

SAVVY SEARCHING STARTS WITH BROWSING

SAVVY SEARCHING

Péter Jacsó

*The University of Hawaii
at Manoa
jacso@hawaii.edu*

You can tell savvy searchers by how they start a search. With few exceptions they never start by banging in their query term and pressing the SEARCH key. They first browse the indexes to explore the variations in the spelling of the search terms. This is especially true with searches involving author names, journal titles, author affiliations, company names or product names. The databases that show the most signs of neglect and negligence, such as Mental Health Abstracts and PASCAL have a significant number of errors even in the descriptor field, the most sacred part of the records.

The volume of misspellings and inconsistencies is enormous in many professional databases; and discouraging in almost all of them. In the professional arena the HW Wilson database family and INSPEC are among the few databases with squeaky clean authority files for authors, descriptors and corporate names. In the consumer market, the free Internet Movie Database runs circles around its professional and consumer-oriented counterparts and practically any database when it comes to the accurate and consistent spelling of titles, personal and geographic names. It also provides the most comprehensive and highest quality cross-references from one name format to another. Those who designed and implemented the Internet Movie Database did not miss the class on authority control.

The symptoms

In systems that allow the users to browse the indexes created from the journal name, author name, corporate name and product name fields, one can easily spot some of the inaccurate and/or inconsistent entries in the index. To save space I use a simple example from the Pascal database (File 144) that shows 8 variants of the title of the journal *Library Acquisitions: Practice and Theory*. (If I had a dollar for every inconsistent and/or incorrect journal title in this database I would not need to write my columns). If you wish to browse the field-specific journal index, these entries nicely congregate and you may be able to include all of them in one fell swoop in your query.

144	1	LIBR. ACQUIS., PRACT THEORY
144	596	LIBR. ACQUIS., PRACT. THEORY
144	3	LIBRARY ACQUIS. PRACT. THEORY
144	36	LIBRARY ACQUIS., PRACT. THEORY
144	1	LIBRARY ACQUIS.: PRACT. THEORY
144	1	LIBRARY ACQUIS, PRACT. THEORY
144	537	LIBRARY ACQUISITIONS. PRACTICE AND THEORY
144	3	LIBRARY AQUIS., PRACT. THEORY

However, if the entries are not adjacent, then the problems are more serious. If users look up the company name 'John Wiley & Sons' in the Dun's Electronic

Directory and finds one record for the company, it is not likely that they would also look up the variant 'Wiley, John & Sons' or 'Wiley John & Sons' - on the off chance. But users would be advised to because there is an entry for this publisher under both other formats. Longer, more complex names have more variations, with even more scatter. The adjacent ones are just an eyesore, but the ones scattered far away are easy to miss.

Those who want to look up my name in the *Information Science Abstracts* database may find 81 records under AU=Jacsó, Peter, but may miss 18 others hiding under the incorrect version AU=Jasco, Peter as there are intervening names such as Jacson. In case of multiple database searching the problem is more aggravating. The variations on the name of the Libyan president Muammar Qadhafi (or is it Gadhafi or Kadhafi, or Khadafi, or Khadaffi) are quite numerous. Authority files were exactly devised to solve such problems, but to develop and use them costs money for the datafile producers who happily shift the burden of responsibility to the searchers.

The implications

Nearly 20 years ago Martha E Williams and Laurence Lannom warned us of the serious lack of standardisation of the journal title data element in databases [1]. They tested 8 databases and developed 4 measures to show the extent of the lack of standardisation.

The same measures can be applied to the company name, product name and author name fields as well, although the work would be much more tedious because the population of these terms is much larger than that of the journal names.

In the past 18 years I have not seen any systematic study to gauge the extent or the implications of this problem. Even without formal research, however, you can imagine what it means when someone misses half of the pertinent records because of inconsistent or wrong spelling of the journal name, as is often the case. In the extreme case of the DIALOG implementation of the *Ulrich's* database, 99 per cent of the records that should have included the name of the abstracting/indexing service *Excerpta Medica* had the name incorrectly spelled as *Exerta Medica* in 5653 records as of 1 June 1999 (see Figure 1).

