1994

Full-Text Retrieval: Systems and Files

Carol Tenopir
University of Tennessee - Knoxville

FULL-TEXT RETRIEVAL:
SYSTEMS AND FILES

Carol Tenopir

INTRODUCTION

Much of the development in the first 30 years of library automation has been in solving the problem of identifying relevant sources. Automation of the library's card catalog provides a finding tool for the library's collections. The books, journals, films, and other materials located through the catalog still mostly reside in their original form, with no direct connection to the automated finding tool.

Most of the early development in electronic publishing was also aimed solely at identifying information sources. Secondary publishers, notably publishers of indexing/abstracting serials, were the first to provide their resources in electronic form. Throughout the 1970s and much of the 1980s, indexing/abstracting (bibliographic) databases were predominant in the online database world. The first CD-ROM databases for libraries were many of these same bibliographic files. Traditionally (and still) bibliographic databases are the most widely used type of electronic resource in libraries.

Advances in Library Automation and Networking, Volume 5, pages 43-71.

ISBN: 1-55938-510-3

Starting in the mid-1980s, due to great increases in disk storage capacities and better document conversion techniques, full texts of certain types of documents became more widely available. In the 1990s, full-text databases (files) are the most rapidly growing type of commercially available database. Better text-retrieval software is leading to more locally created full-text databases as well. Perhaps in this decade we will at last electronically solve the document delivery problem as well as the document location problem.

TYPES OF COMMERCIALLY AVAILABLE FULL-TEXT DATABASES

Libraries have access to a variety of full texts on a pay-per-use basis via commercial online information systems or at a one-time purchase rate or fixed fee subscription rate on CD-ROM. A few published texts are available on diskette or magnetic tape as well. Just as with printed publications, electronic full-text publications are not all the same, so it is helpful to define them and to categorize them by document type.

A commonly accepted definition for full text is given in the *Directory of Online Databases*: "Full Text [databases] contain records of the complete text of an item, for example, a newspaper article, a specification, a court decision, or a newsletter."[1] They are categorized as a subset of "Source Databases. Those that contain original source data, the full text of original source information, or materials prepared specifically for electronic distribution."[2]

A further elaboration is given in a user's guide to the *Directory of Online Databases*:

> This classification is assigned to databases that contain the complete text of published works (e.g., journal articles, specifications, court decisions, newspaper items)—regardless of whether charts, footnotes, illustrations, or other such enhancements in the original work are included. In most cases, the Full Text database corresponds, in whole or significant part, to a publication. However, Full Text also covers databases of "items" that have been prepared expressly for online distribution and that have no corresponding printed publication.[3]

Still, although it is changing, the vast majority of current full texts are just electronic counterparts of printed publications.

All of these definitions of full text exclude complete texts of directories, which are categorized as directory or referral databases, but include a variety of other types of documents. Full-text databases can be categorized into 10 general types.[4] These categories are not always mutually exclusive, nor does every text neatly fit into a category, but they help to differentiate the complexity of a general term such as *full text*.

1. *Statutes, Court Decisions, and Other Primary Legal Documents.* The intended audience is most often legal experts. Materials included here are decisions from U.S. federal and tax courts, laws and legislation from each U.S. state, and various legal materials from many other countries. Length and characteristics of the documents vary widely, but certain key elements are particularly important for searching. These include such things as name of the judge, names of the defendants, the case name and number, cases cited as precedent, date, court, location, and other particulars.

The major online systems in the United States that provide these materials are LEXIS and WESTLAW. CD-ROM regional subsets are offered by a growing number of companies. For example, Michie Company provides New Mexico Law on CD-ROM, and Info-One International and Diskrom Australia offer a variety of Australian case law online and on CD-ROM.

2. *Other Government Documents, Patents, Regulations, and Other Official Publications.* The intended audience varies; patents are often accessed by experienced patent searchers. Each publication has its own characteristics, varying from the short requests for proposals in *Commerce Business Daily* to detailed regulations in *Code of Federal Regulations*. Some cover many different topics; others are more narrowly focused. Important information elements include issuing agency, date, patent or contract numbers, type of document, and subject. LEXIS has patents; NEXIS offers many U.S. documents, such as the *Code of Federal Regulations* and the *Federal Register*. DIALOG has *Commerce Business Daily*. West provides a government contracts file on CD-ROM.

3. *News Releases and Other Unpublished Information Intended for Subject Experts, Other Business People, or Journalists.* Most cover one narrow topic, and the length is rarely more than two pages. Press releases are issued by a government agency or company to announce new products or give new information. They usually

include name and phone number of a contact person. Unpublished reports may also be summaries of new developments or information about an organization. Important access points include date, issuing company, contact person, and subjects. These are most often part of another database and are available primarily online on systems such as DIALOG, NEXIS, and CompuServe.

4. *Newspaper Articles.* Intended audiences include people of all ages who are interested in current events. Frequent users include news professionals and reference librarians. Newspaper databases include articles, columns, and other selected features on a variety of topics. Length varies from several paragraphs to long feature stories. Often articles are written with important information in the first paragraph. Most newspaper databases do not include everything found in the printed equivalent; typical exclusions include such things as advertisements, classified listings, weather forecasts, sports scores, syndicated columns, and stories taken from wire services. Major national and international papers, such as *Wall Street Journal, Pravda*, and *The Times*, are available, as are local dailies or weeklies, such as the *Fresno Bee, Allentown Morning Call*, and others, in whole or in part. Major suppliers of newspaper databases online are NEXIS, VU/TEXT, DIALOG, Dow Jones News/Retrieval, and InfoGlobe search service. On CD-ROM, major suppliers include NewsBank, UMI/Data Courier, and DIALOG.

5. *Newswire Services.* Intended audiences are the same as those of newspapers. Newswire stories vary in length, but are often concise summaries of major news events. Timeliness is critical and they are often updated frequently. Slightly different versions of the same story may therefore occur several times in the same database. Backfiles may not be kept online for long on some systems. Many major newswires from all over the world are available online, including Associated Press, Reuters, Tass, and Kyodo English Language News. Major online systems for newswires include DIALOG, NEXIS, NEWSNET, Dow Jones News/Retrieval, and CompuServe.

6. *Newsletters.* Intended audiences are subject experts or information professionals, usually within a corporate or research environment. Each newsletter has its own style and language, and a database may consist of a single newsletter or many. They are subject specific and often highly technical or of interest only within the target industry. Information may be very time sensitive.

Important access points include date, newsletter title, and subject. Hundreds of newsletters are available online, from many different industries. The two major online systems for newsletters are NEXIS and NEWSNET.

