December, 1988

Users and Uses of Full Text Databases

Carol Tenopir, *University of Tennessee - Knoxville*

Available at: https://works.bepress.com/carol_tenopir/63/
Keywords: Full Text Databases, Retrieval System Performance, Search Strategies, End-Users.

Abstract: The increasing variety in full text databases and wider availability means a variety of users and uses. Recent research has suggested that full texts are being used for many different reasons, including: document location, fact retrieval, paragraph retrieval, browsing, and document delivery. The studies reported in this paper show that search strategies may need to differ for different uses or for different types of full text. The reported studies focus on full texts of non-technical magazine articles.

INTRODUCTION

The number and types of full text databases are expanding at a rapid rate in the current commercial database market. Hundreds of scholarly journals, popular magazines, newspapers, newsletters, textbooks, reference books, legal decisions, and wire services are available in full text through several different online or CD-ROM search services. Non-technical multi-disciplinary literature is especially interesting because the writing style, topics, and potential users of the full texts vary so much. Most of the research on use and search strategies for full text has focused on technical literature.

This paper reports on two ongoing studies that are examining the possible uses of the popular magazine articles in full text databases and different search strategies that may be used by end users to search them. What makes a search successful is examined from the standpoint of who is searching the text and for what reason. Research is showing that end users use full text databases for many different reasons and successful search strategies may vary with the purpose of the search. In addition, the type of text that is being searched is being examined to see if different styles of writing or different types of literature may require different search strategies.

TYPES OF FULL TEXT DATABASES

A wide variety of full text databases are now available for searching on several online systems. To suggest there is a single type of use or that a single search strategy will work for all types is like suggesting that all printed books in the library are read or used in the same way. Like their printed counterparts, full text databases can be categorized by their major characteristics and subjects. The categories listed below are not always precise, but they are useful to describe in general the major
types of full text (non-numeric) databases that are available on online systems today. Listed with each type are representative databases, major online systems that include databases of the type, and general characteristics of the style or structure of each type of full text literature. Full text databases include the following types:

--statutes, court decisions, and other primary legal documents (available for European and United States materials since the mid-1970s on systems such as LEXIS, JURIS, EUROLEX, and Westlaw. Length and characteristics of the documents vary, but the language is often precise or predictable, with certain recurring important elements such as the name of the judge, the case name, the case citation or number, cases cited as precedent, date, court location, etc. End users are usually legal experts);

--government documents, patents, and other official publications (LEXIS includes many United States documents such as the Code of Federal Regulations and Federal Register. DIALOG provides access to other U.S. documents such as Commerce Business Daily. Each publication has its own characteristics, length, and use. Some of these are quite lengthy with multiple topics treated. Issuing agency, date, and code numbers such as patent number or contract number are important recurring elements);

--newspapers (dailies such as the New York Times or weeklies such as the Manchester Guardian Weekly. Includes both major papers such as the Wall Street Journal and small city dailies such as the Fresno Bee. Systems that provide access to several newspapers include VU/TEXT, DIALOG, LEXIS, Dow Jones News/Retrieval, and InfoGlobe search services. Newspaper articles vary from short new summaries to long feature stories. Topics vary, but style often has a certain consistency. Many news stories are written in the "top-down" style, with important information such as who, what, when, where and how at the beginning. Many newspapers include wire service news that overlap with the category below. Most newspaper databases omit certain parts of the newspapers such as advertisements, classifieds, weather forecasts, sports box scores);

--newswire services (national or international multidisciplinary news services such as Associated Press, United Press International, Reuters, Tass, Kyodo English Language News. Available from many services including DIALOG, LEXIS, NEWSNET, Dow Jones News/Retrieval. Like newspaper stories, articles may have a certain consistent style with important information grouped at the beginning. There is quite a bit of redundancy among and within wire services because they carry slightly different versions of the same story. Many stories are concise summaries of news events);