DIALOGWEB.			
Command Search			
now search database alerts cost logoff help			
E15	4	AI=EXCERP. MED.	(UNTIL 1995)
E16	1	AI=EXCERP. MED.	(UNTIL 19992)
E17	1	AI=EXCERP. MED.	(1983-)
E18	4	AI=EXCERP. MED.	(1988-)
E19	1	AI=EXCERP. MED.	(1988-1992)
E20	1	AI=EXCERP. MED.	(1988-1993; 1996-)
E21	2	AI=EXCERP. MED.	(1992)
E22	59	AI=EXCERP. MED.	(1992-)
E23	2	AI=EXCERP. MED.	(1993)
E24	2	AI=EXCERP. MED.	(1993-1994)
E25	1	AI=EXCERP. MED.	(1993-1995)
E26	1	AI=EXCERP. MED.	(1996-)
E27	2	AI=EXCERP. MED.	(1997-)
E28	68	AI=EXCERPTA BOTANICA	
E29	289	AI=EXCERPTA INDONESIA	
E30	1	AI=EXERP. MED.	
E31	5653	AI=EXERTA MEDICA	

Figure 1: Fatal misspelling of *Excerpta* as *Exerta Medica* in 99% of the records

Is it possible that the producer of *Ulrich's*, the RR Bowker Company - an Elsevier subsidiary after all - does not know how to spell the name of this well known journal published by Elsevier? Nothing is impossible, but in this case Bowker is probably innocent. The error was most likely introduced when loading the records into DIALOG. How? I assume that some of the data elements are coded in the original records. During input, the code (say, ExMed) is converted into full format for the sake of end-users. Someone may have defined the equivalent of the code as *Exerta Medica*. Why do I think so? Because the records have the fully spelled out name correctly in Bowker's own CD-ROM version and in the Ovid version. (For the record: there are many serious errors in the raw records that would not make Madame Ulrich happy).

So savvy and responsible searchers devote a lot of extra time (both their time and computer time) to finding as many of the variations of search terms as they can. Missing 20 to 25 per cent of the relevant records for unpredictable name variations can have very serious implications for the end-user. Missing 80 per cent of records may result in a pink slip for the searcher.

But not even the most savvy and cautious searcher can work wonders if the system does not offer sophisticated browsing and additional options to deploy defensive strategies. The systems vary quite widely with regard to index browsability.

1. WILLIAMS, M E and L LANOM, Lack of standardization of the journal title data element in databases, *Journal of the American Society for Information Science*, 32, (3), 229-233.

Software variations

Some of the professional database services don't have any index browsing capabilities. For example, LEXIS-NEXIS users cannot browse indexes at all. The same applies to the Syracuse University implementation of the ERIC database. In the CARL software the number of browsable field-specific indexes are limited. OCLC's nifty SiteSearch software gives absolute freedom for the implementor to decide which fields to make browsable. SilverPlatter offers a number of field-specific indexes (but not in the DOS version). However, the choice of browsable index fields is not always perfect. I find it difficult to understand why the ISSN index is browsable (the searcher is unlikely to click around until finding a handsome and promising ISSN, you either know the ISSN or you don't). I think that instead of browsing the ISSN index the users would rather have the journal and author name index fields made browsable. Bell & Howell Information & Learning (formerly UMI) has a very modest index browsing ability. You may look up journal titles and subject descriptors (within some constraints), but there are no browsable indexes for author names, company names, and product names.

Ovid has the most comprehensive browsable indexes. Every index that is searchable is also browsable, and many of the fields are both word indexed and phrase indexed. This means that an index entry is generated for each word in the journal name and another single entry for the entire journal name. It is

an extra feature in Ovid that multiple indexes can be combined for browsing. For example, the indexes generated from the title, successor title, former title and parallel title fields can be combined, on the fly, for browsing.

