5-1-1994

Standardization Across Databases

Carol Tenopir
University of Tennessee - Knoxville

ONLINE DATABASES

BY CAROL TENOPIR

Standardization Across Databases

THE OPPORTUNITY OF searching hundreds of databases with the same search software has always been one of the big advantages of online over CD-ROM or locally loaded databases. Searching these many databases simultaneously is an incredible timesaver (and, sometimes, money-saver). The thought of checking a Bluesheet for every database in a OneSearch search statement may be too much even for experienced searchers.

Multifile searching—DIALOG's OneSearch, DataStar's StarSearch, and NEXIS file groupings—provides an efficient way to search through millions of records. Searching dozens or more databases at the same time with one generic strategy reduces preparation time and online time. But the reality remains that these databases still are separate entities. They are created by a host of different database producers, each following different conventions and structures.

Even when different databases have fields in common, they may be called by different names. And if the names of fields are the same, the values in the fields may be entered differently from database to database. It is up to the individual online system to impose some order and consistency (and should not be, as it has always been in some systems, still up to the individual searcher). Online systems are getting better at imposing some standardization in fields, with major efforts this year from Mead Data Central and DIALOG.

Searching single databases

When searching a single online database, searchers are trained to check system documentation (be it DIALOG Bluesheets, BRS AidPages, or the NEXIS Segments function key) before searching for field-specific information. Among other things, the documentation tells which fields are available for searching in that database and which tags are used to represent those fields.

For example, on DIALOG, if you want to restrict a search to software reviews, the Computer ASAP database uses AT for article type, Microcomputer Abstracts uses DT for document type, and Library and Information Science Abstracts (LISA) has no such field. Magazine Index has a named person field (/NP) for personal name subjects, ERIC uses the identifier field (/ID), PsycINFO puts them in the descriptor field (/DE).

Experienced searchers know that each database is an entity unto itself; you can never trust your instincts about field tags without checking the printed or online documentation.

Two ways to standardize

There are two ways online systems can standardize fields. The first is the easiest: standardizing the field name or field tag. A seemingly simple process, even this kind of standardization has not been common to all systems. Some searchers may remember when the year of publication in DIALOG was inconsistent. In most databases it was PY=, and in others it was YR=. DIALOG slowly standardized publication year to PY= in all of its databases, reserving PD= for publication dates that include month and/or day in addition to a year. Many other fields remain to be standardized.

The second type of field standardization—standardizing the values within a field—is more difficult and less common. To standardize field values requires some kind of authority list and a matching process when the file is loaded on the online system. It works best in those fields that have a limited number of values, such as publication year or language.

Standardizing the form of publication years and dates is something that can either be done when a file is loaded or with an equivalency table at the time of searching. Almost a decade ago, DIALOG standardized publication years to four digits (e.g., 1994) and publication dates to six digits (e.g., YYMMDD).

Other fields are more complex and are only rarely standardized systemwide. BRS has a standard authority list for all major languages in the language field of its databases. For example, a searcher can feel confident that English will always be searched as "English" and not as "Engl" or "Eng" and Serbo-Croatian will never be just Serbian.

The continuing saga of SF=

Perhaps the most notorious case of lack of standardization is in the field tag SF=, which has a variety of meanings depending on which database you are using. Reva Basch's "The Secret World of SF=" (*Database*, February 1991, p. 13–19) detailed the inconsistencies with this field on DIALOG. At present, SF= is available in most of DIALOG's more than 400 databases, but its meaning still varies from database to database. Nothing much has improved in the three years since Basch's article, but DIALOG will be tackling the SF= problem starting this year.

One common meaning for SF is "Subfile." Subfile is used to differentiate database divisions based on print equivalents. The online database INSPEC, for example, is made up of records from four different printed indexes/abstracts. Normally, they are all searched together online. If a searcher wants to restrict a search to just Computer and Control Abstracts, for example, SF=C will do it.

SF= means subfile in many other databases, including CAB Abstracts, Books in Print Plus, and Commerce Business Daily. Subfile is not the only meaning of SF=, however.

In many full-text databases, it stands for "Special Features," to identify which articles have photographs, charts, or other graphics available in print. SF= is also used variously to mean Source File, Searchable Feature, and a hodgepodge of other things.

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

DIALOG'S efforts over the years

Although it hasn't yet solved the SF= situation, DIALOG has been working toward some level of field tag standardization since the early 1980s. According to a long-time DIALOG employee, the company realized early on that at least the field tags and display formats for bibliographic fields should be standardized. This led to standard use of TI (title), AU (author), JN (journal name), and PY (publication year) field names and tags. In addition, other field tags common in bibliographic databases, such as DE (descriptors) and ID (identifiers), are used consistently.

