Information Retrieval in Digital Libraries: Bringing Search to the Net

Bruce R. Schatz

A digital library enables users to interact effectively with information distributed across a network. These network information systems support search and display of items from organized collections. In the historical evolution of digital libraries, the mechanisms for retrieval of scientific literature have been particularly important. Grand visions in 1960 led first to the development of text search, from bibliographic databases to full-text retrieval. Next, research prototypes catalyzed the rise of document search, from multimedia browsing across local-area networks to distributed search on the Internet. By 2010, the visions will be realized, with concept search enabling semantic retrieval across large collections.

Immediate access to all scientific literature has long been a dream of scientists. The network information systems needed to support such access have steadily improved as the underlying computing and communications infrastructure has improved. The recent advent of World Wide Web searchers and digital libraries has rekindled popular interest in these issues. However, the problems and components have remained relatively unchanged since the early days of information retrieval. Thus, understanding the evolution of network search technology will place these systems in their proper historical context and aid in understanding their future.

Organized collections of scientific materials are traditionally called "libraries," and the searchable online versions of these are called "digital libraries" (*1*). The primary purpose of digital libraries is to enable searching of electronic collections distributed across networks, rather than merely creating electronic repositories from digitized physical materials. Traditionally, information retrieval has been a task for professional librarians. Trained reference librarians interact with online services of specialized materials and report results to querying scientists. Although public computer networks have long been used to access specialized information services, it has taken the recent rise of the Internet to make literature searching directly available to widespread groups of scientists.

Since the beginnings of online information retrieval more than 30 years ago, the base functionality has remained essentially unchanged. A collection of literature is maintained and indexed, which the user accesses by means of a terminal connected to a server across a network. The user specifies a query by a set of words, and all documents in the collection that contain those words are returned. The fundamental technology for searching large collections is finally changing, so that information retrieval in the next century will be far more semantic than syntactic, searching concepts rather than words (Fig. 1).

Although the software has remained largely unchanged, the hardware has improved dramatically as computers and networks have become faster and cheaper. As a result, the "document" has changed from a citation with descriptive headers to the abstract to the complete multimedia contents, including text, figures, tables, equations, and data. Similarly, the size of organization able to serve a collection to the scientific community has decreased, so that there are now hundreds of thousands of servers distributed around the world, instead of a few hundred at central sites.

Today, the online information retrieval available with Internet Web searchers enables interaction with information sources distributed across the international network. The functionality has changed dramatically from the previous generation of text abstracts retrieved using special purpose public terminals from a single large central computer center to the present generation of multimedia documents retrieved using general purpose personal computers from multiple small distributed file servers. The primary users have correspondingly changed from librarians to scientists. This trend will continue when semantic retrieval makes interactive analysis of digital libraries a fundamental part of scientific research.

Grand Visions

There have been many attempts by writers of science and speculative fiction to describe the universal encyclopedia (*2*) or universal library (*3*). The modern technological era, however, is popularly considered to have begun with a visionary article by Vannevar Bush published in 1945, just as World War II was ending (*4*). The influence of this article owes as much to Bush's fame at the time (he had been director of the Office of Scientific Research and Development, coordinating all U.S. technology efforts during the war) as to the actual article itself.

His vision for what scientists should concentrate on in the postwar era was to build systems for information manipulation. The article discussed the importance of technology to help manipulate all the world's knowledge, although the Memex proposed by Bush concentrated on local manipulation capabilities. A user could navigate a trail through materials of different types from different sources and record this trail for later use. Although many information scientists point to this article as the original vision for the online information retrieval system (*5*), the Memex article actually had no discussion of search or networks; thus, it is more properly the seminal work on hypermedia systems for personal computers (*6*).

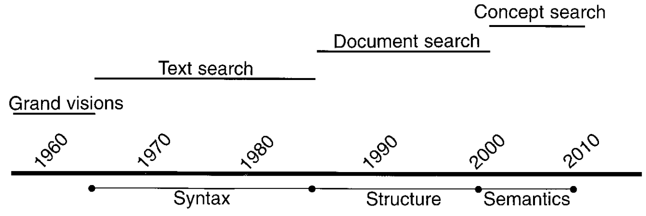


Fig. 1. Rough timeline of the generations of information retrieval in digital libraries. The technology has improved over time, and the generation time is getting shorter as progress in this area speeds up (from 20 to 15 to 10 years for a major generation of functionality, such as syntax to structure to semantics). There is a typical progression within a generation from research prototype to experimental testbed to commercial service.

The author is the Director of the Digital Library Research Program in the University Library and the Research Scientist for digital libraries and information systems at the National Center for Supercomputing Applications, Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. E-mail: schatz@uiuc.edu, http://csl.ncsa.uiuc.edu

A more accurate beginning to the modern era of network information systems is the study carried out in 1961 and 1962 by Licklider. Under the auspices of the Council on Library Resources, he published a report entitled "Libraries of the Future" (*7*), which laid out the research agenda for digital libraries, with surprisingly prescient details, and discussed the extant research systems. The time was ripe for a concrete vision because the mainframe computer business was thriving and the research laboratories were excited by the promise of the interactive systems possible with the new technologies of minicomputers and oscilloscopes.

Licklider scaled and described the technology for what he termed "procognitive" systems. The scale of all the recorded information considered—namely, the "solid" science and technology literature—was estimated as 10^{13} bits or 1 terabyte [(*7*), p. 15]. This was divided into 100 fields and 1000 subfields, so that a subfield's literature was roughly 1 gigabyte (1 billion characters). Even with the explosion in scientific literature that has occurred in the 35 years since, this number is a good estimate of its scale and, accordingly, a good estimate of the existing digital library collections (which are still primarily abstracts rather than full articles).

The functionality of procognitive systems envisioned technology still beyond that which became everyday tools for, first, librarians and then, scientists. Licklider was a prominent (auditory) psychologist, and it was already clear that the appropriate man-machine interface moved far past matching words in documents to matching concepts in the user's mind to concepts in the author's recordings (*8*).

Part of moving past text involved dealing with the structure and classification of complete documents. Such procognitive features include user interaction—with chapters and paragraphs, tables and pictures, abstracts and references, reviews and notes, catalogs and thesauri [(*7*), p. 7]. The research prototypes (small-scale) of the 1980s (Fig. 2) and the experimental testbeds (large-scale) of the 1990s (Fig. 3) have finally achieved this level with structured documents and multiple indexes.

