SEARCHING THE INTERNET

Combining the skills of the librarian and the computer scientist may help organize the anarchy of the Internet

by Clifford Lynch

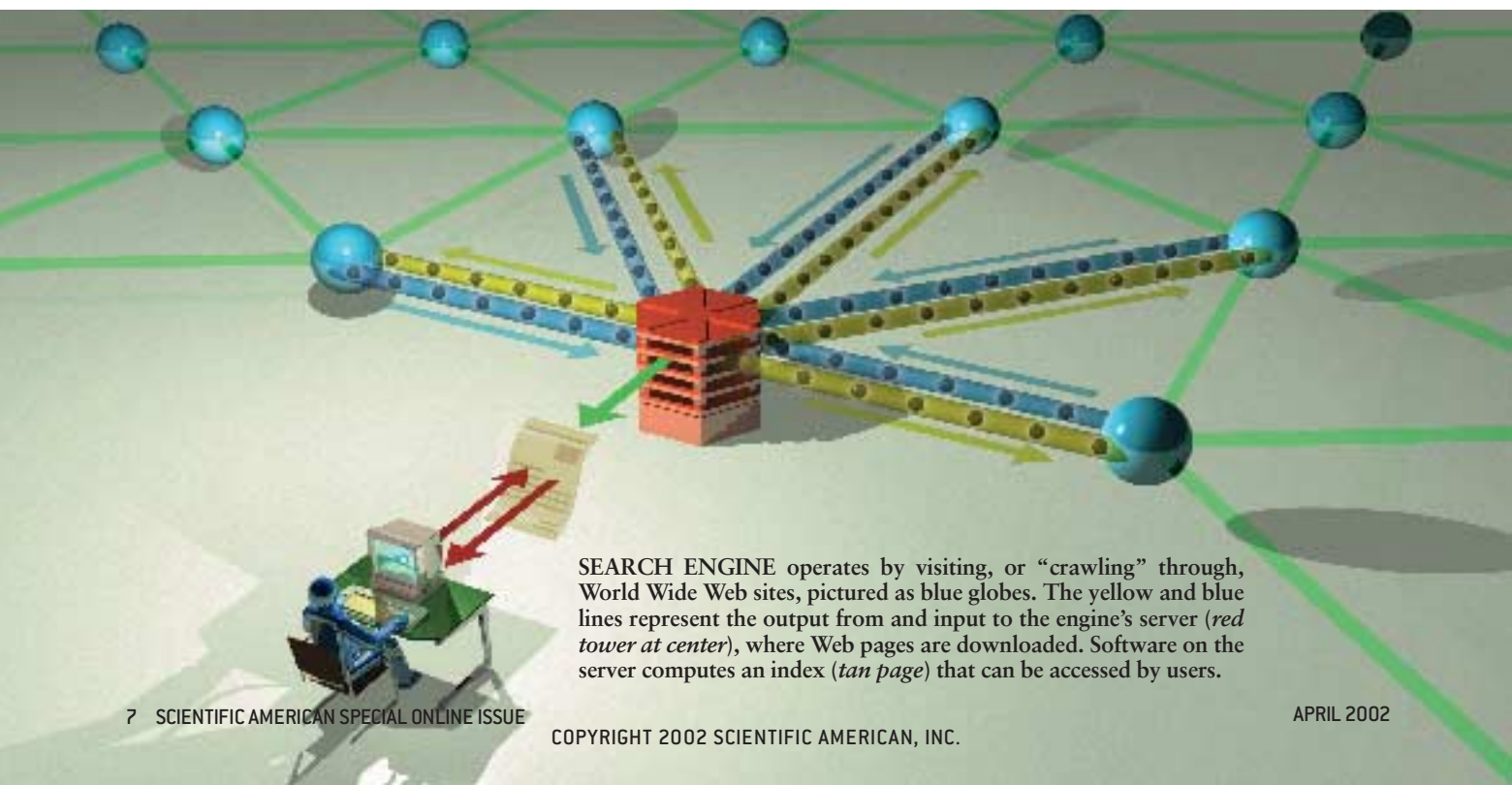
One sometimes hears the Internet characterized as the world's library for the digital age. This description does not stand up under even casual examination. The Internet—and particularly its collection of multimedia resources known as the World Wide Web—was not designed to support the organized publication and retrieval of information, as libraries are. It has evolved into what might be thought of as a chaotic repository for the collective output of the world's digital "printing presses." This storehouse of information contains not only books and papers but raw scientific data, menus, meeting minutes, advertisements, video and audio recordings, and transcripts of interactive conversations. The ephemeral mixes everywhere with works of lasting importance.