7. *Reference Books.* Intended audiences range from school children to researchers, depending on the book. There is much variety in terms of style, length, audience level, graphics included, and useful access points. Many are used primarily for fact retrieval and most are highly structured into short, fairly consistent sections. Encyclopedias are the most widely available type of reference book; the CD-ROM versions of encyclopedias may include graphics, motion, and sound as well as text. Footnotes and cross references are important access points in encyclopedias in addition to subjects. Other reference book databases range from highly technical standard works, such as *Kirk-Othmer Encyclopedia of Chemical Technology* and the *Merck Index*, to Internal Revenue Service tax information, to the Bible, to Quanta's About Cows reference CD-ROM. Online systems with reference books include BRS, DIALOG, NEXIS, STN International, Data-Star, and Dow Jones News/Retrieval. CD-ROM books are often available directly from the book publisher, such as Britannica Software, National Geographic Society, or WorldBook. Others are sold by third parties such as CMC ReSearch, SilverPlatter, and so forth.

8. *Other Books.* A growing number of novels, collections of short stories, and other books are being converted into electronic form. Unlike reference books, these books are intended for reading, with fact retrieval playing a minor role; however, they may be used for text analysis. CD-ROM is the publishing medium of choice for electronic books. Several titles are available from CMC ReSearch Inc., including Shakespeare on Disc. OCLC and G.K. Hall produce an American Authors disc.

9. *Scholarly or Technical Journal Articles.* The primary audience is subject specialists, usually researchers. Most articles are lengthy with many footnotes. Sentences and paragraphs may be long, and language is technical. Printed versions include many tables, figures, and equations that may or may not be in the database version. Abstracts may precede the article. Usually, only the major articles are included in the online version: letters to the editor, book reviews, news stories, and columns from the printed versions may not be in the online version. Subject access is most important, but journal

name, authors, cited authors, and dates are useful as well. Many of the journals from the American Chemical Society are online, as are major medical journals such as the *New England Journal of Medicine, Lancet*, and the *British Medical Journal.* STN International, BRS, and DIALOG are major online systems that provide journals. The ADONIS CD-ROM project provides journal articles from over 400 scientific and biomedical publications. CD-ROM journals may be ASCII text like online versions or image files that look just like the print versions.

10. *Nonspecialist or General Interest Magazine Articles.* Most of these are for the general reader. They are especially useful for students or laypeople for personal or school-related information. The writing style, subjects, and length of articles varies greatly, but most do not include technical language, footnotes, or abstracts. The printed versions contain many photographs, charts, sidebars, and other graphics, most of which are not included online, although captions may be. By far the most important access point is subject, but magazine title and date are useful as well. Several hundred magazines are available, either as stand-alone titles or within a multi-title database such as Magazine ASAP. Online systems that provide access to magazines include BRS, DIALOG, and NEXIS. Information Access Company's Magazine Rack is a popular, inexpensive CD-ROM full text for home users, while a CD-ROM version of their Magazine ASAP is marketed to libraries. UMI/DataCourier and Ebsco also have magazines on CD-ROM.

NUMBERS OF FULL-TEXT DATABASES

According to industry leader Martha Williams, there are now over 2,040 full-text databases, available either online, on CD-ROM, on magnetic tape, on diskette, or on a combination of electronic options.[5] Full text now make up 44% of the total number of *word-oriented* databases.[6] Figure 1 shows how full text has grown in the last decade in both numbers and percent of total databases.

A more meaningful figure may be the number of electronic full-text sources available, whether they are mixed with other sources into one database or available alone as a separate file. It is difficult to tell just how many magazines, journals, newspapers, and so forth are represented in a figure like 2,040. Some titles are available as stand-

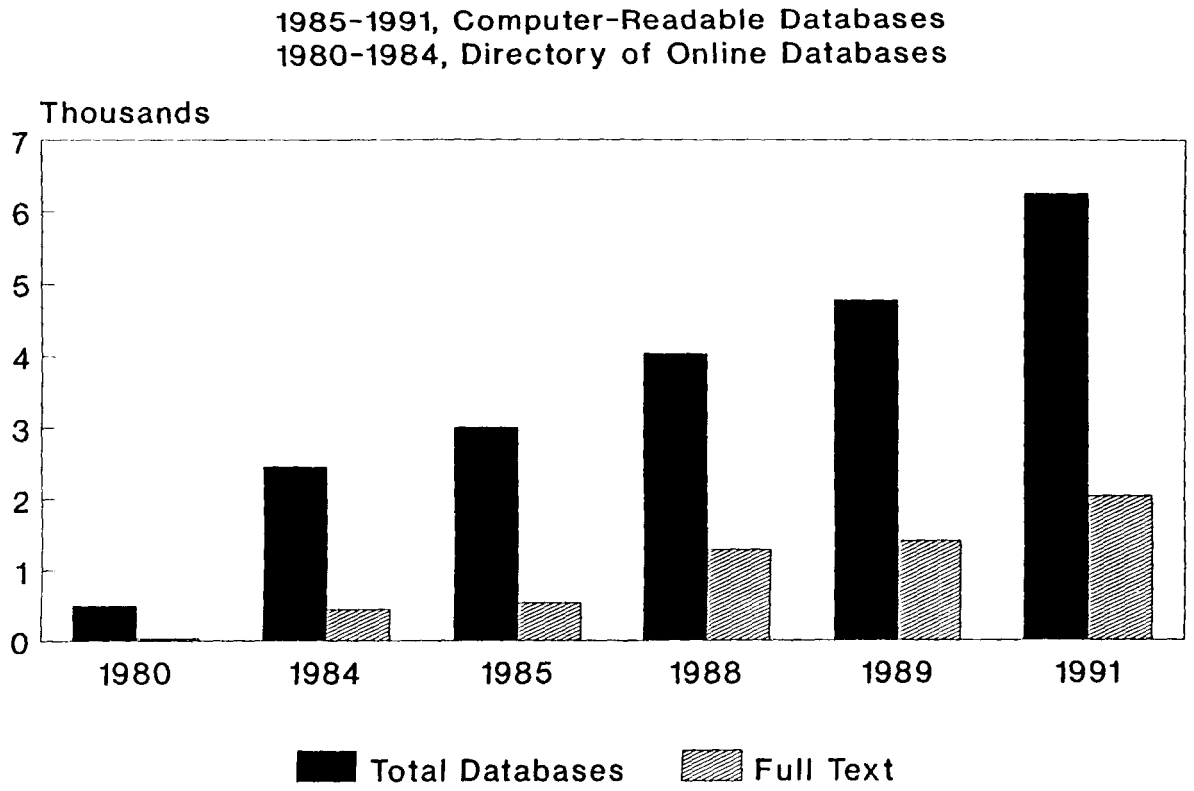

Figure 1. Growth in Full-Text Databases

alone databases (for example, *Harvard Business Review, Los Angeles Times*, and Compton's MultiMedia Encyclopedia); others are part of one large mega-database (for example, Magazine ASAP and Trade & Industry ASAP, ACS Journals Online, and Newsbank) that may combine hundreds of periodical titles. A current trend is to create *hybrid* databases that mix full text with some bibliographic records in one database (for example, ABI/INFORM and PTS PROMPT).[7] Other CD-ROM databases, such as PComm, combine full-text articles and news with directory information and computer programs.