--newsletters (specialized, often industry-specific newsheets or wires. Many are available on LEXIS and NEWSNET. The language tends to reflect the interests and technologies of the industry, most do not mix subjects. They are often highly technical or specific to their industry. Primary users would be subject experts. Information may be extremely time-sensitive with a mixture of numeric and textual information such as in Platt's Oilgram).
--reference books (encyclopedias are widely available online and on CD-ROM from many sources. Each has its own fairly consistent writing style that includes embedded cross references and citations for further reading at the end. In addition to encyclopedias, many other reference books are online and on CD-ROM. These include such standards as the Merck Index, Kirk-Othmer, Mental Measurements Yearbook, the Bible, Gray's Anatomy, Drug Information FullText available on systems such as DIALOG, BRS, NEXIS, STN International, DATA-STAR, Telesystemes-Questel and Dow Jones News/Retrieval. Reference books are typically used for fact retrieval and most are highly structured in a consistent pattern);

--directories (many directories are online, including Books in Print, Marquis Who's Who, the Electronic Yellow Pages, Peterson's College Database, database directories, business directories, etc. on various systems including DIALOG, BRS, NEXIS, STN International, Dow Jones News/Retrieval, Telesystemes-Questel, DATA-STAR. Like other reference books, directories are most often used for specific fact retrieval. The writing style tends to be concise and highly structured. Information in each field may be brief such as a name or address);

--scholarly or technical journals (articles from professional society or other journals including many of the American Chemical Society journals on STN International and BRS, Harvard Business Review on DIALOG and BRS and medical journals such as the New England Journal of Medicine, Lancet, and the British Medical Journal in the BRS Biomedical Complete Text Library. Articles tend to be lengthy with many footnotes. Sentences and paragraphs may be long, although styles vary with the contributor. Abstracts may be included before texts, but these are not always in the online versions. Some journals make extensive use of graphs and charts that provide important information to the article, but are not always included in the online version. Book reviews, letters-to-the-editor, and short articles may be omitted from the online journal);

--nontechnical or general interest magazines (a variety of nonscholarly popular magazines are available on NEXIS and in Magazine ASAP [TM] on BRS, DIALOG, and NEXIS).

Since nontechnical/general interest magazines are the focus of the studies reported here, it is useful to look at this type of full text database in more detail. The over 100 magazines on Magazine ASAP can be further subdivided into the following categories: news [i.e., Time and Newsweek], business [i.e., Forbes and Money], hobby [i.e., Popular Photography and Popular Mechanics], political/commentary [i.e., New Republic and Nation], women's [Ladies' Home Journal and Redbook], entertainment [i.e., Sports Illustrated, Teen and Rolling Stone], and science [i.e., Science and Psychology Today]. The writing style and length of articles varies tremendously among the magazines, but most are aimed at the general non-specialist reader. Most do not include footnotes or abstracts. The printed counterparts tend to contain many photographs; the electronic versions include just the textual portions with picture caption headings. (Graphics are beginning to become available for some CD-ROM full text databases, but are not yet available in most online versions.)

What is included in a full text database is not always clear-cut or consistent. With magazine or journal databases for example, some
databases consist of a single title, while others (like Magazine ASAP) include the texts from many sources that are searched together. For titles other than those in Mead's version of Magazine ASAP, the NEXIS system allows the user to select a single title for searching or to group several titles together. Harvard Business Review Online on BRS and DIALOG consists of articles from that journal only. Magazine ASAP, Trade & Industry ASAP, and McGraw-Hill Publications Online each combine articles from many different magazines on different subjects. Effective search strategy formulation will be complicated by this variation if successful strategies depend on writing style or literature type.

3 RELATED RESEARCH ABOUT USERS AND USES

Much of the research on full text databases through early 1984 was summarized in my chapter in the Annual Review of Information Science and Technology (Ref. 1). Some of these early studies looked at the users and uses of full text, especially a series of studies conducted by the Books and Journals Division of the American Chemical Society (ACS) with BRS from 1979 through 1982 (Refs 2-6). The positive reaction of chemist end-users to a prototype ACS full text journal article database led ACS to make their journals available online. ACS found that chemists used the full text database for a variety of purposes, especially to find factual information imbedded in the texts, to find information that is peripheral to the main focus of the article, or as a bibliographic retrieval aid. Most scientists did not use the database for extended browsing or reading—instead they preferred to locate articles online and then go to the printed copy. The lack of graphics and poor terminal quality accounted for some of this preference.