DIALOG has quite comprehensive sets of browsable indexes but most of them are only phrase indexed. Finding the widely scattered index entries even within a single database can be an arduous task. The number of variants of the Journal of the American Society for Information Science is a challenge even for the most savvy searcher.

Luckily, DIALOG has a unique database family: the Finder database family for journal names, company names and product names. The Journal Name Finder database includes the names in the journal name fields of all the databases in both word indexed and phrase indexed format. A carefully formulated search that contemplates the possible abbreviations can help the searcher to find almost all the possible variants, and yields a list like shown in Figure 2.

Software can help to ease the pain of creating defensive strategies that contemplate numerous variants for a name, but it is still extra work for the searcher and sophisticated index browsing options are still the exceptions rather than the rule. It is not likely to change when most of the software reviews focus almost exclusively on the search functions and ignore the browse capabilities.

© Péter Jacsó

Figure 2: Journal Name Finder brings up a comprehensive list of variant names

FILE NUMBER	RECORD COUNT	JOURNAL NAME
1	11	AMER SOC INFORM SCI
1	7	J AMER SOC INFO SCI
1	14	J AMER SOC INFORM SCI
1	1260	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
1	5	JOURNAL OF THE AMERICAN SOCIETY OF INFORMATION
2	1399	J. AM. SOC. INF. SCI. (USA)
3	581	J. AM. SOC. INF. SCI. (USA)
4	818	J. AM. SOC. INF. SCI. (USA)
7	2491	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
8	88	J AMER SOC INFORM SCI
8	859	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
15	1471	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
21	1	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
34	434	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
36	1	AMERICAN SOCIETY FOR INFORMATION SCIENCE, JOUR
37	32	AMERICAN SOCIETY FOR INFORMATION SCIENCE, JOUR
38	8	J. OF THE AM. SOC. FOR INFORMATION SCI
39	1	J. OF THE AM. SOC. FOR INFORMATION SCI
41	1	J. AM. SOC. INF. SCI
47	23	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
49	3	AM SOC INFO SCI J
49	34	AM SOC INFO SCIENCE J

cont'd over...

FILE NUMBER	RECORD COUNT	JOURNAL NAME
49	2	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
61	1734	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
73	25	J. AM. SOC. INF. SCI.
73	24	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
75	811	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
76	1	J. AM. SOC. INF. SCI.
78	193	J AM SOC INF SCI
78	193	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
103	7	J. AM. SOC. INF. SCI.
103	2	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
119	2	JASIS: JOURNAL OF THE AMERICAN SOCIETY FOR INF
119	1	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
144	962	J. AM. SOC. INF. SCI.
144	362	J. AMER. SOC. INFORM. SCI.
144	861	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
144	245	JOURNAL OF THE SOCIETY FOR INFORMATION DISPLAY
148	537	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
149	10	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
151	68	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
159	1	J AM SOC INF SCI
163	2	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
202	3	JOURNAL OF AMERICAN SOCIETY FOR INFORMATION SC
202	1854	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
239	255	J. AM. SOC. INF. SCI.
239	7	J. AM. SOC. INF. SCI. J.
248	3	J. AM. SOC. INF. SCI.
420	656	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
438	1219	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
439	33	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
440	1282	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO
485	5	JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATIO

The author

Péter Jacsó

Péter Jacsó is an associate professor of the Information and Computer Sciences Department at the University of Hawaii. He is the recipient of the 1998 Pratt-Severn/ALISE National Faculty Innovation Award for his innovative use of information technology in curriculum design and coursework. He is a frequent speaker at national and international conferences. His columns are published in *Database*, *Information Today*, *Computers in Libraries* and *Link-Up*. He has received various awards for his writings, including the 1998 Louis Shores/Oryx Press Award of the ALA Reference and User Services Association for his discerning database reviews.