Part of moving past text involved dealing with the content and classification of complete subfields. Such procognitive functions were more like knowledge manipulation than information retrieval, in performing searches and doing analysis of multiple types of information across multiple domains of knowledge. At the heart of these functions was switching vocabulary and mapping knowledge across subjects—for example, "Convert all Nyquist diagrams in set A to Bode plots" and "Transfer W. Smith's AJAX simulation to my Experiment C database as simulation 2" [(*7*), p. 31]. The research prototypes of the 1990s are finally approaching this level of functionality, so that the commercial systems of the 2000s will likely perform semantic retrieval with vocabulary switching.

Licklider's book was written in the heady optimism of artificial intelligence at the Massachusetts Institute of Technology (MIT) and surrounds in the early 1960s, when many researchers felt that general-purpose knowledge manipulation would shortly be feasible. This knowledge-based technology proved to be primarily useful in information retrieval for providing specialized rules in specialized domains (*9*). However, many of the fundamental technologies of network information systems were pioneered during this era (*10*).

The same time period also saw great activity in information retrieval research, which, in contrast to the work above, concentrated on statistical analysis of the text in the documents, such as word frequency. This approach has the advantage of relying primarily on the documents themselves and thus is immediately applicable across subject domains. Classic work such as the document vector space clustering of Salton (*11*) defined and tested the algorithms in great detail. The difficulty in mass realization was that the computers of the 1960s could only run these algorithms on a few hundred documents. Only now, after 35 years, can the supercomputers of today begin to effectively simulate procognitive systems at the scale of scientific literature using statistical information retrieval.

Text Search: Bibliographic Databases

The first attempts at realization of the grand visions, during the 1960s, centered around text search of technical citations. The content was the text of a bibliographic citation of a journal article, which included the title, author, journal, and keywords of the referenced article. A search query was matching specified words to words in the fields of the citation.

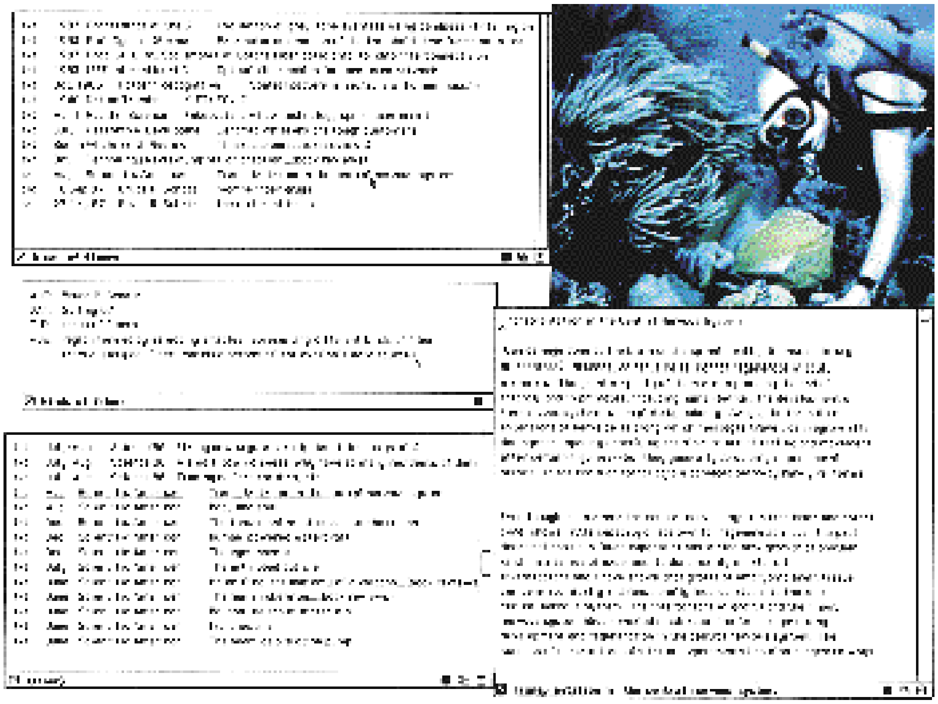


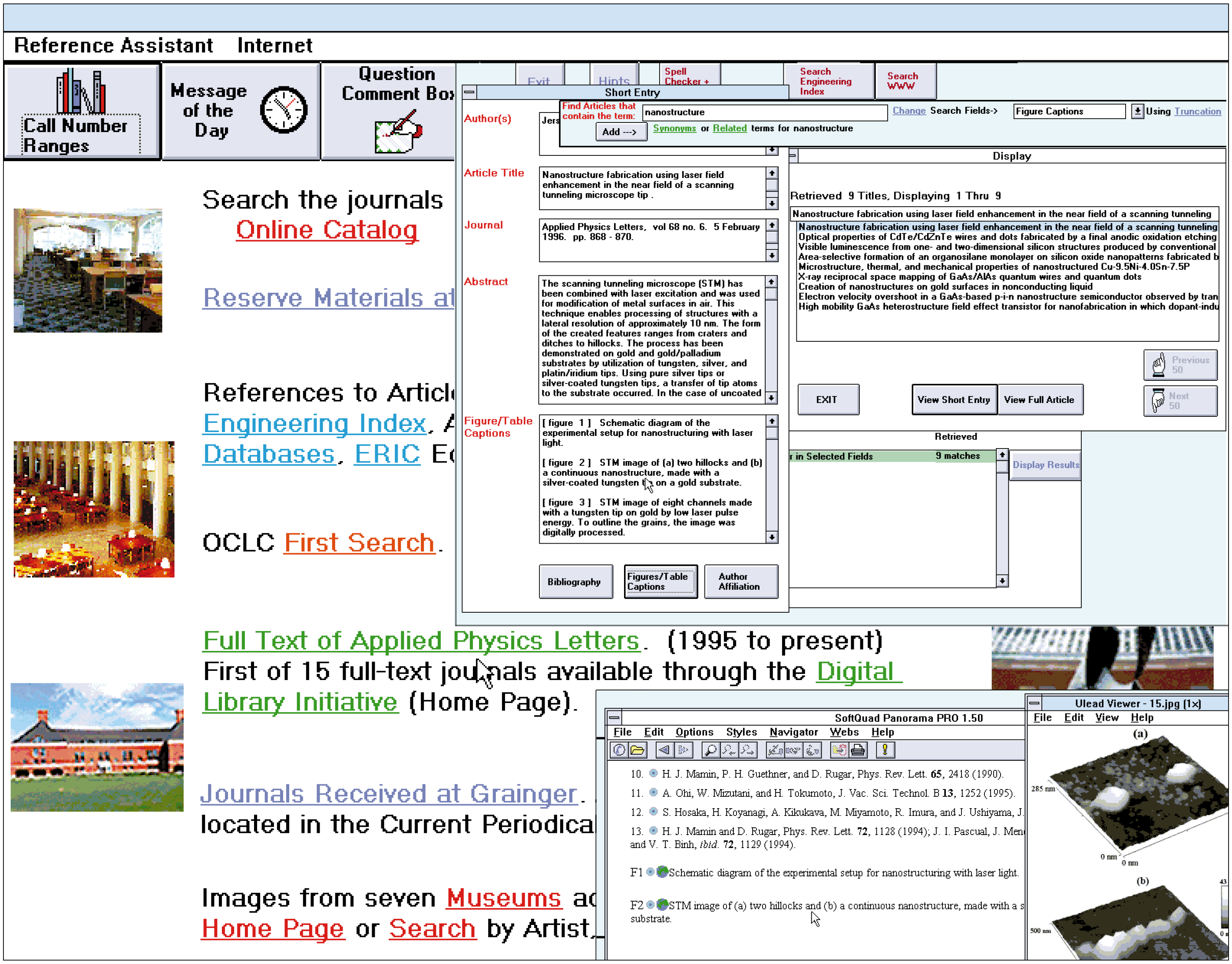
Fig. 2. Telesophy project (1986): a research prototype of the 1980s. The user has issued a broad query (lower left) for "fiber," with the goal of gathering instances of different kinds of fibers. This query was sent out to all the sources within the information space, and the full texts of documents with matching text were returned. After scrolling through the results, the user has located a desired magazine article, which appears from the title to discuss nerve fibers, and zoomed into it. The full text appears in its own window (lower right). A related image also appears (upper right); this picture was linked to the article, and the image displayer was automatically invoked when the link was followed. The window at the upper left illustrates grouping and sharing: It is a region of pointers to located objects, which were copied into the region after being located during a search. The items returned include journal abstracts, magazine articles, and movie reviews. The picture is also an object in the system and thus can be contained in regions, as can the note in the middle left that was entered on the fly. This dynamically created region can be saved and later retrieved by searching on its title (the picture similarly had all text associated with it automatically indexed). All of the pointers in the region are live.

