10-1993

Full Text Searching On Major Supermarket Systems: DIALOG, NEXIS, and DataStar.

Carol Tenopir
University of Tennessee - Knoxville

Sharon Berglund

Recommended Citation

Carol Tenopir and Sharon Berglund. "Full Text Searching On Major Supermarket Systems: DIALOG, NEXIS, and DataStar." Database 16 (5) (October 1993): 32-42.

Full-text searching on major supermarket systems: DIALOG, Data-Star, and NEXIS

Authors: Carol Tenopir and Sharon Berglund
Date: Oct. 1, 1993
From: Database(Vol. 16, Issue 5.)
Publisher: Information Today, Inc.
Document Type: Product/service evaluation
Length: 3,502 words

Abstract:
The largest online vendors of full-text sources to general libraries, DIALOG, Data-Star and NEXIS, are described and compared. DIALOG offers the largest number of full-text sources and the best set-building search techniques, but does not offer automatic searching of inconsistent text such as plurals or spelling equivalencies as do the others. Data-Star has the strongest European focus and information on chemistry, pharmacy and medicine. NEXIS has the most user-friendly search language and superior display features and the greatest number of unique titles.

Full Text:
Although Fulltext Sources Online lists 19 online systems that offer significant numbers of full-text sources [1], just a few systems dominate use in general (nonlegal) libraries. DIALOG, BRS, and NEXIS were the most used online systems for full-text searching in the general library and information center marketplace for many years [2]. In 1993, BRS announced they would concentrate only on the medical information market, thus effectively removing them from the general library marketplace. DIALOG and NEXIS now dominate use, but other supermarket systems are vying for the third place position of BRS in the general full-text market. Closest is Data-Star, the European online system recently acquired by DIALOG's parent Knight-Ridder, and now second only to DIALOG in general full texts. Figure 1 shows how many fulltext sources are available on the systems covered by Fulltext Sources Online. (Westlaw, the system with the most full-text sources, is targeted to the legal market so is excluded from this discussion.)

Martha E. Williams has been tracking the online database industry for years. Her figures show that in only seven years, the number of full-text sources has grown nearly 600% (Figure 2). In her annual "State of Databases Today," Williams shows that full-text databases now make up 47% of the database total [3] (Figure 3).

The general interest systems share many full-text sources, yet each also has unique offerings. The commands and features useful for full-text searching also have many similarities, but vary in power and implementation in the systems. Searchers need to compare and select which is the best source when each offers the same texts, and also how to change a successful search strategy when going from system to system for unique offerings. The similarities, differences, and full-text features of the three most used supermarket systems for full text are examined here.

DIALOG

DIALOG has been the most used online system in libraries and information centers for many years and is familiar to most professional online searchers. This is not likely to change in the near future, because all the graduate schools of library and information science accredited by the American Library Association offer students a chance to learn DIALOG [4].

In the late 1960s to early 1970s, DIALOG was developed as a bibliographic retrieval service- It was not until 1983 that DIALOG added its first full-text database (Harvard Business Review Online). Since then, DIALOG has launched an aggressive campaign to add more full text. With the purchase of DIALOG by Knight-Ridder in 1988, DIALOG began adding all of the Knight-Ridder newspapers in full text, in addition to many others.

DIALOG now offers the most general interest full-text databases of any commercial online system. Approximately 120 of DIALOG's databases are full text in whole or part, as seen in Figure 4. Some full-text databases on DIALOG consist of a single text source, such as the Los Angeles Times or Harvard Business Review. Others are made up of hundreds of sources, such as Magazine ASAP or Trade & Industry ASAP. Taken together, DIALOG's full-text databases provide access to over 2,600 full-text sources. These sources include journal articles, magazine articles, newspapers, newswires, reference books, newsletters, some government documents, reports, conference papers, and more.

Since DIALOG was originally designed for bibliographic database searching, most of the full-text features have been added in the last few years. These now include both search and display features as shown in Table 1. As an intermediary system, the DIALOG software puts most of the burden of inconsistent texts on the searcher. Unlike Data-Star and Mead, there are, for example, no features for automatic singulars/plurals or automatic spelling equivalencies.

