



SAVVY SEARCHING

h-index using
Web of Science

The pros and cons of computing the h-index using Web of Science

Péter Jacsó

University of Hawaii, Laie, Hawaii, USA

673

Abstract

Purpose – The purpose of this paper, the fourth in a series, is to discuss the pros and cons of the h-index in the Web of Science (WoS).

Design/methodology/approach – The paper looks at the content and software advantages and disadvantages of WoS from the perspective of calculating the h-index as a single measure of published research output and influence at the individual researcher level.

Findings – The paper finds that there are notable similarities and differences between WoS and Scopus, and that any WoS edition has a unique and important feature. As opposed to other cited reference enhanced databases, WoS always includes all the cited references for every record created, irrespective of the publication year.

Originality/value – The paper provides insights into the advantages and disadvantages of WoS from the perspective of calculating the h-index.

Keywords Indexing, Databases, Information retrieval, Research

Paper type Research paper

Introduction

The background to this analysis was laid out in the introductory overview (Jacsó, 2008b), which discussed the plausibility of computing the h-index of scholarly publishing productivity and the impact of researchers using reference enhanced databases. It presented essential database content and software features for calculating a reasonable h-index for researchers, using examples from a number of citation enhanced databases. This was followed by reviews of the advantages and limitations of Google Scholar (GS) (Jacsó, 2008c) and Scopus (Jacsó, 2008d) for computing the h-index.

This fourth part of the series looks at the content and software advantages and disadvantages of Web of Science (WoS) from the perspective of calculating the h-index as a single measure of published research output and influence at the individual researcher level. Discussion and comments about methodological issues related to determining and/or estimating the vital statistics of the databases for computing the h-index (total number of records, percentage of the total number of cited reference enhanced records, their distribution across the years and among document types) have also been included.

Those readers who are not familiar with the three major multidisciplinary reference enhanced systems will find indepth reviews for general background information in several papers (Dess, 2006; Fingerman, 2005, 2006; Goodman, 2004; Goodman and Deis, 2007; Jacsó, 2007a, c, 2008a; Mayr and Walter, 2007; Myhill, 2005; Neuhaus and Daniel, 2008; Neuhaus *et al.* 2006, 2008; Noruzi, 2005; Robinson and Wusteman, 2007; Tenopir, 2005; Vine, 2006; White, 2006; Wleklinski, 2005). Evaluation of these sources,



with emphasis on their citation analysis capabilities or related matters such as breadth of source coverage, are discussed (among others) in Bauer and Bakkalbasi (2005), de Moya-Anegón *et al.* (2007), Gavel and Iselid (2008), Jacsó (1998, 2005a, b, 2006, 2007b), Meho and Yang (2007b), Yang and Meho (2006) and Walters (2007). Papers addressing the theoretical and practical aspects of the h-index itself for the purpose of evaluating and comparing the publishing performance of scholars, beyond the seminal work of Hirsch (2005), include Bar-Ilan, 2006, 2008; Bar-Ilan *et al.*, 2007; Bornmann and Daniel, 2005; Cronin and Meho, 2006; Glänzel and Persson, 2005; Harzing and van der Wal, 2008; Iglesias and Pecharroman, 2006; Jeang, 2007; Norris and Oppenheim, 2007; Oppenheim, 2007; Sanderson, 2008 and Vinkler, 2007. Those papers that compare two or more systems are of particular interest.

Content issues

WoS is available in many different versions (editions) and always needs clear identification by qualifying its name with years of coverage as it depends on what content the library chooses to license. The century edition is the most comprehensive, covering publications from 1900 in the sciences, 1956 in the social sciences and 1975 in the arts and humanities. It is to be understood that within these timeframes for the components, it is the library that decides the start of the time span. For example, my library at the University of Hawaii chose 1980 for each component. Another library may choose only the sciences component, but going back to, say, 1900 or 1945, depending on a variety of reasons. These may include the disciplinary composition, the weights or importance of the courses offered by the university or of the research fields, or grant applications applied for by research and/or teaching faculty researchers at the parent institutes. The availability at the library of other, discipline-specific, cited reference enhanced databases that can be used directly or indirectly for calculating the h-index, such as the Astrophysics Data System (ADS), the Physics Review Online Archive (PROLA) or the CSA Illumina implementation of the PsycINFO database, may also influence the choice of what edition is licensed and used (There are several implementations of PsycINFO and other databases of the American Psychological Association, but for various reasons the other editions by Ebsco, Ovid, OCLC and Dialog do not make it possible to calculate the h-index and none of them reports the h-index automatically. At least, the CSA version allows the export of records into RefWorks with the citedness count so users can calculate the h-index).