The hardware capability of the central mainframes of this period largely determined the functionality of the information retrieval system. Disk space mandated the collection of citations rather than the complete text of articles. The output device was a paper teletypewriter, mandating a short display, such as a title or citation. The network was a telephone line, used as one of the first packet-switched networks. Thus, the interface goal was to enable the specification of a precise query to retrieve a particular set of items from the citation database and return them to be printed on the teletypewriter terminal. The items were not yet actual documents but pointers to physical documents.

These systems were well suited to generating a bibliography—for example, exhaustively printing all articles that contained the keywords "information retrieval" and "computer networks" for the references of a paper. Although the systems were intended for searching and locating desired items, their slow speed and precise queries limited their effectiveness in browsing, and their primary users were professional librarians generating bibliographies for scientists.

The collections of citations that the online systems handled became known as "bibliographic databases." Generated by the abstracting and indexing industry, bibliographic databases were extended over time to include searchable abstracts of the articles. These databases—such as MEDLINE in biology and medicine or Inspec in electrical engineering and computer science—are still the primary coverage for the scientific literature today (*12*).

Prototypes of online systems for searching bibliographic databases were built in research laboratories in the early 1960s, and the most successful of these evolved through the large-scale experimental testbed phase into commercial services (*13*). For example, the RECON system developed by Summit at Lockheed Palo Alto Research Laboratory for NASA was in the research prototype phase in 1965 (*14*), became an experimental testbed at Lockheed using the ERIC (education) database in 1969, and had become the Dialog online system by 1972. Dialog Information Systems, now owned by Knight-Ridder, is still the most



Search the journals

Online Catalog

Reserve Materials at

References to Articl

Engineering Index, A

Databases, ERIC E

OCLC First Search.

Full Text of Applied Physics Letters. (1995 to present)

First of 15 full-text journals available through the Digital Library Initiative (Home Page).

Journals Received at Grainger.

located in the Current Periodica

Images from seven Museums ac

Home Page or Search by Artist

Find Articles that contain the term: nanostructure

Change Search Fields-> Figure Captions Using Truncation

Add --->

Synonyms or Related terms for nanostructure

Author(s): Jers

Article Title: Nanostructure fabrication using laser field enhancement in the near field of a scanning tunneling microscope tip .

Journal: Applied Physics Letters, vol 68 no. 6. 5 February 1996. pp. 868 - 870.

Abstract: The scanning tunneling microscope (STM) has been combined with laser excitation and was used for modification of metal surfaces in air. This technique enables processing of structures with a lateral resolution of approximately 10 nm. The form of the created features ranges from craters and ditches to hillocks. The process has been demonstrated on gold and gold/palladium substrates by utilization of tungsten, silver, and platin/iridium tips. Using pure silver tips or silver-coated tungsten tips, a transfer of tip atoms to the substrate occurred. In the case of uncoated

Figure/Table Captions:

[figure 1] Schematic diagram of the experimental setup for nanostructuring with laser light.

[figure 2] STM image of (a) two hillocks and (b) a continuous nanostructure, made with a silver-coated tungsten t on a gold substrate.

[figure 3] STM image of eight channels made with a tungsten tip on gold by low laser pulse energy. To outline the grains, the image was digitally processed.

Retrieved 9 Titles, Displaying 1 Thru 9

Nanostructure fabrication using laser field enhancement in the near field of a scanning tunneling

Nanostructure fabrication using laser field enhancement in the near field of a scanning tunneling
Optical properties of CdTe/CdZnTe wires and dots fabricated by a final anodic oxidation etching
Visible luminescence from one- and two-dimensional silicon structures produced by conventional
Area-selective formation of an organosilane monolayer on silicon oxide nanopatterns fabricated b
Microstructure, thermal, and mechanical properties of nanostructured Cu-9.5Ni-4.0Sn-7.5P
X-ray reciprocal space mapping of GaAs/AlAs quantum wires and quantum dots
Creation of nanostructures on gold surfaces in nonconducting liquid
Electron velocity overshoot in a GaAs-based p-i-n nanostructure semiconductor observed by tran
High mobility GaAs heterostructure field effect transistor for nanofabrication in which dopant-indu

Retrieved

r in Selected Fields 9 matches

10. H. J. Mamin, P. H. Guethner, and D. Rugar, Phys. Rev. Lett. **65**, 2418 (1990).

11. A. Ohi, W. Mizutani, and H. Tokumoto, J. Vac. Sci. Technol. B **13**, 1252 (1995).

12. S. Hosaka, H. Koyanagi, A. Kikukava, M. Miyamoto, R. Imura, and J. Ushiyama, J.

13. H. J. Mamin and D. Rugar, Phys. Rev. Lett. **72**, 1128 (1994); J. I. Pascual, J. Men and V. T. Binh, *ibid.* **72**, 1129 (1994).

F1 Schematic diagram of the experimental setup for nanostructuring with laser light.

