THE LINEAR FILE

Guest Editorial
by Péter Jacsó
School of Library and Information Studies
University of Hawaii

A Proposal For Database
"Nutrition And Ingredient" Labeling

Ten years ago, on these very same pages of *DATABASE*, Jeff Pemberton called for exposing the problems of dirty data. Jeff, the need exists more than ever. The problems seem to get worse as the numbers of unsuspecting, casual users of online and CD-ROM databases keep growing, and hastily compiled databases have not only dirty data but also too many white spots, i.e., no data where data is expected. In this, and the April 1993 issue of *DATABASE*, I show a few techniques to explore systematically the skeletons in the cupboard (see feature article beginning on page 14), but I think that we need something else, too. The new FDA food-labeling rules gave me the idea.

By the time you read this guest editorial, the FDA regulations for the implementation of the Nutrition Labeling and Education Act of 1990 should be finalized, including the content and formats of the labels that will appear on foods beginning in May 1993. The purpose of the new food labeling legislation is to educate the consumer in unambiguous and standard format about the ingredients and nutritional values of the foods, and to put an end to unscrupulous claims, deceptive ads, and false statements so prevalent in the food industry. I feel that database users also deserve similar protection and safeguards.

While savvy database users, using some advanced search techniques, often are able to call the bluffs of the data file producers and database publishers, it is crying over spilt milk to learn about the deficiencies of the database *after* the CD-ROM license fee has been paid, or hundreds of dollars were spent searching online databases falling far short of the promises made in glossy ads, promotional pamphlets, user guides, or the cover sleeve of CD-ROM jewel boxes. It is akin to learning in the hospital emergency room that your heart attack has been triggered by your consciously selected low-cholesterol foods, which were in fact high in saturated fat, but the food producer concealed the latter with euphemistic and misleading claims on the packaging.

The misleading database claims are made as hazy statements, locker-room boasts, hypes and euphemisms in the best case, and are straight-faced shams, and deceptive statements in the worst case. The crows mostly relate to various aspects of the database coverage, content, and accessibility. These are quantifiable, and the file producers do (or should) know them precisely. They should publish such data along with the claims. Fair players have nothing to lose.

→

I created a sample label of a bibliographic database (the fictitious MYOB) merely for illustration. Some categories on this label may apply only to bibliographic databases; directory and full-text type databases may require somewhat different categories. The point is to start the ball rolling. (The database nutrition and ingredient label, of course, may be an addendum to the blue pages, yellow pages or aid pages of online services, to the database catalogs, a "certificate" freely accessible in online directories, or an external post-it card on CD-ROM databases not requiring the opening of the jewel box.) The label (at least the most essential part of it) should be also the first or second introductory screen for casual users.

COVERAGE CROWS

The size of the database is usually declared but it is often loosely stated, or simply false. I am not talking about peanuts, but significant exaggerations. A publisher claims to have 1.4 million citations in its CD-ROM database. In reality, there are only 903,000 records. It is their online database, which combines two of their CD-ROM products, that has, indeed, 1.4 million records. This may be a nonchalant oversight, but it definitely misleads the user.

The recommended form of size specification is the exact number of records as of a given date, and the monthly (bimonthly, quarterly, etc.) increase. The increase should reflect the actual update frequency of the database, not the adding of the records to the in-house database of the file producer.

The period coverage, i.e., the time span of the database, is often "expanded" by sprinkling the database with records from the year that is claimed to be the start date. Declaring the number of records for the first few years allegedly covered, and for the last few years is much more informative. From the sample label, it is clear that the coverage of the early years is much more shallow than that of the current period on which the increase statement is based.

The other end of the period coverage, i.e., the currency of the database, is also often blurred as I pointed out in the June 1992 issue of DATABASE. The date in the header of an online database, or on the introductory screen of a CD-ROM database is not necessarily the date of the indexed publications, but the update/reload date of the database. In a library and information science database even the most recent records refer to six month old publications, and the largest portion is for articles published nine to ten months before. It is hopeless to search this database for current articles, and the user should be made aware of it. The recommended form of declaring currency is time-lag statistics based on the most recently added batch of records, i.e., 7.2% of the records describe publications published in the same quarter, 9.4% published in the previous quarter, etc.

The width of coverage claims are often grossly exaggerated by quoting only the number of total journals indexed without regard to the spread and depth of coverage. File producers often count journals under their former and current titles, and drop journals after a few issues. Declaring the percent of active titles of the total titles makes it clear how many titles ceased publication, or were dropped by the file producer.

It is easy to inflate the width of coverage by including a few items from a large number of journals. Declaring the concentration factor can reveal this trend. In the sample database the 90/56 concentration factor shows that 90% of the citations are from 56% of the titles. Such a value indicates that many journals may have only skin-deep coverage. Alternatively, or additionally, it may be specified what percent of the sources are indexed cover-to-cover, selectively, and occasionally (or exceptionally).

