5-1-1991

The Same Databases on Different Systems

Carol Tenopir
University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

The Same Databases on Different Systems

EXCEPT FOR differences in cost and search commands and some minor differences in field tagging, most searchers assume that the same database on different online systems will still be the same. ABI/INFORM should be ABI/INFORM whether it is on BRS, DIALOG, NEXIS, or Data-Star. We are taught early on in our careers that the database producer is responsible for content; each online vendor simply takes the content from the producer and loads it according to its online system rules and procedures.

Experienced searchers know this is not always true. The same database on different online systems may be dramatically different. Differences in database loading and searching capabilities can lead to different results when searching a database on different systems. Decisions made by database producers and online vendors or just plain old errors may actually result in different versions of the same database. Usually these differences are not publicized; sometimes they are not even known by the database producer! Sometimes the only way to find them is by accident.

Unexplained results

Recently, one of my students was comparing how controlled vocabulary descriptors are treated on various online systems. She searched for the same ABI/INFORM descriptors in the NEXIS, BRS, and DIALOG versions of the database. (ABI/INFORM is available on nine online systems.) As expected, many of the searches retrieved the same number of records, but NEXIS often had fewer postings. Some of the difference in results was not explainable by variations in word or phrase parsing—articles found on DIALOG and BRS apparently should have been retrieved in NEXIS, but were not.

Puzzled, we returned to NEXIS to check each missing item. The articles retrieved on DIALOG and BRS simply were not in the NEXIS version of ABI/INFORM. Finally, we discerned a pattern to the missing items. Although they came from different journals, there was a clustering of dates in September 1988, May and August 1990, and October and November 1990. A subsequent search by journal title and date of the missing articles showed that no records from those issues were present in the NEXIS version of ABI/INFORM.

To make a long story short, we asked the folks at ABI/INFORM to help. They discovered, to their chagrin, that five weeks of records in 1988 and two weeks in 1990 had never been loaded by NEXIS (the October and November 1990 records subsequently were). According to ABI/INFORM, "there is indeed a need for better checking of loading procedures by vendors." A recheck of NEXIS two months later showed that the missing tapes had still not been loaded.

The fullness of full text

Tape-loading schedules differ from vendor to vendor. Other differences result from agreements between database producers and vendors, not all of which are publicized. Several years ago Ruth Pagell, librarian at the Lippincott Library of the Wharton School, examined the different versions of two of Information Access Company's full-text ASAP files.[1] She compared both Magazine ASAP and Trade & Industry ASAP as loaded on BRS, DIALOG, Dow Jones, and Mead Data Central (NEXIS). This is a case where some differences are acknowledged by the producer, yet are not spelled out in detail.

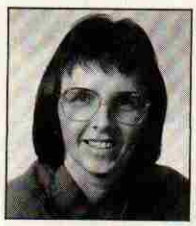
Different versions are acknowledged in that the databases on three systems are given different version numbers. DIALOG just has Magazine ASAP and Trade & Industry ASAP, the Mead versions are called Magazine ASAP II and Trade & Industry ASAP II, and BRS has version III. Each version includes a different number of periodicals, due to licensing agreements with the publishers and the online systems. In 1987, 14 titles were available only on the DIALOG ASAP versions.

Surprisingly, Pagell found that not all of the publishers of the original magazines were aware that their titles were in one ASAP version but not in another. *Forbes* was found only on DIALOG ASAP, even though *Forbes* had a nonexclusive license with IAC. Pagell opened what was called a Pandora's Box by the editor of *Database* when she used detective work to discover that IAC left out *Forbes* on the BRS and Mead versions at its competitor's (DIALOG) request.

Another difference that is confusing to searchers is that different versions of ASAP may have different dates of coverage for the same magazine titles. Pagell found, for example, that a March 1987 update of Trade & Industry ASAP on DIALOG included *Financial World* issues through January 6, 1987. The March 1987 update on BRS included *Financial World* only through 1985.

Differences in coverage dates for full-text newspapers or magazines are not unusual and are not limited to ASAP files. A look through *Fulltext Sources Online*[2] shows that many full-text magazines carry different coverage dates depending on the system. *Time*, for example, is carried 1986 to the present on Vu/Text, 1981 to the present on NEXIS, 1989 to the present on Dow Jones, 1983 to the present on DIALOG Magazine ASAP, and 1983–1988 on BRS Trade & Industry ASAP.

To further complicate matters, the time lag between publication date in print and date of availability online may vary as well. The *Washington Post* is available on a variety of online systems but with six different starting dates ranging from 1977 on NEXIS to 1988 on BRS. The time lag on these systems for the *Washington Post* varies from 24 hours to five weeks.



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

Differences in processing and loading

Sometimes differences in search results occur because of the way online vendors process and load the information they receive from the producers. DIALOG puts all subject-related fields into a "Basic Index," which is searched by default whenever unqualified search terms are entered. BRS, on the other hand, puts all of the fields in a database record into a single index.