In short, the Net is not a digital library. But if it is to continue to grow and thrive as a new means of communication, something very much like traditional library services will be needed to organize, access and preserve networked information. Even then, the Net will not resemble a traditional library, because its contents are more widely dispersed than a standard collection. Consequently, the librarian's classification and selection skills must be complemented by the computer scientist's ability to automate the task of indexing and storing information. Only a synthesis of the differing perspectives brought by both professions will allow this new medium to remain viable.

At the moment, computer technology bears most of the responsibility for organizing information on the Internet. In theory, software that automatically classifies and indexes collections of digital data can address the glut of information on the Net—and the inability of human indexers and bibliographers to cope with it. Automating information access has the advantage of directly exploiting the rapidly dropping costs of computers and avoiding the high expense and delays of human indexing.

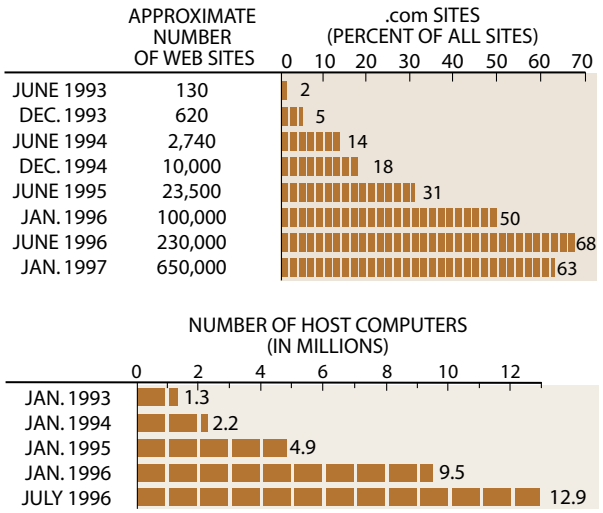
But, as anyone who has ever sought information on the Web knows, these automated tools categorize information differently than people do. In one sense, the job performed by the various indexing and cataloguing tools known as search engines is highly democratic. Machine-based approaches provide uniform

BRYAN CHRISTIE



SEARCH ENGINE operates by visiting, or "crawling" through, World Wide Web sites, pictured as blue globes. The yellow and blue lines represent the output from and input to the engine's server (*red tower at center*), where Web pages are downloaded. Software on the server computes an index (*tan page*) that can be accessed by users.

and equal access to all the information on the Net. In practice, this electronic egalitarianism can prove a mixed blessing. Web "surfers" who type in a search request are often overwhelmed by thousands of responses. The search results frequently contain references to irrelevant Web sites while leaving out others that hold important material.



GROWTH AND CHANGE on the Internet are reflected in the burgeoning number of Web sites, host computers and commercial, or ".com," sites.

