10-1-1993

# The New Generation of Online Search Software

Carol Tenopir
*University of Tennessee - Knoxville*

# ONLINE DATABASES

BY CAROL TENOPIR

# The New Generation of Online Search Software

MOST OF TODAY'S database searchers spend a lot of time learning the commands of a variety of systems, how to formulate queries, and the correct use of Boolean operators. Even with end user systems—whether online, CD-ROM, or locally loaded databases—reference librarians report an increased need for bibliographic instruction. Why is something that makes research so much faster so complicated?

Part of the problem is that most of today's online systems and many CD-ROM systems operate with essentially the same software developed for the first online systems 20 years ago. Although improved and sometimes rewritten in more modern programming languages, the major systems still reflect first-generation search techniques. They rely on exact match Boolean logic, structured commands or menu choices, and convoluted input syntax, features that may be advantageous to experienced searchers, allowing them to control the search process, but unsatisfactory for end user systems.

## Software improvements

"Innovations in Text Retrieval Software" (Online Databases, *LJ*, June 1, 1992, p. 94, 96) discussed the many improvements in text retrieval software that evolved from the information retrieval research laboratories into the software market for end users. More innovative search techniques such as natural-language input, relevance or word frequency ranking, and automatic thesaurus features are appearing in commercial products. A little more than a year ago, most innovations in retrieval software were available only for in-house databases—in software packages such as Topic, Personal Librarian, and ZyINDEX.

The new generation now has spread to the commercial online and CD-ROM environment. Thus far, the Personal Librarian (PL) software and Westlaw Is Natural (WIN) are two of the most successful representatives of the online second generation.

PL is still available as a stand-alone program, but it is now used as the search software for many CD-ROM products and as a search engine for several online hosts. WIN provides an alternative method for searching Westlaw's many online legal databases. This month, I will profile Personal Librarian; next month I will focus on WIN.

## Personal Librarian

PL, known originally as SIRE, was developed over a decade ago by Matthew Koll and his colleagues at Syracuse University as an experimental retrieval system. It was first offered in 1983 as a commercial software product for creation of microcomputer-based in-house databases. (Personal Library Software, Inc., 2400 Research Blvd., Rockville, MD 20850; 301-990-1155)

PL was the first commercial application to offer a number of search features. Natural-language input is perhaps the most obvious to users. Instead of entering a search statement in the correct and stilted Boolean operator form as with other systems, users can input any sentence that describes their information need.

Thus the statement "I need information about the effect of last year's hurricanes in Florida and Hawaii on tourism" will work as an input string. PL does not use any artificial intelligence or other techniques for interpreting the meaning of the statement, nor does it match words to a thesaurus. Instead it just eliminates stop words from the string, then ORs together all the remaining words. A concise statement loaded with meaningful words will thus work best.

At first glance, a Boolean OR between every word may seem like sure disaster in full-text databases. It will retrieve many documents, but it works because the PL software uses relevance ranking. Retrieved documents are ranked in order of likely relevance; users can browse through documents until the relevance diminishes or their information need is met.

PL's relevance ranking works by a complex formula that takes into account how many of the words occur in each document, how many times each word occurs, each document's length, and how often each term occurs in the entire database compared with how many times it occurs in each document. Tests show that although the formula is not perfect, it does predict likely relevance much of the time.

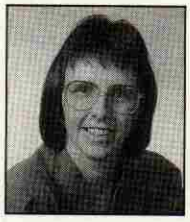
Another notable search feature is the ability to use a relevant document as a query. When a "good" one is found, users can request similar documents. PL examines word occurrence in the relevant document to search for documents with similar word frequencies.

Although PL has no standard thesaurus features, it will locate words that occur frequently with another word. A user can expand any search term to locate related terms in the database and then search the additional terms automatically.

PL is still known as a package for creating in-house databases, with DOS, Windows, Mac, UNIX, and VMS versions. Stand-alone and networked versions make it popular with both small companies and large corporations, including Apple Computers and Unisys. In the last year or two, the larger information industry has taken note of its strengths as well.

## CD-ROM products

You may not even realize you are using PL when you purchase a CD-ROM product from the U.S. Government Printing Office or from a company such as Grolier. That's because it has been adapted as the search engine by a variety of CD-ROM developers, many of whom put their own interface



Carol Tenopir is Associate Professor at the School of Library and Information Studies, University of Hawaii at Manoa, Honolulu

onto the powerful PL search engine. According to Richard Black, VP of Business Development at Personal Library Software, "several CD-ROM developers use our search engine buried deep in the bowels of their products with their own custom interfaces." The licensing agreement, however, does require a copyright statement on the disc and on the packaging.