The total number of titles available in full text is thus difficult to estimate, but, thanks to the efforts of a small publishing company, it can be estimated for online versions of serials such as magazines, journals, newspapers, newswires, and newsletters. BiblioData's 1992 edition of *Fulltext Sources Online* includes listings for over 3,000 serial titles available via nineteen online systems.[8] (This is up from 1,700 titles listed in the 1989 edition.)

Each source title is listed separately in the directory, even if on some systems it is part of a single multi-title database. Thus, the listing for *Financial World* shows it as part of the multi-title database Trade & Industry ASAP on BRS and DIALOG, part of the Business Library on Dow Jones, part of the Dow Jones file on DataTimes, but available separately on NEXIS (FINWLD) and Reuters (FINWO). Each entry also gives coverage of the title on each system, because there may be great variation. *Financial World* is included from January 1983 to present on BRS, DIALOG, and Mead (NEXIS); from January 1985 to present on Dow Jones; from September 1986 to present on Reuters; and from January 1987 to present on DataTimes.

Just because a title is listed in *Fulltext Sources Online* doesn't mean that everything in the print equivalent will be found online. According to the editor, "Fulltext means that complete articles are found online. It does not mean that a periodical is found cover-to-cover in the database. Database producers often choose to include only the most meaningful articles."[9] The editor also indicates that although:

> one would like to think that a journal or newspaper described as being available online in fulltext is available cover-to-cover ... [t]his is rarely the case. Coverage policies differ widely from one periodical to another and from one database vendor to another.

> In no case do online periodicals reproduce advertisements that appear in the original. None, to our knowledge, reproduces long tabular material such as pages of stock quotations. Because of copyright restrictions, most newspapers exclude wire service stories and syndicated columns online, limiting themselves to items written by in-house staff. In addition, many newspapers omit letters to the editor, editorials, obituaries and filler material. Magazines often omit announcements, index to advertisers, book reviews, notices, corrections, information for authors, classified ads and meetings calendars.[10]

Even when an article from a print issue is included, rarely is the entire article online. Orenstein explains:

> The notion of a complete article has its pitfalls. Given the current state of technology, few articles are provided online with their tables, charts, graphs, illustrations or photographs intact. These things are sometimes noted or described, and published if short. In general, the term "*fulltext*" means that the entire *text* of an article is available online.[11]

A random sample of entries in the 1989 *Directory of Online Databases* showed that approximately 68% of the full text listings were for periodicals, with 5% for monographs and 27% for all other categories.[12] A source such as *Fulltext Sources Online* thus describes a majority of online full-text resources, but it is safe to estimate that there are at least 1,500 additional full-text titles.

BiblioData's newer *Newspapers Online* is an alphabetical directory of North American daily newspapers available online or on CD-ROM.[13] The first edition (1992) includes detailed information on electronic availability of 138 papers. Only general-focus papers that are published at least five times a week are included. Online or CD-ROM versions must contain "substantially all articles of the newspapers ... ; that is, the contents of the newspapers are available electronically virtually 'cover-to-cover.' Newspapers in which only a few articles are selected for inclusion in a file (such as the Business Dateline collection) are not [included in the directory]."[14]

ONLINE SYSTEMS FOR FULL TEXT

Fulltext Sources Online covers the serial titles available on nineteen online systems that include large amounts of full text. All of the online systems also offer other types of information in addition to full text,

including bibliographic citations, abstracts, directories, statistics, financial information, or stock prices.

The online systems are available worldwide, although they are headquartered in several different countries, as can be seen in Table 1.[15]

Newspapers Online also includes papers online via CompuServe (Columbus, Ohio).

In the library/information center market, use is dominated by just two systems—DIALOG and Mead Data Central, as can be seen in Figure 2.[16] A total of just five of the full-text systems (DIALOG, Mead, BRS, STN, and Westlaw), along with two bibliographic systems (NLM and Orbit) account for almost all online searching in libraries and information centers.

Table 1. Online Systems with Full-Text Databases

UNITED STATES
BRS Information Technologies (A division of Maxwell Online, headquartered in McLean Virginia)
Burrelle's Broadcast Database (Livingston, NJ)
DataTimes (headquartered in Oklahoma City)
DIALOG Information Services Inc. (located in Palo Alto, California, and owned by Knight-Ridder)
Dow Jones News/Retrieval (Princeton, NJ)
Mead Data Central (LEXIS/NEXIS) (Dayton, Ohio)
NewsNet (Bryn Mawr, PA)
Nikkei Telecom (Nihonkeizai Shimbun America) (New York)
STN International (A joint venture of the U.S. Chemical Abstracts Service [Columbus, Ohio], German FIZ Karlsruhe, and Japan Information Center of Science and Technology)
VU/TEXT (located in Philadelphia, owned by Knight-Ridder)
WESTLAW (West Publishing Company, St. Paul, MN)
CANADA
Info Globe (Toronto, Ontario)
Infomart Online (Don Mills, Ontario)
QL Systems Ltd. (Kingston, Ontario)
EUROPE
Data-Star (Switzerland, with offices in U.S. and U.K.)
G.CAM/EDD (France)
Genios Wirschaftsdatenbanken (Germany)
Reuters Ltd. (United Kingdom)
FT PROFILE (United Kingdom)