The edition of WoS can significantly influence the h-index of the subjects evaluated, depending on the scientific age of the researchers whose publications are analysed, i.e. when they started publishing scholarly papers, what percentage of the journals in which they published were classified into one or more of the three main disciplinary components of WoS, and how consistently.

For the statistics and examples in this review, the WoS Century edition covering the 1900-2008 period was used in the first week of July, unless otherwise noted.

It is essential to know that any WoS edition has a unique and important feature. As opposed to other cited reference enhanced databases, WoS always includes all the cited references for every record created, irrespective of the publication year.

This is in contrast with Scopus, which includes cited reference information only for records of papers published from 1996 onward (There is a negligible set of nearly 7,300 records enhanced by cited references for papers published before 1996.). In GS, this is a

moot question because it takes items from any warm body of a cooperating partner's collection and from websites scraped by the searchbots of GS, including simple abstracting databases and single papers posted on the web. In some situations this is advantageous because GS practically always has a master record to hang any citations onto, and if it has none (or GS does not recognise it), it creates one from the cited references in any source document. Often, it can collate different records gathered from different sources, and with the "[citation]" label it indicates records that are just extracted cited references with no master records on their own. It shows the two types of records intermixed and does not allow filtering. This is in contrast with Scopus and WoS, and with the other systems (ADS, PROLA, PsycINFO, etc.) that use only genuine master records in accruing citations and calculating the h-index.

The size of the database and its reference enhanced subset

Only three cited reference enhanced databases have multidisciplinary coverage. The total number of records is reported automatically at the very bottom of the page by WoS for any valid search that produces a hit, telling how many hits were found from how many records. The database size is easy to determine in Scopus using the `py BEF 2009` command. On July 4, 2008 Scopus had 34.7 million master records and WoS had 40.5 million unique master records. The WoS figure given here eliminates the overlap between the three components, i.e. the ones that are added to more than one WoS component. Without this, the total number of WoS master records would be 42.4 million. Scopus covers a somewhat longer time span than WoS as it has more than 62,000 records for documents published before 1900 (not shown on the graph below), and has abstracts for many more records (24.6 million) than WoS (13 million). This is important for:

- retrieving more records about a search topic through finding matching query words in the abstract unavailable in the title or the index terms; and
- learning more about the items in the results list.

The reason for the much larger proportion of records with abstracts in Scopus is that in WoS abstracts were added only from 1989 for the sciences, 1992 for the social sciences and 2000 for the arts and humanities – if there was an abstract in the original paper. Scopus was primarily built from indexing/abstracting databases of Elsevier, such as EMBASE, GEOBASE and ECONBASE, and abstracts were created by staff members even if the source document did not have one.

However, from the perspective of citation searching and especially the h-index calculation, what matters most as value-added information is the number of records enhanced by cited references and the total number of cited references.

Subset of records enhanced by cited references

The former is easy to determine in the Dialog version of the three citation index databases of the Institute for Scientific Information (ISI), because Dialog has a command to limit the searches to such records. The start year of the time span of the ISI databases hosted on Dialog, however, is limited to 1972, 1974 and 1980 for the social sciences, sciences, and arts and humanities, respectively. Nevertheless, as the proportion of cited reference enhanced records are stable within each of the three major disciplinary categories across the years, knowing the total number of records for the

earlier periods in the three databases of WoS makes it possible to estimate the size of the cited reference enhanced subset fairly well.

For Scopus the number of records enhanced by cited references can be determined using test queries where the most common lead characters and numbers are used with truncation and then combined in Boolean OR operations.