Crawling the Web

The nature of electronic indexing can be understood by examining the way Web search engines, such as Lycos or Digital Equipment Corporation's AltaVista, construct indexes and find information requested by a user. Periodically, they dispatch programs (sometimes referred to as Web crawlers, spiders or indexing robots) to every site they can identify on the Web—each site being a set of documents, called pages, that can be accessed over the network. The Web crawlers download and then examine these pages and extract indexing information that can be used to describe them. This process—details of which vary among search engines—may include simply locating most of the words that appear in Web pages or performing sophisticated analyses to identify key words and phrases. These data are then stored in the search engine's database, along with an address, termed a uniform resource locator (URL), that represents where the file resides. A user then deploys a browser, such as the familiar Netscape, to submit queries to the search engine's database. The query produces a list of Web resources, the URLs that can be clicked on to connect to the sites identified by the search.

Existing search engines service millions of queries a day. Yet it has become clear that they are less than ideal for retrieving an ever growing body of information on the Web. In contrast to human indexers, automated programs have difficulty identifying characteristics of a document such as its overall theme or its genre—whether it is a poem or a play, or even an advertisement.

The Web, moreover, still lacks standards that would facilitate automated indexing. As a result, documents on the Web are not structured so that programs can reliably extract the routine information that a human indexer might find through a cursory inspection: author, date of publication, length of text and subject matter. (This information is known as metadata.) A Web crawler might turn up the desired article authored by Jane Doe. But it might also find thousands of other articles in which such a common name is mentioned in the text or in a bibliographic reference.

Publishers sometimes abuse the indiscriminate character of automated indexing. A Web site can bias the selection process to attract attention to itself by repeating within a document a word, such as "sex," that is known to be queried often. The reason: a search engine will display first the URLs for the documents that mention a search term most frequently. In contrast, humans can easily see around simpleminded tricks.

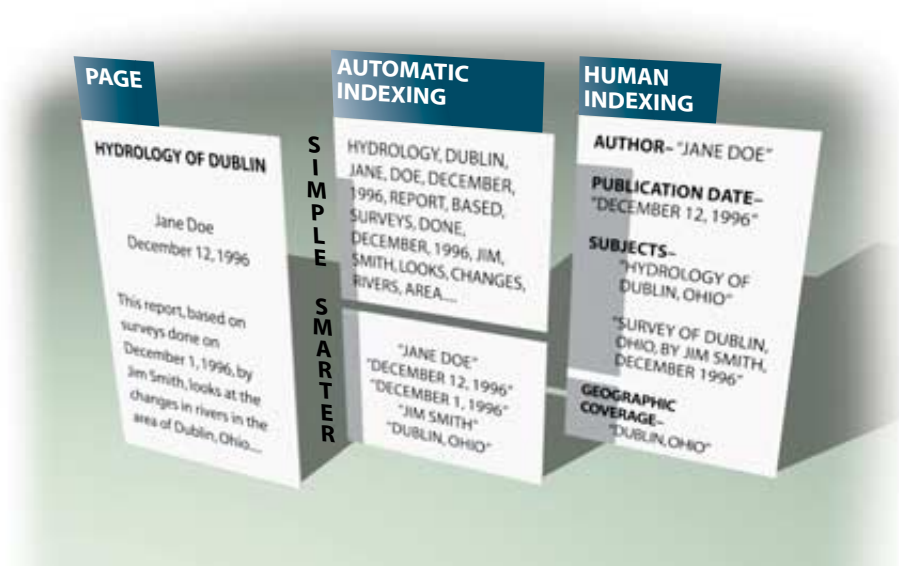
The professional indexer can describe the components of individual pages of all sorts (from text to video) and can clarify how those parts fit together into a database of information. Civil War photographs, for example, might form part of a collection that also includes period music and soldier diaries. A human indexer can describe a site's rules for the collection and retention of programs in, say, an archive that stores Macintosh software. Analyses of a site's purpose, history and policies are beyond the capabilities of a crawler program.

Another drawback of automated indexing is that most search engines recognize text only. The intense interest in the Web, though, has come about because of the medium's ability to display images, whether graphics or video clips. Some research has moved forward toward finding colors or patterns within images [*see box on next two pages*]. But no program can deduce the underlying meaning and cultural significance of an image (for example, that a group of men dining represents the Last Supper).

