

péter's picks & pans



Péter Jacsó
University of Hawaii

ISI Web of Science, Scopus, and SPORTDiscus

My two picks, the newest version of ISI Web of Science (and ISI Web of Knowledge) and the new Scopus database, represent the state of the art in indexing/abstracting databases. Not only are the databases huge and interdisciplinary, but both are also endowed with cited references, make superb use of citation indexing, and facilitate citation searching, which reduces the need to rely on controlled vocabulary. The pan is SPORTDiscus, which is huge (for a single discipline database), but in an unhealthy way, due to the duplicate and erroneous records that are not removed even after correction. It contains a befuddling thesaurus, which illustrates how frustrating controlled vocabulary searching can be with an inappropriate tool.

...the databases ... are
also endowed with
cited references,
make superb use of
citation indexing, and
facilitate citation
searching...



the picks

ISI WEB OF SCIENCE (WOS)

This summer, Thomson Scientific's ISI Web of Knowledge (WoK) platform and ISI Web of Science (WoS) service got an "interfacelift" that would make Botox aficionados green with envy. The databases also received an overall rejuvenation treatment, with a couple of substantial brain-booster features. One improvement, the Analyze feature, allows users to rank a results set of up to 2,000 records by author, institution name, source title, publication year, subject category, language, or document type, and to see the results in a tabular format along with a histogram. The distribution figures instantly and clearly show the most productive authors, institutions, journals, and subject categories for the topic of your search in a way that would have been appreciated by Bradford and Lotka (who formulated theories about the scatter-distribution of information in scholarly journals and the concentration of authors writing about a specific subject). I can only wish for one more enhancement—an option for analyzing the set by cited journals, cited authors, and cited year.

The other very important cerebral option is the Related Items feature. This now shows not only the list of papers that have common citations with any of the records selected by the user, but also gives the number of common (shared) citations, which can also be displayed. These features will make even the most technophobe users sit up and pay attention. ISI Web

of Science demonstrates the ideal combination of power of use and ease of use. The acquisition of BIOSIS by Thomson and the licensing of scholarly databases, such as INSPEC,

FSTA, CAB Abstracts, and PsycINFO, present especially promising potential if integrated with WoS on the much-enhanced ISI Web of Knowledge platform.

SCOPUS

I tested only the beta version of Elsevier's new mega database, Scopus (the official launch of the product is scheduled for November 2004). It has about the same number of records as WoS has for the 1975-2004 time period (approximately 27 million), fused from four sources: the abstracting/indexing (A/I) databases of Elsevier and MEDLINE, the ScienceDirect archive of the Elsevier journals, some open access Web components of Scirus, and its partner publishers' archives. Enhancing the Elsevier A/I databases with cited references must have been a massive project. The result shows splendidly in the novel presentation format that I have been awaiting for many years.

The grid format of the short results list is conducive to quick scanning and, most importantly, the results matrix prominently shows the number of times the paper was cited by journal articles whose records appear in one or more of the component resources of Scopus. The grid can be re-sorted almost instantly by any of several bibliographic data elements, including the citedness count of the items. Sorting by citedness count (and some other elements) is also possible in WoS. However, the citedness count is not displayed in the WoS short results list, only in the detailed record format. This layout in Scopus will motivate users to choose the most-cited documents. Obviously, an article published 4 years ago will likely have been cited much more often than an article published a year ago, but it could be offset by selecting another column to show the adjusted (relative) citedness count.

Scopus also has an automatic feature similar to the Analyze option of WoS. You can hide the analysis panel, but it is worth every inch of the screen real estate, as it provides distribution figures of the results set similar to the ones produced by WoS' analyze feature—very informative, instant statistical snapshots of the demographics of the search topic. True, you can't alphabetically sort the entries of this snapshot in Scopus, but you can keep expanding the list if you want to see more entries. You can also exclude some entries from any category (i.e., a document type or a subject area not relevant to you).