The difference between WoS and Scopus in terms of the total number of cited reference enhanced records and their distributions across the entire time spectrum from 1900 to 2008 is very apparent from the graph below. It is to be noted that here I disagree with the estimate that appears in Scopus's most current marketing information, where it claims that it has 15 million records enhanced by cited references (Scopus, 2008). In my current test, the number of cited reference enhanced records was 12,169,625 (Figures 1 and 2).

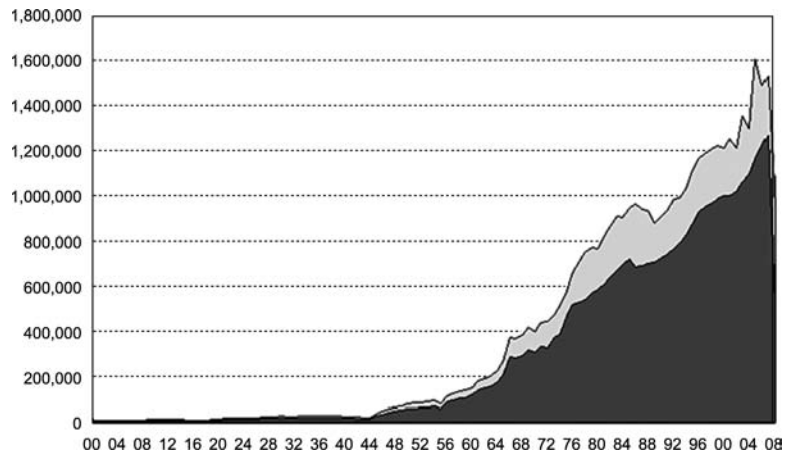


Figure 1.
Number of records enhanced by cited references in WoS

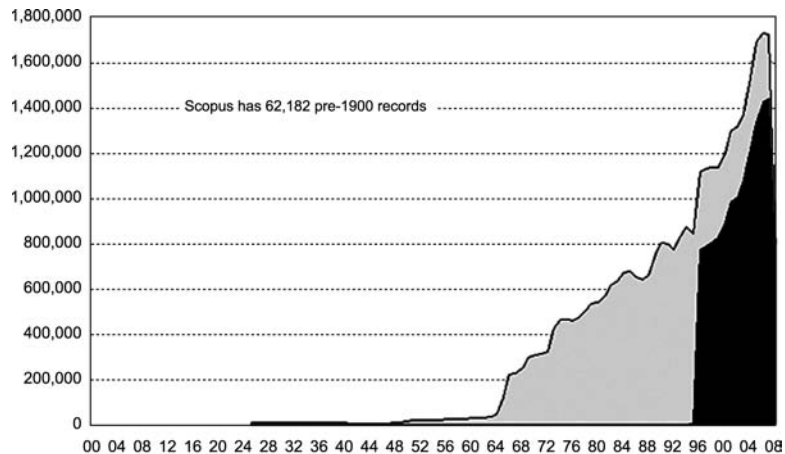


Figure 2.
Number of records enhanced by cited references in Scopus

Key:
Light area = total number of records/year
Dark area = records enhanced by cited references/year

Total number of cited references

The total number of cited references can also be estimated, although not as precisely as the number of records enhanced by cited references. It is still good enough for at least a ballpark estimate.

Once again, the Dialog subset of the three component citation databases shows the distribution of records by the number of cited references in a browsable index, and can be uploaded with a somewhat tedious process into a spreadsheet for further analysis, as I did it. The tedium comes from the limitation that only 50 entries are shown at once (actually 49 for technical reasons), so a lot of copy and paste actions are needed to do this, but it was worth it in order to get a reliable, long-term average for the average number of cited references in each of the three components. For the rest of the years not covered by the Dialog implementation, the average was used for each category.

All the cited references in the same source documents are included in both Scopus and WoS, as opposed to the practice of the late e-psyche database. Its creators, two veterans of the information industry, kept cutting more and more corners, including increasingly selective addition of references to the master records. It got to the point where only the first ten references were included. Considering that many journals list the references alphabetically by author, the effect of this was significant (www2.hawaii.edu/~jacso/extra/cj-03/eps/eps.htm).