Directory database field tags posed more of a problem, since there is such a variety of fields in these databases with little consistency from database to database. Important directory fields such as mailing address, phone number, sales, and company name are now consistently tagged in DIALOG directory databases.

DIALOG'S standardization push

DIALOG will begin a renewed and expanded push in 1994 toward standardization. Powerful search features such as OneSearch, Duplicate Detection, Rank, and Target added over the past few years make databases' inconsistent use of field tags and inconsistent field values glaringly obvious. Searchers have asked DIALOG to make standardization of field names a high priority.

Database standardization will begin slowly and will continue this year and next. You may have already noticed DIALOG's first standardization step. Beginning in May, DIALOG will provide consistent file names and consistent copyright identification on records from all databases.

Generic field designs

DIALOG is working toward generic recommended designs for each type of database on the system. Database producers will be asked to follow the standard field designs if possible, but DIALOG will also be imposing some of the standardization after-the-fact at the database loading end.

Full-text magazines, journals, newspapers, and newswires will be the first to be moved toward standardization. In addition to the already standard field names such as (TI), (AU), text, (JN), publication date (PD), etc., the full-text field standard will clear up the most common problems reported by searchers.

A document type field (DT=) will be standard, as will special features (SF=) and a language field (LA=). Geographic names at various levels (city, state, country) will be extracted from the descriptor or identifier fields to go into a geographic name field. Each database will have either an abstract, extract, or lead paragraph field that will put a concise content summary at the beginning of each full-text record. No matter which of these three is present, a user can search them together with a new general search code (XP).

Delving into content

DIALOG will also go beyond mere field name standardization and delve into field content. At the time of loading, author names will be automatically inverted, with a comma separating the last and first names. Periods will be inserted after initials if the file doesn't use them. Journal names and languages will be spelled out.

All of these changes will be made retrospectively on each file, so changes will progress file by file. DIALOG's standardized content will be searchable in each database along with whatever form is used by a particular database producer, so old methods of searching will still be available.

Company directories will be next to have a generic field structure and standardization. Company names will be phrase-indexed with and without initial articles (e.g., Crown Company/ The, as well as The Crown Company). Standard Industrial Classification (SIC) codes that come from the U.S. government SIC list will have the two-digit tag US added at the front of the code number, so U.S. searchers will not inadvertently retrieve codes with different meanings from other countries.

DIALOG display formats will also be standardized. Already this year, format 9 is being used consistently as the full-record format. Format numbers 1–8 will be standardized by type of database as much as possible. Unique formats will still exist in some special databases, but these will have format numbers of 11 or greater.

Long-term benefits

Standardization also has long-term benefits that may not be so obvious to searchers. Standard formats and ways to store data help behind the scenes and make it easier for programmers to create new search features and to develop new products.

Standardization on this scale will be a lengthy process, since each file to be standardized must be completely reloaded so changes will cover the entire file. Changes will occur gradually and, for some files, not for awhile. Databases that are not full-text news and journals or company directories (including bibliographic, other types of full text, and other directories) won't be worked on until later.

NEXIS and uniform segmentation

Mead Data Central was one of the first systems to offer easy multifile searching but one of the last to work on field standardization. Starting in the summer of 1993, Mead began standardizing the field names (called segments by Mead) in NEXIS. Starting with its news sources, such as newspapers, newswires, newsletters, and news magazines, NEXIS implemented "uniform segmentation" for more than 1000 NEXIS files. Financial and legislative sources followed.

NEXIS has always made multifile searching easy with preset subject-related or document-type groupings. Large groupings, e.g., OMNI or Current, search all the NEXIS files simultaneously, or the many smaller groupings search all of the files covering a topic (e.g., insurance) or a type of literature (e.g., magazines).

Until uniform segmentation began, a segment (field) search was often incomplete. Grouping mixed databases that used different segment names, but the Segment function key wasn't set up to show all the variations when a searcher was in a group file. Some databases used Byline, some used Author. For source name, some used Publication as the segment name, some used Source.

With uniform segmentation, all NEXIS files will now use Byline for the author of an article. Instead of having to enter "byline (smythe) or author (smythe)," searchers now can be assured that the search "byline (smythe)" will be complete in all files. Similarly, Publication is the new standard segment name.

Better late then never

Such standardization efforts are welcome (even if they are long overdue). The trend of searching multiple online databases simultaneously will continue. Doing away with the necessity of checking system documentation for field tags and field formats is one step in the right direction.