Internet are solicited to assist in making the Internet Society robust, productive and structured to meet the needs of its members.

Internet Society

The Internet is a collection of cooperating, interconnected, multiprotocol networks which supports international collaboration among thousands of organizations. Internet Society seeks to foster the voluntary interconnection of computer networks into a global research, development, and information infrastructure. The Internet Society does not operate the Internet. Internet operation continues to be a collaborative activity which the Society seeks to facilitate. The Society provides assistance and support to groups and organizations involved in the use, operation and evolution of the Internet. It provides support for forums in which technical and operational questions can be discussed and provide mechanisms through which interested parties can be informed and educated about the Internet, its function, use, operation and the interests of its constituents.

Membership

Internet Society is a professional membership organization with voting individual members and non-voting institutional members. There are several classes of institutional members. The society publishes a newsletter on a regular basis and holds an annual meeting to which all members and other interested parties are invited. The topics of the annual meeting vary, but focus on current research in networking, Internet functionality and growth, and other interests of the Society constituency.

Membership dues vary according to class of membership. The amounts of these dues and the basis on which they are set are determined by the Board of Trustees of the Society and may be revised from time to time as provided in the By-Laws.

Charter

The Society is a non-profit organization and operated for academic, educational, charitable and scientific purposes among which are:

A. To facilitate and support the technical evolution of the Internet as a research and education infrastructure and to stimulate involvement of the academic, scientific and engineering communities, among others in the evolution of the Internet.

B. To educate the academic and scientific communities and the public concerning the technology, use and application of the Internet.

C. To promote scientific and educational applications of Internet technology for the benefit of educational institutions at all grade levels, industry and the public at large.

D. To provide a forum for exploration of new Internet applications and to foster collaboration among organizations in their operation and use of the Internet.

Activities of the Society

1. Support for Internet Technical Evolution

The Internet Architecture Board (IAB) has been concerned with the development and evolution of architectures supporting the use of multiple protocols in a networked environment. The Internet Society has incorporated the IAB and its functions into the operation of the Internet Society. The Internet Society works with other interested organizations to support and assist efforts to evolve the multiprotocol Internet. The Internet Society looks to the Internet Engineering and Research Task Forces to stimulate networking research and facilitate the evolution of the TCP/IP protocol suite and the integration of new protocol suites into the Internet architecture. The Internet Society works actively with parties and organizations interested in fostering improvement in the utility of the Internet for its constituent users.

2. Meetings and Conferences

Internet Society convenes an annual meeting ("INET") and organizes and facilitates workshops and symposia, jointly with other organizations where appropriate, on specific topics of interest to the Society membership. The annual meeting addresses issues of global and regional importance to the evolution and growth of the Internet.

3. Information and Infrastructure Services

The Internet Society publishes the quarterly Internet Society News providing members with information about the international activities of Internet constituents. In addition, the Society also provides assistance to and support for organizations responsible for maintaining the databases crucial to Internet function (e.g., the Domain Name System, X.500 Directory Services, etc.) and organizations concerned with the security of the Internet (e.g., the Software Engineering Institute Computer Emergency Response Team (CERT)). The Society assists in the development of educational, advisory and informative materials of use to Society members. Where appropriate, the Society organizes or supports activities which aid in the coordination among the organizations operating components of the Internet.

The Society refers members to appropriate parties involved in operating the various parts of the Internet where they may be helpful with specific questions. Where possible, the Society seeks to provide access to its information on-line, but also offers hard copy and, perhaps eventually, CD-ROM-based information resources.

Plans

The initial organizers of the Internet Society include the Corporation for National Research Initiatives (CNRI), EDUCOM, Reseaux Associees pour la Recherche Europeenne (RARE) and the Internet Architecture Board. Computer networking has become a critical infrastructure for the research and development community and has the potential to become the basis for world-wide collaboration and cooperation in every field of human endeavor. The Internet Society seeks to solidify, enhance and encourage further international collaborative networking. Individuals joining the Society during its formation have received special recognition as Society pioneers and have been instrumental in shaping the early agenda of Society activities. Institutional members are represented in the Advisory Council of the Internet Society where their advice and counsel is sought to refine Internet Society objectives and activities.

APPENDIX

A Brief History of the Internet and Related Networks

Introduction

In 1973, the U.S. Defense Advanced Research Projects Agency (DARPA - now ARPA) initiated a research program to investigate techniques and technologies for interlinking packet networks of various kinds. The objective was to develop communication protocols which would allow networked computers to communicate transparently across multiple, linked packet networks. This was called the Internetting project and the system of networks which emerged from the research was known as the "Internet." The system of protocols which was developed over the course of this research effort became known as the TCP/IP Protocol Suite, after the two initial protocols developed: Transmission Control Protocol (TCP) and Internet Protocol (IP).

In 1986, the U.S. National Science Foundation (NSF) initiated the development of the NSFNET which, today, provides a major backbone communication service for the Internet. With its 45 megabit per second facilities, the NSFNET carries on the order of 12 billion packets per month between the networks it links. The National Aeronautics and Space Administration (NASA) and the U.S. Department of Energy contributed additional backbone facilities in the form of the NSINET and ESNET respectively. In Europe, major international backbones such as NORDUNET, EBONE, EuropaNet and others provide connectivity to over three hundred thousand computers on a large number of networks. Commercial network providers in the U.S. and Europe are beginning to offer Internet backbone and access support on a competitive basis to any interested parties.

"Regional" support for the Internet is provided by various consortium networks and "local" support is provided through each of the research and educational institutions. Within the United States, much of this support has come from the federal and state governments, but a considerable contribution has been made by industry. In Europe and elsewhere, support arises from cooperative international efforts and through national research organizations. During the course of its evolution, particularly after 1989, the Internet system began to integrate support for other protocol suites into its basic networking fabric. The present emphasis in the system is on multiprotocol interworking, and in particular, with the integration of the Open Systems Interconnection (OSI) protocols into the architecture.

Both public domain and commercial implementations of the roughly 100 protocols of TCP/IP protocol suite became available in the 1980's. During the early 1990's, OSI protocol implementations also became available and, by the end of 1992, the Internet had grown to include some 11,000 networks in over seventy countries, serving over 1,700,000 host computers used by over 5,000,000 people.

A great deal of support for the Internet community has come from the U.S. Federal Government, since the Internet was originally part of a federally-funded research program and, subsequently, has become a major part of the U.S. research infrastructure. During the late 1980's, however, the population of Internet users and network constituents expanded internationally and began to include commercial facilities. Indeed, the bulk of the system today is made up of private networking facilities in educational and research institutions, businesses and in government organizations across the globe.

The Coordinating Committee for Intercontinental Networking (CCIRN), which was organized by the U.S. Federal Networking Council (FNC) and the European

Reseaux Associees pour la Recherche Europeenne (RARE), plays an important role in the coordination of plans for government-sponsored research networking. CCIRN efforts have been a stimulus for the support of international cooperation in the Internet environment. Reseau IP Europeenne (RIPE) has mounted a very successful grass-roots effort in Europe to extend Internet in Europe.

Internet Technical Evolution

Over its fifteen year history, the Internet has functioned as a collaboration among cooperating parties. Certain key functions have been critical for its operation, not the least of which is the specification of the protocols by which the components of the system operate. These were originally developed in the DARPA research program mentioned above, but in the last five or six years, this work has been undertaken on a wider basis with support from Government agencies in many countries, industry and the academic community. The Internet Activities Board (recently re-named the Internet Architecture Board) was created in 1983 to guide the evolution of the TCP/IP Protocol Suite and to provide research advice to the Internet community.

During the course of its existence, the IAB has reorganized several times. It now has two primary components: the Internet Engineering Task Force and the Internet Research Task Force. The former has primary responsibility for further evolution of the TCP/IP protocol suite, its standardization, and the integration of other protocols into Internet operation (e.g., the Open Systems Interconnection protocols). The Internet Research Task Force continues to organize and explore advanced concepts in networking under the guidance of the Internet Architecture Board and with support from various research sponsoring agencies.

A secretariat has been created to manage the day-to-day function of the Internet Architecture Board and Internet Engineering Task Force. IETF meets three times a year in plenary and its approximately 80 working groups convene at intermediate times by electronic mail, teleconferencing and at face-to-face meetings. The IAB meets quarterly face-to-face or by videoconference and at intervening times by telephone, electronic mail and computer-mediated conferences.

Two other functions are critical to Internet function: Publication of documents describing the Internet and the assignment and recording of various identifiers needed for protocol operation. Throughout the development of the Internet, its protocols and other aspects of its operation have been documented first in a series of documents called Internet Experiment Notes and, later, in a series of documents called Requests for Comment (RFCs). The latter were used initially to document the protocols of the first packet switching network developed by DARPA, the

ARPANET, beginning in 1969, and have become the principal archive of information about the Internet. At present, the publication function is provided by an RFC editor.

The recording of identifiers is provided by the Internet Assigned Numbers Authority (IANA) who has delegated one part of this responsibility to an Internet Registry which acts as a central repository for Internet information and which provides central allocation of network and autonomous system identifiers, in some cases to subsidiary registries located in various countries. The Internet Registry (IR) also provides central maintenance of the Domain Name System (DNS) root database which points to subsidiary distributed DNS servers replicated throughout the Internet. The DNS distributed database is used, inter alia, to associate host and network names with their Internet addresses and is critical to the operation of the higher level TCP/IP protocols including electronic mail.

There are a number of Network Information Centers (NICs) located throughout the Internet to serve its users with documentation, guidance, advice and assistance. As the Internet continues to grow internationally, the need for high quality NIC functions increases. Although the initial community of users of the Internet were drawn from the ranks of computer science and engineering, its users now comprise a wide range of disciplines in the sciences, arts, letters, business, military and government administration.

Related Networks

In 1980-81, two other networking projects, BITNET and CSNET, were initiated. BITNET adopted the IBM RSCS protocol suite and featured direct leased line connections between participating sites. Most of the original BITNET connections linked IBM mainframes in university data centers. This rapidly changed as protocol implementations became available for other machines. From the beginning, BITNET has been multi-disciplinary in nature with users in all academic areas. It has also provided a number of unique services to its users (e.g., LISTSERV). Today, BITNET and its parallel networks in other parts of the world (e.g., EARN in Europe) have several thousand participating sites. In recent years, BITNET has established a backbone which uses the TCP/IP protocols with RSCS-based applications running above TCP.

CSNET was initially funded by the National Science Foundation (NSF) to provide networking for university, industry and government computer science research groups. CSNET used the Phonenet MMDF protocol for telephone-based electronic mail relaying and, in addition, pioneered the first use of TCP/IP over X.25 using commercial public data networks. The CSNET name server provided an early example of a white pages directory service and this software is still in use at numerous sites. At its peak, CSNET had approximately 200 participating sites and international connections to approximately fifteen countries.

In 1987, BITNET and CSNET merged to form the Corporation for Research and Educational Networking (CREN). In the Fall of 1991, CSNET service was discontinued having fulfilled its important early role in the provision of academic networking service. A key feature of CREN is that its operational costs are fully met through dues paid by its member organizations.

Introducing the Current Trustees and Officers of the Internet Society

After the chartering of the Internet Society late in December 1991, the three initial trustees, Kenneth King, Juergen Harms and Robert Kahn unanimously elected an interim Board of Trustees, shown below. The interim board is responsible for initial operation of the Internet Society including approval of budgets, appointment of officers and election of an expanded Board whose members will serve for periods ranging from one to three years. At annual intervals, nominations for Board positions will be opened and elections held to refresh approximately one-third of the Board of Trustees each year. The first such election by the individual Internet Society members takes place in Spring 1993, with the elected officials taking office in July 1993.

Internet Society Board of Trustees

Charles Brownstein
<cbrownst@note.nsf.gov>

Vinton Cerf, President
<vcerf@cnri.reston.va.us>

Lyman Chapin
<lyman@bbn.com>

Ira Fuchs
<fuchs@pucc.princeton.edu>

Robert Kahn
<rkahn@cnri.reston.va.us>

Tomaz Kalin
<kalin@ijs.ac.mail.yu>

Kenneth King
<kmk@educom.edu>

Lawrence Landweber, Vice President
<lhl@cs.wise.edu>

Frode Greisen, Treasurer
<frode.greisen@uni-c.dk>

Kees Neggers
<neggers@surfnet.nl>

Geoff Huston
<g.huston@aarnet.edu.au>

Michael Roberts, Executive Director
<roberts@educom.edu>

Anthony Rutkowski, Vice President
<amr@sprint.com>

Introducing the Charter, Founding, and Organizational Members

CHARTER MEMBERS

CORPORATION FOR NATIONAL RESEARCH INITIATIVES
EDUCOM
RESEAU ASSOCIEES POUR LA RECHERCHE EUROPEENNE

FOUNDING MEMBERS

ADVANCED NETWORK & SERVICES
APPLE COMPUTER CORPORATION
AT&T
AUSTRALIAN ACADEMIC AND RESEARCH NETWORK
BELL COMMUNICATIONS RESEARCH
BOLT BERANEK AND NEWMAN
CISCO SYSTEMS
COALITION FOR NETWORKED INFORMATION
CORPORATION FOR OPEN SYSTEMS INTERNATIONAL
CORPORATION FOR RESEARCH & EDUCATIONAL NETWORKING
DEFENSE INFORMATION SYSTEMS AGENCY
DIGITAL EQUIPMENT CORPORATION
EUROPEAN ACADEMIC RESEARCH NETWORK
EUROPEAN LABORATORY FOR PARTICLE PHYSICS
FREEPORT-MCMORAN
HUGHES AIRCRAFT COMPANY
INTERNATIONAL BUSINESS MACHINES CORPORATION
INTEROP COMPANY
ISRAELI INTER-UNIVERSITY COMPUTATION CENTER
LAWRENCE LIVERMORE NATIONAL LABORATORY

MCI COMMUNICATIONS CORPORATION
MICROSOFT CORPORATION
NATIONAL LIBRARY OF MEDICINE
NORDUNET
NOVELL, INC.
NYSERNET, INC.
PROTEON, INC.
SIEMENS AG
SOFT-SWITCH, INC.
SOFTWARE ENGINEERING INSTITUTE
SPRINT
3COM CORPORATION
UNIVERSITY OF NOTRE DAME
UNIVERSITY OF WASHINGTON
U S WEST COMMUNICATIONS, INC.
UUNET TECHNOLOGIES
WELLFLEET COMMUNICATIONS INC.

ORGANIZATIONAL MEMBERS

ELECTRONIC FRONTIER FOUNDATION
NYNEX SCIENCE & TECHNOLOGY, INC.
TENON INTERSYSTEMS
VEDA DATA SYSTEMS, INC.

Terms of Membership

Charter and Founding Organizations provide the Internet Society with vital financial support. This substantial and early support has made the founding of the Internet Society possible. For-profit Founding members commit to a total of $20,000 during the 1992 and 1993 period and $10,000 per year thereafter. Non-profit, Founding government and educational institutions commit to half that amount. Regular for-profit and non-profit organizational members commit to $10,000 and $5,000 per year respectively. There is also provision for start-ups to become regular members during their first three years at a cost of $1,000 per year. Organizations interested in participating in this program should contact Vinton Cerf at the Internet Society secretariat.

Internet Society Advisory Council

The Organizational Members of the Internet Society provide a representative and an alternate to an Advisory Council. The Society is grateful to these individuals and their organizations for the commitments they have made to its success. The names of the Advisory Council members are listed below:

Rick Adams UUNET
Ed Albrigo (alt.) CORPORATION FOR OPEN SYSTEMS
Guy Almes ANS
Stephen An WELLFLEET COMMUNICATIONS INC.
Michael Arnold FREEPORT-MCMORAN
Cliff Bamford MICROSOFT
William Biagi CORPORATION FOR OPEN SYSTEMS
Daniele Bovio (alt.) EARN
David Brandin INTEROP COMPANY
George Buchanan HUGHES AIRCRAFT COMPANY
Lee Caldwell NOVELL
Michael Carter APPLE
Steve Cisler (alt.) APPLE
Avi Cohen IUCC
Sam Coleman (alt.) LAWRENCE LIVERMORE NATIONAL LAB
James Conklin (alt.) CREN
Michael Conn MCI COMMUNICATIONS
Hans Deckers EARN
Rashmi Doshi (alt.) NYNEX SCIENCE & TECHNOLOGY, INC.
Mathew Dovens (alt.) MCI COMMUNICATIONS
Robin Erskine AARNET
David Farber ELECTRONIC FRONTIER FOUNDATION
Francois Fluckiger CERN
Ira Fuchs CREN
Dain Gary (alt.) SOFTWARE ENGINEERING INSTITUTE
Cary Giese (alt.) U S WEST
William Grant U S WEST
Terence Gray UNIVERSITY OF WASHINGTON
Erik Grimmelmann AT&T
Phill Gross ANS
Roger Gulbranson (alt.) UNIVERSITY OF NOTRE DAME
Anita Holmgren TENON INTERSYSTEMS
Steve Holmgren (alt.) TENON INTERSYSTEMS
E.W. Bud Huber (alt.) HUGHES AIRCRAFT COMPANY

Sascha Ignjatovic VEDA DATA SYSTEMS, INC.
Ole Jacobsen (alt.) INTEROP COMPANY
Ron Johnson (alt.) UNIVERSITY OF WASHINGTON
Walter Johnston NYNEX SCIENCE & TECHNOLOGY, INC.
Mitch Kapor (alt.) ELECTRONIC FRONTIER FOUNDATION
Anthony Lauck DIGITAL
Donald Lindberg NATIONAL LIBRARY OF MEDICINE
James Luckett (alt.) NYSERNet
Richard Mandelbaum NYSERNet
Olivier Martin (alt.) CERN
Daniel Masys NATIONAL LIBRARY OF MEDICINE
Stuart Mathison SPRINT INTERNATIONAL
Jeff Mayersohn BOLT BERANEK AND NEWMAN
David McDonald (alt.) FREEPORT-MCMORAN
John Mullarvey (alt.) AARNET
Catherine Muther CISCO SYSTEMS
Dan Nessett LAWRENCE LIVERMORE NATIONAL LAB
Camillo J. Pasquariello DEFENSE INFORM. SYSTEMS AGENCY
Andrew Partan (alt.) UUNET
Janet Perry (alt.) NOVELL
Paul Evan Peters CNI
Rich Pethia SOFTWARE ENGINEERING INSTITUTE
John Pickens 3COM
Werner Sammer SIEMENS AG
Thomas Schwarcz (alt.) SIEMENS AG
Paul Severino (alt.) WELLFLEET COMMUNICATIONS INC.
Robert Shahan (alt.) IBM
David Sincoskie BELLCORE
Donald Spicer UNIVERSITY OF NOTRE DAME
Leonard Swatski (alt.) DEFENSE INFORM. SYSTEMS AGENCY
Dave Thompson (alt.) MICROSOFT
Michael Thurk (alt.) DIGITAL
Paul Toldalagi PROTEON
Nicholas R. Trio IBM
Paul Tsuchiya (alt.) BELLCORE
Peter Villemoes NORDUNET
Richard West (alt.) CNI
Michael D. Zisman SOFT-SWITCH, INC.

Introducing the Internet Society Secretariat

During its initial period of operation, the Internet Society secretariat occupies space provided by the Corporation for National Research Initiatives and by EDUCOM:

Internet Society
1895 Preston White Drive, Suite 100
Reston, VA 22091 USA
+1 703 648 9888
+1 703 620 0913 FAX
isoc@isoc.org

Britt Jackman - Secretarial Support
bjackman@cnri.reston.va.us

Cynthia Matthews - Individual Membership Services
isoc@isoc.org (preferred)
cmatthew@cnri.reston.va.us

John Stewart - Technical Support
jstewart@cnri.reston.va.us

Theresa Weigler - Organizational Membership Services
tweigler@cnri.reston.va.us

In addition to these, two EDUCOM staff members are also assisting in the operation of the Society:

EDUCOM
1112 16th Street NW, Suite 600
Washington, DC 20036
+1 202 872 4200
+1 202 872 4318 FAX

Michael Roberts - General planning and operation
roberts@educom.edu

Elizabeth Barnhart- Conference/Publication Support
barnhart@educom.edu

Library-Oriented Lists and Electronic Serials

By Charles W. Bailey, Jr.

1.0 Computer Conferences

Computer conferences are becoming an increasingly important form of communication for librarians. There are a growing number of computer conferences of interest to librarians on BITNET and Internet.