At the same time, the way information is structured on the Web is changing so that it often cannot be examined by Web crawlers. Many Web pages are no longer static files that can be analyzed and indexed by such programs. In many cases, the information displayed in a document is computed by the Web site during a search in response to the user's request. The site might assemble a map, a table and a text document from different areas of its database, a disparate collection of information that conforms to the user's query. A newspaper's Web site, for instance, might allow a reader to specify that only stories on the oil-equipment business be displayed in a personalized version of the paper. The database of stories from which this document is put together could not be searched by a Web crawler that visits the site.

A growing body of research has attempted to address some of the problems involved with automated classification methods. One approach seeks to attach metadata to files so that indexing systems can collect this information. The most advanced effort is the Dublin Core Metadata program and an affiliated endeavor, the Warwick Framework—the first named after a workshop in Dublin, Ohio, the other for a colloquy in Warwick, England. The workshops have defined a set of metadata elements that are simpler than those in traditional library cataloguing and have also created methods for incorporating them within pages on the Web.

Categorization of metadata might range from title or author to type of document (text or video, for instance). Either automated indexing software or humans may derive the metadata, which can then be attached to a Web page for retrieval by a crawler. Precise and de-



BRYAN CHRISTIE

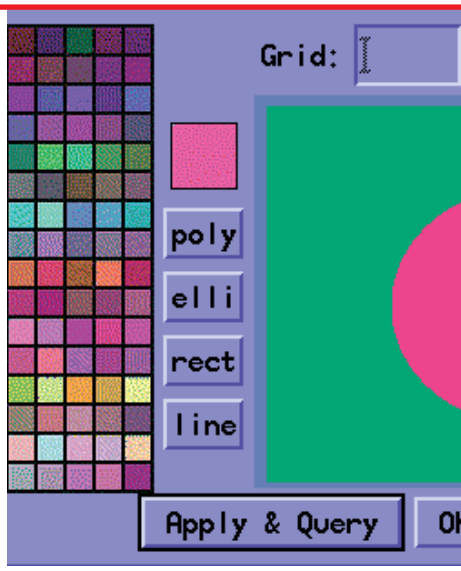
AUTOMATED INDEXING, used by Web crawler software, analyzes a page (*left panel*) by designating most words as indexing terms (*top center*) or by grouping words into simple phrases (*bottom center*). Human indexing (*right*) gives additional context about the subject of a page.

tailed human annotations can provide a more in-depth characterization of a page than can an automated indexing program alone.

Where costs can be justified, human indexers have begun the laborious task of compiling bibliographies of some Web sites. The Yahoo database, a commercial venture, classifies sites by broad subject area. And a research project at the University of Michigan is one of

Finding Pictures on the Web

by Gary Stix, *staff writer*



The Internet came into its own a few years ago, when the World Wide Web arrived with its dazzling array of photography, animation, graphics, sound and video that ranged in subject matter from high art to the patently lewd. Despite the multimedia barrage, finding things on the hundreds of thousands of Web sites still mostly requires searching indexes for words and numbers.

Someone who types the words "French flag" into the popular search engine AltaVista might retrieve the requested graphic, as long as it were captioned by those two identifying words. But what if someone could visualize a blue, white and red banner but did not know its country of origin?

Ideally, a search engine should allow the user to draw or scan in a rectangle with vertical thirds that are colored blue, white and red—and then find any matching images stored on myriad Web sites. In the past few years, techniques that combine key-word indexing with image analysis have begun to pave the way for the first image search engines.