F2 STM image of (a) two hillocks and (b) a continuous nanostructure, made with a s substrate.

Fig. 3. Illinois DLI project (1996): an experimental testbed of the 1990s. A search query (upper right) uses the structure information in the documents to match "nanostructure" only in figure captions. Matching documents are retrieved across the network, and a summary version is displayed. The full article has been displayed (lower right) in a separate SGML viewer (labeled "SoftQuad Panorama"); the window has been scrolled down to display the references and figure captions at the end. Overlaid on the right-end side of this window is the result of following the figure link to display the image. The background of this composite screendump shows through on the left and illustrates the integration with the online services of the Engineering Library at the University of Illinois, for example, the online catalog and bibliographic databases.

widely used online search system in libraries for generalized scientific literature.

Specialized services also arose in this same time period (*15*). For example, COLEX was developed by Systems Development Corporation (SDC) for the Air Force in 1965, and a system based on this experience became available as SDC's ORBIT in 1968. Concurrently, the National Library of Medicine (NLM) developed a collection called MEDLINE containing citations to medical literature within a local batch system. In 1968, NLM began an experimental online retrieval system called AIM-TWX with a subset of MEDLINE as the document collection. The search retrieval, using software called ELHILL, was a version of ORBIT developed by Cuadra at SDC for NLM. By 1971, ELHILL had evolved into the nationally available MEDLARS system, accessible across public packet-switched networks, which contains MEDLINE and now many other databases. MEDLARS is the most widely used specialized online service (*16*).

Text Search: Retrieval Technology

The basic technology for searching bibliographic databases is still the primary method for large-scale information retrieval; when there is a large collection (a million documents) to be searched, the retrieval methods used today are those developed for bibliographic databases 30 years ago. These text retrieval methods rely on indexing the documents so that selected items can be quickly retrieved (*17*). A user sends a query from his terminal across the network to a server. Software at the server searches the index, locates the documents matching the query, and returns these documents to the user terminal.

To support retrieval by word matching (find all documents containing the word "fiber"), an inverted-list index is built. The documents are scanned for words, omitting a few noise words (such as "the" and "of"), and a list is built for every word. These lists are called "inverted" because for each word they contain pointers to the documents that contain that word. The index consists of the inverted lists in alphabetical order by word. It can be used for fast search of a specified word by scanning the index for that word and then using the attached document pointers to retrieve the matching documents. This word-matching search often uses word stemming to increase its retrieval effectiveness: Words are shrunk to a canonical form, so that, for example, "comput" represents "computer," "computers," and "computing." If multiple words are specified, the resulting sets of documents can be merged (logical AND results in an intersection; OR yields the union).

As computers became more powerful, the scale of documents for information retrieval became greater. That is, as it became technologically and economically feasible to provide faster networks and larger disks, it became possible to store and retrieve more than just a citation. First, the abstract was added, and this is the economic level that today remains the standard for scientific literature. Then, video terminals became the mode of display, so that text could be viewed more rapidly than with teletypewriters. This led to the extension from abstracts to so-called "full text." An online full-text article contains all words within an article but excludes nontextual materials such as figures, tables, and equations.

The searching technology also increased in scope while staying fundamentally the same in function. Because there was now a full article instead of an abstract (20 versus 2 kilobytes), there were more words per document. Individual words thus became less discriminating in searches, and phrases became more useful. Internally, this change in focus implied that Boolean operators became less useful (for example, finding "fiber" and "optics" anywhere in the same document often happens coincidentally), whereas proximity operators became more useful (for example, "fiber" within two words of "optics" finds such intended phrases as "fiber network optics").

To implement these new proximity operators, additional information was needed in the index. An inverted-list index contains all of the words along with pointers to all documents containing each word. To compute proximity, the word position within the document is also specified. When proximity search is desired, the modified word lists can be intersected as they were for Boolean ANDs, followed by comparison of the word positions within the same document.

Full-text retrieval was driven by demands in the professions, particularly law. Bibliographic retrieval was pioneered in medicine, where it might be argued that abstracts were satisfactory for identifying the content of an article. This was less so in law, and the Lexis system of U.S. court records provided in 1973 the first large-scale commercial system demonstrating the practicality of full-text documents. Mead Data Central extended this service to Nexis, which includes the full text of large-circulation magazines and newspapers. Today, full text is common for the majority of popular materials.

Document Search: Multimedia Browsing

By the 1980s, full-text search had become established commercially in online retrieval systems. This same era saw the initial deployment of bitmapped personal workstations and local-area networks in research laboratories. This technology made it possible to provide new functionality to the established ideas of text search, most notably in the areas of multimedia documents and distributed browsing.

As the computer model changed from central shared mainframes to distributed personal workstations, it profoundly changed information retrieval from text search to document search. As the research workstations of the 1980s turned into the personal computers of the 1990s and Internet access became widespread, the research systems of the 1980s based on full-text technology became the Internet services of the 1990s. Thus, full-text search coupled with multimedia browsing is today available to average scientists for their everyday needs.

The increased speed of both the workstations and the networks brought an expansion in both the basic document and the basic retrieval. Multimedia slowly became possible, so that pictorial materials, such as graphics, images, and videos, could be included in the documents and accessed from collections across the network. For example, interactive display of color pictures from remote sources became technologically feasible.

The increased speed across the network meant that multiple sources could be searched within a single query while still maintaining effective user interaction for the return of results. Multiple collections could be stored in physically distributed locations, yet searched as a single, logically coherent collection. This was an interactive realization of the transparent information gateway technology pioneered in the 1970s (*18*) and commercialized in the 1980s (*19*). In the new computer environment, the network speeds enabled federation across sources to be done dynamically.

More profoundly, a different style of interaction became possible with the increased speeds. Rather than search, where a detailed query is made and comprehensive results returned, browsing enables broad queries to be used to quickly scan for appropriate sections of a digital library. This style resembles using the card catalog to locate a particular section of a physical library, then browsing those shelves in search of suitable materials. The underlying search mechanism is the same—full-text proximity—but any results returned can be scanned much more quickly. This approach changes the character of the interaction from an exact database search to a loose navigation attempting to identify a desired set of materials. When distributed items are linked together, the navigation takes on the character of "hopping" from one document to another.

The style of multimedia browsing combined with distributed search is the main theme of Internet information services today. Its historical antecedents are in the research systems of the previous decade. A premier example is the Telesophy system (Fig. 2), a research prototype designed and built by Schatz at Bellcore in the mid-1980s (*20*). Within the then-small community of Internet insiders, Telesophy was regarded as the forerunner of the future Net of worldwide information spaces (*21*), which seemed far in the future at the end of the project in 1989, even though its mass realization proved to be only 5 years away.