PL is now the search engine for several popular CD-ROM products. Among the notable titles are the U.S. Code from the U.S. House of Representatives, the Guinness Multimedia Disc of Records from Grolier, the American Memory Project from the Library of Congress, the Laws of Washington and Oregon from CD-Law, a series of British newspapers from the Financial Times (including the *Economist*, the *Financial Times*, the *Daily Telegraph*, the *Independent*, and others), and McCarthy's compilation of full-text business articles from European business newspapers and magazines. Both text-only and multimedia CDs use PL.

### Online expansion

PL has expanded into the online arena as well. One of the first online services to use PL as its search engine is Washington Alert, the online system from Congressional Quarterly. Washington Alert developed its own search interface to work with the PL search engine.

Washington Alert debuted with the PL search engine almost five years ago. The service includes approximately 20 databases that together provide comprehensive coverage of Congress and Congressional actions. Notable databases include the full text of the *Congressional Quarterly Weekly Report* and other newsletters; full texts of all bills and all committee reports; bill-tracking; and information about roll call floor votes, members of Congress, committee actions, and schedules.

America Online, the popular end user online service, announced in May 1993 that PL would be used as the search engine for an improved version of its online system. Again, America Online uses its own interface, but the search powers that make retrieval work are now from PL.

America Online is phasing in PL on its system, database by database. The first database searchable with PL is America Online's member directory. This database gets more use than you may expect, since sending and receiving messages is America Online's most popular service. America Online will continue to phase in the conversion of its full-text databases. It now offers news, sports, weather, stock market, and software databases, in addition to its popular E-mail and conferencing functions.

America Online also serves as a gateway to many databases on other online systems in addition to being a resident host. The gateway databases will continue to use the search engines of the systems that house them, but with the America Online interface. This may be confusing to users, since the connections are made transparently.

DataTimes announced in June 1993 that PL would replace BASIS as its newspaper archives software for minicomputer systems. (It did not announce it will be using PL for its commercial online system.) DataTimes has two businesses—as a vendor of the online system and as a provider of internal library systems to newspapers. DataTimes had been selling micro versions of PL for Macintosh and Windows for its newspaper library systems clients. Starting in June, it will also offer the minicomputer version of PL to replace BASIS.

Dow Jones News/Retrieval (DJNR) may be the best-known online system to convert to PL. This past spring, rumors began to surface that DJNR would switch from IBM equipment to DEC computers and from IBM STAIRS to PL. Although the "official" announcement was delayed, PLS and Dow Jones have not denied the rumors.

Informed sources say PL has been selected as DJNR's next generation search engine, with an interface to be developed by Dow Jones. Conversion has not yet begun, but expect something in 1994.

### DIALOG buys into PLS

DIALOG Information Services, Inc. is involved with Personal Library Software in a slightly different vein. DIALOG announced in July a significant minority investment in Personal Library Software. It does not plan to replace the DIALOG software with the PLS system, but DIALOG will be represented on the PLS Board of Directors and the two will jointly develop new products.

One of the first applications areas is likely to be with DIALOG's CD-ROM products. PLS has been successful with multimedia CD-ROMs, an area in which DIALOG has done little to date. We can expect some multimedia and image/document management systems from DIALOG with Personal Library Software.

DIALOG and PLS is a marriage of market presence with forward-looking technical expertise. According to Patrick Tierney, president and CEO of DIALOG, "PLS has excellent technology, strong performance, and knows where the information business is heading. Together, our companies will develop synergistic products that will integrate text and image-based information from internal and external databases and deliver mission-critical data directly to users' desktops."

DIALOG denies plans to abandon its current CD-ROM search software, a program that has gotten consistently positive reviews. How the current system and PL will interact and whether they will be used for separate databases or different types of applications is still unclear. Perhaps more than anything, this announcement shows that Personal Librarian has arrived as a search engine and company to be taken seriously in the larger information industry marketplace.

### Why now?

Why is PLS attracting so much attention in the information industry now after a decade of existence? PLS VP Black speculates that "things are changing very quickly in the entire information industry. Information vending and information retrieval are moving out of the hands of professional searchers and into the public domain. That's the impetus." The time is finally right for innovations that make software easier to use and go beyond the techniques we've been using since the early days.

### Online/CD-ROM '93

I will moderate a session at the Online/CD-ROM '93 meeting in Washington, D.C., November 1, that will debate the relative effectiveness of "old-fashioned" command-based Boolean logic systems with natural-language and relevance ranking systems. Speakers will represent both sides of the issue.

---

Next Month: More on the new generation of online systems with a look at Westlaw Is Natural (WIN).