Sample Database Label

Name: MYOB	Producer: CSUCSUSA
Number of records (06/92) 251,523	Quarterly increase : cca. 3,400 records

PERIOD COVERAGE

From: 01/69

Year	Records
1969	4,470
1970	5,035
1971	4,935
1972	5,659
1973	6,932
1974	5,932
1975	5,960
1976	6,372
......	......
1990	13,040
1991	11,198

CONTENT

Journal articles	71.9%
Conference papers	15.9%
Conference proceedings	7.1%
Books	3.0%
Dissertations	1.6%
English language documents	85.0%
Average length in characters	
Record w/o abstract	215
Abstract	305
Average descriptors/record	3

TIME LAG IN LAST UPDATE

same quarter	7.2%
-1 quarter	9.4%
-2 quarters	6.2%
-3 quarters	32.8%
-4 quarters	17.9%
>1 year	26.5%

ACCESS POINTS

Title	100.0%
Conference title	23.0%
Author	100.0%
Author affiliation	15.2%
Journal name	71.9%
ISBN	2.8%
Named Person	2.0%
Descriptor	99.2%
Publisher	27.5%
Editor	22.7%
Publication year	99.7%
Document type code	49.4%
Abstract	22.6%

SOURCE COVERAGE

Width and spread of coverage	
Active titles	87.2%
Concentration factor	90/56
Depth of coverage	
Cover-to-cover	70.0%
Selectively	25.0%
Occasionally	5.0%
Geographic source of publications	
USA	82.0%
Britain	11.0%
Netherland	7.0%
Italy	3.5%
Germany	3.4%

● Abstracts are added to records from June, 1984
● Foreign language articles w/o English summary have no descriptors
● Some additional access points are available occasionally.

162

International and worldwide are favorite adjectives of database publishers for coverage if there is at least a Canadian title indexed, or if records for a few dozen non-U.S. companies, products, or persons are included in a database. If this moniker is used, the percent of non-U.S. titles, companies, etc., contributing the most records should be also specified.

CONTENT CROWS

To perceive the relevance of a database, it is useful to know its composition as precisely as possible. It is easy to inflate the size of a database by including skimpy bibliographic records about brief news items, letters to the editors, short product announcements, etc. There is nothing wrong with such items, or with the inclusion of thousands of records about recipes or book reviews in a database as long as the user is made aware of it, and can approach the database bearing this in mind. Similarly, inclusion of a few dozen citations for dissertations and the claim that the database has records for dissertations can make the users believe that they have the equivalent of Dissertation Abstracts (for the subject field at least). The four to five major categories of source documents and their share of the total number of records should be listed on the label.

A significant number of records citing non-English language documents may be a blessing or an annoyance, depending on the user's preference. If there are many foreign language materials indexed, their share should be indicated, possibly listing the percent of records in the major languages.

Bibliographic databases often mix records with and without abstracts, but may suggest through their names as if all records would have an abstract. Others magnanimously blur the difference by mentioning that many records carry an abstract. The recommended form is to specify exactly what percent of the records have abstracts.

A related issue is the length of the abstract. I have seen single-sentence abstracts in too many databases, which did nothing more than paraphrase the title, or at best, augment the title. A data file producer recently started to add annotations to its records, but these are six to seven word annotations, which can convey extremely limited information about the content of the original articles. Declaring the average length of abstracts on the label, as it is done by Wilson and UMI in their manuals, is more informative.

The average number of topical and other subject headings may be useful for many users to know. In a general periodical index, for example, only one subject heading is assigned (admittedly, often subdivided), but the users are not advised about this

convention, and may significantly limit their search results if they restrict the query to the subject field. The average record length is also a useful quantitative indicator, and is a welcome addition to Bowker's latest product catalog.

ACCESS POINT CROWS

The most misleading claims are related to the access points of the databases. The user guide for a who-is-who database of library and information professionals boasts that "never before has it been possible to easily search and combine more than 30 fields." What it fails to mention is that out of these 30 data elements only three to four are available in every second record. If this is not misrepresentation, then I do not know what is. True, *thou canst search but thou wouldst not find*. The label should precisely specify in what percent of the records are found some of those proudly advertised data elements. In the sample, the less than 50% availability of document type code would warn the user not to use it.

CONCLUSIONS

All of this information is readily available for the data file producers, and often for the database publishers who process the files. Of course, it is not always in their best interests to disclose the information. While the FDA has the power to enforce the declaration of nutrition and ingredient information, file producers and database publishers can only be requested to volunteer such information. However, the National Federation of Abstracting and Indexing Services, the Association of Independent Information Professionals, ASIS, ALA and other influential bodies, together with trade journals could form a powerful lobby to, a) develop and disseminate guidelines, b) persuade producers and publishers to comply voluntarily, and c) advocate customers to ask for such information.

There are many other qualitative characteristics of databases left for users to do the rest of their homework when selecting a database. Their educated choices should be aided by making available objective and quantitative information about the databases by the file producers. This informative labeling would well complement the evaluative "seal of approval" certificate to be awarded by reputable user groups and critics on the basis of qualitative evaluations as discussed by Reva Basch in *Database Searcher* (October 1990).

Communications to the author should be addressed to Péter Jacsó, Visiting Associate Professor, School of Library and Information Studies, University of Hawaii, 2550 The Mall, Honolulu, HI 96822; 808/956-5817; Fax 808/956-5835; BITNET—jacso@uhunix.bitnet.

SECOND INTERNATIONAL SYMPOSIUM: NATIONAL SECURITY & NATIONAL COMPETITIVENESS: OPEN SOURCE SOLUTIONS Proceedings, 1993 Volume I - Link Page