Data-Star

Until March 1993, when it was acquired by Knight-Ridder (a U.S. company that also is the parent of DIALOG), Data-Star was owned by RadioSuisse. Although accessible in North America, Data-Star had been a European-owned and European-based company for its almost twelve years of existence. According to a 1991 interview with then Managing Director of DataStar, Heinz Ochsner, a decade ago in 1981 RadioSuisse "was in the business of managing and reselling access to packet switching from Switzerland to United States networks" [5]. In an effort to incorporate some value-added services for their European customers, RadioSuisse entered into an agreement with Predicasts, acquired a license to use BRS software in Europe, and began the Data-Star service. Data-Star developed into the premier European online provider.

Until 1991, the majority of DataStar's income (90%) was from its European clientele. At that time DataStar began expanding its market thrust, with the hope of reaching a goal of having 50% of its revenue come from an expanded U.S. market. The full effect of Knight-Ridder's acquisition of Data-Star is yet to be known, but for now Dialog Information Services and Data-Star will remain separate online services.

Data-Star currently has over 50 separate full-text databases, most of particular value for their European focus. They are also strong in company intelligence, chemistry, pharmacy, and medicine. In all, the Data-Star service databases provide access to over 2,100 full-text sources. A breakout of the types of databases offered by Data-Star is shown in Figure 5.

NEXIS

NEXIS is a companion system to Mead Data Central's long-established LEXIS system. LEXIS grew out of the Ohio State Bar Association OBAR legal reference system, developed in the late 1960s. LEXIS was first commercially available in 1973, offering just Ohio and New York codes and cases, the United States Code, and selected federal cases. Today it includes materials from all fifty U.S. States and many other countries, plus specialized state and federal legal libraries. Along with its major competitor Westlaw, LEXIS is used by every law school and nearly every law library and legal firm in the U.S.

NEXIS was added in 1979 to complement the legal texts on LEXIS. NEXIS is a news system in a broad sense; it includes full text of magazines, journals, newspapers; and newswires and some reference books. Databases are counted differently than on DIALOG and Data-Star, as nearly every title can be searched separately (e.g., Journal of Small Business Management), grouped on an ad hoc basis by the searcher (e.g., Journal of Small Business Management (JSMBUS) with New Product News (NPRODN)), or searched in designated groupings (e.g., MAGS or PAPERS). In addition, all NEXIS publications can be searched simultaneously with the OMNI designation. As of 1993, over 1,400 sources are accessible on NEXIS. LEXIS and NEXIS together provide access to over 40 million documents.

LEXIS and NEXIS were rounded for a different audience and for a different purpose than DIALOG and Data-Star. End-users have always been the primary focus and the systems have always concentrated on full text. Although NEXIS includes a handful of bibliographic and directory databases, the bulk of sources are full text. The Mead (LEXIS/NEXIS) search language is therefore considered by many to be friendlier, with a greater range of search features that are geared for effective full-text searching.

OVERLAPPING DATABASES

A search of the online version of Orenstein's Fulltext Sources Online indicates that the three systems in question have some unique, as well as many overlapping, sources among their database collections. According to Orenstein, 446 full-text sources appear in all three of the systems.

Though these titles are the same, they may be organized differently on the different systems. On NEXIS, for example, nearly all titles appear as their own database. On DIALOG and Data-Star, on the other hand, most of the full-text sources, with the notable exception of newspapers, appear in large multisource files, such as Information Access Company's Trade & Industry ASAP, Magazine ASAP, and Health Periodicals.

Fulltext Sources Online can be used to identify and further characterize these overlapping titles. It provides dates of coverage for each source title on each system, something that has been shown to vary considerably, depending on when a system first loads a full-text title or the terms of the licensing agreement with the primary publisher [6].