Field	Source Title	Record Count	% of 1224	Bar Chart
<input type="checkbox"/>	INTERACTING WITH COMPUTERS	121	9.9 %	■
<input type="checkbox"/>	INTERNATIONAL JOURNAL OF HUMAN-COMPUTER STUDIES	60	4.9 %	■
<input type="checkbox"/>	BEHAVIOUR & INFORMATION TECHNOLOGY	49	4.0 %	■
<input type="checkbox"/>	INTERNATIONAL JOURNAL OF HUMAN-COMPUTER INTERACTION	35	2.9 %	■
<input type="checkbox"/>	ERGONOMICS	34	2.8 %	■
<input type="checkbox"/>	HUMAN-COMPUTER INTERACTION	30	2.5 %	■
<input type="checkbox"/>	COMPUTERS IN HUMAN BEHAVIOR	26	2.1 %	■

Instant result analysis from Web of Science is highly user friendly.

Date	Document (Sort by relevance)	Author(s)	Source Title	Cited By
1. 2000	Gene ontology: Tool for the unification of biology Abstract + Refs View at Publisher	Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M.,	Nature Genetics 25 (1), 25-29	521

The novel format and content of the short result matrix of Scopus packs a mighty punch.



the pan

SPORTDISCUS

I paid homage to sports before the Athens 2004 Olympic Games by working out on SPORTDiscus (the Silver-Platter version happened to be Ovid's Free Resource of the Month in July). I tested SPORTDiscus, which is available through DataStar, Dialog, and EBSCOhost as well as Ovid, on several software platforms. Suffice it to say that it takes a lot of sweat to wade through the duplicates (and in the case of Dialog's implementation, the triplicates, quadruplicates, and quintuplicates) in the database. It also requires intense exercise to make heads or tails of choosing the least-erroneous, least-inconsistent, and least-uninformative records.

Numerous duplicate records exist, demonstrating how inconsistently the descriptors and even the classification codes are assigned, not to mention the absolute lack of authority control and unification for author and journal name formats and typography. In April 1996, Uri Miller criticized the SPORT database, as it was then called ("The SPORT Database: Some Comments," *Online & CD-ROM Review*, 20(2):pp. 67-74). Today, the database has slid further downhill by adding records from a variety of other databases in a quest for bragging rights as the largest sport database. This move has made controlled vocabulary searches even worse.

The Sport Thesaurus represents the biggest hurdle and is the bane of subject searching in SPORTDiscus. The thesaurus design flies in the face of many of the basic tenets of thesaurus construction, showing, among other signs, a severe case of geographic illiteracy. Not even Ovid's smart and elegant thesaurus-handling software can put it into a medal contender position. Dialog is knocked out by Sport Thesaurus' idiosyncrasies and makes it look even worse than it already is. This thesaurus appears to have started out as a faceted classification tool, then tried to become a keyword-based,

poly-hierarchical thesaurus, but has ended up as neither fish nor fowl.

Let's start with the small problems that wouldn't make anyone lose much sleep, but indicate the thesaurus' less-than-ideal attitude and aptitude. You don't need to be an avid reader of *National Geographic* to know that Bangladesh, Nepal, Pakistan, Sri Lanka, and India should not be listed under the entry for South East Asia, which is the region south of China and east of India. Burma (or Myanmar) fits that geographic criteria, but it is not listed in the thesaurus, although Burma is used as descriptor in records. The same is true for East Timor. Brunei is also part of that region, but it is not a descriptor in the Sport Thesaurus. Thus, an item titled "Sport Participation in Brunei" is assigned the descriptor *Malaysia*. Close enough geographically, but the Sultan of Brunei, one of the richest guys in the world, may not be delighted with this descriptor choice. Even his money cannot buy everything.

Finding some countries by browsing the thesaurus is an orienteering challenge, especially in Dialog's Guided Search implementation. It makes Indonesia appear as a broader term for the Balkan Region. It does not list Yugoslavia as a preferred term or even as an entry term although it is listed among the narrower terms of both Eastern Europe and the Balkan Region and as a broader term for Slovenia. The irony of the fact is not lost on me that Slovenia became independent more than a decade

before some of the other federated states of the former Yugoslavia. Still, it is the only one of them showing the relationship in the thesaurus.

These geographic gaffes are dwarfed by the incomprehensible policy of designating narrower-broader term relationships in the thesaurus. It has dire, large-scale consequences for indexing and especially for controlled-vocabulary-based retrieval efficiency on all the SPORTDiscus software platforms.

For example, descriptors like *strategy*, *recuperation*, *blocking*, and *serve* (why not *serve*?), are listed as narrower terms for many sports without any qualification. But these words are not specific to, say, "volleyball," as opposed to such terms as "beach volleyball," "bump pass," or "spiking." Still, all these are lumped together as narrower terms under the main term "volleyball." This is not only cognitively confusing, but it makes the result set utterly irrelevant when one uses the explode function for the main term in any of the host software, which searches all the narrower terms of the main term ORed together.