Telesophy literally means "wisdom at a distance." The goal was to make the manipulation of knowledge as easy and transparent as telephony made the transmission of sound (*22*). The prototype was built (1985 to 1986) using the relatively new personal bitmapped workstations, local-area networks, and custom full-text search software (*23*). It enabled many sources of information to be searched across the network and then grouped into units of knowledge. From the retrieval standpoint, multimedia multisource information retrieval was supported. That is, a user could issue a query from his workstation, this query would be propagated to all of the sources on the network, and the results propagated back to the workstation and merged for display. In the prototype, there were some 30 information sources with different search servers on different machines in the network, with a range spanning bibliographic databases, full-text databases, book catalogs, still images, line graphics, and video clips. For creating knowledge, edited versions of query results could be stored for later retrieval, or links between items could be added on the fly.

Document Search: Distributed Technology

The interaction style and system architecture pioneered in the Telesophy system are illustrative of that available on the Internet today. The interaction style combined both searching and browsing, combining the full-text search of the previous generation computer model with the interactive navigation now possible with the present generation. The system architecture supported transparent access to distributed sources, for federated search and for interactive navigation, with the prototype implementation providing careful optimization for fast response across wide-area networks.

The interaction style used search to retrieve a broad selection of relevant items and then browsing to navigate from retrieved items to related ones. The analogy was to go to a section of a library that contained relevant materials related to each other, scan the titles on the spines of the books to locate desired ones, then open a few books to get pointers for which sections to search for next. The digital library, however, contained distributed multimedia documents, so that the system provided transparent network information retrieval. The user did not have to know about different interfaces for different document types because uniform commands were supported on all objects, and the user did not have to know about different accesses to different physical locations because fast access was supported on all sources.

The search used a full-text matching scheme similar to that of the existing online bibliographic systems. The distributed model, however, enabled the search to retrieve a selection of related items from any or all of the information sources. Typically, the results were displayed as one-line summaries, which could be zoomed into, to display the full object for selected items. Component-style type switching enabled different displayers to be invoked for different objects. Thus, different media types—such as text, image, graphics, and video—could all be displayed as appropriate for that type. The speed of the network and the workstation enabled broad queries to be issued, because several hundred items could be quickly fetched and displayed at the summary level.

The browsing was modeled on the few hypermedia systems that had been built by that time, notably NLS (*24*) and the SDMS (*25*). All items in all sources were contained within a single logical information space, which consisted of interlinked information units. An information unit was an object with a standard set of headers, which contained structural information such as author and title, as well as link information to related objects. Thus, any document could be linked to any other document, and these links could be followed at any time.

Information units could contain collections of other units, in addition to encapsulating an external data item. These composite units supported the grouping and sharing of knowledge. A group of units could be formed by creating a region and copying (pointers to) units into it. For example, the results of a query were just a knowledge region, and any item in the region could be zoomed into, thus following links dynamically across the network. Because regions were themselves information units, the summaries and the links were the same as those for documents (*26*).

A typical session with the Telesophy prototype involved retrieving and grouping multimedia items from distributed sources (Fig. 2). A demonstration would include searching across sources of different types, saving a selection of retrieved information into a knowledge region, and then retrieving the selected units again by locating the region and zooming into them. Retrieving a saved multimedia search item in real time by following a live link across a wide-area network was quite novel in 1986.

The architecture of the Telesophy prototype provided for distributed federation. Any or all sources could be searched with a single query, and the results from the various servers were merged dynamically. There were distributed servers for the indexes and for the objects (information units), which could be arbitrarily combined. Thus, an object could be in multiple indexes and an index could reference multiple sources; this organization enabled queries and regions to be handled uniformly. All information units had a canonical set of fields, which supported metadata, such as author and title, in addition to links. A significant attempt was made to map all sources into the canonical set, creating such difficulties as determining who is the author of a movie: the writer, director, producer, or star. This level of structural federation enabled uniform queries to be issued for search, and uniform summaries to be generated for display of query results, even for nontextual materials such as images and videos.

The implementation of the Telesophy prototype demonstrated its scalability. Some 300,000 items from some 30 sources were placed into the information space, and each data item was transformed into a uniform information unit. The object framework, inspired by the language Smalltalk (*27*), could handle dynamic types and user-defined types; for example, bulletin boards were handled by doing a query on-the-fly, and electronic mail was automatically imported so that it could be searched. Careful optimization of the caching and other internal features of the prototype enabled the speed to approximate that of a physical library across a local-area network and between company buildings at different sites across a wide-area network. The new commercial hardware enabled the new research software to demonstrate that browsing multimedia sources distributed across a network was technically feasible and a style of interaction complementary to search of a central collection.

Search on the Internet

Throughout the 1990s, search has finally come to the Internet. Today, Web searchers are reaching functionality equivalent to that of the commercial online systems of 25 years ago, but with many more users (streamlined user interface and a wider range of materials). The driving force behind these re-inventions has been the information superhighway, as evidenced by the huge increase in the number of users of the Internet (growing from 1 million to 25 million in the past 5 years) and by the federal initiatives in National Information Infrastructure (NII).

The first software package to significantly bring search capability to the Internet was WAIS (Wide Area Information Server) (*28*). Inspired by the Telesophy system (*29*), this software provided an indexing program and a search engine. Information providers could take their collection and index it for full-text proximity search. A provided client could then be used to interface with the search engine server, enabling users to access the collection over the Internet.

When Gopher, the first widely used browser, became available, it quickly replaced the native WAIS client. Searchable archives then appeared on the Internet; a user could traverse a Gopher menu to a file representing a collection, then issue a query request that would be sent to the appropriate WAIS search engine. Several hundred searchable collections appeared in the early 1990s, including many in molecular biology as part of the genome projects.

As the Gopher technology was displaced by the World Wide Web, which could handle hyperlinks within documents, browsers became even more widely available. The NCSA (National Center for Supercomputing Applications) Mosaic browser in particular made an enormous impact on the scientific community (*30*) and became the first Internet program to make an impact on the general public at large. It was inspired by the Telesophy system, as one of several attempts by NCSA to make the functionality of Telesophy widely available to the scientific community (*31*). A standard part of the Mosaic interface was a search query, which could be linked via a gateway to a variety of search engines. Although WAIS still predominated, other information retrieval (Z39.50) and database retrieval (SQL) gateways appeared.

The number of information sources on the Web has grown astronomically in the 3 years since the introduction of Mosaic. Although some of these sources are indexed, most are merely unorganized collections of documents. This has created an enormous problem of locating desired documents on the Net. The first wave of solutions has already led to the creation of major information services on the Internet and spawned a rapidly growing commercial industry. These services have reproduced the evolution of the online services at a greatly accelerated pace.