A study by Elsevier Science Publishers reinforced the preference for reading printed journals if they are available (Ref. 7). Some scientists feared they would miss something if they relied only on searching online, others wanted the graphic material. Full texts were used to locate interesting articles, to find particular experimental methods used in a study, or to search the footnotes or references. In both the ACS and ESP studies, end users felt they did not need to be computer experts in order to find useful information.

These studies looked at subject-expert users of scholarly journal articles. Popular literature and its users have not yet been studied in detail. Two recent articles by Pagell (Ref. 8 and Ref. 9) are a beginning look at the uses of the Magazine ASAP database. She found that not everything in the printed journal is included in the database version. Not only are graphics left out, but entire sections of the magazines are omitted from the online file. For some magazines only feature articles and columns are included, dates of coverage vary, and journal titles included in Magazine ASAP vary according to what online system loads the file. She concluded that Magazine ASAP is "useful for individuals looking for the full text of selected articles" but the electronic versions "cannot serve as substitutes for the journals themselves."

Certain categories of use emerge from these studies. Full text databases are being used for: document location (like bibliographic databases are used), fact retrieval, paragraph or peripheral material retrieval, citation searching, and known-article retrieval. They have potential but are less frequently used for browsing or substitutes for printed journals. One of the studies described in this paper is exploring all potential uses of a full text magazine article database by academic end users.
Effective search strategies for full text databases have been studied by several researchers. Love (Ref. 10) described ways to achieve precision with the ACS journals. He compared two methods of linking more than one concept in a topic: searching for concepts linked with the Boolean AND operator (common strategy for bibliographic databases) with searching for concepts linked with the SAME paragraph proximity operator. He concluded that precision can be obtained with SAME paragraph operator, but not when concepts are found anywhere in the text as the Boolean AND allows. Blair and Maron (Ref. 11) found high precision but low recall in searching full texts of legal documents when many different concepts were linked with AND in the same search.

Tenopir (Ref. 12) and Ro (Ref. 13 and Ref. 14) studied optimum search techniques for Harvard Business Review Online, comparing the results from searching texts using the SAME paragraph operator with searching on descriptors, titles, and words from abstracts. Each field (except titles) contributed unique relevant documents, but as was expected text searching contributed the most. Text paragraph searching also had the lowest precision. Some of the low precision with full text searching seemed attributable to the imprecise nature of words in the field of management as covered by Harvard Business Review. Harvard Business Review articles are closer to popular magazine literature than research literature. In addition, only 2 or 3 concepts were linked, while Blair and Maron often were much more restrictive, linking 4, 5, or 6 concepts. Word frequency ranking algorithms improved precision.

Most search strategy experts recommend searching all kinds of full text using the SAME paragraph operator (or specifying approximately 20-40 intervening words between concepts.) The efficacy of this advice for general interest magazine article searching is being tested in one of the studies described below.

5 STUDY OF SEARCH STRATEGY TECHNIQUES

An ongoing study started by the author is comparing results achieved when using different search strategies on full texts [Ref. 15]. Queries posed by undergraduate students or public library patrons at library reference desks are being searched by a professional searcher on the Magazine ASAP database on DIALOG. A majority of the questions searched to date are current event topics of the "document locator" type. The requestors want to locate enough relevant documents to write a term paper or become knowledgeable about a current topic. These topics include such things as how do liquor laws affect the liquor industry, information about the morals of TV evangelists, information on the Pacific fishing rights, and how have microcomputers been used with preschool children. Some questions are of the fact retrieval type (for example, can you get AIDS from mosquitoes?). (If the users were doing their own searches some of the questions might be of the browsing type.)

