6-1-1995

ASCII Full Texts

Carol Tenopir
University of Tennessee - Knoxville

ASCII Full Texts

FULL TEXT ONLINE is nothing new. For more than 15 years online systems such as NEXIS, DIALOG, BRS (now CDP Online), and others have offered full texts of a variety of journals, magazines, newspapers, documents, and books.

The availability of and demand for full text has accelerated rapidly in the last five years. Now, full text is the most common type of database online. According to Martha Williams's introduction to the 1995 edition of the *Gale Directory of Databases*, full text databases now make up almost half (49 percent) of word-oriented databases; bibliographic and directory databases now represent only about one quarter. Even with the long-time bibliographic publisher H.W. Wilson entering the full-text arena in 1995, the number of full-text databases is sure to grow.

What does full text mean?

In the past, full text always meant ASCII text that was fully searchable and downloadable. It meant text only—with all charts, graphs, photos, and the like excluded. But "full text" no longer means only ASCII-searchable text. The image files that are widely available on CD-ROM and Internet look much better than text alone. New image compression techniques and graphical display programs such as Adobe's Acrobat are allowing online display of text that retains the aesthetic features of print while including graphics.

Even though image files may look better, they still have some severe limitations. Therefore, the ASCII text databases that may seem to be old-fashioned are still the most common full texts online. Image files take up much more disk storage space than ASCII, are much slower for online transmission, and require more sophisticated computers, display terminals, printers, and modems. The entry of some new major players in the ASCII full-text market shows ASCII isn't dead yet. Still, not all ASCII is the same.

Carol Tenopir is Professor at the School of Library and Information Science, University of Tennessee at Knoxville. Her E-mail address is tenopir@utkux.utk.edu

ASCII searchable

Full text that is binary encoded according to the American Standard Code for Information Interchange (ASCII), then made searchable, is still the most common type found online—on commercial systems and the Internet. NEXIS, Dow Jones News Retrieval, and DIALOG are three of the commercial leaders in full-text periodicals, with hundreds of ASCII newspapers, magazines, journals, newswires, and more.

The almost 4000 periodical titles available on the commercial online systems listed in *Fulltext Sources Online* (BiblioData., s-a) are all ASCII-searchable texts. In addition, many Internet sites have noncopyrighted ASCII fully searchable texts, such as Project Gutenberg in Illinois with its hundreds of text-only book classics.

ASCII text takes only a byte per character, so it can be sent quickly down telecommunications pipelines and takes up little disk storage space. But just having text in ASCII form does not make it searchable. Online systems typically have made ASCII full texts searchable by creating machine indexes (dictionary files) of every work in the texts (except stopwords).

Fully indexed ASCII text means searchers can look for words wherever they occur in a text and can download texts or portions of texts for further manipulation (within the confines of the copyright law). This method often retrieves more documents than a bibliographic search and also makes these texts useful for much more than just locating documents on a topic or supplementing print collections.

More than document delivery

Searchable ASCII text offers more than a document delivery solution. It can be used to search for a needle in a haystack (e.g., "Who are the treacle eaters in Alice in Wonderland?"); to trace the origin of words (e.g., "When did the mass media first start using the term 'information superhighway' and how many times has it been used in major newspapers every year since?"); to answer factual questions (e.g., "Who carried the flag for the United Kingdom in the 1988 Olympics?"); for public relations (e.g., "What magazines or newspapers have mentioned my client lately?"); or to find documents where the subject is too new or arcane to be indexed.

On the other hand, searching full text may lead to an unacceptable number of records retrieved, including too many false drops. Replacing Boolean ANDs with proximity operators (e.g., within a certain number of words, within the same paragraph, or within the same sentence) may help solve the false drop problem.

When relevant items are retrieved, the entire document may be useful, but often just a paragraph or even a sentence is all that is needed. In the case of tracking word usage, only the number of postings is needed. In these cases, the fact that ASCII text is unaesthetic and without graphics doesn't matter.

ASCII, not searchable

It used to be safe to assume that an ASCII full-text database would be fully searchable. What was lost in aesthetics and graphics was made up in part by the enhanced power of full-text searching. That assumption is no longer valid. Several new products offer ASCII text that has not been indexed.

Unindexed text is displayable and downloadable but not searchable. It can be used for quick (and sometimes inexpensive) document delivery, but it cannot be used for the other purposes described above. Nonsearchable ASCII saves the overhead space required by indexing every word in a text. Access relies on a corresponding bibliographic database tied to the full texts. Searching the bibliographic database rather than words from the texts may cut down on false drops and is less reliant on a variety of proximity operators.

Not surprisingly, this option is most often selected by publishers of indexes who also offer full text or by libraries locally loading tapes of full texts. When

UMI made full texts available on FirstSearch late last year, it made the decision to make the texts displayable or downloadable only. These ASCII full texts are accessed for bibliographic databases such as Periodical Abstracts, ABI/INFORM, Business Dataline, and ArticleFirst. When a user does a bibliographic search and finds relevant articles, those with a corresponding ASCII text may be ordered online.