For example, Lycos (*32*), one of the first major Web searchers, is much like a bibliographic database service, except that the abstracts are generated by a program, called a Web crawler, rather than by a human indexer. The collected abstracts are full text indexed and served from a computer center of file servers, similar to the architecture of Dialog. The rapid evolution of the Web has even made the transition to indexing the full text of documents already—for example, Alta Vista (*33*).

Better search requires better organization. The difficulty with search on the Web at present is that HTML (hypertext markup language) documents are largely unstructured, and HTTP servers just point to files containing these documents. Good-quality professional search requires repositories, which are organized collections of objects with indexes that support search and viewers that support display. Handling distributed repositories has become perhaps the major issue in digital library research (*34*). The question is how to record the structure of the objects in the repositories and to use this structure to guide federated search.

Documents with the same level of structure as the scientific literature are just beginning to appear in the Net. For example, the National Science Foundation (NSF)–Advanced Research Projects Agency (ARPA)–NASA Digital Library Initiative (DLI) is considered to be the flagship research effort of the federal NII program (*35*), and one of the DLI projects is concentrating specifically on scientific literature. The University of Illinois project is constructing a large-scale testbed with tens of thousands of documents from journals and magazines in science and engineering, in a production stream direct from major publishers, for thousands of users distributed around the Big Ten midwestern universities (*36*).

The structure of the documents is marked up in SGML (standard generalized markup language) (*37*), which specifies tags that mark the subparts, including full text, sections, figures, tables, equations, references, and abstracts. Federation of queries across sources is accomplished by a canonical set of metadata and tags, much as in the Telesophy system. Because the project is based in a major engineering library, the SGML repository search is integrated with other library services such as catalogs and thesauri. A typical session in the Illinois DLI testbed searches and displays structured documents from distributed repositories of scientific literature (Fig. 3). Its functionality in the late 1990s will be the first generation of experimental testbeds to approximate information retrieval on the structure of complete documents as envisioned by Licklider in the early 1960s.

Bringing search to the Net will require the development of server software with complete packaging, much as the development of client software with complete packaging brought browsing to the Net. The evolution from research to the Net to commercial products will be repeated for information search of distributed repositories in the next 5 years, much as it occurred in the last 5 years for information browsing of distributed documents. The multimedia information retrieval available in the Telesophy system in 1986 is thus a good indicator of the functionality that will be available in 2001, after the digital library technology brings search to the Net.

Toward Concept Search

Issuing a structured search transparently to distributed repositories across the Net is close to the level of syntactic functionality envisioned by Licklider in the procognitive systems, but it is far from the level of semantic functionality that would allow automatic translation of terminology across subfields. The current approach to semantic translation involves human experts serving as intermediaries. For example, indexers in the abstracting and indexing industry are special librarians who tag every document with terms chosen from a subject thesaurus. These terms are often not in the document; for example, an article that mentions only the term "Unix" would be tagged as being about "operating systems," so that a search of the index terms will be a better approximation of the concepts in the documents. Such metadata for databases provides a more conceptual classification for search purposes. Similarly, when scientists need to search across subject domains into unfamiliar areas, a human intermediary such as a reference librarian can often translate the terms in one subject area into similar terms within another.

"Vocabulary switching" is the name within information science to describe this problem (*38*). A user wishes to specify items (phrases within documents) using their vocabulary but search the repository (documents within collections) of another subject with another vocabulary. The different domains contain similar concepts described with different terminology. A system that performed vocabulary switching would automatically translate terms across domains.

This translation enables scientists to find information effectively outside of their specialties. Vocabulary switching is an important part of what the "official" report on the Digital Library Research Agenda called the "grand challenge of digital libraries," semantic interoperability (*39*).

Large-scale simulations on supercomputers have indicated a crack in the semantic barrier (*40*). Using a week of dedicated computer time on the HP Convex Exemplar at NCSA (and 10 days of CPU time overall), Chen, Schatz, and colleagues generated concept spaces for 10,000,000 journal abstracts across 1000 subject areas covering all of engineering and science (*41*) (Fig. 4). Concept spaces are generated by statistical text analysis techniques that operate independently of subject matter. Earlier research using molecular biology literature, by the same researchers, had shown concept spaces to be effective for interactive term suggestion and vocabulary switching (*42*).

As with the current wave of distributed multimedia browsing, better computer technology implementing the same retrieval technology implies a new wave of functionality. The statistical techniques used for concept spaces, such as term co-occurrence, are standard algorithms from information science research of the 1960s (*43*), and vocabulary switching systems are a research topic from the 1970s (*44*). With the machines of the 1990s being a million times faster, these techniques have become feasible on real collections.

This vocabulary switching computation is the largest ever in information science, but it is just a forerunner of the routine computations in the foreseeable future, given the rapid evolution of processors and parallelism. It is also notable that the scale of the collection (scientific literature) is roughly the same as that envisioned by Licklider (*45*), in addition to the functionality (semantic retrieval) being roughly the same. So the computation represents the first concrete step in scalable semantics required for the realization of the procognitive systems envisioned 35 years earlier.

Automatic indexing with scalable semantics will be necessary in the world of a billion repositories in the next century. These indexed collections will move beyond the subjects and quality of large professional repositories (medicine and engineering) into small community repositories (worms and Smalltalk) and into personal repositories (your documents and e-mail). Concept spaces and vocabulary switching will need to be part of the fundamental infrastructure if digital libraries are to support correlations between information sources at all of these levels.

The Interspace is the world of the next century, where the visions of 1960 will finally be realized after 50 years (*46*). It will again take 10 to 15 years for the research prototypes (*47*), just now demonstrating the first realizations of the procognitive systems envisioned by Licklider in 1960, to reach widespread commercial usage, perhaps by 2010. The first major revolution of the Net Millennium will come when the information infrastructure supports routine vocabulary switching. Then scientists will be able to break the bondage of their narrow specialties and effectively utilize the whole of scientific information in their research.