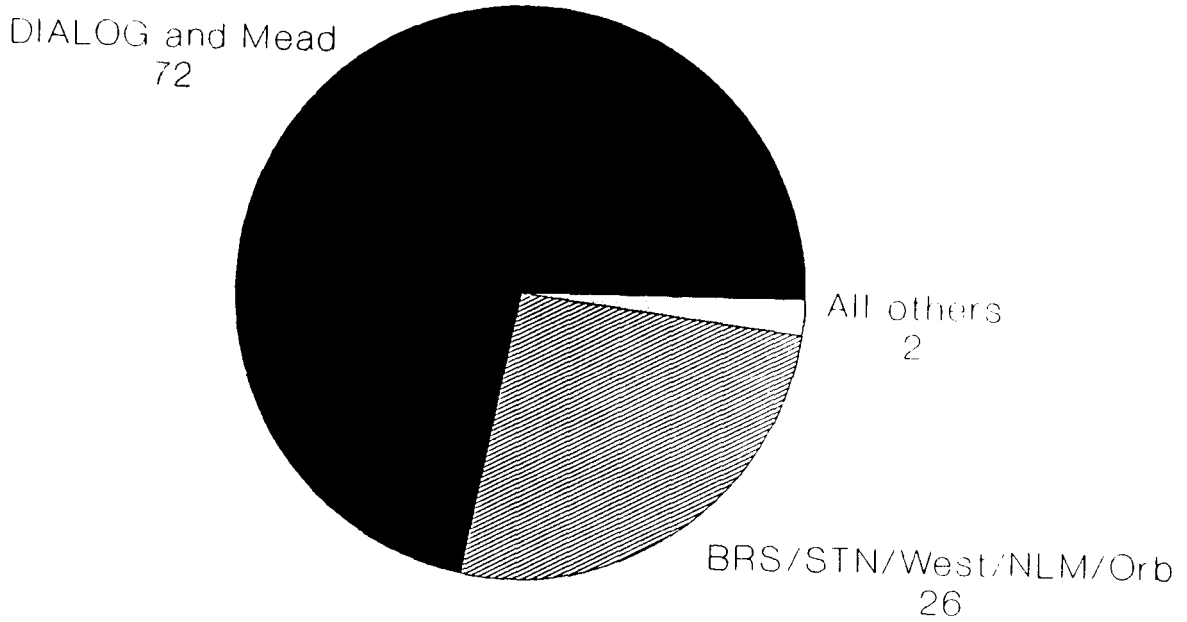

Figure 2. Online Use in Libraries and Information Centers

Many of these online systems originally were developed for bibliographic databases. The search and display features were designed for shorter indexing/abstracting records. When full text was added, they found it necessary to add other search and display features to enhance full-text searching.

CD-ROM SYSTEMS FOR FULL TEXT

The arrangements for creating and marketing commercially available CD-ROM databases are not so clear-cut as the established online model. CD-ROM databases can be sold under a vendor-database producer arrangement just like online systems, but often a database producer arranges for a CD-ROM software company to provide just the software, maintaining marketing and distribution rights. Or, producers may develop proprietary software and do all of the development, marketing, and maintenance themselves.

CD-ROM full-text databases developed and sold under the vendor model are available from such vendors as DIALOG, SilverPlatter, Microsoft, and OCLC. Database producers who have developed their own proprietary software include UMI (ProQuest), Newsbank, Oxford University Press, Information Access Company (InfoTrac), and many others. Database producers who have used other companies for CD-ROM development or CD-ROM software include Grolier (Online Computer Systems), Ebsco (Fulcrum), Bowker Electronic Publishing (Online Computer Systems), and CMC ReSearch (DiscPassage).[17]

With full-text databases, the arrangements may be complex. For example, the H.W. Wilson Company has its own CD-ROM software, WILSONDISC, which it uses for bibliographic files. Wilson contracts with UMI to provide a full-text database on ProQuest, which is a companion file to Wilson's Social Science Index. Other such cooperative arrangements are likely to be forthcoming.

The Directory of Portable Databases (October 1991 edition) lists 934 total CD-ROM databases, 319 (34%) of which are full text, in whole or in part.[18] These are of all types. Much of the innovation in multimedia is taking place with CD-ROM full-text databases because the development costs are lower and, unlike online files, they do not have the limitations of transmitting over the telephone lines. After examining four CD-ROM directories, Nicholls reported 341

multimedia CD-ROM titles (17% of the total) as of June 1991, up from only 40 titles in 1989.[19]

There may be great variation in the appearance and the search features available in CD-ROM full-text files. They may be machine-readable ASCII text like their online counterparts, where every word of the text is a potential search term and texts or portions of text can be downloaded and transferred into word processing or database programs. Ebsco's full text version of Magazine Articles Summaries, Information Access Company's full-text InfoTrac files, and Ziff-Davis's Computer Select are ASCII searchable text.

Or, CD-ROM versions may be combination files that include graphics with ASCII text. In these cases, the text is usually fully searchable, while the graphics can be displayed with relevant text portions. Grolier's Encyclopedia is of this type. The American Chemical Society/OCLC Chemical Online Retrieval Experiment (CORE) project of full-text chemical journals combines ASCII text plus graphics on magnetic media with page images on an optical jukebox.[20]

Finally, CD-ROM versions may just be page-image files. Scanned image files do not support full-text searching, but are usually meant to be used in conjunction with a bibliographic database for document delivery. Once an article is located through the bibliographic database, the CD-ROM scanned image articles can be located and retrieved using a control number. UMI ProQuest texts are of this type: Business Periodicals Ondisc carries page images of journals that correspond to the ABI/INFORM bibliographic CD-ROM file, and General Periodicals Ondisc corresponds to Periodical Abstracts.

FULL TEXT ON DISKETTE OR TAPE

Some full-text databases in ASCII form are now available for lease or purchase on magnetic media. Floppy diskette or magnetic tape are popular options for distributing bibliographic data for use in libraries. The library must then load the data from the diskette or tape onto the hard drive of their local micro- or mainframe computer. In the case of diskette distribution, microcomputer search and retrieval software may come with the text subscription or purchase. Databases on magnetic tape are normally sold without retrieval software; it is up to the subscribing library to load the tapes using

compatible and appropriate search software. With bibliographic databases, this is often the same software that is used for the library's online public access catalog.

Diskette distribution is reserved for small files or those that require frequent updates. It is usually a low-cost alternative, with many full-text files costing under $100. Of 545 diskette databases listed in *The Directory of Portable Databases*, 154 (28%) are classified as full text, in whole or in part. They are of all types, ranging from the *U.S. Code of Federal Regulations*, to the Bible, to selected works of Benjamin Franklin.

Magnetic tape databases, to be loaded on a library's large computer, are as yet mostly bibliographic, directory, or numeric. *The Directory of Portable Databases* lists 45 full-text databases out of 430 tape titles (10.5%). This small number is probably because full-text databases take up so much computer space and require extensive computer resources for loading, maintaining, and searching. A library may not be able to support large full-text databases on the same system that runs their OPAC, bibliographic files, and other library operations. Many government publications are available on tape, including documents from the Internal Revenue Service, Securities and Exchange Commission, Department of Defense, and others.