Computer conferences are commonly called "lists." The software used to support these lists is typically referred to as the "list server."

1.1 Lists That Use Eric Thomas's List Server

Many BITNET lists utilize a list server that was developed by Eric Thomas.

Please note that some of these BITNET lists also have Internet addresses, which are not shown here.

(Entries that appear in this document for the first time or have been changed are marked with an asterisk.)

ADVANC-L@IDBSU Geac Advanced Library System
AFAS-L@KENTVM African American Studies and Librarianship
ALF-L@YORKVM1 Academic Librarian's Forum
ARCHIVES@INDYCMS Archives and Archivists List
ARIE-L@IDBSU RLG Ariel Document Transmission System
ARLIS-L@UKCC Art Libraries Association of North America
ASIS-L@UVMVM American Society for Information Science
ATLAS-L@TCUBVM Data Research ATLAS Users
AUTOCAT@UVMVM Library Cataloging and Authorities Discussion Group
BI-L@BINGVMB Bibliographic Instruction
BIBSOFT@INDYCMS Discussion of Software for Citations and Bibliographies
BRS-L@USCVM BRS/Search Users
BUSLIB-L@IDBSU Business Librarians
CALL-L@UNBVM1 Canadian Academic Law Libraries List*

CARL-L@UHCCVM CARL Users
CDPLUS-L@UTORONTO CDPLUS Software User Group
CDROMLAN@IDBSU CD-ROM LANs
CDS-ISIS@HEARN UNESCO's CDS/ISIS Text Retrieval Software
CHMINF-L@IUBVM Chemical Information Sources
CIRCPLUS@IDBSU Circulation and Access Services
COLLDV-L@USCVM Library Collection Development List
CONSALD@UTXVM Committee on South Asian Libraries and Documentation
COSNDISC@BITNIC Consortium for School Networking
CWIS-L@WUVMD Campus-Wide Information Systems
ELDNET-L@UIUCVMD ASEE Engineering Libraries Division
ELEASAI@ARIZVM1 Open Library/Information Science Research Forum
ELLASBIB@GREARN Library Automation in Greece
ENDNOTE@UCSBVM EndNote/EndLink Users Forum
EXLIBRIS@RUTVM1 Rare Books and Special Collections Forum
FEMINIST@MITVMA ALA Social Responsibility Round Table Feminist Task Force
FISC-L@NDSUVM1 Fee-Based Information Service Centers in Academic Libraries
GEONET-L@IUBVM Geoscience Librarians and Information Specialists
GOVDOC-L@PSUVM Government Documents
ILL-L@UVMVM Interlibrary Loan
IMUG-L@OHSTVMA Innopac Music Users Group
INDEX-L@BINGVMB Indexer's Discussion Group
INFO+REF@INDYCMS Information + Referral List
INNOPAC@MAINE Innovative Interfaces Users
INT-LAW@UMINN1 Foreign and International Law Librarians
JESSE@ARIZVM1 Open Library/Information Science Education Forum
LABMGR@UKCC Academic Microcomputer Lab Management
LIBADMIN@UMAB Library Administration and Management
LIBEVENT@USCVM Library Events in Southern California
LIBEX-L@MAINE Exhibits and Academic Libraries Discussion List*
LIBMASTR@UOTTAWA Library Master Bibliographic Database
LIBPER-L@KSUVM Library Personnel Issues
LIBPLN-L@QUCDN Library Planning
LIBRARY@INDYCMS Libraries and Librarians
LIBREF-L@KENTVM Discussion of Library Reference Issues
LIBSUP-L@UWAVM Library Support Staff
LM_NET@SUVM School Library Media & Network Communications
MAPS-L@UGA Maps and Air Photo Forum
MEDLIB-L@UBVM Medical and Health Sciences Libraries

MLA-L@IUBVM Music Library Association
MULTILIS@ALBNYVM1 multiLIS Users
NETTRAIN@UBVM Internet/BITNET Network Trainers Discussion List
NOTIS-L@TCSVM NOTIS Users
NOTISACQ@CUVMB NOTIS Acquisitions Discussion Group
NOTMUS-L@UBVM Notis Music Library List
NOTRBCAT@INDYCMS Rare Book and Special Collections Catalogers
NYSO-L@UBVM MLA New York State/Ontario Chapter Discussion List
OFFCAMP@WAYNEST1 Off-Campus Library Services List
PACS-L@UHUPVM1 Public-Access Computer Systems Forum
PACS-P@UHUPVM1 PACS-L Publications Only
PRO-CITE@IUBVM The Personal Bibliographic Software Discussion List
RLIN-L@RUTVM1 RLIN Users*
SAVEIT-L@USCVM SAVEIT Users*
SERIALST@UVMVM Serials Users Discussion Group
SLAJOB@IUBVM Special Libraries Association Employment Opportunities*
SPILIB-L@SUVM SPIRES Users
UNICRN-L@PSUORVM SIRSI/UNICORN Automated Library Systems
USMARC-L@MAINE USMARC Advisory Group Forum
VETLIB-L@VTVM2 Veterinary Medicine Library Issues and Information*
VPIEJ-L@VTVM1 Publishing E-Journals: Publishing, Archiving, and Access
Z3950IW@NERVM Z39.50 Implementors Workshop

To subscribe to a list, send the following e-mail message to LISTSERV@NODE, where NODE is the part of the address after the "@" character:

SUBSCRIBE List First_Name Last_Name

(If you are not on BITNET, ask your computer center how to address a message to the desired BITNET node.)

For example, Jane Doe sends the following e-mail message to LISTSERV@UHUPVM1 to subscribe to PACS-L:

SUBSCRIBE PACS-L Jane Doe

You can obtain a directory of list server documentation by sending the following e-mail message to the list server: INFO ?

Once you know the name of the desired documentation file, send another INFO command to the list server to obtain the file. For example, to get a file that describes searching the message database of a list, send the following command: INFO DATABASE.

1.2 Lists That Use Anastasios Kotsikonas's List Server

Anastasios Kotsikonas's list server is being used to support a growing number of Internet lists.

To subscribe to a list, send the following e-mail message to LISTSERV@NODE, where NODE is the part of the address after the "@" character:

SUBSCRIBE List First_Name Last_Name

(If you are not on Internet, ask your computer center how to address a message to the desired Internet node.)

To obtain list server documentation, send the following e-mail message to the list server: HELP.

CNI-ANNOUNCE@CNI.ORG CNI's News Announcement Network
CNI-ARCHITECTURE@CNI.ORG CNI's Architectures and Standards Working Group
CNI-BIGIDEAS@CNI.ORG CNI's Big Ideas Forum
CNI-COPYRIGHT@CNI.ORG CNI's Copyright and Intellectual Property Forum
CNI-DIRECTORIES@CNI.ORG CNI's Directories and Resource Information Services Working Group
CNI-LEGISLATION@CNI.ORG CNI's Legislation, Codes, Policies, and Practices Working Group
CNI-MANAGEMENT@CNI.ORG CNI's Management and Professional and User Education Working Group
CNI-MODERNIZATION@CNI.ORG CNI's Modernization of Scholarly Publication Working Group
CNI-PUBINFO@CNI.ORG CNI's Access to Public Information Working Group
CNI-TEACHING@CNI.ORG CNI's Teaching and Learning Working Group
CNI-TRANSFORMATION@CNI.ORG CNI's Transformation of Scholarly Communication Working Group
PALS-L@KNUTH.MTSU.EDU PALS System
PUBLIB@NYSERNET.ORG Public Libraries and the Internet*
VISIONS@LIBRARY.SDSU.EDU Strategic Visions Steering Committee Electronic Discussion Forum on the Future of Librarianship

1.3 Other Lists

These lists use diverse software. Contact the person who sponsors the list you are interested in to get further information about how the list software works.

ACRLNY-L (Listings of Library Jobs and Events)
Send the following command to LISTSERV@NYUACF:
SUBSCRIBE ACRLNY-L First_Name Last_Name.

AGRIS-L (Agricultural Information)
Send the following message to LISTSERV@IRMFAO01:
SUBSCRIBE AGRIS-L First_Name Last_Name.

ALEPHINT (ALEPH Library System Users)
Send the following command to LISTSERV@TAUNIVM:
SUBSCRIBE ALEPHINT First_Name Last_Name.

CADUCEUS (History of Medicine Collections Forum)
Send a subscription request to Inci Bowman:
IBOWMAN@UTMBEACH.

Conservation DistList (Conservation of Archive, Library, and Museum Materials)
Send a subscription request to Walter Henry:
WHENRY@LINDY.STANFORD.EDU.

CORMOSEA (Committee on Research Materials on Southeast Asia)
Send a subscription request to Kent Mulliner:
MULLINER@OUVAXA.CATS.OHIOU.EDU.

DYNIX_L (DYNIX Users)
Send a subscription request to:
DYNIX_L-REQUEST@SBU.EDU.

EASTLIB (Committee on East Asian Libraries)
Send the following command to LISTSERV@MENTO.OIT.UNC.EDU:
SUBSCRIBE EASTLIB First_Name Last_Name.

IAMSLIC (International Association of Aquatic and Marine Science Libraries and Information Centers)
Send the following command to LISTSERV@UCSD.EDU:

SUBSCRIBE Your E-Mail Address IAMSLIC.

KATALIST (Discussion on Library Systems and Databases--In Hungarian)
Send the following command to LISTSERV@HUEARN:
SUBSCRIBE KATALIST First_Name Last_Name.

LAW-LIB (Law Librarians)
Send subscription requests to: LAW-REQ@UCDAVIS.EDU.
Contact Elizabeth St. Goar for technical questions:
ESTGOAR@UCDAVIS.EDU.

LIB_HYTELNET (HYTELNET Program)
Send subscription request to Peter Scott: SCOTT@SKLIB.USASK.CA.

LS2K (LS/2000 Users Group)
Send subscription request to: LS2K-REQUEST@CC.UTAH.EDU.

PAMnet (Physics, Astronomy, and Mathematics Librarians)
Send a subscription request to Joanne Goode:
JMGOODE@UKCC.UKY.EDU.

STUMPERS-L (Difficult Reference Questions)
Send the following message to ROSLIBREFRC@CRF.CUIS.EDU:
SUBSCRIBE STUMPERS-L Your E-Mail Address.

1.4 Diane Kovacs's Directory

Diane Kovacs and her colleagues have compiled a useful directory of lists in many subject areas.

To retrieve this directory, send the following e-mail message to LISTSERV@KENTVM or LISTSERV@KENTVM.KENT.EDU:

GET ACADLIST README F=MAIL
GET ACADLIST FILE1 F=MAIL
GET ACADLIST FILE2 F=MAIL
GET ACADLIST FILE3 F=MAIL
GET ACADLIST FILE4 F=MAIL
GET ACADLIST FILE5 F=MAIL
GET ACADLIST FILE6 F=MAIL

GET ACADLIST FILE7 F=MAIL

For further information about the directory, contact Diane Kovacs: DKOVACS@KENTVM.

2.0 Electronic Serials

There are a growing number of library-related electronic journals and newsletters available on BITNET and Internet.

ACQNET (The Acquisitions Librarian's Electronic Network)
Send a subscription request to Christian Boissonnas:
CRI@CORNELLC.

ALA Washington Office Newsline (ALAWON)
Send the following e-mail message to LISTSERV@UICVM:
SUBSCRIBE ALA-WO First_Name Last_Name.

ALCTS NETWORK NEWS (Association for Library Collections and Technical Services)
Send the following message to LISTSERV@UICVM:
SUBSCRIBE ALCTS First_Name Last_Name.
For a list of back issue files, send the message:
INDEX ALCTS.

The Arachnet Electronic Journal of Virtual Culture*
Send the following message to LISTSERV@UOTTAWA:
SUBSCRIBE ARACHNET First_Name Last_Name.

Citations for Serial Literature
Send the following message to LISTSERV@MITVMA:
SUBSCRIBE SERCITES First_Name Last_Name.
For a list of back issue files, send the message:
INDEX SERCITES.

Consortium Update (SPIRES)
Send a subscription request to: HQ.CON@STANFORD.

Current Cites
TELNET MELVYL.UCOP.EDU; Enter command: SHOW CURRENT CITES.
(Also distributed on PACS-L.)

Further information: David F. W. Robison, DROBISON@UCBLIBRA.

Hot Off the Tree (HOTT) (Excerpts and Abstracts of Articles about Information Technology)
TELNET MELVYL.UCOP.EDU; Enter command: SHOW HOTT.
Further information: Susan Jurist, SJURIST@UCSD.EDU.

Information Networking News
Sent to CDROMLAN subscribers. See above.

IRLIST Digest (Information Retrieval List Digest)
Send the following message to LISTSERV@UCCVMA:
SUBSCRIBE IR-L First_Name Last_Name.
For a list of back issue files, send the message:
INDEX IR-L.

Issues in Science and Technology Librarianship
Send a subscription request to: ACRLSTS@HAL.UNM.EDU.

Journal of Academic Media Librarianship
Send the following message to LISTSERV@UBVM:
SUBSCRIBE MCJRNL First_Name Last_Name.

LIBRES: Library and Information Science Research Electronic Journal*
Send the following message to LISTSERV@KENTVM:
SUBSCRIBE LIBRES First_Name Last_Name.

MeckJournal
Send a subscription request to MECKLER@TIGGER.JVNC.NET.

Network News
Send the following message to LISTSERV@NDSUVM1:
SUBSCRIBE NNEWS First_Name Last_Name.

Newsletter on Serials Pricing Issues*
Send the following message to LISTSERV@GIBBS.OIT.UNC.EDU:
SUBSCRIBE PRICES First_Name Last_Name.

Public-Access Computer Systems News
Sent to PACS-L and PACS-P subscribers. See above.

For a list of back issue files, send the following message to LISTSERV@UHUPVM1: INDEX PACS-L.

The Public-Access Computer Systems Review
Sent to PACS-L and PACS-P subscribers. See above.
For a list of article files, send the following message to LISTSERV@UHUPVM1: INDEX PACS-L.

Michael Strangelove has compiled a helpful directory of electronic serials. To retrieve this directory, send the following e-mail message to LISTSERV@UOTTAWA or LISTSERV@ACADVM1.UOTTAWA.CA:

GET EJOURNL1 DIRECTRY F=MAIL
GET EJOURNL2 DIRECTRY F=MAIL

For further information, contact Michael Strangelove: 441495@ACADVM1.UOTTAWA.CA.

Revised 1/11/93

Projects

The Clearinghouse for Networked Information Discovery and Retrieval

Several user-friendly client-server tools have been developed within the last couple of years for locating and retrieving information residing on computers reachable over the Internet. Among them, the Wide Area Information Server (WAIS), the Internet Gopher and archie have become popular, and the WorldWide Web (WWW) is poised for popularity once suitable client software is developed and distributed. WAIS, archie and Gopher present overviews of where information of interest is likely to reside and then assist the user in locating specific information objects. WWW permits a user to thread a path through the network by selecting tagged hypertext items.

While focused on the particulars of the evolution of wide area information systems like WAIS, The Clearinghouse for Networked Information Discovery and Retrieval (CNIDR) will work closely with the developers of other tools toward achieving compatibility and consistency, and, to the extent possible, convergence of the tools. Specific activities will provide a central focus and forum for networked information discovery and retrieval (NIDR) tools and will minimize the divergence of individual implementations by providing a repository for collecting, evaluating, and distributing protocol-compliant releases and enhanced versions.

CNIDR participates in standards and policy associations such as the Internet Engineering Task Force and the Coalition for Networked Information to further consensus among developers and explore appropriate uses of networked information. CNIDR also actively promotes the use of networked information discovery and retrieval tools at many national and international conferences and workshops to inform and educate end users.

CNIDR is sponsored by a three-year grant from the National Science Foundation (NSF) and by the Center for Communications at MCNC in the Research Triangle Park, NC.

For more information about the Clearinghouse contact:

George Brett, Director
Electronic mail: ghb@concert.net Telephone: 919-248-1886
Jane Smith, Assistant Director
Electronic mail: jds@concert.net Telephone: 919-248-9213

Jim Fullton, Technical Manager
Electronic mail: fullton@concert.net Telephone: 919-248-9247
Fax: 919-248-1405

As some of you may know, the U.S. National Science Foundation has funded the creation of the Clearinghouse for Networked Information Discovery and Retrieval.

One of the responsibilities of CNIDR is to promote the use and development of tools for discovering and using networked information. We have been specifically charged with the task of providing an evolutionary path for the public-domain version of WAIS, the Wide Area Information Server. Therefore, we have been paying close attention to the remarks and "wish lists" provided by the network community, and are well into a project to try to fulfill those wishes.

We are in the early beta test stages of a new generation of WAIS-like systems that build on the heritage of WAIS as released by Thinking Machines, but with quite a bit of extensibility and new features. Specifically, and with the help of many others, we're creating a Z39.50-92 (soon to be 93) server with full backwards compatibility with current WAIS clients and indexes, features such as boolean search, formatted searching on fielded text information, thesaurus based synonym substitution, stemming, and a search engine integration toolkit.

While the server maintains compatibility with current WAIS clients, it will also allow searches and retrievals from any Z39.50 client capable of recognizing the BIB-1 or INFO-1 attribute sets. Clients developed for this system will also be able to interact with library automation systems within the constraints of the attribute sets. We also hope to provide a tool to make various Z39.50 server maintenance tasks easier.

We hope to support the following Z39.50 features:

1) BIB-1 and dynamic INFO-1 ssearches

2) ES-1 element sets, which allow even more retrieval flexibility than the current WAIS server.

3) Access and resource control - essentially, this will allow for secure servers and servers that can accept billing information from the user.

4) EXPLAIN. Smart clients can use the EXPLAIN facility in conjunction with their preferred attribute set to provide customized search options for the user.

Basic search engine features (configurable): Synonym lookup; Stemming; Formatted field searches; Boolean search logic (to be added before release)

We have not yet begun to create the search engine toolkit as we think we need more user input, and the clients are rudimentary at best. We do hope to have the server running and ready for beta-test by February, although alpha code will be available much sooner. Clients will take somewhat longer as user interface development can be quite time consuming. The entire system will, of course, be freely available for use and modification by the network community.

On the client side, work is progressing on clients for the Macintosh and Unix systems, with a PC client in the wings.

So, when will this wonderful thing be ready? We actually have a server running that answers both WAIS and Z39.50-1992 queries on the same port, but returns results in a non-standard transfer syntax. We are in the process of changing this to support INFO-1 generic data records, and hope to have that completed soon.

We have had many comments from interested folks asking us to make these plans known to the rest of the network community, and at the risk of announcing vaporware, are doing so.

How can you help? We have already had quite a bit from many people, most notably John Kunze at the University of California at Berkeley and Clifford Lynch at UCOP. John created most of the server code and designed the INFO-1 attribute set as well as the ES-1 element set structures. I have made the server modifications to support free-text searching and WAIS compatibility, as well as making the necessary changes to the WAIS search engine. Members of Joan Gargano's staff at the University of California at Davis are now working on X Windows clients.

We are looking for folks willing to work on clients and run beta-test servers with their current WAIS information bases. Once our transfer syntax code is firmed up, trial servers will be made available for this purpose.

As technical manager for the Clearinghouse (as of December 1), I will help coordinate development tasks and incorporate any of your contributions intoreleases (with attribution, of course). We also have staff who will provide documentation and support for final releases.