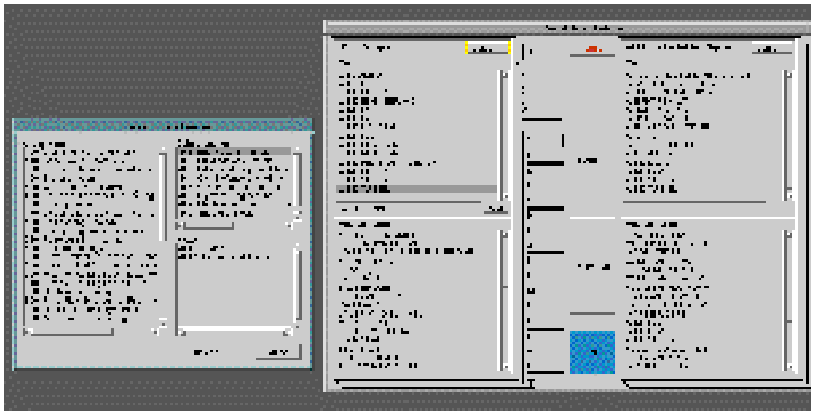


Fig. 4. Interspace project (1996): a research prototype of the 1990s. A sample screen from the vocabulary switching experiment comprising 1000 community repositories with 10 million abstracts across all of science and engineering (computed on 5 years of Compendex and Inspec data using 10 days of supercomputer time on the NCSA HP Convex Exemplar in the spring of 1996). The classification window (left) is the human-indexer categorization showing the subject hierarchy down to "401.1 Bridges." The switching window (right) shows the use of concept spaces for vocabulary switching across community repositories. When a common term occurs across repositories, the system can automatically switch across spaces. The example illustrates a civil engineer who is designing a bridge and is interested in searching ocean engineer literature for similar effects of fluid dynamics on long structures. "Wind Tunnel" is a term common between the repositories for "Bridges" and "Marine Drilling Rigs." The list of related terms on structural stability and fluid dynamics is different because the co-occurrence frequency is different within the two collections. The user can match corresponding terms across spaces, such as "suspension bridges" versus "compliant towers." This match is illustrative of the underlying semantics that suspension bridges have steel cables tethered to prevent swaying in the air, whereas compliant towers are marine rigs with a floating platform tethered to the ocean floor by steel cables that steady the oil drill from swaying in the water.

REFERENCES AND NOTES

1. The term "digital library" has recently displaced the more traditional "electronic library" [E. Fox *et al.*, *Commun. ACM* **38**, 23 (April 1995)]. The advent of computer networks built upon optical fibers made the term "electronic" seem inappropriate because the fibers carry light, not electricity; however, the term "digital" sometimes has the unfortunate connotation of "digitization." There are many aspects of digital libraries that are important to complete systems, such as intellectual property and permanent archiving issues. The discussion here concentrates on the search-and-display issues most relevant to bringing search to the Net.
2. H. G. Wells, *World Brain* (Methuen, London, 1938).
3. J. L. Borges, "The Library of Babel," reprinted in *Labyrinths: Selected Stories and Other Writings* (New Directions, New York, 1964), pp. 51–58.
4. V. Bush, *Atl. Mon.* **176**, 101 (July 1945) [reprinted in (*48*)].
5. M. E. Lesk, *The Seven Ages of Information Retrieval*, in *As We May Think: A 50th Anniversary Celebration of Bush's Vision*, MIT, October 1995. Available at http://www-eecs.mit.edu/AY95-96/events/bush/index.html
6. T. H. Nelson, "As We Will Think," in *On-line 72 Conference Proceedings* (1972), vol. 1, pp. 439–454 [reprinted in (*48*)].
7. J. C. R. Licklider, *Libraries of the Future* (MIT Press, Cambridge, MA, 1965).
8. J. C. R. Licklider and W. E. Clark, *Proc. Am. Fed. Inf. Processing Soc.* **21**, 113 (1962).
9. Y. Wilks, Ed., special issue on Natural Language Processing, *Commun. ACM* **39** (January 1996).
10. Licklider went on to found the Information Processing Techniques Office (IPTO) at ARPA, where he funded a number of revolutionary systems projects concerned with information retrieval across remote networks. These include the first demonstrations of network information retrieval. For example, Bourne in 1963 demonstrated a prototype online retrieval system at Stanford Research Institute (SRI) operating on a local terminal from a remote computer at SDC (C. Bourne, personal communication based on SRI internal reports). The most notable ARPA IPTO project was NLS (oNLine System) in the 1960s at SRI. D. Engelbart was the visionary behind NLS, which was the first and still is almost the only complete system for navigating information spaces (*24*). All documents were divided into paragraphs, which could be arbitrarily linked and followed to other documents across the network. This system inspired many pioneers of later generations. The most immediate outcome of Licklider's book was project INTREX at MIT [C. Overhage and R. Harman, Eds., *INTREX: Report of a Planning Conference on Information Technology Experiments* (MIT Press, Cambridge, MA, 1965)], which discussed what would today be called community systems with federated repositories. Such systems in the 1960s never developed past the demonstration phase, because of the limitations of hardware and software technology. The realization of personal workstations and local-area networks was pioneered at the Xerox Palo Alto Research Center during the 1970s, including many of the ideas from NLS. The first operational prototypes of community systems were built in the 1980s, after the advent of commercial workstations in research laboratories.