A growing number of specialized machine-readable texts are available in the humanities for text analysis research. They are outside the scope of this chapter. For more information on this topic, see a recent article by Price-Wilkin.[21]

FULL-TEXT BOOK LIBRARIES

Undoubtedly, the number of full-text sources available through commercially available online systems and on CD-ROM and other distribution media will continue to grow. Some of the most exciting developments are happening at the network level, where much of the experimentation and innovation in large online textual database creation is taking place.

Several notable projects are converting large numbers of books and other texts into electronic books for access over Internet or other online networks. Project Gutenberg strives to convert and provide access to 10,000 important out-of-copyright books by the year 2000.

Public Access Xanadu is Theodor Nelson's vision of a hypertext network of online documents.[22] The U.S. Marine Corps' Online Books project provides an online hypertext library of the complete Marine Corps University's warfare collection.[23]

Library involvement in such projects will provide new levels of access to all types of materials. Libraries can take a leading role in other cooperative ventures to make electronic publications more widely available.

TRUE ELECTRONIC JOURNALS

Electronic journals that are not derived from print journals have existed—at least experimentally—for quite some time and hold much promise for the future. At least 15 refereed electronic journals are currently available on BITNET and other networks.[24]

In 1991, the American Association for the Advancement of Science and OCLC launched *The Online Journal of Current Clinical Trials*, an online refereed research journal. The journal publishes peer-reviewed research articles that include charts, tables, and graphs; it has videotext-quality typeface. Articles can be searched and downloaded, and the system offers hypertext links to Medline abstracts. If successful, this journal could lead to a new standard for original online full texts.

Non-refereed full texts are also becoming an important part of the research process. This text-based invisible college is accessible through list servers and similar systems on networks such as BITNET and Internet, and it can be a part of a library's text-based services. However, these informal full texts are outside the scope of this chapter.

CREATION OF FULL-TEXT DATABASES

Most full texts commercially available today are still byproducts of printed publications. Few are available only in electronic form. These texts are converted into machine-readable form in several main ways, including keyboarding, scanning with OCR conversion, and direct purchase. Some online system databases receive direct feeds from wire services.

Keyboarding

Keyboarding means hiring data entry clerks to key the text directly into the computer. For in-house databases, it is the most common method when the documents going into a database are generated locally. Surprisingly, until quite recently this was also the most common method used by commercial database publishers to convert printed publications into full-text databases. Such rekeying can be labor intensive, but, with verification, accuracy rates can be high and costs low. Throughout the 1980s it remained less expensive to pay overseas data input operators to rekey the text portions of magazines, journals, and books directly from the printed source than to pursue more automated options. Keyboarding is still used by commercial database producers to convert texts that are printed on thin newsprint, that include many special characters, or that use unusual typefaces. Keyboarded databases usually include text only.[25]

Scanning

In the 1990s, the most common method for creating textual databases from printed publications has changed to scanning. Scanning is used to create image files from many different types of original texts. Scanned image files reproduce the document in page image, but do not produce fully searchable text. Images are compressed, but, even so, have large storage requirements if the resolution is to be acceptable.

For example, OCLC estimates the page image storage requirements for one year of a full-text database that contains 10,000 journals (5,390,000 pages) to be approximately 540 more gigabytes than would be required for ASCII text, graphics, and indexing.[26] On viewing, software must decompress an image file, and response time may suffer.

To create ASCII text from a scanned text involves first scanning the printed document, then converting the characters with optical character recognition (OCR) software.[27] The newest generation of OCR software and scanning equipment, including the latest Calera and Kurzweil machines, produce high accuracy rates at about the same costs as offshore keying and lower costs than in-house keying. Combination image/ASCII files retain graphical material in compressed image files that are linked to appropriate sections of the ASCII text.[28]

Direct Purchase

Another textual database building option is to purchase typesetting tapes directly from the print publisher. Almost all publishers create their printed publications with computer typesetting programs that dictate page layout, arrangements, and typefaces. Since the text is in machine-readable form, such typesetting files can be converted to a format compatible with the database system software. This option is not used as often as the other two simply because publishers use a wide variety of typesetting software and markup conventions. Conversion programs are required for each different package, and, if a publisher changes the format, programs must be rewritten by the database producer. Although a markup standard exists (SGML—Standard Generalized Markup Language), it is not yet in widespread use.[29]

Wire Service Feeds

Some wire services provide a direct online feed to the online systems that load their information. Wire service feeds may come in 24 hours a day, just as they do to major newspapers, and they are used for online databases that are updated frequently. Online system conversion software is created to allow the feeds to be immediately loaded into the online system files.

Storage Considerations

Full-text databases, especially if they include images, have massive storage requirements. The OCLC Office of Research estimates just *one page* of mixed text and graphics for an ACS chemistry journal can take between 8.5 to 100 kilobytes of storage. SGML-coded ASCII text takes the least (8.5 KB); SGML text plus extracted graphics takes 10 KB; text, graphics, and indexing take 25 KB; and page images take 100 KB.[30]

Careful design of software for machine indexing and compression is needed, especially for CD-ROM full-text databases, to ensure acceptable response times when searching this inherently slow medium. Design of software for CD-ROM full texts is discussed in Witten, Bell, and Nevill.[31]

SEARCH AND DISPLAY REQUIREMENTS FOR FULL TEXT

Many of the commercial online systems developed their software with bibliographic databases in mind. Almost all were developed 10 or more years ago, with a continuing patchwork of revisions as full-text databases got larger and more plentiful. A user of commercial online systems may not have much to say about desired features for full-text databases—the user must accept what is offered by the system. Still, several of the systems have a number of features that enhance full-text searching; these features have been identified through a combination of research and experience. Almost all of these features have been tested at one time or another.[32]

There may be more variation and more innovation in CD-ROM software. CD-ROM software is not constrained by the bandwidth limitations imposed by online systems' communications lines. In addition, CD-ROM developers can often provide search or display features geared to one specific text, such as the National Geographic Mammals Encyclopedia or the Compton's MultiMedia Encyclopedia. The tyranny of tradition that inhibits online development may not hold true in the CD-ROM world, although neither may the common sense that comes with years of experience. Jacso discusses in detail criteria and methods for evaluating CD-ROM software.[33]

Software for databases created from texts distributed on diskette or magnetic tape or from in-house texts have even more variation. The purchaser has more choices of search and display features, but also must carefully evaluate each package. Criteria and methods for evaluation are detailed in Tenopir and Lundeen.[34]

Software features for full-text databases can be separated into three levels: (1) minimal for all textual databases—all of these features should be present; (2) important for full-text databases—most of these features should be present; and (3) useful for better text retrieval—some of these features may be present to enhance search and retrieval.

Level One Software Features

Level one features have become de facto standards for textual files, whether they are bibliographic or full text. They are accepted features on all of the major online search systems and almost all of the CD-ROM systems. They are available on many software packages for

in-house databases, although they may not be present in software created for library automation rather than for information retrieval.

