

Web of Science, Scopus,



by François Libmann

and Classical Online: Philosophies of Searching

Elsevier's introduction of Scopus in November 2004 clearly had Thomson's Web of Science in its sights. As direct competitors, the two products share interesting features, probably because of Elsevier's "user-centered design" approach to Scopus' development. It relied heavily on the opinions of more than 300 searchers and information professionals, people thoroughly familiar with Web of Science who had adapted their ways of working and search procedures to the Thomson product. Web of Science is an archetype of an information product, one that sets the standard for scientific searching.

Web of Science, available on the ISI Web of Knowledge platform, is essentially made up of three different databases that can be searched individually or together:

- Science Citation Index Expanded, covering more than 6,000 journals
- Social Sciences Citation Index, covering more than 1,800 journals
- Arts & Humanities Citation Index, covering more than 1,100 journals

With a few exceptions, all documents (articles as well as letters to the editor) are included (cover to cover), and the producer claims that no issue is missing. The whole database contains 38 million references.

In Science Citation Index Expanded, the coverage of several journals, such as the *British Medical Journal*, the *Journal of the American Chemical Society*, *Nature*, and *The Lancet*, goes back to 1900. References in the social sciences date from 1956; arts and humanities date from 1975.

COVERAGE COMPARISON

Scopus' coverage is larger than Web of Science's in terms of the number of titles—it covers 15,000. However, Scopus

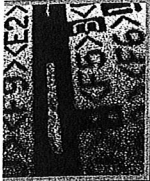
doesn't go back as far (1966 for many journals, earlier for some others), which is why the number of documents abstracted in Scopus is "only" 29 million. (In January 2007, Scopus announced that it will add the complete backfiles of journals published by Springer, Nature Publishing Group, the American Physical Society, and the American Institute of Physics, expanding its coverage by 7 million records.) Web of Science includes all of the references in the original articles, but Scopus' date only from 1996.

Scopus, however, includes conference proceedings, a document type not in Web of Science. Although the ISI Proceedings database is available on the ISI Web of Knowledge platform dating from 1990, cited references are available only for documents published after 1998. In disciplines where there are many conferences, the fact that they are not in the same database with the articles will have an impact in terms of bibliometric studies.

SEARCHING POSSIBILITIES

Web of Science offers a quick search as its first screen, which encourages a topic search and allows date range and database selections. From the results screen, you can refine results by subject category, source title, document types, authors, publication years, countries/territories, institutions, and languages. Advanced search allows you to build a search query using field tags for your search terms. Another specific screen is devoted to citation searching. Moreover, it is possible to write a query on chemical structures. Accessing the primary document is facilitated as much as possible with a series of links, most of them specific to individual subscribing institutions.

Although conceived with a similar "inspiration" and aimed at the same kind of users, Web of Science and Scopus are not identical. Our tests at BASES, along with other published comparisons, show several differences, which are briefly summarized here:



Although several comparisons between Web of Science and Scopus exist, none give a clear advantage to one over the other. In fact, the suggestion is usually to subscribe to both.

- Automatic truncation is available in Scopus but not in the Web of Science.
- AND is the default operator in both systems (until recently in Web of Science, it was adjacency).
- The only proximity operator in Web of Science is SAME for all terms occurring in the same sentence. Scopus offers more possibilities, but they are quite limited.
- In Web of Science, the number of documents contained in a single results set is limited to 100,000. For some complex searches, this is inconvenient.
- In Scopus, the first results from the various rankings are automatically displayed, and a "more" button allows you to see the next ones. You can change the default setting to display 20, 50, 100, or 200 search results per page. In Web of Science, you need to click the "Analyze" button to get these rankings. The number of documents that Web of Science can analyze was recently upgraded from 2,000 to 100,000.
- There are fewer analysis criteria in Scopus. Source, author, publication year, type of document, and subject are common to both services, but Web of Science also offers a ranking by country, institution name, and language.
- In Web of Science, the cited reference search screen offers three boxes: cited authors, titles of the cited publications, and publication years of the cited documents. Indexes for authors and title abbreviations are also available. You can retrieve citations for which some information is lacking. With Scopus, you can also search in the title of the cited article and, using advanced search, in each of the components of the cited reference.
- In Scopus, the administrator can input the list of the publications to which the institution subscribes so that users can seamlessly access full-text articles. In Web of Science, Thomson's help desk sets up these parameters.

WEB OF SCIENCE AND CLASSICAL HOSTS

Although several comparisons between Web of Science and Scopus have been published or have been the subject of conference presentations, none give a clear advantage to one over the other. In fact, the suggestion is usually to subscribe to both. What's more, there have been no real comparisons with the more traditional databases available through the classical hosts of Dialog and STN. And how does Google Scholar compare?

A few test searches, excluding patent information, reveal interesting anomalies. This is not a truly scientific study based on a statistically significant number of queries. (How many would that be, anyway?) These queries are designed to generate a limited number of answers to ease making comparisons, which could introduce bias.