11. G. Salton, Ed., *The SMART Retrieval System: Experiments in Automatic Document Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1971).
12. M. E. Williams, *Science* **228**, 445 (1985).
13. F. W. Lancaster and E. G. Fayen, *Information Retrieval On-Line* (Melville, Los Angeles, 1973).
14. R. K. Summit, in *Proceedings of the 22nd National Conference of the Association on Computing Machinery* (Thompson, 1967), pp. 51–56.
15. C. T. Meadow, *Database* (October 1988), p. 23.
16. D. B. McCarn and J. Leiter, *Science* **181**, 318 (1973).
17. G. Salton and M. McGill, *Introduction to Modern Information Retrieval* (McGraw-Hill, New York, 1983).
18. R. S. Marcus and J. F. Reintjes, *IEEE Trans. Syst. Man Cybern.* **12**, 116 (March–April 1982).
19. M. E. Williams, *J. Am. Soc. Inf. Sci.* **37**, 204 (1986).
20. B. R. Schatz, in *Proceedings of IEEE Globecom †87* (IEEE, New York, November 1987), pp. 1181–1186.
21. For example, as a member of the Internet Research Task Force, I was one of the few members of the generation after the pioneers invited to speak at the 20th Anniversary Symposium for the ARPANET at the University of California at Los Angeles in 1989. My talk, "Telesophy: Towards World-Wide Information Spaces," although full of technical details and projections, seemed grand and futuristic at that point (August 1989).
22. B. R. Schatz, ''Telesophy,'' *Bellcore TM-ARH-002487* (August 1984).
23. ———, in *Proceedings of the 5th IEEE International Conference on Data Engineering* (IEEE, New York, 1989), pp. 188–197.
24. D. C. Engelbart and W. K. English, in *Proceedings of the Fall Joint Computer Conference* (AFIPS Press, New York, 1968), vol. 33, part 1, pp. 395–410.
25. C. F. Herot, *Assoc. Comput. Mach. Trans. Database Syst.* **5**, 493 (1980).
26. An inspiration for knowledge regions was T. Nelson, who designed a grand system called Xanadu to handle all the world's knowledge as a single hyperliterature across multiple collections. His unimplemented treatise, *Literary Machines* (1981), contained many suggestions for building new documents by annotating and linking parts of old.
27. A. Goldberg and D. Robson, *Smalltalk-80: The Language and Its Implementation* (Addison-Wesley, Reading, MA, 1983).
28. B. Kahle *et al.*, *Electron. Networking* **2**, 59 (spring 1992).
29. B. Kahle, personal communication. Kahle developed the WAIS software at Thinking Machines with funding from Apple Computer and later started WAIS Inc., which was purchased by America Online.
30. B. R. Schatz and J. B. Hardin, *Science* **265**, 895 (1994).
31. The two predominant Web browsers are derived from Mosaic: Netscape Navigator was built by the original developers after they left NCSA, and Microsoft's Internet Explorer has at its core a licensed version of Enhanced Mosaic which is produced by Spyglass as the official commercial distributor of NCSA Mosaic. Historically, Telesophy played a role in Mosaic as well, because I have been the scientific advisor for information systems at NCSA since 1989, and Mosaic was one of several attempts at NCSA to reproduce the functionality of Telesophy for the general scientific community.
32. Lycos is a spin-off company from digital library projects at Carnegie-Mellon University. See http://www.lycos.com/
33. Alta Vista was a project, now a service, from Digital Equipment Corporation's Research Laboratories. See http://altavista.digital.com/
34. B. Schatz and H. Chen, *IEEE Comput.* **29**, 22 (May 1996).
35. The May 1996 special issue of *IEEE Computer* contains overview articles from all six DLI projects. See http://www.computer.org/pubs/computer/dli/
36. B. Schatz *et al.*, *Computer* **29**, 28 (May 1996).
37. E. van Herwinjnen, *Practical SGML* (Kluwer, Boston, 1994).
38. F. W. Lancaster, *Vocabulary Control for Information Retrieval* (Information Resources Press, Arlington, VA, 1986).
39. C. Lynch and H. Molina-Garcia, Eds., "Interoperability, Scaling, and the Digital Libraries Research Agenda," 22 August 1995. Available at http://www.hppc.gov/reports/report-nco/reports/iita-dlw/main.html. The Information Infrastructure Technology and Applications (IITA) group is the highest level technical committee of the Federal NII Program.
40. S. Nadis, *Science* **272**, 1419 (1996).
41. H. Chen *et al.*, *IEEE Trans. Pattern Anal. Mach. Intell.* **18**, 771 (1996).
42. H. Chen, J. Martinez, T. Ng, B. Schatz, *J. Am. Soc. Inf. Sci.* **48**, 17 (1997).
43. P. B. Kantor, *Annu. Rev. Inf. Sci. Technol.* **29**, 53 (1994).
44. R. T. Niehoff, *J. Am. Soc. Inf. Sci.* **27**, 3 (1976).
45. The vocabulary switching computation used bibliographic abstracts from Compendex (engineering and science) and Inspec (electrical engineering and computer science). Compendex has 40 broad subject classes (for example, computer science) and 600 class codes total. Inspec is narrower and deeper than Compendex, and the computation included about 150 classes at its highest level, the same as the lowest level of Compendex. Because Inspec has roughly 2500 classes all together, the collection spanned in total about (600/150)2500 = 10,000 community repositories across all of science and engineering. This size is similar to that calculated by Licklider, who stated 100 fields and 1000 subfields, because communities are the next deeper level (for example, Smalltalk is a community within the subfield of programming languages, within the field of computer science). A typical community repository in this computation or in the previous molecular biology computations has 5000 documents, at 20 kilobytes per document for full text. The size of a subfield literature is thus 10 times this, 1 gigabyte, just as computed by Licklider. The vocabulary switching computation thus spanned a representative set of all scientific literature (it used abstracts, not documents, and a sample of communities, so it did not compute the complete literature in toto).
46. B. R. Schatz, "Information Analysis in the Net: The Interspace of the Twenty-First Century", a CIC Forum White Paper for *America in the Age of Information: A Forum*, Committee on Information and Communications (CIC) of the National Science and Technology Council, July 1995. Available at http://www.hpcc.gov/cic/forum/CIC_Cover.html. The CIC is one of nine committees reporting directly to the Science Adviser to the President of the United States.
47. B. R. Schatz, "Building the Interspace," http://csl.ncsa.uiuc.edu/interspace.html
48. J. M. Nyce and P. Kahn, *From Memex to Hypertext: Vannevar Bush and the Mind's Machine* (Academic Press, San Diego, CA, 1991).

49. I thank the members of the DLI project at the University of Illinois in general and the Interspace project in particular, especially H. Chen, K. Powell, and C. Herring. C. Bourne, who was a pioneer in the early days of online information retrieval, carefully reviewed the historical details and suggested many corrections. L. Smith and P. Cochrane also kindly helped with the periods that predated my direct experiences. K. Powell helped with preparation of the figures. Support was provided through NSF-ARPA-NASA DLI grant IRI-94-11318COOP and my NSF Young Investigator award IRI-9257252 in science information systems.

Mathematical and Computational Challenges in Population Biology and Ecosystems Science

Simon A. Levin,* Bryan Grenfell, Alan Hastings,
Alan S. Perelson

Mathematical and computational approaches provide powerful tools in the study of problems in population biology and ecosystems science. The subject has a rich history intertwined with the development of statistics and dynamical systems theory, but recent analytical advances, coupled with the enhanced potential of high-speed computation, have opened up new vistas and presented new challenges. Key challenges involve ways to deal with the collective dynamics of heterogeneous ensembles of individuals, and to scale from small spatial regions to large ones. The central issues—understanding how detail at one scale makes its signature felt at other scales, and how to relate phenomena across scales—cut across scientific disciplines and go to the heart of algorithmic development of approaches to high-speed computation. Examples are given from ecology, genetics, epidemiology, and immunology.

Mathematical and computational approaches to biological questions, a marginal activity a short time ago, are now recognized as providing some of the most powerful tools in learning about nature; such approaches guide empirical work and provide a framework for synthesis and analysis (*1*, *2*). In some areas of biology, such as molecular biology, the advent has been recent but rapid—for example, as an adjunct to the analysis of nucleic acid sequences or the structural analysis of macromolecules. In population biology, in contrast, the marriage between mathematical and empirical approaches has a century-long history, rich in tradition and in the insights it has provided. Statistics and stochastic processes, for example, derive their origins from biological questions, as in Galton's invention of the method of genetic correlations and Fisher's creation of the analysis of variance to study problems in agriculture (*1*). Branching processes were developed to describe genealogical histories, and even such