Jim Fullton
Clearinghouse for Networked Information Discovery and Retrieval
Phone: (919)-248-9247 E-Mail: fullton@concert.net

archie
Search & retrieve anonymous FTP
Source: ftp.sura.net:/pub/archie
or ftp.mcgill.ca:/pub/archie
Demo: archie.sura.net, login as archie
Mailing List: Comments, bug reports to: archie-l@cs.mcgill.ca
To add or delete a site or correct site info:
archie-admin@cs.mcgill.ca
Contact: Alan Emtage bajan@bunyip.com or bajan@cc.mcgill.ca
Peter Deutsch peterd@bunyip.com or peterd@cc.mcgill.ca

gopher
Hierarchical browser for numerous operating systems
Source: boombox.micro.umn.edu:/pub/gopher
Demo: consultant.micro.umn.edu, login as gopher
Mailing list: Questions or comments: gopher@boombox.micro.umn.edu
To join gopher-news mailing list, send e-mail to:
gopher-news-request@boombox.micro.umn.edu
Usenet: comp.infosystems.gopher
Contact: please use gopher@boombox.micro.umn.edu to communicate with the gopher development team

hytelnet
Hypertext interface to telnet to selected sites for numerous operating systems
Source: access.usask.ca:/pub/hytelnet
Demo: access.usask.ca, login as hytelnet
Mailing list: Comments: scott@sklib.usask.ca
To join the updates mailing list, send e-mail to:
scott@sklib.usask.ca
Contact: Peter Scott scott@sklib.usask.ca
Earl Fogel (unix and VMS software) fogel@sask.usask.ca

libraries
Menu-driven scripted telnet connections to library OPACS and other selected resources
Source: sonoma.edu: /pub/libs.sh
Demo: vax.sonoma.edu, login as OPAC
Contact: Mark Resmer Mark.Resmer@sonoma.edu

libtel

Menu-driven scripted telnet connections to library OPACS and other selected resources (UNC-CH enhancements added)
Source: ftp.oit.unc.edu:/pub/doc/libtel.unix
Demo: bbs.oit.unc.edu, login as bbs
Contact: Paul Jones Paul_Jones@unc.edu

prospero
Networked file system
Source: cs.washington.edu: /pub/prospero.tar.Z
Mailing list: For more info send e-mail to: info-prospero@isi.edu
Contact: Clifford Neuman bcn@isi.edu

WAIS
Relevance feedback search and retrieval
Source: quake.think.com:/wais (some clients and servers, references to others)
sunsite.unc.edu: /pub/wais (most of the available clients and servers)
Demo: quake.think.com, login as wais
Mailing list: Moderated, digested list - send e-mail to:
wais-discussion-request@think.com
Technical discussion, unmoderated - send e-mail to:
wais-talk-request@think.com
Usenet: comp.infosystems.wais
Contact: George Brett ghb@concert.net

World-Wide Web
Hypertext interface to internet resources (links embedded in docs)
Source: info.cern.ch:/pub/www
Demo: info.cern.ch, no login or password
Note: WWW documentation is updated dynamically, and is available through the demo
Mailing list: Bugs and suggestions: www-bug@info.cern.ch
Contact: Tim Berners-Lee timbl@nxoc01.cern.ch

Project Envision: A User-Centered Database from the Computer Science Literature

NSF Project IRI-9116991 9/15/91 - 9/15/94

Principal Investigators Edward A. Fox, Lenwood S. Heath, Deborah Hix

Department of Computer Science, Computing Center Virginia Polytechnic Institute and State University (VPI&SU or Virginia Tech) Blacksburg, VA 24061-0106 Common Telephone 703/231-6931 FAX 703/231-6075 (CS Dept.) Common Internet: PI@vtopus.cs.vt.edu for PI=fox,heath,hix Direct to E. A. Fox: 703/231-5113, fox@fox.cs.vt.edu

This project involves managing the computer science literature by developing and evaluating a hypermedia database and system. Human-computer interaction and information storage and retrieval research are being applied to develop a new paradigm of user-centered database development and access. Computer science literature provided by the Association for Computing Machinery is being transformed to better serve the needs of a wide spectrum of computing practitioners and researchers. Based on interviews with computer professionals and analysis of existing publications, "objects" used in computing are being identified so that a tailored object-oriented database including algorithms, algorithm descriptions, bibliographic records, full texts of articles, reviews, data, page images, audio and video tapes can be stored and appropriately linked for searching and browsing. New minimal perfect hashing and graph partitioning algorithms will provide efficient access to the large, distributed graph of objects.

Information Infrastructure Project — Science, Technology and Public Policy Program

John F. Kennedy School of Government
Harvard University

Request for Information:
Policy Research on Information Infrastructure

The Information Infrastructure Project is developing a research-in-progress database on policy issues related to the emerging digital information infrastructure, especially as embodied in evolving Internet/NREN. The database will cover:

- design, management, and commercialization of the Internet and the emerging National Research and Education Network
- diffusion of scientific and technical information and knowledge
- electronic publishing and scholarly communication
- international aspects of the Internet
- national initiatives in internetworking outside the U.S.
- computer-based collaboration and cooperative work, especially across organizational boundaries
- pricing of data networks and digital information
- intellectual property
- commercial and industrial internetworking (concurrent engineering, CALS, EDI)
- access to and dissemination of government information
- network security and privacy issues
- common carrier issues and regulation of enhanced services
- network liability and First Amendment issues

- transformation of libraries and other information institutions
- public funding and procurement of elements of information infrastructure
- application of antitrust law to information resources
- economics of networked information environments
- Internet institutions and standards development processes

A major objective in developing this database is to identify potential presenters and authors for future workshops and publications. We are especially seeking rigorous work presented in a concise, lucid manner that enables policy-makers (who are typically not computer scientists, lawyers, economists, sociologists, etc.) to better understand the issues.

We plan to make the database publicly available and will consider submissions as permission to publish, unless we are notified to the contrary. We reserve the right not to list contributions and to edit those that we use. Listed contributors will have special access to early versions of the database.

If you wish to participate, please send a brief abstract of your project. You may send materials in any form, but we need the abstract if your project is to appear in the published version. Please use the following template, if possible, but we will accept submissions in your own format:

Project Title:
Institution:
Collaborating Institutions:
Principal Investigator(s):
Project Director(s):
Other Contact Person:
Address:
Address:
City:
State or Province:
Zip (Postal Code):
Country:

Telephone:
Fax:
Email:
Funding Source (optional):
Abstract:

Please send project information to:

gowlland@husc8.harvard.edu (gowlland@husc8.bitnet)

Or by U.S. Mail to:

Brian Kahin
Director, Information Infrastructure Project
Science, Technology and Public Policy Program
John F. Kennedy School of Government, Harvard University
79 John F. Kennedy St.
Cambridge, MA 02138
[kahin@hulaw1.harvard.edu
617-495-8903
Fax: 617-495-5776]

Information Infrastructure Sourcebook

The Information Infrastructure Project at Harvard's Kennedy School of Government announces the publication of the Information Infrastructure Sourcebook. The Sourcebook is designed to provide planners and policymakers a single volume reference on efforts to define and develop policy for a national information infrastructure. It includes historical policy documents, private sector vision statements and position papers, program and project descriptions (all sectors), landmark reports and pending legislation.

In assembling the Sourcebook, we have looked for documents that have had or are likely to have an impact on policy development, that are formal in nature, and that deal with information infrastructure at a general rather than topic level. In general, material is reproduced as is, although we have excerpted in the interests of space, relevance, balance, and consistency. Because of changes in technology, markets, programs and policies, the Sourcebook will of necessity require supplementation or revision on a regular basis. So we will look for new or updated contributions whenever they become available. Please direct any suggestions for additional material to:

James Keller
Project Coordinator
Information Infrastructure Project
Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138
tel: (617) 496-4042
e-mail: kellerj@ksg1.harvard.edu

The Sourcebook is available for $40, including Priority Mail postage and handling. If you would like to receive a copy of the Sourcebook, please send a check for $40 payable to Harvard University to:

Graceann Todaro
Center for Science and International Affairs
Kennedy School of Government
Harvard University
79 JFK Street
Cambridge, MA 02138

The Table of Contents from the Sourcebook is provided below:

TABLE OF CONTENTS

America's Economic Growth, A New Direction to Build Economic Strength, White House, Washington, DC, February 22, 1993. (excerpts)

Allan D. Bromley, The National Research and Education Network Program: A Report to Congress, OSTP, Washington, DC, December, 1992. (Table of Contents and Executive Summary)

Committee on Physical, Mathematical and Engineering Sciences, Federal Coordinating Council for Science, Engineering and Technology, Office of Science and Technology Policy, Grand Challenges 1993: High Performance Computing and Communications, a Supplement to the President's Fiscal Year 1993 Budget, Washington, DC, 1992.

Committee on Physical, Mathematical and Engineering Sciences, Federal Coordinating Council for Science, Engineering and Technology, Office of Science and Technology Policy, Grand Challenges: High Performance Computing and Communications, a Supplement to the President's Fiscal Year 1992 Budget, Washington, DC, 1991.

U.S. Congress, Public Law 102-94 - High Performance Computing Act of 1991, Washington, DC, December 9, 1991.

U.S. Congress, Senate Committee on Commerce, Science and Transportation - Report 102-57, High Performance Computing Act of 1991, U.S. Government Printing Office, Washington, DC, May 16, 1991.

Executive Office of the President, Office of Science and Technology Policy, The Federal High Performance Computing Program, OSTP, Washington, DC, September 8, 1989. (Executive Summary)

Office of Science and Technology Policy, Executive Office of the President, A Research and Development Strategy for High Performance Computing, Washington, DC, November, 1987.

Vision Statements and Position Papers

CALS Vision: Economic Growth Through Worldwide Enterprise Integration, CALS Industry Steering Group, National Security Industrial Association, undated.

Cable Television Laboratories, Inc., Cable's Role in the "Information Superhighway", Boulder, CO, 1993.

Computer Systems Policy Project, Perspectives on the National Information Infrastructure: CSPP's Vision and Recommendations for Action, Washington, DC, 1993.

Council on Competitiveness, Vision for a 21st Century Information Infrastructure, Washington, DC, May, 1993.

Electronic Frontier Foundation, The Open Platform: A Proposal by the Electronic Frontier Foundation for a National Telecommunications Infrastructure, Cambridge, MA & Washington, DC, 1992.

Information Industry Association, Telecommunications Infrastructure Objectives and Implementation Principles, Washington, DC, April, 1993.

Mary Gardiner Jones, The Consumer Interest in Telecommunications Infrastructure Modernization, Consumer Interest Research Institute, January 21-24, 1990.

Robert Kahn, National Information Infrastructure Components, Serials Review, Spring and Summer, 1992.

National Cable Television Association, Cable Television and America's Telecommunications Infrastructure, Washington, DC, 1993.

Telecommunications industry CEOs, Policy Statement on NII, March 24, 1993.

Program & Project Descriptions - all sectors

Robert Aiken, Hans Werner Braun, Peter Ford & Kimberly Claffy, NSF Implementation Plan for Interagency Interim NREN, National Science Foundation, May 1, 1992.

The CENDI Group, an overview, undated.

EINet - Stepping into the Electronic Marketplace: a Technical Overview, Microelectronics and Computer Technology Corporation, undated.

Federal Research Internet Coordinating Committee, Department of Energy, Program Plan for the National Research and Education Network, Washington, DC, May 16, 1989.

First Cities - Interactive Multimedia Information: Where, When and How You Want It, Microelectronics and Computer Technology Corporation, undated.

The National Initiative for Product Data Exchange, NIST, an overview, November, 1992.

The National Science Foundation, Program Solicitation: Network Access Point Manager, Routing Arbiter, Regional Network Providers, and Very High Speed Backbone Network Services Provider for NSFNet and the NREN Program, Washington, DC, National Science Foundation, DC, May 6, 1993.

Technology Reinvestment Project, Program Information Package for Defense Technology Conversion, Reinvestment and Transition Assistance, Arlington, VA, March 10, 1993. (Introduction and excerpts)

Reports

Computer Science and Technology Board, Commission on Physical Sciences, Mathematics and Resources, National Research Council, The National Challenge in Computer Science and Technology, National Academy Press, Washington, DC, 1988. (Chapter 3: The Promise of Infrastructure)

Computer Science and Telecommunications Board, National Research Council, National Collaboratories, National Academy Press, Washington, DC, 1993. (Executive summary and Chapter 1)

"The Future Federal Information Infrastructure: The Responsibility of the Federal Information Resources Management Community", excerpted from the Information Resources Plan of the Federal Government, Office of Management and Budget, November, 1992.

National Research and Education Network Review Committee of the Computer Science and Technology Board, Commission on Physical Sciences, Mathematics and Resources, National Research Council, Toward a National Research Network, National Academy Press, Washington, DC, 1988. (Executive Summary)

National Telecommunications and Information Administration, The NTIA Infrastructure Report - Telecommunications in the Age of Information, Executive Summary, U.S. Department of Commerce, October, 1991.

Office of Technology Assessment, U.S. Congress, Critical Connections: Communication for the Future, U.S. Government Printing Office, Washington, DC, January, 1990. (Chapter 1 - Summary)

Office of Technology Assessment, U.S. Congress, Global Standards - Building Blocks for the Future, U.S. Government Printing Office, Washington, DC, March, 1992. (p. 26-30)

Office of Technology Assessment, U.S. Congress, Helping America Compete - The Role of Federal Scientific and Technical Information, U.S. Government Printing Office, Washington, DC, July, 1990. (Chapter 1 - Summary)

Office of Technology Assessment, U.S. Congress, Informing the Nation, U.S. Government Printing Office, Washington, DC, October, 1988. (Chapter 1 - Summary)

Panel on Information Technology and the Conduct of Research, National Academy of Sciences, Information Technology and the Conduct of Research: The User's View, National Academy Press, Washington, DC, 1989. (Executive Summary)

Proceedings of the NREN Workshop, Monterey, California, September 16-18, 1992 (excluding appendices), Computing Research Association/EDUCOM/IEEE U.S. Activities Board, with assistance from the National Science Foundation Grant NCR-921671.

Steven R. Rivkin and Jeremy D. Rosner, Shortcut to the Information Superhighway: A Progressive Plan to Speed the Telecommunications Revolution, Progressive Policy Institute Policy Report No. 15, July 1992. (Introduction and Part four: a Progressive Strategy to Build the Fiber-Optic Network)

Proposed Legislation

U.S. Congress, Communications Competitiveness and Infrastructure Modernization Act of 1993 - H.R. 1504, Washington, DC, March 29, 1993.

U.S. Congress, To Establish a System of State-based Electronic Libraries - S.626, Washington, DC, March 22, 1993.

U.S. Congress, High Performance Computing and High Speed Networking

Applications Act of 1993 - H.R. 1757, Washington, DC, April 21, 1993.

U.S. Congress, Local Exchange Infrastructure Modernization Act of 1993 - H.R. 1312, Washington, DC, March 11, 1993.

U.S. Congress, National Competitiveness Act of 1993 - S.4, Washington, DC, January 21, 1993.

Interactive Accessibility: Breaking Barriers to the Power of Computing

Virginia Tech Research Project Announcement National Science Foundation (NSF) Institutional Infrastructure Award

Co-Principal Investigators
Dr. Roger W. Ehrich, Computer Science
Dr. Edward A. Fox, Computer Science
Dr. H. Rex Hartson, Computer Science
Dr. Deborah S. Hix, Computer Science
Dr. Robert C. Williges, Industrial and Systems Engineering

1. Project Description

The major goal of this interdiciplinary research activity is to improve interactive computing by removing two major barriers people face when using computer systems. First, usability barriers exist because the human-computer interface (HCI) is often poorly designed. Second, conceptual barriers exist because information access systems are poorly integrated. A consulting service and three inter-related research laboratories which are unique to HCI activities will be established under this grant.

THE INTERACTION CONSULTING SERVICE (DIRECTOR: DR. EHRICH) will be developed primarily as an HCI outreach program and will provide practicum experience for graduate students. This service will help the two departments, the university, and the industrial community to learn, adapt, and apply the newest HCI design, development, and evaluation technologies. This service will also help clients implement designs on diverse computing platforms. The Interaction Consulting Service provides new opportunities for practical application of HCI knowledge and skills as well as new sources of data regarding usability methods research.

THE INFORMATION ACCESS LABORATORY (DIRECTOR: DR. FOX) is an applications laboratory for studying multi-media information access for research and education. This lab will provide high volume storage, document capture, electronic publishing, imaging, and presentation services. The aim is to facilitate prototyping of new types of interactive information access systems that test our theories with real-life, realistic size collections.

THE INTERACTION TECHNOLOGY LABORATORY (DIRECTORS: DRS. HARTSON AND HIX) will house state-of-the-art equipment relating to HCI. A commitment to research in interactive systems necessitates a demonstration laboratory with representative up-to-the minute technology in computing platforms, operating systems, networked client-server technology, interactive applications software, and user interface development tools.

THE USABILITY METHODS RESEARCH LABORATORY (DIRECTOR: DR. WILLIGES) will extend and augment the existing HCI facilities in Industrial and Systems Engineering. This combined facility will be networked and will serve as the core of interdiciplinary research in HCI. An electronic conference area with telecommunications capabilities and two usability evaluation rooms will be built aroud a highly instrumented central control room containing video and computing equipment for conducting interactive experiments. Research in this laboratory will develop cost-effective and efficient methods for conducting formative and summative usability evaluation.

2. Importance to Virginia Tech and the Commonwealth of Virginia
A. Provides advanced computer equipment and laboratory facilities.
B. Improves undergraduate and graduate education.
C. Facilitates interdiciplinary undergraduate and graduate research.
D. Encourages industrial and other outside research support.
E. Compliments other major interactive computing activities on campus.
F. Supports University educational, researcg, and service objectives.
G. Serves National needs in advancing information technology.

3. Level of effort
A. Five year duration (July 1993 - June 1998)
B. Funding for equipment, maintenance, facility renovations, and personnel
C. Project Total: $2,279,810 (NSF - $1,375,000 and Virginia Tech - $904,810)

Project JANUS - Large-scale Imaging and Text-Retrieval Project

Introduction

JANUS is a five-year project of Columbia Law Library to develop an electronic, or 'virtual', library that combines digital conversion and storage of document text and graphics, massively parallel supercomputing, and advanced user-friendly search and retrieval software. When completed in 1996, JANUS will be a 'library of the future' and provide unprecedented electronic access nationwide to large-scale databases comprised of full text, graphics, images, data, sound and video materials. JANUS is a joint venture of Thinking Machines Corporation (TMC) of Cambridge MA, the world's largest manufacturer of highly parallel supercomputers, and Columbia Law Library, the nation's third largest collection of legal materials. JANUS uses the world's leading parallel supercomputer, The Connection Machine System (CM-2)*, on loan from TMC and installed at the Law School Library. Columbia and TMC are jointly developing retrieval software for use on large scale databases.

A 'Virtual' Library

Advances in optical scanning allow Project JANUS to convert and store in digital form an exact copy of a document page by page. Project JANUS is developing software that correctly interprets text that may have been corrupted during the optical character recognition process. By 1996, Columbia Law Library expects to convert about 10,000 to 12,000 volumes a year to electronic storage and expects a continued increase to its collection of new documents available only in electronic form.

JANUS' optical scanning process also converts and stores an exact copy of the document into a file that can be searched and manipulated by software developed by Project JANUS. JANUS' unique and powerful ability to search for and retrieve specific information from large scale text, graphics, image, sound and video databases best demonstrates the 'library of the future.' From any computer terminal connected to the JANUS system, a user seeking specific information on a subject can type in a word, phrase or sentence in ordinary English and receive in seconds specific references listed in order of relevance. The user has access to the exact copy of complete text at the position most relevant to the query and to relevant images and sounds. The user can save retrieved items or sections of items as needed. The user can continue searching for additional references by using new words or phrases or by highlighting words, phrases or sections of text already retrieved and again receive in seconds more references listed in orders of relevance.

JANUS' Database Software

Project JANUS accomplishes free-text searches by using a non-Boolean natural language algorithm that incorporates frequency-based indexing, relational value weighting, and relevance feedback from the individual user. Using the immense "number-crunching" power of parallel computing and building on TMC's work with the Wide Area Information Server (WAIS), JANUS's unique retrieval system, WAISSeeker, allows users to search the full text of documents for arbitrary combinations of words and, unlike keyword systems, does not require users to have prior knowledge of the discipline they are searching.

JANUS is also developing a Graphical User Interface (GUI) to make all processes easy to use. In addition to being clearly instructed during the feedback and the information retrieval phases of the process, system users will also be able to incorporate retrieved products into their word processing work products.

JANUS' development of WAISSeeker will provide a storage, search and retrieval system of unprecedented power, flexibility and accuracy. Both researchers and the general user will have easy access to large scale databases of complete text, images and sounds and will retrieve items most relevant to individual user queries.

Interconnection, Standards and Copyright Considerations

JANUS will first serve scholars on Columbia's campus. Because JANUS uses the Z39.50 standard as its communication protocol in its WAIS interface, it will be accessible when operational from any remote computer using a WAIS server and, via Internet, can serve users nationally and internationally. Development of a large bandwidth network channel, such as proposed in development of a National Research and Education Network, would allow a large number of users to browse and work in the Columbia Law Library from any NREN connection in the nation.