Level one features are possible because of how the information is structured in most text-retrieval systems. All of the major online system software and most of the CD-ROM software and software for in-house databases rely on the inverted index file structure. Inverted index systems generate a separate dictionary index (or indexes) of all searchable words in each record in a database. Searches are made in the inverted files and postings are reported to the searcher. The system goes to the actual full records from pointers in the index only when a user wants to view a record. Inverted indexes facilitate certain basic search features, including inverted index display, truncation, set building, Boolean logic, proximity operations, and field searching.

The ability to view words in the indexes helps a user to develop a search strategy and to check the consistency of form of entry. Only the words or phrases in the index are available for searching; if, for example, pages and volume information are not put into the index, they cannot be searched even though they may be present in the records. Each system has its own rules for inverted index creation, including how it treats punctuation, but the ability to view the index makes these rules evident to the user.

There may be one inverted index that includes all of the searchable words or phrases from anywhere in the records, or there may be many separate indexes specified by fields. For example, a system may place words from all subject-related fields, such as title, descriptors, and full text, into one index (Basic Index on DIALOG), but create separate indexes for all other fields. For full-text searching, either method is acceptable as long as a user may specify a given field as needed. For example, the searcher should be able to restrict a search to just the date field or just the title field.

Truncation (word stemming) helps resolve some of the problems that arise from inconsistent word forms and entry. Truncation is especially important with full-text files, since much variation can be expected to occur in texts. Minimally, truncation should be user-specified right-hand stemming so that users can retrieve singulars, plurals, and other ending variations (such as *-ed, -ing*, and *-tion*).

Boolean logic is still the common foundation of working retrieval systems and is useful in certain cases for full-text files. At a minimum, systems should offer the three basic Boolean operators: OR, AND,

and NOT. Boolean OR operations are crucial with full texts, especially those that do not also have controlled vocabulary indexing. Different authors may use different words for the same concept; the searcher must be able to specify *Native Americans OR American Indians*. Boolean AND operations are best to link text words with other fields such as publication year, author's name, or journal name. The Boolean NOT operator is useful to eliminate known irrelevant items such as articles by a certain author, specific document types such as fiction, or phrases such as *IBM AND NOT IBM-PC*. Searchers should be able to specify the order of execution of Boolean operators using nesting (parentheses).

With full-text searching, proximity operations are crucial and are recommended instead of the Boolean AND operator to link concepts within texts. Word adjacency is a minimal capability. It may be taken care of automatically when a user enters two or more words separated by blanks, or it may require a special operator such as ADJ. Also desirable are: (1) the ability to specify within a certain number of words (*gone* within 2 words of *wind*), and (2) the ability to specify word order (*gone* must precede *wind*).

Finally, set building is necessary for a search to be truly interactive. Set building allows searches to be modified, narrowed, or broadened. Full-text searches often retrieve many records, so modification is essential. Different systems handle set building in different ways, from creating a set for every term entered (DIALOG), to creating a set just for the product of an entire line of input (Mead and BRS). Regardless of how it is handled, full-text searchers should be able to refer back to previously created sets to modify, narrow, or broaden a search.

Minimum display features for full text require flexibility in the amount of information viewed or printed. Users should be able to view just titles or a small part of a text, selected portions of a text, or the complete textual records.

Level Two Software Features

The second group of software features may or may not be present in software designed for bibliographic databases, but are important features for full-text files. Many are extensions of the basic level one features.

In addition to user-specified right-hand truncation, a full-text system should do a certain amount of right-hand word normalization

automatically. Regular forms of plurals and possessives should be automatically retrieved when singular forms are entered and vice versa. Ideally, a user should be able to turn this off, however. For example, the user may want to retrieve *electronic journals*, but not *electronics journals*.

To achieve better precision, a user should be able to specify how many characters should appear after a right-hand stem. For example, if a user wants only *compute, computes, computer*, or *computers*, but not *computerization*, the user should be able to specify two characters only following the stem *compute*. If all variations are wanted, an unlimited stemming capability should be present.

Left-hand truncation is important only for certain types of documents. A chemistry collection is better searched, for example, if left-hand truncation can retrieve all phenol compounds, such as nitrophenol, dichlorophenol, and so forth, in one statement. Because of the way inverted indexes are created and searched, left-hand truncation is more difficult to achieve, and it is not available on most large commercial online systems. In-house software may offer left-hand truncation by sequential scanning or by creating a separate inverted index with the words spelled backwards.

The last truncation variation useful for full texts is the ability to do internal truncation. One variation on this capability is *wild card* replacement, which requires a one-to-one relationship of characters. For example, *M*N* will retrieve man or men, but not every word that starts with *M* and ends with *N*. True internal truncation allows any number of characters to appear between the specified letters. This is especially useful for retrieving spelling variations (for example, *colour* or *color*). Ideally, the user should have the choice of either wild card or internal truncation.

Many common spelling variations should be handled automatically by the full-text system, rather than requiring the user to search for all possible variations. Simple things such as differences between British and American spelling are easily identified in dictionaries and should be automatically searched by a full-text system that includes American and British Commonwealth documents.

In addition, standard abbreviations should be automatically matched to the spelled out versions. Months, days, years, and numbers can be normalized to allow automatic retrieval of *January 1988* if a user inputs *1/88*. More complex equivalencies such as automatic matching of abbreviations to spelled out versions of

government agencies and organizations (for example, *F.B.I.* equals *Federal Bureau of Investigation*) are offered by Mead. Software for in-house full-text databases should allow the creator to specify equivalencies and build term-synonym dictionaries.

As mentioned earlier, proximity operations that allow a searcher to specify the relationships between words are crucial in full-text searching. In addition to the minimum capabilities of word adjacency and searching within a specified number of words, software for full text should recognize some of the grammatical structure of texts.

Searching for words within the same grammatical paragraph or the same sentence allows a searcher to take advantage of the inherent structure of texts. Presumably an author will put words that represent intersecting concepts within a sentence or paragraph. Consequently, searching within those units will allow more precision. Ideally, a user should be able to search for words within a specified number of paragraphs or sentences, although this feature is not so commonly available in online or CD-ROM systems.

Some full texts, such as research reports or newswire stories, have inherent structure and writing style. For these types of documents, there are other useful proximity operators, such as the ability to specify which portion of a document (e.g., both words within the lead paragraph or within the conclusions). For books, specifying within a chapter is useful.