Even when coverage dates do not vary, the number of articles included in each online version of the overlapping title may vary considerably, as has been demonstrated by Pagell and others [7]. One version may include letters-to-the-editor and short articles, another may not. One version may count sidebars as separate articles, another may exclude them. Many of the overlapping titles are those that provide selective coverage online, rather than cover-to-cover inclusion.

The overlapping sources run the gamut of subject matter, as the selected sample of titles shows in Table 2. Every overlapping title was searched on each online system to identify how many articles are online. Table 2 shows that the number of articles included for each title often is quite different from system to system.

UNIQUE SOURCES

Unique sources can be calculated for these three by comparing which sources appear in just one of the three services, or uniqueness can be calculated across all 19 services covered by Fulltext Sources Online. When comparing just the big three supermarket services with each other, there are 364 unique sources on DIALOG, 280 on NEXIS, and 27 on Data-Star.

The relative figures for uniqueness change when comparing all 19 of the systems. NEXIS has by far the most unique titles with 204. These include unique ethnic newspapers and journals, as well as ABC television transcripts, legal periodicals, and local newspapers. DIALOG has 23 unique titles (mostly business, banking, tax, and local newspapers), and Data-Star has 20 unique titles (primarily European newspapers and business sources).

SEARCH FEATURES

Effective full-text searching relies on a combination of powerful search features and useful display features. Table 1 compares the search and display features that assist full-text searching on the three systems. All offer Boolean logic (at least AND, OR, and NOT), proximity operators, field qualification, set building, and truncation. They differ in how much power is provided with each of these features, and the specific commands to accomplish each. Only NEXIS and Data-Star offer language enhancement features such as automatic singular/plural searching.

PROXIMITY OPERATORS

Powerful proximity operators are even more of a necessity with full-text searching than with bibliographic or directory searching. Using a Boolean AND to link concepts in lengthy full texts can result in an unwieldy number of documents and an overwhelming number of false drops. Often full-text databases do not include descriptors, so proximity operators must be used to create ad hoc phrases for searching. Proximity operators that recognize the grammatical structure of documents (sentences and paragraphs) are useful, as are those that allow the user to specify and alter the distance or relationship between words.

Table 1 shows the major proximity operators for each of the three major general-interest systems. None offers a complete range of proximity operators. (Westlaw, the legal full-text system, probably offers the widest range of proximity operators.) DIALOG allows a user to specify within a paragraph, within a specified number of words in either direction, or word adjacency. NEXIS provides the latter two. Only Data-Star allows searching for words within the same grammatical sentence or same paragraph in addition to word adjacency. Data-Star does not, however, allow users to specify word order or how many words can intervene.

In addition to specified proximity operators, default values for a blank between words differ for each system and can affect search results. In NEXIS a blank between words defaults to word adjacency. This is the most user-friendly for full-text searching, especially when searching for a specific title or string of words that may appear anywhere in an article. Searching for "New York" in NEXIS will retrieve the phrase as intended anywhere in the document.

Blank spaces in DIALOG's full command system default the search to phrase indexed fields (usually descriptors or identifiers), effectively excluding the complete text from the search. If "New York" does not appear in the descriptor field in a DIALOG full-text database, it will not be retrieved without specifying (w) in place of the blank.

On Data-Star, if no proximity operator has yet been used in a statement, a blank defaults to the Boolean OR (a potential disaster in full-text searching). New York will look for New OR York! If a blank is used after another specified proximity operator in a Data-Star search statement, the blank will default to the previously used operator.

SET BUILDING

DIALOG is by far the most powerful of the three in set building. Since the searching of full texts is often a highly interactive, often almost serendipitous process, powerful set building is an advantage. With its Ss operator, DIALOG creates a set for every word entered and allows the searcher to combine or recombine words in any way. In a search for articles about trade agreements between Beijing and Hong Kong, a DIALOG searcher may start narrow, as illustrated:

SS (Peking or Beijing)(s)Hong(w)Kong(s) trade(w)agreement?

```
s1 888 Peking s2 2198 Beijing s3 4811 Hong s4 5063 Kong s5 4551 Hong(w)Kong s6 9972 trade s7 3322 agreement? ? s8 254
trade(w)agreement? ? s9 4 (peking or beijing)(s)hong (w)kong(s)trade(w)agreement?
```

If there are too few documents, the search can easily be broadened by recombining the sets to widen the relationship of the trade agreement concept (e.g., S s6(10N)s?) or by adding other synonyms.