THE LOTUS EFFECT

The first query was about the lotus effect as applied to the ceramic field, especially to tiles and bricks. This application was developed in Germany. The lotus effect, derived from the plant of the same name, is about surfaces that are so hydrophobic that the water drops on it remain almost totally spherical. Consequently, they do not adhere to the surface. They roll off, dragging with them the dust that is on the surface. Today, it is possible to manufacture materials or coatings with these properties.

The search was in English without any date limitation, with the AND operator between the expression lotus effect and various terms (ceramics, tile, or brick) truncated when appropriate.

This strategy led to zero answers in Web of Science, three in Scopus (1,750 in the Web Results portion of Scopus), and 73 in Google Scholar. Broadening the search to lotus effect by itself resulted in 37 hits in Web of Science and 43 in Scopus. In Dialog, 12 answers were retrieved after eliminating duplicates, but only 11 were relevant because one was a bibliographic reference to a patent. The STN search retrieved roughly the same number as in Dialog, plus three references in KOSMET, a database not available in Dialog.

Looking more precisely at these results, among the 11 references retrieved on the various databases of Dialog, three were actually present in Web of Science, but two lacked abstracts, and the abstract of the third one was shorter. Similarly, in addition to the two references retrieved from Scopus, four others were also present but not retrieved because their abstracts were different or missing.

Examining the first 10 references from Google Scholar, the hits were from the PASCAL and Ceramic Abstracts databases, which is logical because of the agreements between Google and the producers of these databases (INIST and CSA). Two links were broken, and one other document appeared because the word textile was misspelled as tex-tile.

The other documents from Google Scholar were retrieved because the search ran in the full text of the article, not just in the title, abstract, and keywords. This dilutes the relevancy of the results, as there may be only a passing mention of the technology. Two can be read on the screen, but it is impossible to print them.



Two Google Scholar hits were also retrieved on Dialog and Scopus but not in Web of Science. One document, without any publication date, was available only in Google Scholar.

CYSTADANE SEARCH

The second test search explored an orphan drug used for the treatment of homocystinuria: cystadane, the trade name for betaine. The search retrieved 16 references in Dialog databases after eliminating duplicates and non purely scientific references, about the same number in STN—one in Web of Science, seven in Scopus, and 20 in Google Scholar. When comparing the results obtained in Dialog and Scopus, beyond the seven references retrieved with the word cystadane in Scopus, eight others were there but not retrieved. Similarly, eight other references were in Web of Science, but only one was retrieved with the keyword.

In most cases, these additional items had no abstracts or far less detailed indexing than that available on Dialog or STN. EMBASE.com, for example, sports a drug-name field—a field nonexistent in Scopus and Web of Science—both on Dialog and on STN.

Among the 13 references present in Scopus and Web of Science, five were not cited by any other document and three by only one or two. Retrieving these articles when navigating using citations forward or backward is, therefore, unlikely. However, one article found in Scopus, published in 2001, has been cited 147 times.

TWO VERY DIFFERENT PHILOSOPHIES

Dialog and STN services are descendants of the traditional scientific documentation systems. The scientific databases they host are often computerized versions of printed abstracts bulletins, which were disseminated long before the databases' birth.

Traditionally, abstracts are quite detailed. In some databases such as COMPENDEX (since 1970) and FSTA, the percentage of references with an abstract is greater than 95 percent; it even reaches 99 percent in INSPEC for records after 1969. Moreover, these databases have very comprehensive indexing.

In most cases, there is no mention of the citations that are present in the original article. This is the unique feature of ISI citation databases, which pioneered the concept of citation searching. In Dialog and STN, referenced citations are usually very much abbreviated (the article title is not present, and the publication title is greatly compressed) so that the sole possible usage is for statistics.

A few other producers, such as Chemical Abstracts on STN, have begun to add citations, but the functionalities offered differ greatly from what is possible on Web of Science or on Scopus.

PROS AND CONS OF CITATION SEARCHING

Searching via citations in the scientific literature databases brings several advantages. Theoretically, the invisible college concept is far from being meaningless. Searching by

navigating from one article to others intellectually linked to it is clearly of interest. Practically, the intellectual effort needed to use citation searching for relevant information is much more limited than identifying relevant keywords or thesauri terms and then combining them in an optimal manner.

Citation search appeals to those uninterested in professional information-retrieval techniques. Researchers and students who avoid asking an information professional for help, either because they wish to be independent or because no one is available, also embrace citation searching, with its high level of serendipity.

For initial searches on a large subject, citation searches lead quickly to the experts and to the most well-known or "classical" articles and authors published in what are supposed to be the most famous journals. Another important use of citation searching is for very targeted searches written by recognized experts.