The average in each of the three categories determined from the Dialog implementation were used to estimate the total number of cited references for Scopus. While it is true that the composition of the source documents is not the same within the three major categories of Scopus and WoS, this is still a plausible estimate based on the actual number of references from a sample of 33 million records on the Dialog platform.

My estimation is that there are more than 700 million cited references in WoS and about half as many in Scopus. This is smaller than the ratio of the difference in total number of records enhanced by cited references because WoS has more than three million records for papers in the arts and humanities category (with 83 percent of the records having about eight cited references on average), roughly one third of those in the sciences and social sciences categories. In Scopus, the cited reference enhanced subset in the arts and humanities category is merely 61,000 out of the fewer than 110,000 records.

It is impossible to get any reliable quantitative information from GS. It is remarkable how GS can get away with its hit and citation counts when it becomes obvious that it cannot even reproduce them a few minutes later when the same query is repeated. Even the most naïve GS users must suspend their awe of its features when a Boolean OR operation manages to reduce (!) the original set (Figure 3).

Simple queries by time periods that show reasonable hit counts in any scholarly or just professional database provide the GS searcher with absurd hit counts, as the above test searches show. Once again, larger numbers of hit counts for shorter time periods than for longer ones is as clear an indication of serious deficiencies in the GS software as the mishandling of the simplest Boolean OR operations. It is not possible to get even a ball-park estimate for the size of the database or the number of records by years available in GS.

Estimating the percentage of records enhanced by cited reference in GS is also impossible because it does not show the number of cited references even for the master records retrieved from the publishers' digital archives, let alone the bibliographic data



Figure 3.
Boolean OR operator
reducing the original set in
Google Scholar

for the cited references. It does use this information internally to calculate the citedness counts, but as I demonstrated earlier, the matching process is done very poorly and GS does not refrain from including obvious phantom citations.

In contrast, I have never seen in WoS and Scopus negative matches reported as positive matches, let alone phantom citations where there is nothing in the citing documents that could be interpreted as a reference to the purportedly cited work. One of the many examples is my paper about GS published in *Online Information Review* in mid-2005, which was cited – according to GS – in a dissertation written a year earlier than my paper was published and several months before GS was announced.

As for the disciplinary composition of WoS, it has covered the social sciences and especially arts and humanities much more broadly and deeply than Scopus. While the total number of records in the sciences category is practically identical in WoS (33.25 million) and Scopus (33.60 million), these represent 79 and 97 percent in WoS and Scopus, respectively. There are 5.6 million records for the social sciences component of WoS (13 percent), while in Scopus there are just above one million records (3 percent) for the disciplines belonging to the category of social sciences and fewer than 109,000 records (0.31 percent) for arts and humanities. In WoS, there are 3.6 million records (8 percent) for arts and humanities papers.

Norris and Oppenheim (2007) pointed out that in their test Scopus yielded 5.4 percent more citations for their sample. It must be understood, however, that the sample – as the authors point out – was of fairly recent papers. Dess (2006) found a smaller 1.2 percent advantage for Scopus, but once again the sample was of post-1995 documents. Dess emphasises that for researchers who were active in publishing before 1996, the web database is the only appropriate resource to use. I agree with this opinion, especially because, as implied in Hirsch's article, the h-index for a researcher was meant to be calculated for the scientific lifetime of the researcher.

The above-mentioned distribution of records among the three major categories has considerable impact on the h-index of researchers in the social sciences and the arts and humanities. WoS produces higher h-indexes than Scopus for the vast majority of papers and researchers in these two major disciplinary areas who started publishing before 1996. This is partly attributable to the fact that Scopus ignores by principle not only all the pre-1996 publications in the automatically reported Author Details page, but also the citations they received from papers published after 1995.