Although these prototypes suggest possibilities for the indexing of visual information, they also demonstrate the crudeness of existing tools and the continuing reliance on text to track down imagery. One project, called WebSEEk, based at Columbia University, illustrates the workings of an image search engine. WebSEEk begins by downloading files found by trolling the Web. It then attempts to locate file names containing acronyms, such as GIF or MPEG, that designate graphics or video content. It also looks for words in the names that might identify the subject of the files. When the software finds an image, it analyzes the prevalence of different colors and where they are located. Using this information, it can distinguish among photographs, graphics and black-and-white or gray images. The software also compresses each picture so that it can be represented as an icon, a miniature image for display alongside other icons. For a video, it will extract key frames from different scenes.

A user begins a search by selecting a category from a menu—"cats," for example. WebSEEk provides a sampling of icons for the "cats" category. To narrow the search, the user can click on any icons that show black cats. Using its previously generated color analysis, the search engine looks for matches of images that have a similar color profile. The presentation of the next set of icons may show black cats—but also some marmalade cats sitting on black cushions. A visitor to WebSEEk can refine a search by adding or excluding certain colors from an image when initiating subsequent queries. Leaving out yellows or oranges might get rid of the odd marmalade. More simply, when presented with a series of icons, the user can also specify those images that do not contain black cats in order to guide the program away from mistaken choices. So far WebSEEk has downloaded and indexed more than 650,000 pictures from tens of thousands of Web sites.

Other image-searching projects include efforts at the University of Chicago, the University of California at San Diego, Carnegie Mellon University, the Massachusetts Institute of Technology's Media Lab and the University of California at Berkeley. A number of commercial companies, including IBM and Virage, have crafted software that can be used for searching corporate networks or databases. And two companies—Excalibur Technologies and Interpix Software—have collaborated to supply software to the Web-based indexing concerns Yahoo and Infoseek.

One of the oldest image searchers, IBM's Query by Image Content (QBIC), produces more sophisticated matching of image features than, say, WebSEEk can. It is able not only to pick out the col-

several efforts to develop more formal descriptions of sites that contain material of scholarly interest.

Not Just a Library

The extent to which either human classification skills or automated indexing and searching strategies are needed will depend on the people who use the Internet and on the business prospects for publishers. For many communities of scholars, the model of an organized collection—a digital library—still remains relevant. For other groups, an uncontrolled, democratic medium may provide the best vehicle for information dissemination. Some users, from financial analysts to spies, want comprehensive access to raw databases of information, free of any controls or editing. For them, standard search engines provide real benefits because they forgo any selective filtering of data.

The diversity of materials on the Net goes far beyond the scope of the traditional library. A library does not provide quality rankings of the works in a collection. Because of the greater volume of networked information, Net users want guidance about where to spend the limited amount of time they have to research a subject. They may need to know the three "best" documents for a given purpose. They want this information without paying the costs of employing humans to critique the myriad Web sites. One solution that again calls for human involvement is to share judgments about what is worthwhile. Software-based rating systems have begun to let users describe the quality of particular Web sites [see "Filtering Information on the Internet," by Paul Resnick, page 62].

Software tools search the Internet and also separate the good from the bad. New programs may be needed, though, to ease the burden of feeding the crawlers that repeatedly scan Web sites. Some Web site managers have reported that their computers are spending enormous amounts of time in providing crawlers with information to index, instead of servicing the people they hope to attract with their offerings.