With this range of proximity operations, full set building (i.e., the ability to modify previously created sets to narrow or broaden a search) provides the most interactivity. For example, a searcher who retrieves too little from linking words within the same sentence should be able to easily respecify within the same paragraph. If too many false drops arise from words within the same paragraph, a user should be able to limit the search to words occurring only in the conclusion or introduction paragraphs. Marchionini found with school students that, to be most successful, full-text searching should allow maximum flexibility and interactivity.[35]

Displaying lengthy full texts requires a wider range of display features than displaying bibliographic information or short texts. Most important is the ability to display only the portions of the texts that contain the search terms. Such KWIC (Key-Word-in-Context) features are common in full-text software, with variations on how much text is displayed surrounding the search terms. Most systems display a 25 to 50 word text window; ideally, users should be able

to enlarge the window as they view any document. Search terms should be highlighted.

To facilitate browsing and reading of electronic texts, users must be able to move around fully in retrieved documents. This means paging back and forward, enlarging KWIC windows at will, viewing next paragraphs or sentences, and skipping to specified sections or pages of documents. This is more common with CD-ROM and in-house software than with online systems.

Level Three Software Features

All of the software features mentioned so far are widely available in online, CD-ROM, and in-house systems. Although not every variation mentioned is available on every system, most features are widely available on systems that cater to full-text retrieval. Level three features, on the other hand, are less common. They provide a shopping list of special features that enhance full-text search and retrieval, but that are rarely all available on any one system.

Hypertext links are being widely implemented on CD-ROM full texts, especially encyclopedias and similar reference books. Hypertext provides an alternative to traditional indexing by building links between related concepts, documents, parts of documents, or files. In a hypertext encyclopedia, for example, a user may be able to view related articles that contain further information about a topic discussed in the current article, or the user may be able to view related pictures. Hypertext is popular in the Macintosh environment with the HyperCard program and, on a larger scale, is an important part of the ACS/OCLC CORE Superbook project.[36]

Relevance feedback involves using a user's judgment of a document's relevance to find additional similar documents. A user may mark useful documents, and the relevance feedback system will use some algorithm to find other documents that are "like" those. One simple relevance feedback method is to use word frequency of the relevant documents to locate documents that contain a certain percentage of the same words with similar occurrence rates.[37]

Many of these extended features provide ways to increase precision of searches. Word occurrence information with ranked output allows the most potentially relevant documents to be displayed first. The most basic version of this feature is to have the software calculate how often the search terms occur within each document. The

documents with the most occurrences will be displayed first. More complex algorithms for ranking output exist, but research has indicated that all algorithms seem to work equally well for full-text retrieval.[38]

Other modified Boolean search methods serve to increase recall using a variety of partial-match techniques. One simple method is to allow partial membership in a set; documents will be retrieved even if all concepts linked by ANDs are not present. Those with all concepts may be displayed first, followed by documents in descending order by how much of the search query they contain. Sound-alike searching allows items to be retrieved that contain words similar to input search terms that are not exact matches. For example, if a user searches on the name *Brown*, a sound-alike system may also retrieve *Browne* and *Braun.*

Other partial-match methods, although not yet widely available in commercial systems or software, are being added to some software for the creation of in-house databases. These include fuzzy sets, probabilistic retrieval, and vector space retrieval.[39] Belkin and Croft discuss all of these search methods and more.[40]

Display with full text is still at its infancy. Due to the limits of standard phone lines, online full-text files are almost completely ASCII text only. They are unaesthetic, without the typefaces, spacing, and page layout that make printed texts so attractive. (*The Online Journal of Clinical Trials* is a noteworthy exception.) Without graphics, sidebars, photographs, or special characters they include only part of what makes many texts valuable.

CD-ROM and in-house systems do not have the limitations of online systems and should offer many more attractive display features. At setup, the librarian should be able to specify type styles, size, and colors for full-text display. Texts should be displayed with sections and subsections set apart through appropriate headers and spacing. Users should be able to move between sections.

Finally, images are becoming a part of many CD-ROM texts and are supported by some in-house software packages. Hypertext links between text and images create a hypermedia database that makes the most efficient use of the power of ASCII text and linked image files. Due to more efficient compression algorithms, multimedia products can now also include moving images as well as still images. An extended discussion of compression for CD-ROM full texts is beyond the scope of this chapter.[41]

All search, retrieval, and display features must, of course, be balanced with ease of use and appropriateness for the users. Sometimes there is a trade-off between power and ease of use in any software package. Features that are confusing or of no use to the users of a system hinder rather than help the retrieval process.

SOFTWARE FOR IN-HOUSE DATABASES

There are a growing number of powerful and friendly software packages that allow libraries to create their own full-text databases with texts generated internally or from texts leased or purchased from publishers. Table 2 shows a partial list of these. Even more so than online or CD-ROM full-text systems, software for creating in-house databases has great variety. Not all packages are useful for all types of texts, so it is useful to categorize them according to their general characteristics.[42]

Structured Text-Retrieval Packages require fields and field characteristics to be specified before records are added to the file. Usually there is a configuration module that is used for this initial setup. Field structure allows for faster and more precise searching,

Table 2. Selected Software for In-House Textual Databases

askSAM (for micros)
Basisplus (for minis and mainframes)
BRS/Search (for micros, minis, and mainframes)
CAIRS-TMS Information Retrieval Package (for micros and minis)
Concept Finder (for micros and minis)
Concordance (for micros)
Excalibur (for micros and minis)
FolioViews (for micros)
Fulcrum (for micros)
GOfer (for micros)
Hypercard (for the Macintosh)
IBM/STAIRS (for mainframes)
Inquire/TEXT (for mainframes)
Lotus Magellan (for micros)
Personal Librarian (for micros and minis)
Sonar Professional (for the Macintosh)
Topic (for micros, minis, and mainframes)
ZyINDEX (for micros)

and it provides more control over formatting of output. Structured text-retrieval packages are traditionally used for bibliographic data and are the most common type in use in libraries. Commercial online systems are structured, so they are familiar to online searchers. Of the packages listed in Table 2, Personal Librarian is of the structured type.

Unstructured Text-Retrieval Packages are less common in a library setting. They require no initial setup and accept incoming text files without any field structure. These packages often recognize inherent structures of text, such as sentences and paragraphs, but they do not recognize fields. This means they lack the capabilities to search, sort, or customize output based on field criteria, but they are good for free-text searching of texts. Unstructured text-retrieval packages are especially suited to managing existing unfielded word processing files or files downloaded from a variety of incompatibly fielded databases. If texts already exist in machine-readable form, they can be used to create a full-text retrieval system very quickly. Unstructured text-retrieval packages include ZyINDEX and Lotus Magellan.