Source coverage

The source coverage of Scopus is much wider than that of WoS. In WoS, the number of active journals covered is less than 10,000 while in Scopus this number is above 15,000. Then again, this number in itself may not be fair because in Scopus the retrospectivity and depth of coverage of journals is very different from those in WoS. Scopus started to cover thousands of journals only in the past few years which have been covered in WoS for decades.

Looking at a few examples from the library and information science field clearly shows these differences of late pick-ups, such as for *EContent* (which WoS started to cover in 1999, Scopus from 2000), *Serials Librarian*, *Libri*, *Interlending & Document Supply* and *Library Quarterly* (all of which are covered in Scopus from 1996 and in WoS from the 1980s).

There are also inexplicable gaps in Scopus where journals' coverage started in the 1970s, but then was suspended or neglected for over a decade, then picked up again. For the very significant differences in the depth of coverage, it is only a partial answer that Scopus does not include records for book reviews, as many journals with modest book reviews or no book reviews at all still show much more shallow coverage in Scopus than in WoS. There are late pick-ups, gaps and especially journals with only temporary coverage in WoS also, but not to the extent seen in Scopus.

Undoubtedly, WoS is much more conservative and selective in choosing journals, because if they are not cited to the extent of the average journal in the subject category, there is not much reason to include them and add to the very significant expense of indexing journals.

In a sudden turn, and responding to the criticism of not covering regional journals, in 2008 WoS started to add an additional 700 journals to its stable. I have no doubt that journals about, say, camel breeding are important for some countries, but it seems that covering the journal dedicated to this topic (as has been done by WoS) would suffice for 99 percent of its customer base, especially along with the occasional coverage of breakthrough developments in camel breeding by more universal journals of animal husbandry and by multidisciplinary journals.

This enhancement may be a good test to see to what extent the content of such regional journals and serials dealing with topics of modest worldwide interest would increase the utility of a cited reference enhanced database from a broader perspective. These journals are undoubtedly important for the given country, but not necessarily for researchers working and teaching elsewhere in the subject area.

With that said, I am the first to criticise WoS for not picking up journals of worldwide interest in the library and information science field, especially when they are high quality open access journals with a constructed impact factor that is higher than the median in the category. My examples include *D-Lib Magazine* and *First*

Monday, both of which are excellent journals and are very well covered by Scopus, but not by WoS.

Document types

As for document types, Scopus has much better coverage of conference papers than WoS. Comparison is difficult as the two systems use different document type categories with only a few in common (article, note, and review). WoS does not even have the document type “conference paper”, or “conference proceedings”, even though it has tens of thousands of records that are conference papers. Searching for the term “proceedings” in the source name field does not help as there are many journals with that word in their title including the *Proceedings of the National Academy of Sciences* and the proceedings of the various royal academies of the UK.

In Scopus, there are 3.25 million documents with conference paper designation and nearly half of them are enhanced by cited references. Actually, in the Dialog subset of ISI data there are records for more than four million meeting abstracts taken mostly from journals, but less than 10 percent have cited references because such abstracts usually do not include the references cited in the full paper. There is a separate product on the Web of Knowledge platform, ISI Proceedings, dedicated to conference papers, but it does not have the Citation Report feature.

As for monographic books, WoS does not have master records for this document type, while Scopus identifies 20,000 documents as books, but only 27 of them are enhanced by cited references, so they do not matter from the h-index perspective as citing documents. These master records can still be useful because they can accrue the citations the books received from documents that have cited references in Scopus, whereas in WoS all references to monographic books by definition are orphan references because they do not have master records. These can be identified through cited reference searching, but there is no h-index calculation option for such search results. (The implications of this are explained in the section on software issues.) It is to be noted that Scopus recently added more global source type categories, limited to five types – journals, conference proceedings, book series, books and reports. The last two are same as the ones in the document type category.

Software issues

WoS was the first database to incorporate the h-index and got it right as a first-timer. There are still some features that I would like to see changed and in some cases even the adaptation of solutions introduced by Scopus would be useful.

WoS offers an option right on the results list that shows the h-index and several other bibliometric measures, called “Citation Report”. The name offers enough of a clue, although “Citation Measures” would be more suggestive. WoS automatically calculates the measures on the set created by the query. It is a very good idea to show immediately the total number of source records for the items matching the query along with the total number of citations received and the average number of citations per item. These give an instant idea of the performance of the researcher.