To address this issue, Mike Schwartz

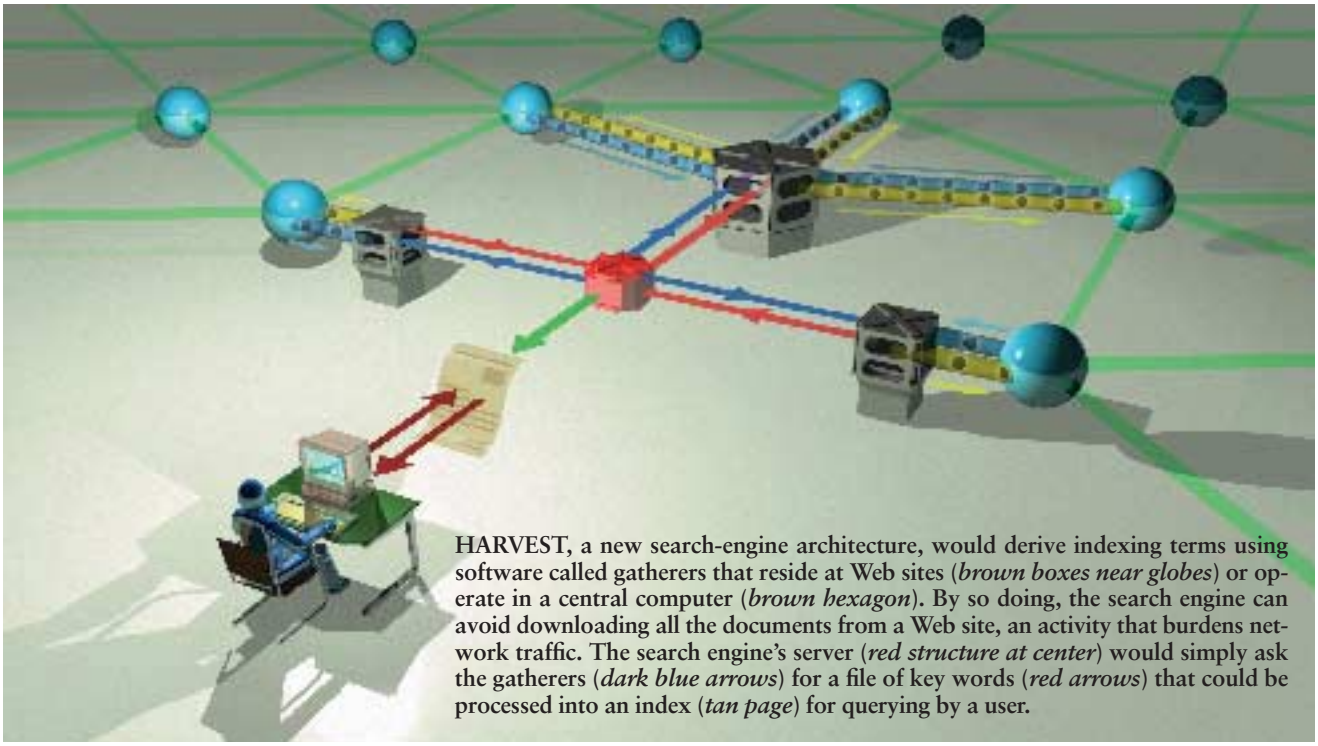
IBM CORPORATION/ROMTECH/COREL

ors in an image but also to gauge texture by several measures—contrast (the black and white of zebra stripes), coarseness (stones versus pebbles) and directionality (linear fence posts versus omnidirectional flower petals). QBIC also has a limited ability to search for shapes within an image. Specifying a pink dot on a green background turns up flowers and other photographs with similar shapes and colors, as shown above. Possible applications range from the selection of wallpaper patterns to enabling police to identify gang members by clothing type.

All these programs do nothing more than match one visual feature with another. They still require a human observer—or accompanying text—to confirm whether an object is a cat or a cushion. For more than a decade, the artificial-intelligence community has labored, with mixed success, on nudging computers to ascertain directly the identity of objects within an image, whether they are cats or national flags. This approach correlates the shapes in a picture with geometric models of real-world objects. The program can then deduce that a pink or brown cylinder, say, is a human arm.

One example is software that looks for naked people, a program that is the work of David A. Forsyth of Berkeley and Margaret M. Fleck of the University of Iowa. The software begins by analyzing the color and texture of a photograph. When it finds matches for flesh colors, it runs an algorithm that looks for cylindrical areas that might correspond to an arm or leg. It then seeks other flesh-colored cylinders, positioned at certain angles, which might confirm the presence of limbs. In a test last fall, the program picked out 43 percent of the 565 naked people among a group of 4,854 images, a high percentage for this type of complex image analysis. It registered, moreover, only a 4 percent false positive rate among the 4,289 images that did not contain naked bodies. The nudes were downloaded from the Web; the other photographs came primarily from commercial databases.