Combination Text-Retrieval Software are becoming the most common option and offer the advantages of each of the other two categories. They support defined field structure in addition to unstructured text. Thus, standard bibliographic information, such as author, title, date, and source, can be put into specified fields that can then be searched, sorted, or output. Textual portions of documents can be treated as unstructured text to be searched with more powerful search features. Combination packages include Concept Finder, Concordance, and Topic.

Libraries can use these packages to manage their locally generated texts, such as pathfinders, bibliographies, reports, manuals, and curricular materials. Optionally, within the provisions of copyright regulations, they can download full texts from online or CD-ROM systems or lease machine-readable versions of full texts directly from publishers for loading on an in-house system.

CONCLUSION

Whether they are accessed through a commercial online search service, leased or purchased on CD-ROM, magnetic tape, or diskette,

or created locally, full-text systems and files are becoming an important part of library services. A bibliographic database without full text support solves only half of the retrieval problem. Paper-based collections take up space and are getting too expensive; paper-based document delivery systems waste natural resources and provide documents that cannot be manipulated. Full-text files and systems will increasingly help solve the information retrieval problem.[43]

NOTES

1. *Directory of Online Databases*, vol. 12, nos. 3 and 4 (Detroit: Cuadra/Gale, 1991), viii.
2. *Directory of Online Databases*, vii.
3. *Online Database Selection: A User's Guide to the Directory of Online Databases* (New York: Cuadra/Elsevier, May 1989), 18.
4. Carol Tenopir and Jung Soon Ro, *Full Text Databases* (New York: Greenwood Press, 1990); and Carol Tenopir, "Users and Uses of Full Text Databases," in *Proceedings of the International Online Meeting, London, December 1988* (Oxford: Learned Information, Ltd., 1988), 263-270.
5. Martha E. Williams, "The State of Databases Today: 1992," in *Computer-Readable Databases: A Directory and Data Sourcebook*, 8th ed. (Detroit: Gale Research Inc., 1992), xi-xxi.
6. Williams separates databases into "word-oriented, number-oriented, image, audio, electronic services, and software." Word-oriented, which make up 72% of the databases listed in the directory (4,661), include "bibliographic, Patent/Trademark, Full Text, Directory, Dictionary, and other."
7. Carol Tenopir, "Hybrid Databases," *Library Journal* 117 (1 February 1992): 64-66.
8. Ruth M. Orenstein, ed., *Fulltext Sources Online: For Periodicals, Newspapers, Newsletters & Newswires* (Needham Heights, MA: BiblioData, 1992).
9. *Fulltext Sources Online*, iii.
10. *Fulltext Sources Online*, v.
11. Ibid.
12. Tenopir and Ro, *Full Text Databases*, 16.
13. Susan Bjorner, comp. and ed., *Newspapers Online* (Needham Heights, MA: BiblioData, 1992).
14. Bjorner, *Newspapers Online*, 1. Inclusion criteria is given in detail in the front matter.
15. Ruth M. Orenstein, *Fulltext Sources Online*.
16. Information given to me by Martha E. Williams and available in her quarterly report: *Information Market Indicators* (Monticello, IL: IMI).
17. For more examples, see: Paul T. Nicholls, CD-ROM *Collection Builder's Toolkit: 1992 Edition* (Weston, CT: Eight Bit Books, 1991).
18. *Directory of Portable Databases* (Detroit: Cuadra/Gale, October 1991).

19. Nicholls, *CD-ROM Collection Builder's Toolkit.*

20. Lorrin Garson et al., "CORE: The Chemical Online Retrieval Experiment," in *Annual Review of OCLC Research: July 1990-June 1991* (Dublin, OH: OCLC, 1991), 32-33.

21. John Price-Wilken, "Text Files in Libraries: Present Foundations and Future Directions," *Library Hi Tech* 9, no. 3 (1991): 7-44.

22. For more information on several projects and an excellent overview of electronic books, see: Reva Basch, "Books Online: Visions, Plans, and Perspectives for Electronic Text," *Online* 15 (July 1991): 13-23.

23. Bruce Flanders, "On-Line Books: An Advanced Technology Electronic Library System," *Computers in Libraries* 12 (January 1992): 44-47.

24. Michael Strangelove and Diane Kovacs, *Directory of Electronic Journals, Newsletters and Academic Discussion Lists*, 2nd ed. (Washington, DC: Office of Scientific and Academic Publishing, Association of Research Libraries, 1992).

25. Ernest Perez, "Low-Budget, Cost-Effective OCR: Optical Character Recognition for MS-DOS Micros," *Library Software Review* 9 (July-August 1990): 209-217.

26. Stuart Weibel, "The CORE Project: Converting a Large Document Collection for an Electronic Library Project." (Presentation at the American Society for Information Science Annual Meeting, 30 October 1991).

27. Lori Grunin, "OCR Software Moves Into the Mainstream," *PC Magazine*, 30 October 1990, 299-356.

28. Clyde W. Grotophorst, "Keyless Entry: Building a Text Database Using OCR Technology," *Library Hi Tech* 7, no 1 (1989): 7-15.

29. Betsy N. Kiser, "Standard Generalized Markup Language: Why Reference Librarians Should Care," *Reference Services Review* 18 (Fall 1990): 37-40, 52.

30. Weibel, "The CORE Project."

31. Ian H. Witten, Timothy C. Bell, and Craig G. Nevill, "Indexing and Compressing Full-Text Databases for CD-ROM," *Journal of Information Science* 17, no. 5 (1991): 265-271.

32. Carol Tenopir, "Full-Text Databases," *Annual Review of Information Science and Technology* 19 (1984): 215-246; and Tenopir and Ro, *Full Text Databases.*

33. Peter Jacso, *CD-ROM Software, Dataware, and Hardware: Evaluation, Selection, and Installation* (Englewood, CO: Libraries Unlimited, 1992).

34. Carol Tenopir and Gerald W. Lundeen, *Managing Your Information: How to Design and Create a Textual Database on Your Microcomputer* (NY: Neal-Schuman, 1988).

35. Gary Marchionini, "Information-Seeking Strategies of Novices Using a Full-Text Electronic Encyclopedia," *Journal of the American Society for Information Science* 40, no. 1 (1989): 54-66.

36. Dennis E. Egan et al., "Hypertext for the Electronic Library? CORE Sample Results," in *Hypertext '91 Proceedings* (New York: Association for Computing Machinery, 1991), 1-14.

37. Nicholas J. Belkin and W. Bruce Croft, "Retrieval Techniques," *Annual Review of Information Science and Technology* 22 (1987): 109-145.

38. Jung Soon Ro, "An Evaluation of the Applicability of Ranking Algorithms to Improve the Effectiveness of Full-Text Retrieval. II: On the Effectiveness of