JANUS's database control system is designed to provide a datafile tagging and logging structure to protect copyrights and intellectual property. Columbia Law School is collaborating with Simon and Shuster to set standards for copyright protections applicable to wide-scale electronic retrieval and dissemination of information.

Project JANUS 1991-1996: "The Library of the Future"

Project JANUS is nearing completion of Phase One by developing and testing its natural language algorithm and extensively testing the system's ability to retrieve corrupted text. Phase Two will see a limited use production system available to researchers at the Columbia Law Library and via the Internet. JANUS will acquire large-scale databases and develop procedures for supporting multiple users concurrently using WAISSeeker.

In its final phase, JANUS will become a full production system with staffing for the Law School 'virtual' library in place. JANUS will have established relations with publishers to allow for use of copyrighted materials directly in electronic form and will have developed programs to track and verify use of licensed materials electronically.

*Connection Machine, CM-2 and Thinking Machines are trademarks of Thinking Machines Corporation. The CM-2 is equipped with 32,000 processors and has 256 megabytes of main memory and 20 gigabytes (billions of bytes) of hard disk storage.

The Mercury Electronic Library

William Y. Arms
December 16, 1992

Carnegie Mellon has a long term project known as Mercury, to build a large-scale digital library. Mercury began in late 1988/89 and became the Libraries' production system in 1991/92. Today, Mercury is probably the most complete distributed library in the United States and is the basis for a vigorous program of research and development.

One part of Mercury to create a large collection of managed information on-line.

The Libraries are a key contributor, but Mercury is designed to support databases that are managed by independent units, such as departments and research groups, and to interoperate with similar systems elsewhere. Mercury works closely with publishers since the economic and legal issues are a key to the development of a digital library.

The software architecture uses standard protocols and formats. Version 1 uses TCP/IP networking; Kerberos authentication; Z39.58 search syntax; Z39.50 (1988) retrieval protocol; ASN.1 encoding; documents stored as ASCII text or bit-mapped images with Fax G4 compression. It has user interfaces for Unix computers with the Motif window manager and VT100 terminal emulation. A Macintosh user interface is under development. The retrieval protocols have been implemented for OCLC's Newton database system and are being imported to the Fulcrum database software.

Special strengths of the software are the distributed computing and the support for images over networks. All the software building blocks are non-proprietary except the database systems. During the fall, Carnegie Mellon will provide an unsupported distribution of the software. Digital is planning a commercially supported version.

The Mercury plan combines both research and development. The following are the principal components of the current plan: intelligent search and retrieval using methods of computational linguistics and concept based retrieval; personal organization and structuring of information by establishing associative relationships between information objects; tools and algorithms for the capture, storage, manipulation, and recognition of high resolution images over networks; scalability; and interoperation with systems that are developed and managed by others.

NSF Network Information Services Awards

For Immediate Release
January 5, 1993

In cooperation with the Internet community, the National Science Foundation developed and released, in the spring of 1992, Project Solicitation NSF92-24 for one or more Network Information Services Managers (NIS Manager(s)) to provide and/or coordinate (i) Registration Services, (ii) Directory and Database Services, and (iii) Information Services for the NSFNET. As a result of this solicitation, three separate organizations were competitively selected to receive cooperative agreements totalling over $12 million in the three areas of (i) Registration Services, (ii) Directory and Database Services, and (iii) Information Services. Together, these three awards constitute the NIS Manager(s) Project, named the INTERNIC. Network Solutions will provide registration services, AT&T will provide directory and database services, and General Atomics will provide information services. It is important that the three project participants work closely together to provide a seamless interface for users in need of services. For this reason, the three awardees, at the request of the Foundation, have developed a detailed concept and plan to provide this seamless interface called the "INTERNIC" and have agreed to the structuring of their three separate awards as one collaborative project.

Steve Wolff, Director of NSF's Division of Networking and Communications Research and Infrastructure says, "We all feel intuitively that the domestic Internet and the distributed collaboration that it facilitates are rapidly creating a national "workplace without walls". These three awards to geographically dispersed organizations for Network Information Services, which require a high degree of coordination and collaboration, will both exploit and demonstrate the success of the network in enabling such distributed collaboration."

Consistent with FNC guidelines on obtaining reasonable cost recovery from users of NREN networks, the NSF has determined that the INTERNIC Information Services provider may charge users beyond the U.S. research and education community for any services provided. Also, the INTERNIC Directory and Database Services provider may charge a fee for maintenance of special databases, for extensive directory listings and may charge users beyond the U.S.research and education community. Finally, because the registration function provided by the INTERNIC Registration Services applies to domestic and international, commercial and individual users in addition to research and educational users, it is expected that an appropriate registration fee structure will take time to develop. NSF expects to engage in an extensive discussion with the domestic and international Internet community of the motivation, strategy and tactics of imposing fees for these services during the next fifteen months. Decisions will be implemented only after they have been announced in advance and an opportunity given for additional public comment.

Network Solutions will provide registration services as the IP registrar, issue IP numbers worldwide using delegated registries under the guidance of the Internet Assigned Numbers Authority and also register domain names, and track points of contact. Applications for assignment will be accepted via email or facsimile. The information from these assignments will be provided to the directory and database services provider to be made available to the entire Internet community. As a part of the Domain registration efforts Network Solutions will periodically release the top level zone files to be used by all root Domain Name servers.

AT&T will develop and maintain a Directory of Directories, including lists of FTP (File Transfer Protocol) sites, lists of various types of servers available on the Internet, lists of white and yellow page directories, library catalogs and data archives. AT&T will also provide white and yellow pages type Directory Services. Access to these services will initially be provided through several currently popular in-use interface methods while migrating to the use of X.500 technology, the current standard specification for distributed information storage and retrieval. The database services which AT&T will provide include the establishment of Database Services to extend and supplement the resources of the NSFNET, such as databases of contributed materials of common interest to the user community. AT&T will also offer database design, management, and maintenance to institutions and groups for inclusion in the Internet.

General Atomics will provide Information Services acting as the NIC of first and last resort and the NIC of NICs. The INTERNIC information services will include a full-service Reference Desk, a database of comprehensive networking materials called the Info Source, training classes and documentation, and coordination services among all appropriate groups in the community. In keeping with the innovative spirit of the Internet, several new approaches to distributing services will be implemented. Among these innovations is NICLink, a user-friendly hypermedia interface offering access to the Info Source and all the information it contains. NICLink will be distributed on both standard computer diskettes and CD-ROM. Another is the concept of the Info Scout, an individual who will scout out new resources and innovative uses of the network for inclusion in the Info Source.

Network Solutions is a 400-person telecommunications analysis and integration company headquartered in Northern Virginia. Its mission is to support its customers in achieving their missions through the mastery and application of networking technology. Network Solutions currently operates the DDN NIC.

AT&T is the leading provider of global information movement and management products and services. AT&T offers a wide array of data communications services that includes private line, X.25, frame relay, TCP/IP, protocol conversion, and electronic mail services.

General Atomics is a San Diego-based high-technology research and development company and operates CERFnet and the San Diego Supercomputer Center. CERFnet is an Internet network service provider that operates throughout the state of California and nationally. CERFnet was launched in the spring of 1989 with a

$2.8 million grant from the National Science Foundation. The San Diego Supercomputer Center is a five year cooperative agreement funded by the National Science Foundation to support high performance computing.

National Science Foundation Contact
Don Mitchell
(202) 357-9717
dmitchel@nsf.gov

Network Solutions Contact
Mary Bloch
(703) 742-4740
maryb@netsol.com

AT&T Contact
Shelly London
(908) 221-4355
london@attmail.com

General Atomics Contact
Susan Calcari
(619) 455-3900
calcaris@cerf.net

The Open Source Processing Research Initiative

MITRE Corporation 7525 Colshire Dr, McLean, VA 22102

The United States continues to face a rapidly changing world. Issues formerly given relatively low priority, such as international trade and Third World development, are now given much more weight by present government policy makers. Additionally, the Soviet Union's dissolution has not decreased the need for information about that region instead, government officials require detailed information about the former Soviet republics and their actions. Fewer people covering a growing set of issues creates a strong role for Open Source Information, a relatively low-cost and low-risk source of breaking news, reference materials, and scientific developments.

The MITRE Corporation, recognizing the important role that Open Source material will play in the future of the IC, has created the Open Source Processing Research Initiative (OSPRI). The goal of OSPRI is to address the broad spectrum of technical issues that the government faces in attempting to take advantage of Open Source material. These area's need to be addressed due to the large volume of information from multiple sources, in multiple formats, and with a high degree of duplication. Users will quickly grow frustrated with Open Source analysis if the information cannot be provided quickly, with a high degree of relevancy, in an intuitive manner. MITRE sees these issues as a microcosm of document management and information retrieval problems faced in other domains. Therefore, the technologies developed will be applicable across the general problem area.

The underlying concept behind OSPRI is that current information retrieval techniques place a significant burden on the user to develop sophisticated queries and dissemination profiles to locate relevant documents. The OSPRI research team will attempt to relieve this burden by focusing on providing filtering assistants that will generate knowledge about the document, aiding in information retrieval. A simplistic example of this would be to have filtering assistants that understand the part-of-speech for the words in a document and at query execution ensure that a match is in the appropriate part-of-speech so that a search on the word 'cooperative' as a noun would retrieve documents where 'cooperative' occurred in the 'private cooperative' sense. These filtering assistants would encode the "meta data" (data about the data) into the document for use by the query processor. The system would be extensible to any type of filtering assistant one could implement and assistants could use the work of previous assistants, allowing the system to grow with the technology. Examples of other assistants could include ones that locate geographical references and tag the latitude and longitude for use in displaying documents on maps, identifying important individuals, or assignment of topics to components of documents. These would allow the user to make queries such as a request to show all documents that refer to locations in some part of the world (which could be done by gesturing to a map). The OSPRI is developing a proof-of-concept system which includes a range of filtering assistants, user tools, and an information schema.

The OSPRI team's research efforts to develop filtering assistants include methods such as identifying items of interest referenced directly by proper nouns or indirectly by description or allusion. Parallel research is working toward identifying other grammatical elements of sentences, with the purpose of extending these techniques to identify the overall meaning of a body of text.

The OSPRI research staff has constructed a prototype system that demonstrates the feasibility of implementing this approach. The demonstration system developed was intended to prove the feasibility of a number of concepts: using Standard Generalized Markup Language (SGML) as the encoding mechanism, the importance of both the syntactic and semantic structure of a text, and improving information retrieval over standard keyword or boolean techniques. The data set for the demonstration system consists of approximately 350 articles from the 1989 Wall Street Journal (WSJ) and four closed caption transcripts from the MacNeil-Lehrer News Hour and ABC World News Tonight. The OSPRI staff manually examined all of the WSJ articles and video transcripts for these concepts and added appropriate semantic tags where the concepts appeared. The data set contained a non-trivial number of instances of these concepts that could not be effectively retrieved by a keyword search alone.

To provide a baseline to compare enhanced information retrieval techniques from tagged text, the staff conducted an empirical study on the prototype system's data set. Specifically, the staff compared the results of searches using unmodified keyword search terms (such as 'cooperative'), modified keyword terms ('private cooperative'), and keyword modified by syntactic category ('cooperative' as a noun). The study results demonstrate that even rudimentary natural language processing (NLP) techniques available today can significantly improve the performance of keyword search systems. As the OSPRI team brings increasingly sophisticated computational linguistic techniques to bear on information retrieval, it is expected that researchers will demonstrate substantial improvement over conventional methods.

The OSPRI is an umbrella program under which several projects are grouped. Each is described below:

Open Source Research - R. D'Amore

The objective is to extend the state-of-the-art in information retrieval by improving text analysis, mark up, and search tools. The program encompasses research that will extend automated Natural Language Processing (NLP) techniques such as logical representation of semantic meaning, proper noun identification, and categorization. In this project, output from each of these tools will be encoded in the text of the processed document as semantic and syntactic SGML tags, greatly improving current information retrieval, data base generation, indexing, and intelligence dissemination techniques.

Language Processing for Intelligence Applications - M. Vilain

The objective is to provide a working natural language understanding system. The scope of our research activity towards this objective includes two major areas: Robust language processing techniques that provide, in particular, for graceful error recovery in the face of unanticipated linguistic phenomena and self extensible knowledge bases these are necessary to break the knowledge acquisition bottleneck that prevents current language systems from applying in many domains.

Message Content Analysis - E. Lusher

The objective of this research is to develop and test the utility of two methodologies: a first filter that performs broad subject classification, and a second filter that identifies documents relevant to a user request.

Contextual Analysis to Improve Optical Character Recognition Performance - D. D'Amato

The objective is to improve optical character recognition (OCR) system accuracy by developing algorithms to: combine the outputs of two or more OCR systems, use system-specific information, such as character confusion probabilities, and perform domain-specific contextual processing.

Machine Translation of Spoken Dialogues - S. Luperfoy

The objective is to advance the state-of-the-art in dialogue processing by creating discourse solutions to unsolved problems in machine translations, user interface design, and speech recognition.

Open Source Broadcast Subject Classification - T. Albina

The objective is to reduce the amount of speech data to be processed manually by automatically grouping data into topic clusters that can be analyzed more rapidly.

The final product of OSPRI at the end of the year is an Advanced Technology Demonstration of an Open Source System. This demonstration will highlight each of the research areas with a focus on the leverage one obtains by combining the capabilities.

MITRE has been evaluating and selecting sources of material for inclusion in the OSPRI testbed. The primary sources of Open Source Information currently available to MITRE and the government are commercial databases, compact discs-read only memory (CD ROMs), government publishers such as the Foreign Broadcast

Information Service (FBIS), and free or inexpensive academic sources of text. One commercial source of Open Source Information is DIALOG, a comprehensive database system that includes the full text of popular newspapers, as well as journals from many disciplines. More technical material is available through the Defense Research, Development, Testing, and Evaluation (RDT&E) On-line System (DROLS). This database system, developed by the Defense Technical Information Center (DTIC), is maintained to provide on-line access to the DTIC data collection on worldwide research and development, technology advances, and energy acquisition and exploitation. Other databases cover legal, business, and U.S. Federal Government information. Of the other commercial data resources, CD-ROMs, which can provide up to 500 megabytes (MBs) of data per CD-ROM at a price often under $250, are the most cost effective.

Future technical areas to be addressed and improved include: refining the OSPRI semantic analysis algorithms, pre-filtering the mass of available Open Source data to produce a manageable and relevant subset for each user group, integrating multimedia sources and displays, examining data security and tag encryption, and continuing work on a graphical user interface that takes advantage of the increased functionality made possible by the OSPRI NLP techniques. William (Bill) A. Ruh, war@mitre.org, The MITRE Corporation, Workstation System Engineering Center Phone - (703) 883-6529 FAX - (703) 883-3315

Project on Scholarly Communication in the Network Environment

Announcement of Computer Conferences

The Science, Technology and Public Policy Program at Harvard's John F. Kennedy School of Government and the Coalition for Networked Information have undertaken a project to address principles, policy, and practice related to new forms of scholarly communication in the network environment. Funded in part by the National Science Foundation's Program on Ethics and Values Studies in Science and Technology, the Project looks at a set of social, ethical, and legal issues raised in the communication and dissemination of research results. The goal is to help different groups ranging from research teams to academic and professional societies develop appropriate policies and practices.

A draft background paper has been prepared by project director Brian Kahin and is available on the Internet by anonymous ftp from ftp.cni.org using the following sequence of commands:

```
ftp ftp.cni.org
login anonymous
[send e-mail address as password]
cd /CNI/projects/Harvard.scp
get background.txt
exit
```

We are now ready to initiate a set of computer conferences which will bring together diverse disciplinary and service perspectives. The conferences will address six issue areas which are discussed sequentially in the background paper:

1.) JOINT AUTHORSHIP AND OWNERSHIP -- How should jointly authored research be structured and how should publication processes be handled?
list name: OWNERSHIP
moderator: Michael Strait (mstrait@linknet.com)

2.) RIGHTS IN COMPUTER CONFERENCING -- What are reasonable expectations for the handling and reuse of messages and other material posted to groups and mailing lists?
list name: REPOST
moderators: Edward Vielmetti (emv@msen.com)
Steve Cisler (sac@apple.com)

3.) DERIVATIVE AND ITERATIVE WORKS -- What practices should apply to sequenced and variant publications of the same and related work?
list name: DERIV
moderators: Ann Okerson (okerson@umdc.bitnet)
Steven Zink (stevenz@equinox.unr.edu)

4.) CONTROL OF DISSEMINATION -- To what extent should key scholarly resources be controlled by particular scholars or organizations?
list name: RESOURCES
moderator: Doug Greenberg (sdgls@cunyvm.bitnet)

5.) SITE LICENSING -- How will widespread site licensing affect access to information by unaffiliated individuals and small firms and organizations?
list name: SITE-LICENSE
moderators: John Garrett (jgarrett@nri.reston.va.us)
Steve Gilbert (gilbert@educom.edu)

6.) INTERNATIONAL ACCESS -- How should researchers and practitioners in the developing world be assured access to research results?
list name: INTERNATIONAL
moderator: Art St. George (stgeorge@bootes.unm.edu)

Persons interested in participating in one of these conferences should send a single line mail message to
LISTSERV@CNI.ORG as follows:

subscribe [name of list] [your first name] [your last name]

If you wish to participate in more than one conference, put each request on a separate line in the same format. Please also send the moderators background information so they will know who is participating. The moderators will set their own policies and some may choose limit the size of their conference.

We plan to report on these conferences at the November meeting of the Coalition for Networked Information Task Force. There will be an invitational workshop for the most active participants and representative academic organizations in Washington in early 1993. A final report is due in the Spring.

For additional information, contact the list moderators at the listed email addresses. Or:

Brian Kahin
Director, Information Infrastructure Project
Science, Technology and Public Policy Program
John F. Kennedy School of Government, Harvard University
79 John F. Kennedy St.
Cambridge, MA 02138
617-495-8903
Fax: 617-495-5776
kahin@hulaw1.harvard.edu

Paul Peters
Director
Coalition for Networked Information
1527 New Hampshire Ave., NW
Washington DC 20036

202-232-2466
Fax: 202-462-7849
paul@cni.org

For technical assistance:

Craig A. Summerhill, Systems Coordinator
Coalition for Networked Information
1527 New Hampshire Ave., N.W.
Washington, DC 20036
craig@cni.org

Soviet Archive Exhibit

From: usmarc-l@RGUE.LOC.GOV

On June 17th, the Library of Congress will open an exhibit in the James Madison Memorial building of approximately 300 historically significant documents, photographs, and film clips from the Communist Party of the Soviet Union, the KGB, and other Soviet archives.

In addition, the Library will also digitize significant portions of about twenty - five documents, with translations, and distribute them as a national on-line exhibit that begins on the same day as the exhibit opening. The on-line exhibit will contain accompanying text, translations, captions, and graphic images for each of the documents included. The documents are available for personal study by scholars and students, and not for publication. Other exhibit information presented will be available for copying and subsequent use.

The on-line exhibit will be available in three services, all of which will be available on June 17, 1992. America On-line (AOL), a nationally-recognized commercial service, will provide one distribution network for the exhibit, Internet will provide another, and Sovset, an international computer network for Russian and East European studies will provide the third.

Through the use of text, message boards, libraries, "conference rooms," polls and surveys, AOL will assist the Library of Congress in its endeavor to make a major exhibit available to the public through the technology of on-line services. AOL allows for real time communications and the ability to engage in E-mail discussion of the documents. AOL's "The Library of Congress On-line" will mark the first time that a true nationwide dialog on a landmark exhibit has been made possible, and the first time that a simultaneous electronic surrogate exhibition has been presented by a museum or library.

On Thursday June 18th from 1-3pm, Dr. James H. Billington, the Librarian of Congress, Dr. Rudolph G. Pikhoia, chairman of the Committee on Archival Affairs of the Russian Federation, and General Dmitrii Volkogonov, head of the KGB Archive, will participate in an on-line computer forum on America On-line. The Library invites scholars, educators, students, and the general public across the United States to join this discussion of the exhibit and the opening of the Soviet Archives. America On-line will provide free membership software and 5 hours of on-line use without charge upon request for those who are not now members.