Data-Star is next in power and flexibility, building a separate set for each search statement and allowing these sets to be combined or recombined as desired. If the example above was entered in one search statement in Data-Star, only a single set would be built (equivalent to DIALOG's single S command.) If the searcher broke it down into separate steps at input, all sets would be available for recombination, but the individual parts of each set (e.g., trade or agreement) would have to be reentered.

1_: peking or beijing

RESULT 3204

2_: hong adj kong

RESULT 5304

3_: trade adj agreement$1

RESULT 267

4_: 1 same 2 same 3

RESULT 8

NEXIS is the least flexible of the three in set building. NEXIS searchers cannot recombine previous search statements at will, they can be used only as a base from which to narrow a search with the MODIFY or FOCUS keys. All modifications must begin with a Boolean or proximity operator that will apply to the entire previously created set.

peking w/20 hong kong w/20 trade agreement

Your search request has found 7 STORIES through Level 1.

To MODIFY your search request, press the M key (for MODFY) and then the TRANSMIT key.

LANGUAGE ENHANCEMENT/ WORD EQUIVALENCIES

All of the systems allow a user to specify word stemming (truncation). Since full texts contain a variety of word forms and word endings, truncation is crucial for full-text retrieval. On DIALOG and Data-Star, the searcher can specify how many characters maximum should follow the stem, or can designate unlimited truncation. DIALOG does not go beyond this simple word stemming, however, in aiding searchers to retrieve all word variations.

NEXIS and Data-Star are more sophisticated than DIALOG in language enhancement/word equivalency features. Both NEXIS and Data-Star provide automatic singular/plural searching and British/American English spelling equivalencies. In the example above, NEXIS will automatically search for "agreements" when "agreement" is input (and "trades" for "trade"). If plurals had been turned on by the Data-Star searcher above, "agreements" would have been searched automatically.

In NEXIS, these features happen automatically and cannot be disabled by the searcher; the Data-Star searcher can turn the features off and on at will. Sometimes unexpected results occur in NEXIS (such as electronics journals when electronic journals was desired), but in most cases, the automatic feature takes the burden of remembering to truncate from the searcher. This works well for regular plurals, but most irregular plurals still have to be searched with an OR operator in both systems.

The European focus of many of Data-Star's databases makes the British-American spelling equivalent feature important. Neither the DataStar nor the NEXIS equivalencies are foolproof, however, because not every word is included. For comprehensive searches the searcher must still explicitly state every variation.

NEXIS also provides automatic equivalent searching for Chinese romanization schemes, compound words, abbreviations for federal agencies and other organizations, and other common abbreviations, such as days of the week (Mon/Monday) or numbers (3rd/third). In the previous search example, NEXIS retrieved all articles with Beijing OR Peking by just inputting Peking.

Searching acronyms will usually retrieve articles in which the spelled out version occurs, but the opposite does not always hold true. FBI will retrieve Federal Bureau of Investigation on NEXIS, but if the searcher inputs Federal Bureau of Investigation the system will not also search FBI. None of the automatic word enhancements are absolutely complete, so thorough full-text searches still require the searcher's input.

DISPLAY FEATURES

Perhaps because it was developed as an end-user full-text system, NEXIS offers the most useful display features for full-text searching. NEXIS function keys or commands will display citations, keywords in context (KWIC), or full articles. Previous or next pages of the same article can be requested as well. This ability to look at the previous or next page allows browsing on NEXIS much as a user would browse through print.

DIALOG and Data-Star both allow a user to view just citations (or specified portions of a citation), KWIC (called HITS on Data-Star), or full articles. Both allow flexibility within the field structure of a document--a user can view only those fields or combination of fields he/she wishes.