That means words or phrases from the author field, corporate source, etc., will be included in an unqualified subject search on BRS, but not on DIALOG. Searching for "WOOD$" on BRS will retrieve the subjects wood, woods, woodwork, woodlands, etc., but also people named Wood, Woods, and Woodly. "WOOD?" on DIALOG will only get the subjects, because you must specify the author field ("AU=") to search for Wood as an author. Authors on BRS are usually double-posted, however, while they are only phrase indexed on DIALOG. That gives you more ways to retrieve author names on BRS ("WOOD$" or "WOOD-WILLIAM-P." as opposed to only "AU=Wood, William P." on DIALOG).

Due to differences in the way the initial field structures are set up by systems, the same database may have a different number of searchable fields. Editor names are displayable only on the BRS version of Ulrich's International Periodicals Directory, but can be searched on DIALOG. DIALOG allows searching of the Notes field in NTIS, while BRS does not. The BRS version of the Reader's Guide to Periodical Literature has 24 fields; the WILSONLINE version has only 21. Luckily most online systems (with NEXIS a notable exception) publish indispensable "cheatsheets" for each database that outline their field structure. No experienced searcher does without the DIALOG Bluesheets, BRS AidPages, ORBIT QRGs (Quick Reference Guides), etc.

MEDLINE

MEDLINE is another major database that is available on a multitude of online hosts. Frequent MEDLINE searchers in medical libraries may favor the National Library of Medicine online version because of its low cost. Other searchers often access MEDLINE via DIALOG, BRS, Mead, Data-Star, or other systems that offer better word proximity searching for free-text searching.

Janne Hunter, head of the Wistar Institute Library in Philadelphia, recently did an extensive comparison of MEDLINE on BRS, Data-Star, and DIALOG.[3] MEDLINE searchers are heavy users of descriptors because of the excellent indexing with Medical Subject Headings (MeSH). DIALOG and BRS double post descriptors so a multiple-word descriptor is searchable both as a phrase ("diabetes mellitus") and as words ("diabetes" and "mellitus"). Data-Star only phrase indexes them so that a search for the term "DIABETES" will not retrieve articles indexed under diabetes mellitus unless diabetes is mentioned in the title or abstract. Retrieval for the same topic on DIALOG and BRS versions will likely retrieve more documents.

Spelling variations

BRS allows searchers to turn on automatic searching of singulars/plurals and automatic British/American spelling. This means you will sometimes retrieve more documents when free-text searching on BRS, because if you enter the American spelling (i.e., liter, tumor, or fetal) any articles using the British versions (i.e., litre, tumour, or foetal) will also be retrieved. (The feature also works if you begin with British versions.)

NEXIS also offers automatic plurals and automatic British/American spelling. The main difference between the BRS and NEXIS features is that with NEXIS you cannot turn the feature on and off—it is a permanent feature of the system.

A colleague recently compared the British/American feature on BRS and NEXIS and found that for the most part searching on the same database with either British or American spelling will get the same results on BRS and NEXIS. Words that end in "er" or "re" (liter, meter, etc.) or words that contain a "u" near the end in British versions (flavour, catalogue, honour) are translated by both systems. The NEXIS equivalency list seems more comprehensive, however, as only NEXIS translates the term "tyre" into "tire" or "aeroplane" to "airplane." This results in more documents being retrieved on NEXIS than BRS for the same free-text search on the same database.

CD-ROM confuses the picture

When you add CD-ROM to the picture it becomes even more confusing. Psychological Abstracts on CD-ROM (PsycLIT) leaves out the dissertations and technical reports that are included in the online version. All records on the PAIS CD-ROM use subject headings from its 1990 subject headings list, while on DIALOG new records use the new list and older records are indexed with the 1984 subject headings. (This will be fixed sometime when the PAIS file is reloaded on DIALOG.) Compact Cambridge MEDLINE includes records from 1976 to the present; SilverPlatter offers MEDLINE back to 1966.

Just as with online systems, differences in loading procedures and search software can result in differences in search results between CD and online versions of the same database. ERIC on DIALOG's Ondisc CD-ROM version treats the journal name and source field differently than DIALOG's online software. Online, the journal name field (JN=) is pretty clean with only journal names appearing.

Ondisc, there is a mishmash of information in the journal name inverted index. Expand JN= ondisc and you'll retrieve JN= volume numbers, issue numbers, and pages as well as journal names. This is critical, especially when you are downloading from the disc and transferring records to database programs.

What does it all mean? As with many things in database searching—let the searcher beware! Even seemingly simple assumptions may not be true. You may not be getting what you think you are for any database online or ondisc.

I owe thanks for the examples in this column to several people: Wilma Wilke, Marilyn Reppun, Patricia Louis, and Diane Nahl-Jakobovits.

References

1. Pagell, Ruth, "Searching Full-Text Periodicals: How Full Is Full?" *Database*, Oct. 1987, p. 33–36; and "Searching IAC's Full-Text Files: It's Awfully Confusing," *Database*, Oct. 1987, p. 39–45.
2. Orenstein, Ruth M., ed. *Fulltext Sources Online*. BiblioData, s-a.
3. Hunter, Janne A., "MEDLINE on BRS, Data-Star and DIALOG—Does It Make a Difference?" *Database*, Feb. 1991, p. 43–50.