Those with access to Internet can obtain all of the text and graphic files that make up the on-line exhibit, but will not be able to hold real time conferences. The FTP files available from a mini-computer at the Library of Congress will require passwords that the Library will release and publicize on the opening day of the exhibit, June 17. However, the Library hopes that the exhibit will spark lively and informative exchanges of E-Mail and discussions in existing BITNET listserv forums. Library staff will monitor the discussions and seek to answer questions that may arise. The forums that LC will monitor are Sovset (which requires paid membership); Talk.politics.soviet (on BITNET from listserv@indycms under the list name TPS-L); Soc.culture.soviet (on BITNET via listserv@indycms under SCS-L); and RUSSIA (Russia and Her Neighbors List, through BITNET listserv@indycms on CREN and listserv@indycms.iupui.edu on the Internet).

Sovset is accessible over the Internet, CompuServe, and Telenet. Sovset's purpose is to encourage communication, data sharing, and collaboration on projects between research communities in the field. The three basic services offered by Sovset are computer conferencing, a data library, and electronic mail.

Those visiting the exhibit can view it in its electronic form in either the National Demonstration Lab on the first floor of the Madison building or in the Machine Readable Collections Reading Room, which is located on the ground floor of the Jefferson building (LJ G22).

TREC II Guidelines

This rather long message contains all the information needed for running TREC, broken up into convenient sections. It is important to read it all at some time, but particularly to read the introduction and guidelines sections before you make the routing queries. For those of you who participated in TREC-1, please note that the guidelines for query construction have been modified and should be re-read.

Table of Contents

Section 1 -- Introduction

TREC is designed to encourage research in information retrieval using large data collections. Two types of retrieval are being examined--retrieval using an "ad-hoc" query such as a researcher might use in a library environment, and retrieval using a "routing" query such as a profile to filter some incoming document stream. The test design and evaluation procedures are based on traditional information retrieval practices and are further elaborated in section 2. There are 40+ TREC participants, including some very large commercial groups, some very small university research groups, some actual products, and some radically different experimental approaches to information retrieval. The various participants are using a wide variety of indexing/knowledge base building techniques, and are using a variety of approaches to generating search queries. Therefore it is necessary to establish clear guidelines for the TREC task and to develop some methods of standardized reporting to allow comparison. The guidelines (section 3) deal with the methods of indexing/ knowledge base construction, and with the methods of generating the queries from the supplied topics. In general they are constructed to reflect an actual operational environment, and to allow as fair as possible a separation among the diverse query construction approaches.

The standardized reporting will take two complementary forms. The first will be your ranked results for each topic and a list of the actual queries used to produce those results. Section 4 gives the formats for submitting the results and some guidelines for submitting the queries. It is important that you follow the results format, and it is highly recommended that those sites new to TREC submit a sample of results to make sure your formats are correct. If you wait and submit official results in an incorrect format, there is no guarantee we will be able to get back in time to ask for corrections and your results may be rejected!

To make the actual relevance judgments, the top 100 ranked documents for each topic for each submitted set of results will be merged at NIST, randomized, and judged anonymously by relevance assessors. This ensures that the highest ranked documents from each submitted run will be evaluated and builds a large pool of sample documents for judgment. This process should also produce a quality test collection for future retrieval research, since relevance judgments will have been based on the combined output of more than 40 systems, including sites using human search expertise. Each group will get recall/precision tables for their group and the complete list of judged documents for each topic (tagged with relevance status). These can be used for failure analysis purposes, or for additional experiments before the conference. In addition to the traditional recall/precision evaluation, it is important to compare techniques by examining the amount of effort needed to produce various performance levels. The second form of reporting will be a standarized form (section 5) describing features and timing aspects of your system. This form covers many of the issues that you might be presenting in a talk or poster, but insures that everyone reports on certain items. Some of these items are for mostly statistical interest (such as how large is your common word list), but most of these issues are critical to comparing techniques as to the total amount of effort needed to perform searching. This is particularly applicable to timing issues so that we can all compare the benefit of such techniques as manually generating queries versus automatically generating queries, building extensive knowledge bases by hand versus automatic indexing, performing deep parsing of the data versus performing shallow parsing, etc. Please be as honest and consistent as possible as this form is meant to be genuinely informative to the rest of the TREC participants. The deadlines listed below are an expanded version of the ones in the call--the dates have not changed.

Schedule:

- April 1, 1993 -- Third gigabyte of data distributed via CD-ROM, after routing queries (for topics 51-100) are received at NIST. This is the earliest the test documents will be available--you may take more time to create the routing queries but it means less time to work with the new data.
- May 15, 1993 -- 50 test topics (101-150) distributed

- June 1, 1993 -- results from the 50 routing queries and the 50 test topics due at NIST
- July 30, 1993 -- relevance judgments and individual evaluation scores due back to participants
- Aug 30-Sept 1., 1993 -- TREC conference at NIST in Gaithersburg, Md.

Section 2 -- Conceptual View of TREC

The Application

The TREC task is not tied to any given application, and is not concerned with interfaces or optimized response time for searching. However it is helpful to have some potential user in mind when designing or testing a retrieval system. The model for a user in TREC is a dedicated searcher, not a novice searcher, and the model for the application is one needing monitoring of data streams for information on specific topics (routing), and the ability to do ad-hoc searches on archived data for new topics. It should be assumed that the users need the ability to do both high precision and high recall searches, and are willing to look at many documents and repeatedly modify queries in order to get high recall. Obviously they would like a system that makes this as easy as possible, but this ease should be reflected in TREC as added intelligence in the system rather than as special interfaces.

Test Design

TREC has been designed to evaluate system performance both in a routing (filtering or profiling) mode, and in an ad-hoc mode. The test design is based on traditional information retrieval models, and evaluation will use traditional recall and precision measures. A diagram of the test design illustrates the various components of TREC.

(Diagram Deleted)

This diagram reflects the four data sets (2 sets of topics and 2 sets of text) that are provided to participants. These data sets (along with a set of relevance judgments for the 100 training topics) are to be used to construct three sets of queries. Q1 is the set of queries (probably multiple sets) created to help in adjusting your system to this task, create better weighting algorithms, and in general to train the system for testing. The results of this research are used to create Q2, the routing queries to be used against the test documents. Q2 should be topics 51-100, not the full set of training topics. Q3 is the set of queries created from the test topics (101-150) as ad-hoc queries for searching against the combined training documents (D1 + D2). The results from searches using Q2 and Q3 are the official test results, and these queries and their search results are the information that should be sent to NIST. For example, topic 051, one of the training topics, could be run against the training documents as part of Q1, and then developed into a routing query to be run against the test documents as part of Q2. Topic 101, one of the test topics, should be run against the combined training documents (D1 + D2) as part of Q3.

For those TREC participants using a subset of the data (Category B), your training documents would be the Wall Street Journal text on the training CD-ROMs. The subsets of the test documents will be provided later.

Evaluation

The evaluation in TREC will mainly be standard recall/precision tables generated from your results. We will be using the SMART evaluation package, which includes both recall-level and document-level averaging. The package itself (programmed in C) is available from Chris Buckley (chrisb@cs.cornell.edu) and will be very useful to have to get early results or to evaluate additional runs using the same measures. One of the evaluation "improvements" suggested at TREC-1 was to strongly urge all participants to make multiple internal runs to attempt to modularize their various techniques to separate the effects of each technique, and then present the "unofficial" evaluations at the conference in addition to their official results.

Section 3 -- TREC Guidelines

Guidelines for constructing and manipulating the system data structures

The system data structures are defined to consist of the original documents, any new structures built automatically from the documents (such as inverted files, thesaurii, conceptual networks, etc.) and any new structures built manually from the documents (such as thesaurii, synonym lists, knowledge bases, rules, etc.).

1. System data structures can be built using the initial training set (documents D1 + D2, topics 1-100, and relevance judgments). They may be modified based on the test documents D3, but not based on the test topics 101-150. In particular, the system data structures may not be based in any way on the results of retrieving documents for the test topics and having a human look at the retrieved documents (or even just their titles).

2. There are several parts of the Wall Street Journal and the Ziff material that contain manually assigned controlled or uncontrolled index terms. These fields will be delimited by SGML tags, as specified in the read_me files included with the data. Other parts of the TREC data contain no manual indexing. Since the primary focus of TREC is on retrieval and routing of naturally occurring text, these manually indexed terms should not be indiscriminately used as if they are a normal part of the text. If your group decides to use these terms, they should be part of a specific experiment that utilizes manual indexing terms, and their use should be declared.

3. Special care should be used in handling the routing topics. In a true routing situation, a single document would be indexed and "passed" against the routing queries. Since most of you will be indexing the test data set as a complete set, routing should be simulated by not using any test document collection information (such as IDF based on the test collection, total frequency based on the test collection, etc.) in the searching. It is perfectly permissible to use training-set collection information however.

Guidelines for constructing the queries

There are many possible methods for converting the topics we will supply into queries that your system can execute. We have broadly defined three generic methods, based on the kind of manual intervention used. It is expected that all TREC participants will turn in either one or two sets of results for each topic. If two sets of results are submitted, they may correspond to different methods, or if desired, can be variants within the same method. Different methods can be used for the routing tests and the ad-hoc tests. For example, a site might submit results for ad-hoc retrieval for two different runs using automatic query construction (AUTOMATIC), perhaps comparing two different ranking formulae, or may submit results for automatic initial query construction (AUTOMATIC) vs relevance feedback (FEEDBACK). However, for the routing results, they might use one AUTOMATIC technique and one MANUAL technique (see below for definitions of these categories). The intent of the categories is as follows. The first two categories below (AUTOMATIC and MANUAL) are artificial categories. No decent information retrieval system would expect the user to look at 500+ retrieved documents without interacting with the system and at least tweaking the query a bit. But such human-system interaction is difficult to evaluate. In order to evaluate and compare systems, it's helpful to put artificial constraints on the problem, thus the first two categories. The FEEDBACK category loosens those constraints, but at a cost in the ability to accurately compare systems. Please read these category definitions CAREFULLY, and contact the program chair if you have any questions. Several groups in TREC-1 submitted misleading results due to misunderstanding the category definitions. We intend to avoid this in TREC-2.

1. AUTOMATIC -- completely automatic query construction.

ad-hoc queries -- The system will automatically extract information from the topic to construct the query. The query will then be submitted to the system (with no manual modifications) and the results from the system will be the results submitted to NIST. The system is not allowed to use any software, auxiliary files, knowledge bases, etc. that were manually changed after the ad-hoc topics were received. The only exception to this is to correct for formatting problems in the topic or document collections, as officially announced by NIST.

routing queries -- the queries should be constructed automatically using the training topics, the training relevance judgments and/or the training data. The queries should then be submitted to NIST before the test data is released and should not be modified after that point. The unmodified queries should be run against the test data and the results submitted to NIST.

2. MANUAL -- manual query construction.

ad-hoc queries -- The query is constructed in some manner from the topic, either manually or using machine assistance. Once the query has been constructed, it will be submitted to the system (with no manual intervention), and the results from the system will be the results submitted to NIST. There should be no manual intervention after initial query construction that would affect the results. This mimics the performance of a manual system with the user's initial query attempt, before any interactive tweaking of the query occurs. In particular, absolutely no alteration of the query based on examining of retrieved documents is allowed!

routing queries -- the queries should be constructed in the same manner as the ad-hoc queries for MANUAL, but using the training topics, relevance judgments, and training data. They should then be submitted to NIST before the test data is released and should not be modified after that point. The unmodified queries should be run against the test data and the results submitted to NIST.

3. FEEDBACK -- automatic or manual query construction with feedback.

adhoc queries -- The system takes an automatically or manually formed query and returns some documents to the user. After looking at those documents, the query may be reformulated and more documents returned for the new query. This is the most realistic category, but is the hardest to evaluate. It is essential that every document (or document title) looked at for a query be kept track of (in the order in which they were looked at). Evaluation will be based on this ordering, as well as the final set of retrieved documents.

The general algorithm would look like:
1. Form initial query (either automatically or manually)
2. Retrieve some documents
3. Look at a number of documents (or titles), keeping track of the docid's and order in which they looked at.
4. Reformulate the query based on the original query and information from the retrieved documents in step 3
5. Retrieve more documents.
6. Either stop, or go to step 3 again if another iteration of feedback is desired.

What will finally be turned in to NIST is your final list of the top X documents, plus your list of looked-at documents. NIST will then form a new set of X documents by joining your lists, removing possible duplicates, and ranking the looked-at documents ahead of your final list, in the order that they were looked at. (This is the standard "frozen evaluation" method of feedback evaluation). Note that an implication of this approach is that you probably want to look at as few documents from your initial run as is necessary, since your reformulated query should rank documents better than your initial query.

routing queries -- FEEDBACK cannot be used for routing queries.

- ---

Section 4
Results format and Query format

The results should be the top X documents retrieved for each topic. X will be determined at a later date, but will be AT LEAST 500 and might be 1000. This depends on logistics. If your group would have trouble keeping the top 1000 documents, please let us know. The format should be as follows:

030 Q0 ZF08-175-870 0 4238 prise1
030 Q0 ZF08-306-044 1 4223 prise1
030 Q0 ZF09-477-757 2 4207 prise1
030 Q0 ZF08-312-422 3 4194 prise1
030 Q0 ZF08-013-262 4 4189 prise1

etc.

The first column is the topic number. The second column is the query number within that topic. This should always be Q0 for AUTOMATIC and MANUAL, but should change in FEEDBACK as new iterations are made (i.e. the first iteration of feedback would use Q1, the second Q2, etc.). The third column is the official document number of the retrieved document and is the number found in the "docno" field. The fourth column is the rank the document is retrieved, and the fifth column shows the score (integer or floating point) that generated the ranking. This score MUST be in descending order and is important to include so that we can handle tied scores (for a given run) in a uniform fashion (the evaluation routines rank documents from these scores, not from your ranks). The sixth column should be a unique identifier for your group AND for the method used. The example shows the prise system, method 1. Please pick 5 letters followed by a 1, 2 or 3 depending on the method the run matches. If by some weird chance we get the same set of letters for two groups, we will let you know (see below). The width of the columns is not important, but it is important to include all columns and have at least one space between the columns.

Please send in a sample of results from at least two topics as soon as possible if you are new to TREC. These will not be used for anything but to make sure your formats are correct, and to help us set up our programs here to merge the official results. Because there is such a tight schedule in June, any mistakes in format must be corrected now.

- ---

Query format (by Dave Lewis)

Sites taking part in TREC are required to submit the queries they derive from the NIST-supplied topic descriptions. Routing queries are to be submitted to NIST before the test data for routing is received by the site. Ad hoc queries are submitted to NIST along with the top X documents retrieved by the ad hoc query.

The form of the queries submitted will of necessity vary from site to site. There are several somewhat contradictory goals to be met by the submitted queries:

1) The submitted queries should reflect the results of all manual intervention in the query process. For MANUAL querying without feedback the query submitted should be one for which all further system processing should be fully automated. For FEEDBACK querying with relevance feedback, we want two versions of the query--one before any relevance feedback -- i.e., one that would satisfy the MANUAL

restrictions -- and one after all relevance feedback and manual tuning is completed.

2) The submitted query should be as close as possible to the form actually used by your system in ranking documents.

3) The proprietary nature of the knowledge bases used by some systems should be respected to the extent possible.

4) The submitted query should be interpretable in isolation from the site's system to the extent possible.

There is likely to be considerable tension among desiderata 2, 4, and possibly 3. For instance, the submitted query for a system which ranked documents using a linear function might be in the form of two arrays, one of which gives unique integer term numbers for each indexing term and one of which gives real-valued coefficients for the linear function:

"POSITION: 0 1 2 3 4
ARR1: 3783 3838 4677 13999 20033
ARR2: 1.73 -3.8 13.38 2.73 0.47 "

This, however, is not a terribly useful representation from the standpoint of letting other researchers understand what your system is doing.

Now, suppose that this classifier is used to assign scores to documents based on the presence of whitespace-delimited words in the document text. We assume each indexing term corresponds to one such word, and the indexing term takes on a numeric depending on the number of occurrences of the word in a document. Then a more useful representation of the query might be:

"1.73 * Japan - 3.8 * U.S. + 13.38 * auto + 2.73 * export + 0.47 * car"

if these were the corresponding words. Even this form would only be completely interpretable in the context of additional information provided in documentation about the system, such as how it tokenizes text, and what within document term weights are used.

Some techniques, such as neural nets, factor analysis, and methods using knowledge bases, may produce queries that are difficult to represent in a comprehensible form. Sites which have questions about the form in which to submit their queries should contact the designated program committee contact for this issue, David Lewis

(lewis@research.att.com)."

Section 5
System Summary and Timing

General Comments
The timings should be the time to replicate runs from scratch, not including trial runs, etc. The times should also be reasonably accurate. This sometimes will be difficult, such as getting total time for document indexing of huge text sections, or manually building a knowledge base. Please do your best.

I. Construction of indices, knowledge bases, and other data structures
(please describe all data structures that your system needs for searching)

A. Which of the following were used to build your data structures
1. stopword list
a. how many words in list?
2. is a controlled vocabulary used?
3. stemming
a. standard stemming algorithms
which ones?
b. morphological analysis
4. term weighting
5. phrase discovery
a. what kind of phrase?
b. using statistical methods
c. using syntactic methods
6. syntactic parsing
7. word sense disambiguation
8. heuristic associations
a. short definition of these associations

9. spelling checking (with manual correction)
10. spelling correction
11. proper noun identification algorithm
12. tokenizer (recognizes dates, phone numbers, common patterns)
a. which patterns are tokenized?

13. are the manually-indexed terms used?
14. other techniques used to build data structures (brief description)

B. Statistics on data structures built from TREC text
(please fill out each applicable section)
1. inverted index
a. total amount of storage (megabytes)
b. total computer time to build (approximate number of hours)
c. is the process completely automatic?
if not, approximately how many hours of manual labor?
d. are term positions within documents stored?
f. single terms only?
2. clusters
a. total amount of storage (megabytes)
b. total computer time to build (approximate number of hours)
c. brief description of clustering method

d. is the process completely automatic?
if not, approximately how many hours of manual labor?
3. n-grams, suffix arrays, signature files
a. total amount of storage (megabytes)
b. total computer time to build (approximate number of hours)
c. brief description of methods used
d. is the process completely automatic?
if not, approximately how many hours of manual labor?
4. knowledge bases
a. total amount of storage (megabytes)
b. total number of concepts represented
c. type of representation (frames, semantic nets, rules, etc.)

d. total computer time to build (approximate number of hours)
e. total manual time to build (approximate number of hours)
f. use of manual labor
(1) mostly manually built using special interface
(2) mostly machine built with manual correction
(3) initial core manually built to "bootstrap" for completely machine-built completion
(4) other (describe)

g. auxiliary files needed for machine use
(1) machine-readable dictionary (which one?)

(2) other (identify)
5. special routing structures (what?)

a. total amount of storage (megabytes)
b. total computer time to build (approximate number of hours)
c. is the process completely automatic?
d. brief description of methods used

6. other data structures built from TREC text (what?)

a. total amount of storage (megabytes)
b. total computer time to build (approximate number of hours)
c. is the process completely automatic?
if not, approximately how many hours of manual labor?
d. brief description of methods used

C. Data built from sources other than the input text
1. internally-built auxiliary files
a. domain independent or domain specific (if two separate files, please fill out one set of questions for each file)
b. type of file (thesaurus, knowledge base, lexicon, etc.)

c. total amount of storage (megabytes)
d. total number of concepts represented
e. type of representation (frames, semantic nets, rules, etc.)

f. total computer time to build (approximate number of hours)
(1) if already built, how much time to modify for TREC?
g. total manual time to build (approximate number of hours)
(1) if already built, how much time to modify for TREC?
h. use of manual labor
(1) mostly manually built using special interface
(2) mostly machine built with manual correction
(3) initial core manually built to "bootstrap" for completely machine-built completion
(4) other (describe)

2. externally-built auxiliary file
a. type of file (Treebank, WordNet, etc.)

b. total amount of storage (megabytes)

c. total number of concepts represented
d. type of representation (frames, semantic nets, rules, etc.)