The challenges of computer vision will most likely remain for a decade or so to come. Searches capable of distinguishing clearly among nudes, marmalades and national flags are still an unrealized dream. As time goes on, though, researchers would like to give the programs that collect information from the Internet the ability to understand what they see.



HARVEST, a new search-engine architecture, would derive indexing terms using software called gatherers that reside at Web sites (*brown boxes near globes*) or operate in a central computer (*brown hexagon*). By so doing, the search engine can avoid downloading all the documents from a Web site, an activity that burdens network traffic. The search engine's server (*red structure at center*) would simply ask the gatherers (*dark blue arrows*) for a file of key words (*red arrows*) that could be processed into an index (*tan page*) for querying by a user.

BRYAN CHRISTIE

and his colleagues at the University of Colorado at Boulder developed software, called Harvest, that lets a Web site compile indexing data for the pages it holds and to ship the information on request to the Web sites for the various search engines. In so doing, Harvest's automated indexing program, or gatherer, can avoid having a Web crawler export the entire contents of a given site across the network.

Crawler programs bring a copy of each page back to their home sites to extract the terms that make up an index, a process that consumes a great deal of network capacity (bandwidth). The gatherer, instead, sends only a file of indexing terms. Moreover, it exports only information about those pages that have been altered since they were last accessed, thus alleviating the load on the network and the computers tied to it.

Gatherers might also serve a different function. They may give publishers a framework to restrict the information that gets exported from their Web sites. This degree of control is needed because the Web has begun to evolve beyond a distribution medium for free information. Increasingly, it facilitates access to proprietary information that is furnished for a fee. This material may not be open for the perusal of Web crawlers. Gatherers, though, could distribute only the information that publishers wish to make available, such as links to summaries or samples of the information stored at a site.

As the Net matures, the decision to opt for a given information collection method will depend mostly on users. For which users will it then come to resemble a library, with a structured approach to building collections? And for whom will it remain anarchic, with access supplied by automated systems?

Users willing to pay a fee to underwrite the work of authors, publishers, indexers and reviewers can sustain the tradition of the library. In cases where information is furnished without charge or is advertiser supported, low-cost computer-based indexing will most likely dominate—the same unstructured environment that characterizes much of the contemporary Internet. Thus, social and economic issues, rather than technological ones, will exert the greatest influence in shaping the future of information retrieval on the Internet.

The Author

CLIFFORD LYNCH is director of library automation at the University of California's Office of the President, where he oversees MELVYL, one of the largest public-access information retrieval systems. Lynch, who received a doctorate in computer science from the University of California, Berkeley, also teaches at Berkeley's School of Information Management and Systems. He is a past president of the American Society for Information Science and a fellow of the American Association for the Advancement of Science. He leads the Architectures and Standards Working Group for the Coalition for Network Information.

Further Reading

THE HARVEST INFORMATION DISCOVERY AND ACCESS SYSTEM. C. M. Bowman et al. in *Computer Networks and ISDN Systems,* Vol. 28, Nos. 1–2, pages 119–125; December 1995.

The Harvest Information Discovery and Access System is available on the World Wide Web at http://harvest.transarc.com

THE WARWICK METADATA WORKSHOP: A FRAMEWORK FOR THE DEPLOYMENT OF RESOURCE DESCRIPTION. Lorcan Dempsey and Stuart L. Weibel in *D-lib Magazine,* July–August 1996. Available on the World Wide Web at http://www.dlib.org/dlib/july96/07contents.html

THE WARWICK FRAMEWORK: A CONTAINER ARCHITECTURE FOR DIVERSE SETS OF METADATA. Carl Lagoze, ibid.