I would prefer WoS to show automatically the number of citing items (with a link to the existing footnote/help information which explains that the latter may be fewer than the former because a single article may cite more than one of the papers of the author whose research output measure is being searched). The option to display the list of

citing articles either with or without self-citation is great, because one of the reservations against the original h-index idea is that it favours self-citing researchers (Reaching a certain point of productivity, most authors would engage in self-citing by necessity, to provide links to one or more of their former papers about or related to the subject. As long as on average the self-citation rate is not more than 25-30 percent, this is a normal citation behaviour.). The citation report mini-table is wrapped up with presenting the h-index (again with a smart pop-up help to link to for explanation about the measure) (Figure 4).

Quite tellingly, the h-index for Jorge E. Hirsch is just 16 in Scopus. It is based on 170 records in Scopus, as opposed to 186 in WoS, but the primary reason for the huge difference is due to the policy to disregard all the citations received by pre-1996 publications, including those received from papers published after 1995.

It alleviates this serious deficiency and injustice to many of the most productive and influential researchers who started publishing before 1996 that one can just scroll down the list of the author's papers sorted by citedness to get to the point when the citedness count of a paper is larger than or equal to its rank number. In the case of Jorge E. Hirsch, this is still only 33 as opposed to 50 in WoS, but more than twice as high as in the automatically calculated h-index by Scopus (Figure 5).

WoS's graphical side-by-side display of published items in each year and citations in each year is excellent. These provide data for the most recent, maximum 20-year window in a well-designed bar chart format that clearly marks the dynamism and the changes throughout the years. If needed, a smart link can pop-up the charts in a full screen pane for the full time span of yearly published items and yearly citations.

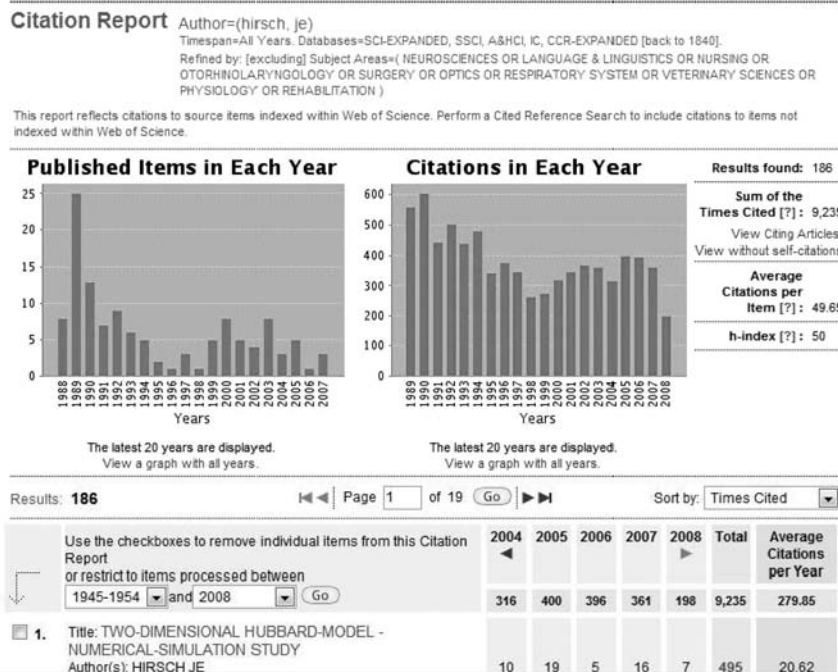


Figure 4. WoS Citation Report with automatically generated h-index for Jorge E. Hirsch