II. Query construction
(please fill out a section for each query construction method used)

A. Automatically built queries (ad-hoc)
1. topic fields used
2. total computer time to build query (cpu seconds)
3. which of the following were used?
a. term weighting with weights based on terms in topics
b. phrase extraction from topics
c. syntactic parsing of topics
d. word sense disambiguation
e. proper noun identification algorithm
f. tokenizer (recognizes dates, phone numbers, common patterns)
(1) which patterns are tokenized?
g. heuristic associations to add terms
h. expansion of queries using previously-constructed data structure (from part I)
(1) which structure?
i. automatic addition of Boolean connectors or proximity operators
j. other (describe)

B. Manually constructed queries (ad-hoc)
1. topic fields used
2. average time to build query (minutes)
3. type of query builder
a. domain expert
b. computer system expert
4. tools used to build query
a. word frequency list
b. knowledge base browser (knowledge base described in part I)
(1) which structure from part I
c. other lexical tools (identify)

5. which of the following were used?
a. term weighting
b. Boolean connectors (AND, OR, NOT)
c. proximity operators

d. addition of terms not included in topic
(1) source of terms
e. other (describe)

C. Feedback (ad-hoc)
1. initial query built by method 1 or method 2?
2. type of person doing feedback
a. domain expert
b. system expert
3. average time to do complete feedback
a. cpu time (total cpu seconds for all iterations)
b. clock time from initial construction of query to completion of final query (minutes)
3. average number of iterations
a. average number of documents examined per iteration
4. minimum number of iterations
5. maximum number of iterations
6. what determines the end of an iteration?

6. feedback methods used
a. automatic term reweighting from relevant documents
b. automatic query expansion from relevant documents
(1) all terms in relevant documents added
(2) only top X terms added (what is X)
(3) user selected terms added
c. other automatic methods
brief description

d. manual methods
(1) using individual judgment with no set algorithm
(2) following a given algorithm (brief description)

D. Automatically built queries (routing)
1. topic fields used
2. total computer time to build query (cpu seconds)
3. which of the following were used in building the query?
a. terms selected from
(1) topic
(2) all training documents
(3) only documents with relevance judgments
b. term weighting

(1) with weights based on terms in topics
(2) with weights based on terms in all training documents
(3) with weights based on terms from documents with relevance judgments
c. phrase extraction
(1) from topics
(2) from all training documents
(3) from documents with relevance judgments
d. syntactic parsing
(1) of topics
(2) of all training documents
(3) of documents with relevance judgments
e. word sense disambiguation
(1) using topic
(2) using all training documents
(3) using documents with relevance judgments
f. proper noun identification algorithm
(1) from topics
(2) from all training documents
(3) from documents with relevance judgments
g. tokenizer (recognizes dates, phone numbers, common patterns)
(1) which patterns are tokenized?
(2) from topics
(3) from all training documents
(4) from documents with relevance judgments
h. heuristic associations to add terms
(1) from topics
(2) from all training documents
(3) from documents with relevance judgments
i. expansion of queries using previously-constructed data structure (from part I)
(1) which structure?
j. automatic addition of Boolean connectors or proximity operators
(1) using information from the topics
(2) using information from the all training documents
(3) using information from documents with relevance judgments
k. other (brief description)

E. Manually constructed queries (routing)
1. topic fields used
2. average time to build query (minutes)
3. type of query builder

a. domain expert
b. system expert
4. data used for building query
a. from training topic
b. from all training documents
c. from documents with relevance judgments
d. from other sources (what?)

5. tools used to build query
a. word frequency list
b. knowledge base browser (knowledge base described in part I)
(1) which structure from part I
c. other lexical tools (identify)

d. machine analysis of training documents
(1) describe

5. which of the following were used?
a. term weighting
b. Boolean connectors (AND, OR, NOT)
c. proximity operators
d. addition of terms not included in topic
(1) source of terms
e. other (brief description)

III. Searching
A. Total computer time to search (cpu seconds)
1. retrieval time (total cpu seconds between when a query enters the system until a list of document numbers are obtained)

2. ranking time (total cpu seconds to sort document list)

B. Which methods best describe your machine searching methods
1. vector space model
2. probabilistic model
3. cluster searching
4. n-gram matching
5. Boolean matching
6. fuzzy logic (include your definition)

7. free text scanning
8. neural networks
9. conceptual graph matching
10. other (describe)

B. What factors are included in your ranking?
1. term frequency
2. inverse document frequency
3. other term weights (where do they come from?)
4. semantic closeness (as in semantic net distance)
5. position in document
6. syntactic clues (state how)
7. proximity of terms
8. information theoretic weights
9. document length
10. completeness (what % of the query terms are present)
11. n-gram frequency
12. word specificity (i.e., animal vs. dog vs. poodle)
13. word sense frequency
14. cluster distance
15. other (specify)

IV. What machine did you conduct the TREC experiment on?
How much RAM did it have?
What was the clock rate of the CPU?

V. Some systems are research prototypes and others are commercial. To help compare these systems:

1. How much "software engineering" went into the development of your system?

2. Given appropriate resources, could your system be made to run faster? By how much (estimate)?

3. What features is your system missing that it would benefit by if it had them?

Research Announcements

Research on Digital Libraries

,

A Joint Initiative of:

NATIONAL SCIENCE FOUNDATION
COMPUTER AND INFORMATION SCIENCE AND ENGINEERING DIRECTORATE

ADVANCED RESEARCH PROJECTS AGENCY
COMPUTING SYSTEMS TECHNOLOGY OFFICE and the
SOFTWARE AND INTELLIGENT SYSTEMS TECHNOLOGY OFFICE

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

PROPOSALS MUST BE RECEIVED AT NSF NO LATER THAN FEBRUARY 4, 1994

NSF 93-141 (NEW)

Digital Library Initiative, FY 1994

INTRODUCTION
The past decade has seen a remarkable expansion in digital networks within the U.S. research and education community from a state where networking was the purview of the privileged few to one where it is considered an essential tool by millions of researcher and educators. Since the mid-1980's with the advent of NSFNET the volume of traffic, the number of interconnected networks and the functionality of the networks has grown and continues to grow exponentially. The entire assemblage of linked networks using the IP communications protocol throughout the world is now referred to as the Internet. It is a network of networks, which within the U.S., links one third of all two year and four year colleges and universities, many primary and secondary schools, public and private institutions, commercial enterprises,

individuals in their homes, and foreign institutions in sixty countries. Information sources accessed via the Internet are the ingredients of a digital library. Today, the network connects some information sources that are a mixture of publicly available (with or without charge) information and private information shared by collaborators. They include reference volumes, books, journals, newspapers, national phone directories, sound and voice recordings, images, video clips, scientific data (raw data streams from instruments and processed information), and private information services such as stock market reports and private newsletters. These information sources, when connected electronically through a network, represent important components of an emerging, universally accessable, digital library.

To explore the full benefits of such digital libraries, the problem for research and development is not merely how to connect everyone and everything together in the network. Rather, it is to achieve an economically feasible capability to digitize massive corpora of extant and new information from heterogeneous and distributed sources; then store, search, process and retrieve information from them in a user friendly way. Among other things, this will require both fundamental research and the development of "intelligent" software. It is the purpose of this announcement to support such research and development by combining the complementary strengths of the participating agencies in basic research, advanced development and applications, and academic/industry linkage.

IMPORTANT CONSIDERATIONS

1. Awards will be made by NSF as Cooperative Agreements for proposals to engage in research, prototype and testbed activities. Awards of up to $1,200,000 a year for up to four years will be made to conduct programs of research and to develop and test elements of a digital library on a significant scale in a distributed environment. It is expected that up to 6 awards will be made under this announcement, depending on the quality of proposals and the availability of funds.

2. Cost sharing of at least 25% is required for all projects submitted in response to this announcement. The proposed cost sharing will be considered in evaluating proposals and will be a condition of any resulting awards (See the Proposal Preparation section, below).

3. Successful proposals will have demonstrated that as part of the research they will digitize a significantly large and important information collection, or use an existing collection, to serve as an experimental platform to demonstrate scale-up potential and as an experimental testbed for the research proposed. These testbeds must also be made accessible for research purposes to individuals who are not part of the

proposal team (see Proposal Preparation, 3, below)

4. Each proposal should, as appropriate to the research focus, include the active participation of the following groups, as relevant (these may be separate organizations or parts of a single organization): (1) client groups (e.g., specific research communities or other users of the information encompassed in the proposal); (2) commercial enterprises that would be involved in the commercialization of a digital library system (e.g., publishers, software houses, stock exchanges, equipment manufacturers, communications companies, etc.); (3) archival establishments, either private or governmental (e.g., libraries, data repositories, clearing houses, government or private information or data services); and (4) relevant computer and other science and engineering research groups (e.g., academic departments, supercomputer centers, industrial laboratories). These groups should be involved as sub-contractors to a single primary proposing academic institution.

5. A requirement for all awardees is that all publications, reports, data and other output from awards must be prepared in digital format and meet requirements for storage, indexing, searching and retrieval in a repository to be set up to capture results of these awards. These requirements will be devised jointly by the group of all awardees and the sponsoring agencies.

RESEARCH TO BE SUPPORTED

It is the purpose of this initiative to provide the funding and leadership for research fundamental to the development of digital libraries. Applicants can propose research in any or all of the following areas:

1-capturing data (and descriptive information about such data) of all forms (text, images, sound, speech, etc.) and categorizing and organizing electronic information in a variety of formats.

2-advanced software and algorithms for browsing, searching, filtering, abstracting, summarizing and combining large volumes of data, imagery, and all kinds of information; and

3-the utilization of networked databases distributed around the nation and around the world.

The types of research related to the above three areas are shown below. The examples are not meant to be exclusive but are meant to be illustrative:

Area 1

--- New research on systems for capturing data of all forms

For example:

* OCR page layout, segmentation and analysis software
* Speech recognition, audio segmentation and analysis software
* File conversion into editable, processable representations
* Broadcast capture & digitization (to generate multimedia data bases)
* Graphics understanding (image, drawing, graph recognition)
* Quality, fidelity maintenance

--- New research on how to categorize and organize electronic information in a variety of formats

For example

* Indexing, interpretation, classification and cataloging
* Multi-lingual indexing on content and citations
* Hypermedia structuring and linking of documents
* Graphical interfaces for knowledge representation
* Browsing technology for large knowledge spaces, news grazing

Area 2

--- New Research fundamental to the development of advanced software for searching, filtering, and summarizing large volumes of data, imagery, and all kinds of information

For example:

* Retrieval theories and models for data, metadata, information, knowledge bases, evaluation methods
* Formal structures of documents and texts, query languages
* Intelligent text processing and document management
* Feature-based image analysis and classification, pattern recognition
* Multi-pass retrieval algorithms (progressive refinement, feedback)
* Updatable indexing systems
* Spatial-temporal feature indexing of video
* Filtering, routing, alerting, selective dissemination of information
* Clustering, summarization, abstracting
* Natural language analysis for data extraction or representation
* Natural language generation systems
* Lexicon, thesaurus, concept space generation
* Adaptive/learning systems: connectionist, neural networks

* Context based pattern matching and retrieval of multimedia data
* Robust matching with noisy data, uncertainty, imprecision
* Fast search, query optimization

- - - Research on visualization and other interactive technology for quickly browsing large volumes of imagery

For example:

* Pictorial feature recognition, image classification
* Human perception (visual, auditory)
* Multi-scale displays, zooming
* Data visualization (e.g., airflow, human genome)
* Interactive visualization control
* Use of simulation to improve visualization/description
* Navigation, hypermedia, retaining serendipity, guides/paths/tours
* Developing metaphors, usable virtual reality environment: suites of objects (peripherals, information types, relations, properties, views) that work together for each domain
* Sustaining rapid performance with regard to moving and manipulating large digital imagery data sets.

Area 3

- - - Research on networking protocols and standards needed to insure the ability of the digital network to accommodate the high volume, bandwidth and switching requirements of a digital library.

For example:

* Network security
* Protocol design
* Data compression
* Ensuring scalability for large orders of magnitude increases in the number of simultaneous users

- - - New research leading to simplifying the utilization of networked databases distributed around the nation and around the world

For example:

* Knowbots/agents/mediators, intelligent gatekeepers
* Federated heterogeneous distributed object-oriented data and information base systems
* Personalized interactive news, magazine, and journal services

* Adaptable systems and services for disabled users, human augmentation
* Authentication, authorization
* Modeling, simulating usage, economics of access
* Collaboration technology with multimedia information interchange, multi-user editing, drawing, storage, retrieval, display, annotation, shared objects

- - - Research on individual and group behavioral, social and economic issues in digital libraries

For example:
* Intellectual property rights
* Privacy and security
* Impact of digital libraries on the conduct of science
* Publishing in a digital environment
* Charging mechanisms for copyrighted documents

PROPOSAL PREPARATION
The proposals must be marked DIGITAL LIBRARIES in the top left hand box, "Program Announcement", on the cover sheet (NSF Form 1207).

Proposals must be prepared according to the instructions given in Grants for Research and Education in Science and Engineering (GRESE: NSF 92-89) except that the following sections are also required.

1. A section following the Project Summary to be entitled, "Executive Summary". This section is limited to five pages. It will form an important first step in the review process and it should therefore be a careful abstraction of the key aspects of the proposal.

2. The Project Description section may not exceed 40 pages. It should be prepared as indicated on Page 4 of GRESE, making sure to explain the specific research and experiments to be conducted. These must be described in sufficient detail to allow merit review of the scientific content of the proposed research and its chance of success.

3. A section of the Project Description, entitled "Testbed Facility", describing the purpose and operation of the experimental testbed platform: its design; development; and management. This should include an explanation of how experiments and tests proposed by individuals or groups who are not part of the proposal will be solicited, evaluated and accommodated. (It is the intention of the

National Science Foundation to encourage applications for supplemental funding in years 2 through 4 to help fund the research of successful individual applicants.)

4. A section of the Project Description entitled, "Expected Accomplishments". This should be a statement and enumeration of the expected accomplishments and the measures you propose for evaluating progress annually for each year of the award duration.

5. A section separate from the Project Description entitled "Organizational Roles". This section should describe the nature of each organization's participation in the proposal with a description of its cost sharing, and the overall management plan for the project.

6. A section entitled, "Cost Sharing". The amount of cost sharing must be shown in the proposal in enough detail to allow NSF to determine its impact on the proposed project. Documentation of availability of cost sharing must be included in the proposal. Only items which would be allowable under the applicable cost principles, if charged to the project, may be made from any non-Federal source, including non-Federal grants or contracts. Contributions from non-Federal sources may be counted as cost
sharing toward Federal projects only once.

PROPOSAL EVALUATION

Proposals will be evaluated by panels of experts. Supplementary mail reviews will be solicited as feasible and necessary to achieve a fair and accurate review of all proposals. Some potentially successful submissions may receive site visits if the panel deems this desirable in order to properly evaluate the proposals. Selection of awards will be made jointly by the sponsoring agencies from the group of proposals receiving the highest ratings from the panel of experts.

Evaluation criteria which are applied to all NSF proposals are those listed in Grants for Research and Education in Science and Engineering, NSF 92-89, October, 1992. Additional criteria for this announcement are:

- inter-institutional aspects which would provide a positive stimulus to the design, development and commercialization of digital libraries;

- a commitment to and plan for operating an experimental testbed platform (see Proposal Preparation, 3, above);

- the extent to which the proposed research is likely to produce results generalizable to other digital libraries.

- cost sharing of at least 25% will be a specific evaluation criterion.

SCHEDULE
Organizations interested in submitting proposals should send an electronic mail message with the following statement: "I am interested in submitting a proposal to the Digital Libraries Initiative in Areas
___(1,2,3)."

Name
Title
Organization
Mailing address

email address

Address the email to:
gbarber@nsf.gov
cc: lrosenbe@nsf.gov

If email on the internet is not available to you then send a postcard with the same information to:

Up to October 25, 1993
Gwendolyn Barber
Room 310
National Science Foundation
1800 G Street, NW
Washington, D.C. 20550

As of October 25, 1993
Gwendolyn Barber
National Science Foundation
4201 Wilson Boulevard
Arlington, VA 22230

This information will be used to notify you of special regional briefings that will be held for this announcement. These briefings will be designed to answer your questions and to help you find potential partners for research consortia under this announcement. It is anticipated that these briefings will occur during November/ December, 1993.

IN ORDER TO BE CONSIDERED UNDER THIS ANNOUNCEMENT, PROPOSALS MUST BE RECEIVED BY NSF NO LATER THAN FEBRUARY 4, 1994

SUCCESSFUL PROPOSERS WILL RECEIVE NOTIFICATION IN JUNE/JULY, 1994.

OTHER
The National Science Foundation (NSF) provides awards for research in the sciences and engineering. The awardee is wholly responsible for the conduct of such research and preparation of the results for publication. The Foundation, therefore, does not assume responsibility for such findings or their interpretation.

The Foundation welcomes proposals on behalf of all qualified scientists and engineers, and strongly encourages women, minorities, and persons with disabilities to compete fully in any of the research and research related programs described in this document.

In accordance with Federal statutes and regulations and NSF policies, no person on grounds of race, color, age, sex, national origin, or disability shall be excluded from participation in, denied the benefits of, or be subjected to discrimination under any program or activity receiving financial assistance from the National Science Foundation.

Facilitation Awards for Scientists and Engineers with Disabilities provides funding for special assistance or equipment to enable persons with disabilities (investigators and other staff, including student research assistants) to work on an NSF project. Contact the program coordinator in the Directorate for Education and Human Resources. The telephone number is (202) 357-7562.

The Foundation has TDD (Telephonic Device for the Deaf) capability, which enables individuals with hearing impairment to communicate with the NSF Information Center about NSF programs, employment, or general information. The telephone number is (202) 357-7492.

This program is described in the Catalog of Federal Domestic Assistance category 47.070

Copies of NSF publication, Grants for Research and Education in Science and Engineering (GRESE: NSF 92-89) are available at no cost from:

National Science Foundation
Forms and Publications, Room 232
Washington, DC 20550

Telephone number: (202) 357-7963 or Fax number: (703) 644-4278.

Publications may also be ordered electronically using the Science and Technology Information System (STIS). The full text can be searched on-line, and copied from the system. Instructions for the use of the system are in NSF 91-10 "STIS Flyer". The printed copy is available from the Forms and Publications Unit. An electronic copy may be requested by sending a message to "stis@nsf" (bitnet) or "stis@nsf.gov" (Internet).

Technical questions about this announcement may be addressed to:

Laurence C. Rosenberg
Deputy Division Director
Information, Robotics & Intelligent Systems Division
National Science Foundation
1800 G Street, NW
Washington, D.C. 20550

Starting October 25, 1993
4201 Wilson Boulevard
Arlington, VA 22230

Electronic Mail address: lrosenbe@nsf.gov.
Telephone: (202) 357-9592
Fax: (202) 357-0320

Privacy Act and Public Burden Statements
The information requested on the application materials is solicited under the authority of the National Science Foundation Act of 1950, as amended. It will be used in connection with the selection of qualified proposals and may be used and disclosed to qualified reviewers and staff assistants as part of the review process and to other government agencies. See System of Records, NSF-50, Principal Investigator/Principal File and Associated Records and NSF-51, Reviewer/Proposals File and Associated Records, 56 Federal Register 54907 (October 23,

1991). Submission of the information is voluntary. Failure to provide full and complete information, however, may reduce the possibility of your receiving an award.

Public reporting burden for this collection of information is estimated to average 120 hours per response, including the time for reviewing instructions. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to:

Herman G. Fleming
Reports Clearance Officer
Division of Contracts Policy and Oversight
National Science Foundation
Washington, DC 20550

and to:

Office of Management and Budget
Paperwork Reduction Project (3145-0058)

Research on Scientific Databases

Special Announcement

Division of Information, Robotics and Intelligent Systems
Directorate for Computer and Information
Science and Engineering

NATIONAL SCIENCE FOUNDATION

PROPOSAL DEADLINE: September 15, 1992

The Division of Information, Robotics and Intelligent Systems (IRIS) in the Directorate for Computer and Information Science and Engineering (CISE) announces a special interest in supporting interdisciplinary research efforts in scientific and engineering databases.

BACKGROUND

Research on the design, development, management, and use of databases has traditionally focused on concepts and requirements critical to business-like environments. However, current database technology falls short of supporting the diverse needs of scientific and engineering applications. New advances in data storage/access technology, knowledge-based systems, and networked computing have brought about the promise to greatly increase the productivity in science and engineering research in this final decade of the 20th century. Scientific databases can be viewed as critical repositories of knowledge, both existing and yet to be discovered. Global change studies, astronomy, human genome mapping, social and economic studies, and engineering design are a few examples of research areas that generate and require access to extraordinarily large amounts of multi-media data forms: numbers, symbols, texts, images and others. Addressing the special characteristics and requirements of scientific and engineering databases will potentially further database technologies and enable a wide range of scientists and engineers to better utilize their data and other computing resources.