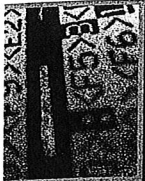
Drawbacks exist, however. As seen in the two sample searches, abstracts are present more often in traditional databases. According to the tests done by Péter Jacsó in 2005 ("As We May Search," *Current Science*, V. 89, No. 9–10, November 2005; www.ias.ac.in/currensci/nov102005/1537.pdf), 67 percent of Scopus records offer an abstract; this percentage is slightly higher than in Web of Science. [Editor's Note: Scopus reports, as of February 2007, that its ratio has increased to 70 percent.]

Thomson Scientific indicates that 90 percent of its records in the clinical medicine literature have abstracts, but this percentage falls to 25 percent in the arts and humanities. This frequent lack of abstracts, plus limited indexing, leads to researchers missing articles that are in the database if they search using a strategy based solely on combining keywords. When these articles are not cited by another article, which happens frequently, the probability of retrieving these articles is low. Does this fallibility in the search artificially increase the volume of documents really available in Scopus or Web of Science?

POTENTIAL PITFALLS

The quite restrictive selection of publications, much more evident in Web of Science than in Scopus, is justified by Bradford's law, which says that the main scientific information is limited to a small number of journals. That does not mean, however, that articles published elsewhere are not of interest. When working on a very narrow subject, the results may be more limited than expected. Taking this a step further, searchers who rely on the most "official" publications have very little chance of finding thoughts or authors outside the scope of dominant theories or of the consensus existing at that time.

Google Scholar is completely hazardous, even though it retrieves some relevant answers. But these answers are not always workable: There are broken links and documents you can't print. Google Scholar refuses to tell searchers what publications are included. Searches are run on the full text, with no attempt to support complex search strategies, so the level of noise is high. However, it is true that Google Scholar is free—that is very important for some people.



The ability to analyze citations on a carefully selected set of journals has another application apart from the search for information.

WEAKNESSES IN CITATION SEARCHING

This information is extracted from Chapter 4, "Les indicateurs bibliométriques et la mesure des performances scientifiques," of the report "Évaluation de la recherche publique dans les établissements publics français" published in December 2002 by the Comité National d'Évaluation de la Recherche (www.cner.gouv.fr/fr/pdf/bib.pdf). (Note this was before Scopus existed.)

- The number of journals taken in ISI is high as an absolute value but low compared to the number of scientific journals published worldwide. American journals dominate, and the English language is overrepresented.
- Articles begin to be cited only 1 year after publication.
- Although it is possible to take into account all the authors of a given article, it is impossible to know how the scientific work has been split between them.
- About 10 percent of the citations can be considered abusive because they belong to a citation network, while 10 percent are autocitations. However, some studies conclude that autocitations do not introduce significant bias in bibliometric work.
- Citations selection might be a problem: Some authors tend to cite only what they consider the most important references; others do not cite the most well-known authors. In other cases, only the authors defending the same position are cited. A researcher might not cite another researcher because of scientific competition.
- A very frequent behavior is to cite only the more recent results.
- Errors in citations can vary from 10 percent to 50 percent depending on the journal.

- Some authors cite other authors to contradict them while others avoid citing negative influences.
- Authors over cite references in journals of the publisher to which they submit their articles or of the supposed reviewer to increase the probability of acceptance of their articles.
- The average number of citations varies significantly from one discipline to another.

NOT AN EXACT SCIENCE

Although searches based on navigation via citations and their analysis has many advantages, they are far from being an exact science.

One can then think that a "classical" search, based on references with abstracts and detailed controlled vocabulary, is still very interesting—if it is possible to use a query language of good quality and if one makes the effort to build a search strategy that is complex enough. But you have to understand that the ability to analyze citations on a carefully selected set of journals has another application apart from the search for information.

Web of Science has, at least in France, become a "must-have" tool for the evaluation of researchers and laboratories. Some people use it only for this purpose. The number of publications—and particularly, the number of citations in a set of journals selected as being the most important—are considered excellent evaluation criteria.

The belief that this usage is a fortress will be hard to conquer by Scopus, particularly since Web of Knowledge signed a 3-year, \$7.2 million contract in September 2006 in France with a consortium including INIST, the Couperin Group, INRA, INSERM, CNRS, the Ministry of Research, and several universities (Lyon, Strasbourg, Paris, Grenoble, Marseille, and Montpellier). It will allow 60,000 searchers and 400,000 students to access the ISI Web of Knowledge platform, and, in particular to Web of Science and ISI Proceedings, some institutions also have access to other parts of the platform.

The competition between Scopus and Web of Science will certainly continue. With their differing philosophical approaches to information gathering and dissemination, particularly the role of controlled vocabulary and intensive indexing, the role of the classical hosts will not diminish. That will be extremely beneficial for scientific researchers and information professionals.

François Libmann (flibmann@bases-publications.com) is editor of BASES and the principal of FLA Consultants, a French information broker.

This article is a translation of the article published in the September 2006 issue (No. 230) of the French newsletter BASES under the title "Web of Science: une philosophie de la recherche." BASES is available online on LexisNexis, Factiva, and EBSCOhost.