What is meant to be a smart operation to pick up all records that contain the main descriptor and its genuinely narrower terms with a single click turns into a senseless operation in SPORTDiscus. It will pick up, for example, all the records with the descriptor *strategy*—many of which have nothing to do with the descriptor

SPORTDiscus
 <1830 to July 2004>

PayPerView Account

Author
 Title
 Journal
 Search Fields
 Tools
 Combine
 Limit
 Basic
 Change Database
 Logoff

#	Search History	Results
1	volleyball/	7050
2	exp volleyball/	24756

Personal Account
 Saved Searches
 Save Search History
 Delete Searches

Users will explode when they see that the majority of "hits" have nothing to do with volleyball.

that was exploded. The result set for volleyball / as a descriptor (indicated by the slash in Ovid) yields 7,050 records. Using the explode button will create a result set of 24,756 records, with more than two-thirds of the results totally irrelevant to volleyball. This happens to thousands of descriptors. It is madness without method. For example, the word shoes is listed as a narrower term of "tennis," "squash," "golf," and "football," but not of "soccer," "basketball," or "boxing."

No matter what term you look up, chances are good you'll find narrower terms that are odd and confusing in one way or another. Basketball is designated as a narrower term of contact sport (mind the singular) along with karate and judo. True, Charles Barkley, Dennis Rodman, and some other players went out of their ways to make it a contact sport peppered with trash talk, but even they were not successful enough to justify this category assignment. Taekwondo is not listed

among the contact sports—and as a descriptor, it's spelled tae kwon do (with no "see reference" from taekwondo), even though it appears twice as many times in the titles and almost four times as often in the abstracts as "taekwondo." Careless attitude shows up everywhere. Muay thai is not a thesaurus term but at least it is an entry term with a cross reference to Thai boxing. Too bad that it is spelled muai thai in the thesaurus. In SPORTDiscus, the word does not occur in any records with this spelling. It appears as muay thai in 22 records (in 50 records on Dialog thanks to its duplicates and triplicates). Ignoring common usage supported by literary warrant is common in this thesaurus.

Omission of terms from the thesaurus is equally inexplicable, not only for country names but also for subject terms such as "ski boots." Again, Dialog adds insult to injury and further debilitates the thesaurus by denying

descriptor status to many terms when looking them up in Dialog's Guided Search mode, or when expanding the Basic Index for such terms as "soccer" and "mixed doubles."

This database is not unlike the female swimmers of East Germany in the 1970s—way overdosed on steroids. (when their coach was asked at a press conference about their oddly deep voices, the coach said that they came to swim not to sing.) With its ballast of duplicates and triplicates to make it look bigger, SPORTDiscus may sink, and its users drown, in infoglut.

Péter Jacsó [jaco@hawaii.edu] is professor of library & information science at the University of Hawaii's Department of Information and Computer Sciences.

Comments? E-mail letters to the editor to marydee@xmission.com.

Information Today, Inc.

invites YOU
to subscribe to

NewsLink

A **FREE** weekly e-mail newsletter, NewsLink is designed to highlight the information that both users and producers of information products and services need to do their jobs as effectively as possible. As a subscriber, on the first of every month you will receive a full-length issue filled with a variety of important information. Each section of NewsLink provides current coverage of news, articles, events, and books pertinent to the library and information industries:

- ◇ **NewsLink Monthly Spotlight**—featuring an original article written by Paula J. Hane on current industry news and trends.
- ◇ **NewsBreaks**—keeps you in tune and up-to-date on the latest industry happenings.
- ◇ **Featured Articles**—provides you with links to articles from the latest editions of Information Today, Inc. publications.
- ◇ **Conference Connection**—between Conference Updates and the Conference Calendar, you will get the latest event information for the library and information fields.
- ◇ **Bookshelf**—takes the opportunity to introduce you to the newest industry-related books.

In addition to the full-length issue, every Monday you will receive NewsLink NewsBreaks, a weekly update for our subscribers that acts as notification of our NewsBreaks as soon as they are posted.

Visit www.infotoday.com to subscribe!



Information Today, Inc.

143 Old Marlton Pike, Medford, NJ 08055