The research needed to make progress in scientific and engineering databases must be drawn from specialists in many disciplines. Therefore, collaboration of CISE researchers with researchers in other scientific or engineering disciplines is strongly encouraged.

Depending on the content of the received proposals, joint evaluation and funding by the relevant scientific and engineering NSF programs outside CISE is planned.

OBJECTIVES

The objective of this announcement is to foster coupling between database technology and scientific or engineering research for the advancement of both. Specifically, the aim is to promote:

1. Stimulation of multi-disciplinary research in scientific and engineering databases that addresses significant, real requirements of an application domain. Understanding of the requirements should be derived through collaboration with the domain scientists or engineers.

2. Expansion of general database technology through addressing the generic requirements of the application domains. Methodologies and tools developed for

scientific or engineering databases should be at a high conceptual level with an aim to enhance capabilities of the next generation of general information systems.

3. Enhancement of scientific infrastructure by making the databases, software, and other sharable resources produced under this initiative available to the research community.

AREAS OF INTEREST

Research is sought on methodologies and tools for the representation and manipulation of very large volumes of scientific or engineering data in highly distributed heterogeneous environments. In this context, research in three interrelated areas is encouraged:

1. Scientific Database Models and Systems. Theoretical foundations for the representation and manipulation of new data types (e.g., temporal, spatial and image data, textual data, spectrum data, engineering design data, materials data, chemical compounds, sequences, graphs, user-defined objects with inheritance and encapsulation, or declarative extensions); metadata management; data/knowledge calibration and validation; and uncertainty handling. System issues include system extensibility; rapid prototyping support; development of user-transparent, multi-level storage management (main memory through tertiary storage); multimedia data indexing; partial match retrieval algorithms; long/parallel/concurrent transaction processing; archiving; and version control. Research in this area must consider the special data characteristics associated with a scientific or engineering discipline.

2. Knowledge Discovery in Scientific Databases. Innovative methods, techniques and tools that provide seamless integration between database management and scientific analytic or engineering design tools. Topics span computing environment transparency; event finding, data examination, selection, analysis and manipulation of temporally or spatially related data (images); data analysis algorithms; scientific visualization; parallel model execution and cross-validation on large volumes of data; automated knowledge acquisition; incorporation of new knowledge into the system; and audit trail provisions.

3. Resource Sharing Environments. Improvement of data access and resource sharing in distributed, networked environments to support collaboration among scientists or engineers. Research in interoperability is critical to browsing, resource (data, bibliographic references or tools) location, access, and joint processing on systems ranging from personal workstations to supercomputers. Topics include

heterogeneous database systems management; effective methods for transport, combination and manipulation of data subsets, images or literature references; development of domain-specific and interdisciplinary lexicons and directories for databases and software; assessment of the data relevance and quality; and evolutionary establishment of self-describing, extensible standards for data exchange.

The topics listed above are not intended to represent the complete set of issues comprising the areas; they are suggestive rather than limiting. Other relevant information may be found in NSF workshop reports, "Scientific Data Management", "Database Systems: Achievements and Opportunities", and "Heterogeneous Databases Systems", published in SIGMOD Record, Vol. 19, No. 4, Dec. 1990, and in the report "Grand Challenges 1993: High Performance Computing and Communications" (available by calling NSF, Division of Information, Robotics and Intelligent Systems, 202-357-9572).

INQUIRIES

Inquiries about this research initiative announcement are welcome and may be directed to:

Dr. Maria Zemankova
Program Director, Database and Expert Systems
National Science Foundation
1800 G Street, N.W. -- Room 310
Washington, D.C. 20550
Telephone: 202-357-9570 Fax: 202-357-0320
Email: mzemanko@nsf.gov / mzemanko@nsf.bitnet

[Editor's Note: Dr. Zemankova left NSF on leave as of Fall 1993.]

Library of Congress

Excerpts from: Workshop On Electronic Texts Proceedings

Edited by James Daly

9-10 June 1992

Library of Congress Washington, D.C.

Supported by a Grant from the David and Lucile Packard Foundation

Acknowledgements

I would like to thank Carl Fleischhauer and Prosser Gifford for the opportunity to learn about areas of human activity unknown to me a scant ten months ago, and the David and Lucile Packard Foundation for supporting that opportunity. The help given by others is acknowledged on a separate page.

19 October 1992

INTRODUCTION

The Workshop on Electronic Texts (1) drew together representatives of various projects and interest groups to compare ideas, beliefs, experiences, and, in particular, methods of placing and presenting historical textual materials in computerized form. Most attendees gained much in insight and outlook from the event. But the assembly did not form a new nation, or, to put it another way, the diversity of projects and interests was too great to draw the representatives into a cohesive, action-oriented body.(2)

Everyone attending the Workshop shared an interest in preserving and providing access to historical texts. But within this broad field the attendees represented a variety of formal, informal, figurative, and literal groups, with many individuals belonging to more than one. These groups may be defined roughly according to the following topics or activities:

- Imaging
- Searchable coded texts
- National and international computer networks

- CD-ROM production and dissemination
- Methods and technology for converting older paper materials into electronic form
- Study of the use of digital materials by scholars and others

This summary is arranged thematically and does not follow the actual sequence of presentations.

NOTES:
(1) In this document, the phrase electronic text is used to mean any computerized reproduction or version of a document, book, article, or manuscript (including images), and not merely a machine-readable or machine-searchable text.

(2) The Workshop was held at the Library of Congress on 9-10 June 1992, with funding from the David and Lucile Packard Foundation. The document that follows represents a summary of the presentations made at the Workshop and was compiled by James DALY. This introduction was written by DALY and Carl FLEISCHHAUER.

PRESERVATION AND IMAGING

Preservation, as that term is used by archivists,(3) was most explicitly discussed in the context of imaging. Anne KENNEY and Lynne PERSONIUS explained how the concept of a faithful copy and the user-friendliness of the traditional book have guided their project at Cornell University.(4) Although interested in computerized dissemination, participants in the Cornell project are creating digital image sets of older books in the public domain as a source for a fresh paper facsimile or, in a future phase, microfilm. The books returned to the library shelves are high-quality and useful replacements on acid-free paper that should last a long time. To date, the Cornell project has placed little or no emphasis on creating searchable texts; one would not be surprised to find that the project participants view such texts as new editions, and thus not as faithful reproductions.

In her talk on preservation, Patricia BATTIN struck an ecumenical and flexible note as she endorsed the creation and dissemination of a variety of types of digital copies. Do not be too narrow in defining what counts as a preservation element, BATTIN counseled; for the present, at least, digital copies made with preservation in mind cannot be as narrowly standardized as, say, microfilm copies with the same objective. Setting standards precipitously can inhibit creativity, but delay can result in chaos, she advised.

In part, BATTIN's position reflected the unsettled nature of image-format standards, and attendees could hear echoes of this unsettledness in the comments of various speakers. For example, Jean BARONAS reviewed the status of several formal standards moving through committees of experts; and Clifford LYNCH encouraged the use of a new guideline for transmitting document images on Internet. Testimony from participants in the National Agricultural Library's (NAL) Text Digitization Program and LC's American Memory project highlighted some of the challenges to the actual creation or interchange of images, including difficulties in converting preservation microfilm to digital form. Donald WATERS reported on the progress of a master plan for a project at Yale University to convert books on microfilm to digital image sets, Project Open Book (POB).

The Workshop offered rather less of an imaging practicum than planned, but "how-to" hints emerge at various points, for example, throughout KENNEY's presentation and in the discussion of arcana such as thresholding and dithering offered by George THOMA and FLEISCHHAUER.

NOTES:
(3) Although there is a sense in which any reproductions of historical materials preserve the human record, specialists in the field have developed particular guidelines for the creation of acceptable preservation copies.

(4) Titles and affiliations of presenters are given at the beginning of their respective talks and in the Directory of Participants (Appendix III).

THE MACHINE-READABLE TEXT: MARKUP AND USE

The sections of the Workshop that dealt with machine-readable text tended to be more concerned with access and use than with preservation, at least in the narrow technical sense. Michael SPERBERG-McQUEEN made a forceful presentation on the Text Encoding Initiative's (TEI) implementation of the Standard Generalized Markup Language (SGML). His ideas were echoed by Susan HOCKEY, Elli MYLONAS, and Stuart WEIBEL. While the presentations made by the TEI advocates contained no practicum, their discussion focused on the value of the finished product, what the European Community calls reusability, but what may also be termed durability. They argued that marking up--that is, coding--a text in a well-conceived way will permit it to be moved from one computer environment to another, as well as to be used by various users. Two kinds of markup were distinguished: 1) procedural markup, which describes the features of a text (e.g., dots on a page), and 2) descriptive markup, which describes the structure or elements of a document (e.g., chapters, paragraphs, and front matter).

The TEI proponents emphasized the importance of texts to scholarship. They explained how heavily coded (and thus analyzed and annotated) texts can underlie research, play a role in scholarly communication, and facilitate classroom teaching. SPERBERG-McQUEEN reminded listeners that a written or printed item (e.g., a particular edition of a book) is merely a representation of the abstraction we call a text. To concern ourselves with faithfully reproducing a printed instance of the text, SPERBERG-McQUEEN argued, is to concern ourselves with the representation of a representation ("images as simulacra for the text"). The TEI proponents' interest in images tends to focus on corollary materials for use in teaching, for example, photographs of the Acropolis to accompany a Greek text.

By the end of the Workshop, SPERBERG-McQUEEN confessed to having been converted to a limited extent to the view that electronic images constitute a promising alternative to microfilming; indeed, an alternative probably superior to microfilming. But he was not convinced that electronic images constitute a serious attempt to represent text in electronic form. HOCKEY and MYLONAS also conceded that their experience at the Pierce Symposium the previous week at Georgetown University and the present conference at the Library of Congress had compelled them to reevaluate their perspective on the usefulness of text as images. Attendees could see that the text and image advocates were in constructive tension, so to say.

Three nonTEI presentations described approaches to preparing machine-readable text that are less rigorous and thus less expensive. In the case of the Papers of George Washington, Dorothy TWOHIG explained that the digital version will provide a not-quite-perfect rendering of the transcribed text--some 135,000 documents, available for research during the decades while the perfect or print version is completed.
Members of the American Memory team and the staff of NAL's Text Digitization Program (see below) also outlined a middle ground concerning searchable texts. In the case of American Memory, contractors produce texts with about 99-percent accuracy that serve as "browse" or "reference" versions of written or printed originals. End users who need faithful copies or perfect renditions must refer to accompanying sets of digital facsimile images or consult copies of the originals in a nearby library or archive. American Memory staff argued that the high cost of producing 100-percent accurate copies would prevent LC from offering access to large parts of its collections.

THE MACHINE-READABLE TEXT: METHODS OF CONVERSION

Although the Workshop did not include a systematic examination of the methods for converting texts from paper (or from facsimile images) into machine-readable form, nevertheless, various speakers touched upon this matter. For example, WEIBEL reported that OCLC has experimented with a merging of multiple optical character recognition systems that will reduce errors from an unacceptable rate of 5 characters out of every 1,000 to an unacceptable rate of 2 characters out of every 1,000.

Pamela ANDRE presented an overview of NAL's Text Digitization Program and Judith ZIDAR discussed the technical details. ZIDAR explained how NAL purchased hardware and software capable of performing optical character recognition (OCR) and text conversion and used its own staff to convert texts. The process, ZIDAR said, required extensive editing and project staff found themselves considering alternatives, including rekeying and/or creating abstracts or summaries of texts. NAL reckoned costs at $7 per page. By way of contrast, Ricky ERWAY explained that American Memory had decided from the start to contract out conversion to external service bureaus. The criteria used to select these contractors were cost and quality of results, as opposed to methods of conversion. ERWAY noted that historical documents or books often do not lend themselves to OCR. Bound materials represent a special problem. In her experience, quality control--inspecting incoming materials, counting errors in samples--posed the most time-consuming aspect of contracting out conversion. ERWAY reckoned American Memory's costs at $4 per page, but cautioned that fewer cost-elements had been included than in NAL's figure.

OPTIONS FOR DISSEMINATION

The topic of dissemination proper emerged at various points during the Workshop. At the session devoted to national and international computer networks, Clifford LYNCH, Howard BESSER, Ronald LARSEN, and Edwin BROWNRIGG highlighted the virtues of Internet today and of the network that will evolve from Internet. Listeners could discern in these narratives a vision of an information democracy in which millions of citizens freely find and use what they need. LYNCH noted that a lack of standards inhibits disseminating multimedia on the network, a topic also discussed by BESSER. LARSEN addressed the issues of network scalability and modularity and commented upon the difficulty of anticipating the effects of growth in orders of magnitude. BROWNRIGG talked about the ability of packet radio to provide certain links in a network without the need for wiring.

However, the presenters also called attention to the shortcomings and incongruities of present-day computer networks. For example: 1) Network use is growing dramatically, but much network traffic consists of personal communication (E-mail). 2) Large bodies of information are available, but a user's ability to search across their entirety is limited. 3) There are significant resources for science and technology, but few network sources provide content in the humanities. 4) Machine-readable texts are commonplace, but the capability of the system to deal with images (let alone other media formats) lags behind. A glimpse of a multimedia future for networks, however, was provided by
Maria LEBRON in her overview of the On-line Journal of Current Clinical Trials (OJCCT), and the process of scholarly publishing on-line.

The contrasting form of the CD-ROM disk was never systematically analyzed, but attendees could glean an impression from several of the show-and-tell presentations. The Perseus and American Memory examples demonstrated recently published disks, while the descriptions of the IBYCUS version of the Papers of George Washington and Chadwyck-Healey's Patrologia Latina Database (PLD) told of disks to come. According to Eric CALALUCA, PLD's principal focus has been on converting Jacques-Paul Migne's definitive collection of Latin texts to machine-readable form. Although everyone could share the network advocates' enthusiasm for an on-line future, the possibility of rolling up one's sleeves for a session with a CD-ROM containing both textual materials and a powerful retrieval engine made the disk seem an appealing vessel indeed. The overall discussion suggested that the transition from CD-ROM to on-line networked access may prove far slower and more difficult than has been anticipated.

WHO ARE THE USERS AND WHAT DO THEY DO?

Although concerned with the technicalities of production, the Workshop never lost sight of the purposes and uses of electronic versions of textual materials. As noted above, those interested in imaging discussed the problematical matter of digital preservation, while the TEI proponents described how machine-readable texts can be used in research. This latter topic received thorough treatment in the paper read by Avra MICHELSON. She placed the phenomenon of electronic texts within the context of broader trends in information technology and scholarly communication.

Among other things, MICHELSON described on-line conferences that represent a vigorous and important intellectual forum for certain disciplines. Internet now carries more than 700 conferences, with about 80 percent of these devoted to topics in the social sciences and the humanities. Other scholars use on-line networks for "distance learning." Meanwhile, there has been a tremendous growth in end-user computing; professors today are less likely than their predecessors to ask the campus computer center to process their data. Electronic texts are one key to these sophisticated applications, MICHELSON reported, and more and more scholars in the humanities now work in an on-line environment. Toward the end of the Workshop, Michael LESK presented a corollary to MICHELSON's talk, reporting the results of an experiment that compared the work of one group of chemistry students using traditional printed texts and two groups using electronic sources. The experiment demonstrated that in the event one does not know what to read, one needs the electronic systems; the electronic systems hold no advantage at the moment if one knows what to read, but neither do they impose a penalty.

DALY provided an anecdotal account of the revolutionizing impact of the new technology on his previous methods of research in the field of classics. His account, by extrapolation, served to illustrate in part the arguments made by MICHELSON concerning the positive effects of the sudden and radical transformation being wrought in the ways scholars work.

Susan VECCIA and Joanne FREEMAN delineated the use of electronic materials outside the university. The most interesting aspect of their use, FREEMAN said, could be seen as a paradox: teachers in elementary and secondary schools requested access to primary source materials but, at the same time, found that "primariness" itself made these materials difficult for their students to use.

OTHER TOPICS

Marybeth PETERS reviewed copyright law in the United States and offered advice during a lively discussion of this subject. But uncertainty remains concerning the price of copyright in a digital medium, because a solution remains to be worked out concerning management and synthesis of copyrighted and out-of-copyright pieces of a database.

As moderator of the final session of the Workshop, Prosser GIFFORD directed discussion to future courses of action and the potential role of LC in advancing them. Among the recommendations that emerged were the following:

* Workshop participants should 1) begin to think about working with image material, but structure and digitize it in such a way that at a later stage it can be interpreted into text, and 2) find a common way to build text and images together so

that they can be used jointly at some stage in the future, with appropriate network support, because that is how users will want to access these materials. The Library might encourage attempts to bring together people who are working on texts and images.

* A network version of American Memory should be developed or consideration should be given to making the data in it available to people interested in doing network multimedia. Given the current dearth of digital data that is appealing and unencumbered by extremely complex rights problems, developing a network version of American Memory could do much to help make network multimedia a reality.

* Concerning the thorny issue of electronic deposit, LC should initiate a catalytic process in terms of distributed responsibility, that is, bring together the distributed organizations and set up a study group to look at all the issues related to electronic deposit and see where we as a nation should move. For example, LC might attempt to persuade one major library in each state to deal with its state equivalent publisher, which might produce a cooperative project that would be equitably distributed around the country, and one in which LC would be dealing with a minimal number of publishers and minimal copyright problems. LC must also deal with the concept of on-line publishing, determining, among other things, how serials such as OJCCT might be deposited for copyright.

* Since a number of projects are planning to carry out preservation by creating digital images that will end up in on-line or near-line storage at some institution, LC might play a helpful role, at least in the near term, by accelerating how to catalog that information into the Research Library Information Network (RLIN) and then into OCLC, so that it would be accessible. This would reduce the possibility of multiple institutions digitizing the same work.

CONCLUSION

The Workshop was valuable because it brought together partisans from various groups and provided an occasion to compare goals and methods. The more committed partisans frequently communicate with others in their groups, but less often across group boundaries. The Workshop was also valuable to attendees--including those involved with American Memory--who came less committed to particular approaches or concepts. These attendees learned a great deal, and plan to select and employ elements of imaging, text-coding, and networked distribution that suit their respective projects and purposes.

Still, reality rears its ugly head: no breakthrough has been achieved. On the imaging side, one confronts a proliferation of competing data-interchange standards and a lack of consensus on the role of digital facsimiles in preservation. In the realm of machine-readable texts, one encounters a reasonably mature standard but methodological difficulties and high costs. These latter problems, of course, represent a special impediment to the desire, as it is sometimes expressed in the popular press, "to put the [contents of the] Library of Congress on line." In the words of one participant, there was "no solution to the economic problems--the projects that are out there are surviving, but it is going to be a lot of work to transform the information industry, and so far the investment to do that is not forthcoming" (LESK, per litteras).

The Library of Congress Information System (LOCIS) is now available over the Internet.

The telnet address is:

locis.loc.gov
140.147.254.3

LOCIS accepts both telnet 3270 and line mode.

LOCIS includes over 15 million catalog records and over 10 million records for other types of information: federal legislation, copyright registrations, Braille and audio, organizations, and selected foreign legal materials.

Searching hours are (all times USA eastern; closed national holidays):

Monday - Friday: 6:30am - 9:30pm
Saturday: 8:00am - 5:00pm
Sunday: 1:00pm - 5:00pm

Printed manuals will be available for sale later this summer and very soon via FTP (ftp seq1.loc.gov /pub/LC.On-line). There will be a LOCIS Quick Search Guide and a LOCIS Reference Manual. LC On-line Internet: lcon-line@seq1.loc.gov Library of Congress

Stan Horwitz Internet: STAN@VM.TEMPLE.EDU Bitnet: STAN@TEMPLEVM

Temple University's Sr. Mainframe Consultant; Manager of the Help-Net and Suggest lists; Listserv Postmaster (Standard disclaimers apply)

CHAPTER 6

Summary and Recommendations

Sketch of a Digital Library Initiative

E.A. Fox

Editorial Note: This subsection lays out my attempt to sketch key aspects of a national initiaative. It must be remembered that NSF normallywill fund research, development, and prototyping efforts such as those which are essential to the success of a broader National Initiative in this area.

A. Purpose

To advance U.S. science and engineering efforts, particularly research, education, and technology transfer, by improving the availability and supporting technology for access to useful information.