The cost of each article depends on what FirstSearch payment plan is in effect in the library. In the pay per search plan, each full text costs five searches (between $2.50 and $4). Subscription plan libraries pay between $11,000 and $18,000 per year for unlimited use of the full texts on top of the subscription price for the corresponding index. (For more information, see "A Second Look at FirstSearch," *LJ*, November 1, 1994, p. 30ff.)

WilsonText

The newest major entrant into the full text market is the H.W. Wilson Company. Wilson debuted its first round of WilsonText products this April, with more to follow later this year and in 1996. The Wilson magazines and their journal products are ASCII, nonsearchable texts, while their biographical products are ASCII, fully searchable.

One of the WilsonText products is being developed to enhance the Wilson Abstracts databases. The first of these, available since April, is Readers' Guide Abstracts (RGA) Full Text Mini Edition. At first it will be available only on CD-ROM, but Wilson plans to make the full-text products available on magnetic tape for local loading and online via Wilsonline later in 1995 and, still later, on FirstSearch. The first CD-ROM version is on the Wilsondisc system; a SilverPlatter CD-ROM version is forthcoming.

RGA Full Text Mini includes the full text of articles in 54 of the 160 titles that are indexed and abstracted. The full text includes a variety of titles, e.g., *America, Architectural Digest, Children Today, Consumer's Digest, Discover, FDA Consumer, Maclean's, PC World, Parents, Road & Track, Saturday Evening Post, U.S. News & World Report*, and *Writer*.

Articles from these magazines may be searched by author, title, publication date, journal name, subject heading, and other standard bibliographic fields. A keyword search will look at words from titles and words from abstracts. Since Wilson abstracts are mostly informative and fairly lengthy, searching keywords will increase recall, but the ASCII full texts are not searchable.

Like other Wilson CD-ROM products, the RGA Full Text Mini Edition on Wilsondisc will be sold rather than leased and can be used on a local area network for no additional charge. Later this year, users will be able to link a search online to Wilsonline to retrieve the most current records at no additional fee. The yearly cost for monthly updates is $1,095; school year updates is $895; and a quarterly updated version is $695. Wilson's main competitor for ASCII text on CD-ROM is EBSCO, which has a version of its Magazine Articles Summaries with ASCII full text.

> When relevant items are retrieved, the entire document may be useful, but often just a paragraph or a sentence is all that is needed

Wilson's second magazine full-text product on CD-ROM will be a Mega edition of RGA, available sometime in the early fall. Mega RGA will be available on tape and online later in the year, and CD-ROM will be available from SilverPlatter and CDP OVID as well. It will provide the full text for 100 periodicals along with abstracting and indexing for 240 titles. A monthly updated version will cost $2,795 per year; school updates will cost $2,195 per year; and a quarterly updated verson will sell for $1,595 per year. Wilson will release additional periodical full-text titles on CD-ROM and tape in 1996, the first of which is likely to be a full-text version of Business Abstracts. It has aready negotiated ASCII rights for 750 titles out of the 3900 titles indexed by the Wilson indexes, so more products will follow. Almost 140 titles are already arranged for Business Abstracts Full Text.

Wilson has not ruled out making the ASCII text searchable. But for now, not indexing the full text allows more articles to be placed on a CD-ROM. The entire RGA Mini Edition can fit on a single disc.

All of the full texts on WilsonText are created by keying in the article texts from print versions of the magazines. Wilson contracts out the keying (which is done twice to reduce error rates), but in-house staffers proofread every keyed text before it is released to customers. According to Debbie Loeding, director of marketing and sales at Wilson, keying was selected over scanning the text portions and then converting to ASCII text with optical character recognition (OCR) software because there are "higher error rates with scanning."

Wilson's biographical series

Wilson will also have some ASCII-searchable text. Wilson Author Biographies (WAB) is the lead product of a proposed series of full-text biographical databases on CD-ROM from Wilson's General Publishing Department. WAB includes biographical and accompanying bibliographies for approximately 4300 authors who lived from 800 B.C. to the early 20th century. It includes five print titles: *British Authors Before 1800*; *British Authors of the Nineteenth Century*; *American Authors 1600–1900*; *European Authors 100–1900*; and *Greek and Latin Authors: 800 B.C.–A.D. 1000*. The second biographical product, due out this summer, will be Current Biography.

Unlike the periodical article products, all of the biographical titles will be fully searchable ASCII text. This keyword searching feature might not be used as much in these highly structured databases, however, as it would be in less-structured periodical article databases. Often a user will just be looking for the biographical entry for a specific author. In addition, the biographies include many enhanced controlled access fields for searching by categories. For example, in WAB, there are categories for genre, nationality, language, gender, century/period, and birthday.

Images?

Wilson is also considering releasing image files. According to Loeding, the ASCII full texts are "a starting point." ASCII is "an easier place to start" and is "more suited to our Readers' Guide customers in terms of numbers of CD-ROM discs required." Publishers may not own all of their images (photographs taken by freelance photographers, for example), so image rights are more complex.

Since one of Wilson's main competitors in this market, UMI, has released successful CD-ROM image products, Wilson has to consider images in addition to ASCII. But ASCII full texts will also be around for quite a while.