Each question was searched 4 different ways. Each search of the full texts in Magazine ASAP looked for words or phrases using the following relationships to link concepts: 1) Boolean AND, 2) within the SAME grammatical paragraph (S); 3) within 10 words in either order (10N); and 4) within 5 words in either order (5N). The AND strategy is the broadest and includes all of the documents retrieved by the other three methods. Likewise, SAME includes all documents retrieved by (10N) and (5N), etc. Absolute recall cannot be measured, but if the AND strategy is defined as 100% recall, relative recall can be compared for the other strategies.
Precision was measured by judging relevance on a three-value scale: 1) not relevant; 2) part of the document relevant (defined here as 10 lines or fewer of relevant information); and 3) relevant (more than 10 lines.)

So far, the wisdom of experienced searchers has been upheld—the SAME paragraph strategy on the average gives the best balance of recall and precision. It is not the best strategy for every query, however. AND retrieved many false drops, but also many partially relevant documents. In at least one case, the one line that was relevant in a document retrieved by AND logic answered the fact retrieval query (one line of an article about AIDS in Science) told the user that he could not get "the disease from mosquitoes." In such cases measures of recall and precision are meaningless. Linking concepts with AND followed by browsing through KWIC portions of the documents retrieved worked the best. Term frequency algorithms would not help.

Surprisingly, (5N) did not always have the highest precision. Writing style of some documents seems to be the reason. For example, news magazine stories reporting on political campaigns resulted in the most false drops with all techniques. Unrelated issues discussed by a candidate or in a party platform are strung together separated by commas and few intervening words, so even a (5N) strategy retrieves these false drops. The ability to use grammatical structure or punctuation in full text searching might improve precision. Term frequency ranking algorithms would probably also help in these cases.

Types of articles as well as type of magazines were often predictors of false drops (or relevant hits). The ability to NOT out (or include) such things as Book Reviews, Movie Reviews, or news stories will often improve precision. Predefined and user-defined subsets of magazines by type would be useful for searchers, because often the false drops came from magazines that were obviously inappropriate. For a question about AIDS, for example, relevant documents came from Science, Science News, and Time. False drops with no relevant articles came from Flying, Datamation, Fortune, and Cycle. Such problems are inherent in a multi-title full text database, but could be overcome with value-added assistance from the database producer.

6 STUDY OF USERS AND USES

The author and Christine Tomoyasu, Business Reference Librarian at Hamilton Library, University of Hawaii are currently conducting a study of users of full text under a grant from the Council on Library Resources, Washington, DC. The study is examining how end users in an academic setting might search and retrieve information from a full text magazine article database and how such databases can best be used in libraries. The study is described in this paper, with further details and preliminary results described in my talk at the 12th International Online Information Meeting.

A total of 12 subjects are participating in the study: four undergraduate students, four graduate students, and four faculty members. Each is being instructed in the use of the Magazine ASAP database on DIALOG using the DIALOGLINK communications software. All have an interest in database searching as evidenced by their use of the library's intermediary search service, but they are not experienced online users. After initial instruction, each subject will search approximately five hours throughout the semester. Topics searched, uses of the database, and strategies used will be up to each person, but assistance will always be available from the online lab monitor. All online strategies and documents researchers.
documents retrieved will be recorded for later analysis by the researchers. Each subject was interviewed before they learned to search to get their ideas of how they might use full text databases. After each online session they will be asked for their opinion on the success of the search. We will analyze search strategies, search heuristics, and documents retrieved and compare each with the users' opinions of success. From the interviews and analysis of the searches we will attempt to discover all of the ways the full text database was used and how the type of use might correlate with search strategy. We will try to discover how both of these correlate with perceptions of success and traditional quantitative measures of success. A next step might be to run the same searches using different strategies to see if results and perceptions of success differ.

7 CONCLUSIONS

The variety of full text databases and types of literature in full text databases that are available are not limited to the primary journals. There are full text databases for secondary journals, books, and other types of literature. We hope that the results of this study will help professionals and users of full text databases to better understand the characteristics of each type of database and how they might use full text databases in the future and their perceptions of the strengths and weaknesses of each. From the interviews and analysis of the searches we will attempt to discover all of the ways the full text database was used and how the type of use might correlate with search strategy. We will try to discover how both of these correlate with perceptions of success and traditional quantitative measures of success. A next step might be to run the same searches using different strategies to see if results and perceptions of success differ.