Note: We launch this in 1993, the 250th anniversary of the birth of Thomas Jefferson, who insisted that the free and vigorous pursuit of knowledge was essential to a democracy.

In modern terms: To develop vehicles and intelligent navigation methods for our nations' information highway and knowledge network.

B. Goals

- To develop a prototype national digital library that will be a proof of concept, showing it can be constructed, and that it yields important benefits.
- To develop digital information bases in key areas, that will be usable and useful, and should assist those working in the selected areas to be more productive.
- To develop improved methods and technologies for accessing (searching, browsing, routing, etc.) digitally encoded information.
- To strengthen U.S. research and, thereby, industry in this area, making it more competitive in the global community.
- To set a clear direction for coordinated development of larger national digital libraries, so that efforts in this area are compatible and systems are interoperable.

C. Objectives

- To leverage the network and computer technology from HPCC by adding content to be transmitted over the network, and by providing methods for more efficient and effective access to that information.

- To learn what methods work best in building a digital library, by observing the efforts and monitoring the progress of several top-rated groups as they work on the construction.

- To pick the best areas for building information bases, maximizing what can be learned in the process and maximizing the benefits to users that result from having the information.

- To learn what methods work best in accessing a digital library, by observing the efforts and monitoring the progress of several top-rated groups as they research this problem. To ascertain what works best as a function of the content (area, level), users (knowledge, experience, skill), and tasks. Key user communities should be included: K-12, undergraduate, graduate, practicioner, unemployed being retrained in new technology.

- To demonstrate the interdependence of information bases, methods for accessing them, and the quality of scientific and engineering research and education.

- To set in motion a self-sustaining government/industry cooperation that will ensure all publicly available information and knowledge becomes easily and universally accessible.

D. When Can We Tell This is Done?

- NSF's role in developing information bases will diminish once there are enough data/information/knowledge bases to support most of the types of important research that is needed, or when other agencies or corporations have taken over this effort. The first will be measured by asking proposers to justify why further development is needed for their research. The second will be observed when bases needed for research are existent or being developed by other groups.

- Development of information bases will be undertaken by government agencies like Library of Congress, National Library of Medicine, National Agriculture Library, or by other groups or corporations charged with that responsibility, or embarking upon this work as part of the free enterprise system.

- Research in these areas should be ongoing, though not necessarily as well supported if good efficiency, effectiveness, and usability has been scientifically verified based on known approaches.

E. Activity Context

- This responds to NSF's long range goal for "investment in human resources."
- The research undertaken deals with central concerns in computer and information science, especially those dealing with data, information, and knowledge — including their accumulation, processing, transformation, access, and utilization — as well as large scale systems to support those processes.
- This relates to other NSF goals: discovery of knowledge, building upon knowledge, ...
- This continues funding that goes back to the 1960's in the area of information access, but which has not been supported at this level in recent years. Thus, the aim is to build upon and discover new knowledge.
- This is an active area of research, supported by branches of NSF like CISE. That support, however
    - is at a relatively low level
    - is not coordinated or applied to help with the development of Digital Libraries

F. Impact Monitoring Plan / Evaluation Plan

- Annual progress reports

- Annual meeting of all funded by the initiative to report on progress and to ensure coordination and interoperability. Part of this will be "testing" of several access systems at a time, controlling the information base, user, and tasks -- to compare efficiency, effectiveness, and usability. Part of this can be "canned" comparisons, and part should be live.

G. Management

- NSFshould be the lead agency. IRISshould be the lead division at NSF. Other agencies, foundations, state governments, etc. should be invited to assist so that maximal leveraging of NSF funds takes place.
- A coordination/standard setting function is needed so that all research projects can share results, especially digital library data and software. Once the scope of this function is agreed upon, all funded groups would be obligated to participate in discussions and to abide by decisions falling in that scope and agreed to by a majority.
- If possible, the prototype information bases should be launched on a rolling basis, so that lessons learned from earlier efforts can be incorporated in work on others.
- If possible, the key areas should be dispersed around NSF, so that all of the important areas are included: hard science, social/behavioral science, K-12 education (preferably, say, building a 10K volume equivalent from out-of-copyright works that are at the core of the K-12 education scene). Another area might be patents, to demonstrate technology transfer, with funding from the Department of Commerce. DARPA might pick some key area to fund, such as materials, based on strategic defense needs.
- A monthly caucus, possibly co-directed by NSF and DARPA,should be held, with all government agencies invited to the entire event, and at least part of the event (e.g., presentations, demonstrations, talks) open. These would help ensure coordination, and lead to spin-off technical or administrative task forces (e.g., to agree on suites of standards).

Included should be: Departments (Agriculture, Commmerce, Defense, Education, Energy, Justice), Libraries (LoC, NAL, NLM), and Agencies (CIA, DARPA, FBI, NASA, NIST), Key contacts include: Larry Hunter, Barry Jacobs, Alexa McCray, Gio Wiederhold.

CHAPTER 7 Bibliography

References

From the June issue of BYTE 1991 in the State of the Art section and in the Solutions Focus section:

State of the Art Section - Managing Infoglut:

Making Knowledge Pay by Christopher Locke, pp. 244-246, 248, 250, 252.

A good overview of the way the new information managing technologies can be put to efficient use in organizations, particularly in support of collaborative projects. The potential for expert systems is down played, while the expanding potential of hyptertext and structured forms of information is emphasized. There is mention of the article "The Dark Side of DIP" (BYTE, April 1991, p. 193) on the increasing need for indexing to help manage document images, which is an article that every information professional should read.

The New Age of Documents by John E. Warnock (CEO of Adobe Systems), pp. 257-260.

Discusses the critical role of universal document interchange, including the tranfer of images and standard fonts, in efficient communication. The author describes the technologies of Matching Fonts, a techniques for transferring fonts between computers with a minimum of information overhead, and Carousel, the Adobe designed format to enable universal document exchange and sharing.

Electronic Books by Louis R. Reynolds and Steven J. DeRose, pp. 263-268.

The practical prospects for the introduction of electronic books are described, as are some of the likely immediate applications of such a medium. A number of large scale electronic book projects are discussed, along with the Motif, Intermedia and DynaText technologies which are presented as examples of what is now available. They also talk about the importance of SGML and other publishing standards - as well as those standards still under development such as the ISO HyTime and the DSSSL (Document Style, Semantics and Specification Language) standards - in implementing this technology.

Search and Retrieval by Earlene Busch, pp. 271-274, 276.

A nice short review of some of the new text retrieval methodologies such as document clustering, probabilistic pattern matching, term weighting and other techniques which replace or enhance Boolean searches in large textual databases.

SGML Frees Information by Haviland Wright, pp. 279-282, 284, 286.

For those of you that have not yet seen an SGML encoded document, this article is a must. It describes the impact SGML will have on efficient text management and describes how they are authored and tagged.

Resource Guide: Search and Retrieval Systems, pp. 287.

Prices and sources for 32 products that allow you to create and search text databases.

Referenced and Related Documents

[Editor's Note: Throughout this Source Book, many entries make references to their origins. This section has supplementary references to other influential publications.]

After the Electronic Revolution, Will You Be the First to Go?: Proceedings of the 1992 Association for Library Collections & Technical Services President's Program, 29 June 1992, American Library Association, 1993

Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, 1989 by Gerard Salton

Capitalizing on the Nation's Knowledge: Making Use of What We Know, A Statement by John Seely Brown before the Subcommittee on Science, Technology, and Space of the Committee on Commerce, Science, and Transportation, United States Senate, 15 September 1989

Excerpt from Introduction:

In its potential to amplify productive and innovative practices in the crucial years of international competition that lie ahead, the National Research and Education Network offers an extremely exciting prospect. Along with its generally recognized possibilities for research and education, we find latent with this high-performance network means to overcome some of the present and future challenges to the nation's entire manufacturing and technological base.

Communications in Support of Science and Engineering: A Report to the National Science Foundation from the Council on Library Resources, pp 5, 7, 9, 10, 18, 24, 25, 31, August 1990

Excerpt from Abstract:

During the summer of 1989, the National Science Foundation asked the Council on Library Resources to explore selected aspects of scientific and engineering communication with the objective of learning more about the relationships between information resources and scientific productivity. Such information is of use in the preparation of the Science and Engineering Indicators series of the National Science Board. The enclosed papers, which constitute the report of the project, are the result of CLR's effort. The topic is a complex one, but it is the Council's hope that what has been done will point the way toward productive areas for investigation.

Draft Preliminary Findings of the Rights for Electronic Access to and Delivery of Information (READI) Project, Robert Ubell Associates for Coalition for Networked Information, September 29, 1992

Excerpt from Objective:

Under contract with the Coalition for Networked Information, Robert Ubell Associates convened a series of three expert panels to explore whether and, if so, how contract law, in the form of licenses and related agreements between creators and users of published works, could be applied within the context of copyright law to ease the flow of networked information. The particular objective was to determine whether it is possible to develop a common set of terms and conditions for managing relationships in the emerging market for networked information.

EDUCOM Update, March/April 1993, Volume 2, Number 2, "EDUCOM Calls for A National Network Not a Federal Network"

Editor's Note: This article can be obtained by sending e-mail to LISTSERV@BITNIC containing the command:
GET HETERICK UPMAR93 EDUNEWS

RG Fiddles and DR Winterbottom, *The Feasibility of Electronic Accession of Archive Material at the Printers: British Library R&D Report 6061*, British Library Research and Development Department, 1991

Excerpt from Background Summary:

This report is the result of a study into the feasibility of acquiring material for storage in an electronic archive during the production process when it is first typeset and is held by the typesetter in electronic form. The premise to be investigated was that it may be feasible to take a copy of the electronic form of the material held by the printer or typesetter rather than converting the printed copy back to an electronic form at a later date. This would result in a copy of the material which contained text in ASCII code, which is compact and which would allow the recipient to use electronic search techniques to enhance the use of the material. A survey of printers was carried out to establish the current state of technology and of practives in the printing industry in the UK to determine the likely technical problems that would be encountered in copying electronic data.

Frakes, William B., and Baeza-Yates, Ricardo, Eds.; *Information Retrieval: Data Structures and Algorithms*, Prentice Hall, 1992

Garrett, John R. and Alen, Joseph S., *Toward a Copyright Management System for Digital Libraries*, Copyright Clearance Center, Inc. March 1992

Excerpt from the Executive Summary:

"What is it about digital information that makes designing a workable copyright management system so important, and so complicated?"

Prepared by the Copyright Clearance Center under an agreement with the Corporation for National Research Initiatives, this Report seeks to define the key issues in copyright management in digital environments, and to describe the legal and institutional contexts in which any workable system must operate.

The Report discusses several assumptions about copyright management of electronic information in detail, outlining and responding to counter-arguments. There is also an extensive analysis of the U.S. and international legal framework for copyright management of digitized information, with a detailed discussion of major issues of concern to both rightsholders and users.

A Hybrid Systems Approach to Preservation of Printed Materials: A Report to the Commission on Preservation and Access, by Don Willis, Commission on Preservation and Access, November 1992

Internet Society, *The Internet Society News*, Winter 1992, vol 1, no. 1

Joint Study in Digital Preservation: Report Phase I: Digital Capture, Paper Facsimiles, and Network Access, The Cornell/Xerox/Commission on Preservation and Access, Project Managers: Anne R. Kenney and Lynne K. Personius, Cornell University, September 1992

Excerpt from Executive Summary:

Cornell University and the Xerox Corporation, with the support of the Commission on Preservation and Access, have collaborated for the past two years in a Joint Study to investigate the use of digital technology to preserve library materials. The primary emphasis of this study has been on the capture of brittle books as digital images and the production of printed paper facsimiles. Of equal interest, however, has been the role of digital technology in providing networked access to library resources, and preliminary work in this area has also been accomplished.

The Joint Study has led to a number of conclusions regarding preservation, access, electronic technology, and the role of the library. In particular, participation in this study has convinced Cornell of the value of digital technology to preserve and make available research library materials. Such digital preservation presents a cost effective alternate to photocopying, and -- subject to the resolution of certain remaining problems -- a potential adjunct or alternative to microfilm preservation. The greatest promise of digital technology as a preservation option is to improve access to materials. Cornell expects to work with others to find ways to resolve the remaining issues surrounding the use of digital technology.

Krol, Ed; *The Whole Internet: User's Guide and Catalog*, O'Reilly & Associates, Inc., September 1992

Methodologies for Intelligent Text Processing and Information Resources Management: Achievements and Opportunities: Report of the NSF Invitational Workshop on Future Directions in Text Analysis, Retrieval and Understanding, Jitender S. Deogun, Chair, NSF & University of Nebraska, 1993

The Organizational Phase of Project Open Book: A Report to the Commission on Preservation and Access, by Donald Waters and Shari Weaver, Library and Administrative Systems, Yale University, Commission on Preservation and Access, September 1992

Excerpt from Introduction:

The Yale University Library is now organized to move ahead with Project Open Book, the conversion of 10,000 books from microfilm to digital imagery. In the final phase of the Project -- the organizational phase -- Yale established a Steering Committee, including several faculty members, and created a project team. In addition, Yale conducted a formal bid process and selected the Xerox Corporation to serve as its principal partner in the project.

In this report we review the purpose and scope of the project, outline the steps taken in this first, organizational phase, and finally, present a summary of the results to date.

Preservation of New Technology: A Report of the Technology Assessment Advisory Committee to the Commission on Preservation and Access, by Michael Lesk, The Commission on Preservation and Access, October 1992

Proceedings of the Workshop on Future Directions in Text Analysis, Retrieval and Understanding, NSF & University of Nebraska, February 1993

Talbutt, Palmer, *Rationalization and Focused Retrieval*, 1992

Technology, Scholarship, and the Humanities: The Implications of Electronic Information, Summary of Proceedings for September 30 - October 2, 1992, The American Council of Learned Societies and The J. Paul Getty Trust, 1993

Excerpt from Foreward:

This volume constitutes a summary record of an important conference. Its brevity is intended to offer a rapid grasp of the main discussion points and conclusions produced by the conference, rather than to convey the spirit and diversity of the event and its participants.

The findings of the working groups are offered as guides to those whose decisions affect the creation and use of electronic resources in institutional, technical, scholarly, and philanthropic settings. Reflecting not just the voice of the scholar nor only that of the librarian or systems specialist, this volume, like the conference, has captured the conjoint thinking of diverse, highly placed experts from the many areas that bear on scholarly computing in the humanities.

Time Magazine, "Cyberpunk!", pp 59-65, February 8, 1993

Time Magazine, "The Electronic Superhighway", pp 50-55, April 12, 1993

A Workshop on Scientific Collaboration and Interaction with Remote, Unique Facilities: Planning a Sondre Stromfjord Upper Atmospheric Research Facility Testbed; A Final Report to the National Science Foundation, University of Michigan, 1992

CHAPTER 8 Glossary

Abbreviations

AI	Artificial Intelligence
ACM	Association for Computing Machinery
ALA	American Library Association
ARL	Association for Research Libraries
ARPA	Advanced Research Projects Agency
BDF	Bibliotheque de France
CCIRN	Coordinating Committee for Intercontinental Networking
CCL	Common Control Language
CISE	Directorate for Computer and Information Science and Engineering, NSF
CIX	Commercial Internet Exchange
CNI	Coalition for Networked Information
CODER	COmposite Document Expert/extended/effective Retrieval System
CNRI	Corporation for National Research Initiatives
CSCW	Computer Supported Cooperative Work
DARPA	Defence Advanced Research Projects Agency

DBMS	Database Management System
DTIC	Defense Technical Information Center
DVI	Digital Video Interactive (TM Intel)
EFF	Electronic Frontier Foundation
GIS	Graphical Information System
HCI	Human-Computer Interaction
HMSO	Her Magesty's Stationery Office (HMSO)
HPCC	High Performance Computing and Communications Initiative
IAB	Internet Architecture Board
IR	Information Retrieval
IRIS	Division of Information, Robotics and Intelligent Systems, NSF
LC	Library of Congress
LEND	Large object-oriented External Network Database
LOC	Library of Congress
NAL	National Agricultural Library
NASA	National Aeronauntics and Space Administration
NECC	National Education Computing Conference
NEH	National Endowment for the Humanities
NEL	National Electronic Library
NIST	National Institute for Standards and Technology
NLM	National Library of Medicine
NREN	National Research and Education Network
NSF	National Science Foundation
NTIS	National Technical Information Service
OCLC	On-line Computer Library Center
OSIS	Office of Science Information Service
RARE	European Reseaux Associees pour la Recherche Europeenne
RLG	Research Libraries Group
STN	Science and Technology Network
TGB	Tres Grande Bibliotheque

TREC Text Retrieval Experiment Conference, sponsored by DARPA and coordinated by NIST, in conjunction with the TIPSTER initiative for routine and ad hoc retrieval from large (e.g., million document) text databases.

WAIS Wide Area Information Server

WWW World-Wide Web

Terms from "Cyberpunk", Time Magazine

From "Cyberpunk", *Time*, February 8, 1993 pages 59-65

Cybernetics - Norbert Wiener of MIT was designing systems for World War II antiaircraft guns when he realized the critical component in a control system, whether animal or mechanical, is a feedback loop that gives a controller information on the results of its actions. He called the study of these control systems cybernetics (from kybernetes, the Greek word for helmsman) and helped pave the way for the electronic brains that we call computers.

Punk - Cyberculture borrows heavily from the rebellious attitude of punk music, sharing with such groups as the Sex Pistols a defiance of mainstream culture and an urge to turn modern technology against itself.

Brain Implants - Slip a microchip into snug contact with your gray matter (a.k.a. wetware) and suddenly gain instant fluency in a foreign language or arcane subject.

Virtual Reality - An interactive technology that creates an illusion, still crude rather than convincing, of being immersed in an artificial world. The user generally dons a computerized glove and a head-mounted display equipped with a TV screen for each eye. Now available as an arcade game.

Internet - The successor of an experimental network build by the U.S. Defense Department in the 1960's, the Internet links at least 3 million computers, many of them university and research related, around the world. Users can connect to the Internet by phone to share information or tap into data banks.

Virtual Communities - Collections of like-minded people who meet on-line and share ideas on everything from politics to punk rock. The global village is full of tiny electronic subdivisions made up of cold-fusion physicists, white supremacists, gerontologists and Grateful Deadheads. Like any other community, each has its own in-jokes, cliques, bozos and bores.

Cyberspace - Science fiction writer William Gibson called it "a consensual hallucination. . . a graphic representation of data abstracted from the bands of every computer in the human system." You can get there simply by picking up the phone.

Computer Virus - The cybernetic analogue of AIDS, these self-replicating programs infect computers and can destroy data. There are hundreds loose in cyberspace, although few are as destructive as the Internet Virus — which is now classified as a "worm" because the writer of the program did not mean to do damage.

The Well - Compared with million-plus-member networks such as CompuServe and Prodigy, the Northern California-based Whold Earth 'Lectronic Link is a tiny outpost in cyberspace. But its 7,000 subsribers include an unusual concentration of artists, activists, journalists and other writers. "It has a regional flavor," says co-founder Stewart Brand. "You can smell the sourdough."

Flame - Sociologists note that without visual cues, people communicating on-line tend to flame: to state their views more heatedly then they would face to face.

Neuromancer - William Gibson's novel is the first to win SF's triple crown — the Hugo Nebula and Philip K. Dick awards — quickly became a cyberpunk classic, attracting an audience beyond the world of SF. Critics were intrigued by a dense, technopoetic prose style that invites comparisons to Hammett, Burroughs and Pynchon. Computer literate readers were drawn by Gibson's nightmarish depictions of an imaginary world disturbingly similar to the one they inhabit.

Turing - British mathematician Alan Turing predicted in 1950 that computers would someday be as intelligent as humans.

Artificial Life - Inspired by the behavior of computer viruses, scientists are wondering how sophisticated a computer program or robot would have to be before you could say it was "alive." One computer-software company, Maxis, has marketed a whole line of simulated anumals, and colonies, cities, train systems and even a planet-like organism called Gaia.

Cryonics - For a price, a terminally ill patient can be frozen — as in the new movie Forever Young — until some future time when a cure has been discovered. Some people save on storage costs by having just their head frozen.

Synaesthesia - From the Greek syn (union) and aesthesia (sensation), synesthesia is a merging of sensory input in which sounds appear as colors in the brain or words evoke a specific taste or smell.

A

B

D

H

O

P

S

V

W