For browsing or viewing long texts, Data-Star and DIALOG fall down because neither allows a user easily to request a previous or next page or paragraph. Data-Star searchers can request viewing of each desired text paragraph p tx(1), but it is not as convenient as the NEXIS function keys.

None of these three systems will sort and display output according to the frequency of search words in the documents. Data-Star will count word frequency, however, and let the user select those documents with the highest occurrence. This feature is printed in an

"occurrence table" (..print oc). The occurrence table displays how often and where each search term appears in each document. This information can be used to help judge relevance or to select appropriate paragraphs to print.

CONCLUSION

Among the online systems, DIALOG, NEXIS, and Data-Star provide access to thousands of general interest full-text resources and dominate general library use of full texts. Other systems (notably DataTimes, Dow Jones News/Retrieval, and FT Profile) should be considered as well.

Although these three systems provide the most access to general, nonlegal full-text sources, none is the perfect system for full text. Ideal search and display features for full text have been described elsewhere, but searchers are still waiting for the best system [8]. A full range of proximity operators that recognize grammatical structure, automatic language enhancement features that can be turned off and on, display features that facilitate browsing and aid relevance judging all enhance fulltext searching. Better full-text features are coming from Westlaw, Dow Jones News/Retrieval, and America Online [9]. Can the "Big Three" afford to lag behind?

REFERENCES

[1] Orenstein, Ruth. Fulltext Sources Online. Needham Heights, MA: BiblioData, quarterly. Available online on Data-Star.

[2] Williams, Martha E., ed. Information Market Indicators. Monticello, IL: IMI, quarterly.

[3] Williams, Martha E. "The State of Databases Today: 1993." In: Gale Directory of Databases. Detroit: Gale Research, 1993.

[4] Tenopir, Carol. "Education for Database Intermediaries: How Library Schools Have Changed (And How They Haven't)." ONLINE 13, No. 6 (November 1989): p. 60.

[5] Pemberton, Jeffrey K. "What's Going on at Data-Star?" ONLINE 15, No. 6 (November 1991): p. 13.

[6] Tenopir, Carol and Katie Hover. "When is the Same Database Not the Same? Database Differences Among Systems." ONLINE 17, No. 4 (July 1993): pp. 20-27.

[7] Pagell, Ruth "Searching Full-Text Periodicals: How Full is Full?" DATABASE 10, No. 5 (October 1987): pp. 33-36.

[8] Tenopir, Carol and Jung Soon Ro. Full Text Databases. Westport, CT: Greenwood Press, 1989.

Basch, Reva. "The Seven Deadly Sins of Full-Text Searching," DATABASE 12, No. 4 (Aug. 1989): pp. 15-23.

[9] Pritchard-Schoch, Teresa. "Natural Language Comes of Age." ONLINE 17, No. 3 (May 1993): pp. 33-43.

THE AUTHORS

CAROL TENOPIR is a professor at the School of Library and Information Studies, University of Hawaii. She is the author of several books and many articles on various aspects of databases and online searching. Her monthly column, "Online Databases," has appeared in Library Journal since/983.

SHARON BERGLUND is a librarian at the Hawaii Medical Library. She was a high school librarian for eight years and is a consultant for the DIALOG CLASSMATE program. She received her MLIS from the University of Hawaii.

Communications to the authors should be addressed to Carol Tenopir, professor, School of Library and Information Studies, University of Hawaii-Manoa, 2550 The Mall, Honolulu, HI 96822; 808/956-5815; DIALMAIL--carol tenopir; BITNET--tenopir@uhunix; Internet--tenopir@ uhunix.uhcc.hawaii.edu and/or Sharon Berglund, Reference Librarian, Hawaii Medical Library, Inc., 1221 Punchbowl Street, Honolulu, Hawaii 96813; 808/536-9302, Ext. 113; Internet---sharonb@uhunix. uhcc.hawaii.edu.