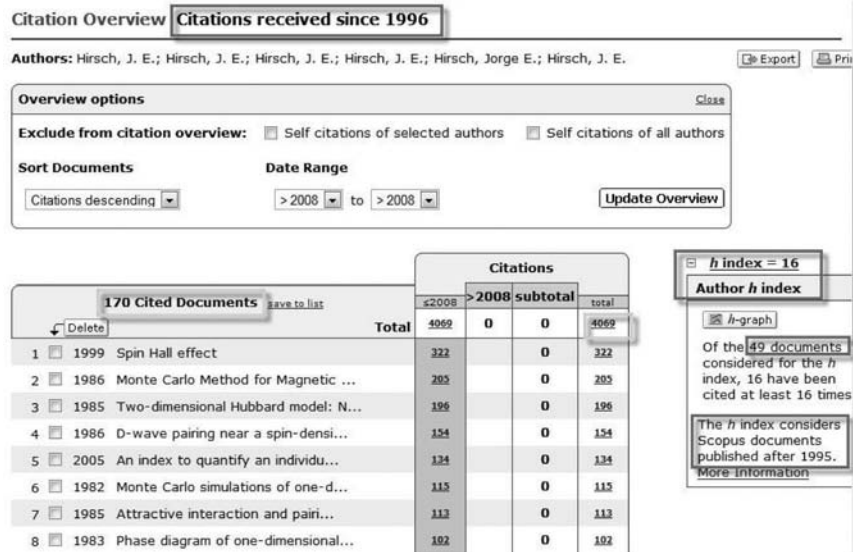


Figure 5. Automatically generated h-index value in Scopus for Jorge E. Hirsch

I would like to be able to personalise this option and see the graphs first for the entire spectrum and its visual representation even if the bar charts would be twice as wide or even wider (and therefore the bars much narrower) in the case of authors who have published for more than 40 years. This panoramic visualisation could then be clicked to get a narrower and closer view.

One of the best features in the Citation Report of WoS is that in addition to the total number of citations received by a paper, it also shows the number of citations received year by year. I wish the results list could also be sorted by this worthy element. Hirsch himself referred to the importance of considering the m-factor for the h-index, which is exactly this figure – the number of times a paper was cited divided by the number of years since it became published and citable.

In Scopus, one needs to know that the Citation Tracking button would bring up the h-index if the user first chose the items on the results list to which the h-index would apply. The motivation probably is that there may be records that should be excluded from the results list, but this approach may seem to be a bit school masterish for some users. A chart can be popped up and the h-index displayed in a small frame. The time span can be specified by the user and the graph redrawn (which may push the h-index frame off the screen and require scrolling). Scopus does not show the citation received per year for each item, but using its very convenient and nifty export feature this can be easily constructed in a spreadsheet.

In a rare moment of self-restraint, GS does not report the h-index, perhaps realising that the software is already way out of its boundaries. It does provide the citations – purportedly – received by each item in the results list. Thus, GS does not perform the h-index calculation, but provides the data (very often incorrect) for some of the third-party h-index generators. GS stopped ranking the results by citedness count in 2007 and it must have had a reason for this, but there is no statement (PR or authorial) about it. There is a statement about ranking the results by relevance, but it is just

“talk”, which becomes obvious when duplicates and triplicates appear widely scattered in the results list from high to low rank position when the results set is sorted using the Publish or Perish software (www.harzing.com/pop.htm).

Reporting of the purported citing documents obviously has the same basic limitations and problems as reporting the number of hits purportedly matching the query terms. GS stretches itself way beyond its abilities in this regard, as to determine the item-level citedness it also has to recognise a combination of matching elements (such as author name, journal name, volume number, starting page number), which it has difficulty with when handling these as single, let alone combined, data elements. The free and smartly designed utility, Publish or Perish, does calculate the h-index from the results served up by GS, but given the millions of records where the parsing of GS produces wrong publication years and author names and the equally deficient matching algorithms produces absurd hit and citation counts, it is like serving stale fish and chips with the best silverware in Buckingham Palace.

Handling orphan and stray references

In WoS, there is a low-key cue to encourage users to go to the Cited Search menu option to (among other things) look up references that could not be matched against a master record either because there is no master record for the cited paper (I call these “orphan references”) or because the cited reference is erroneous or at least differs from the key bibliometric element(s) in the master record (I call these “stray references”).

For calculating a reasonable h-index for any researcher, it is essential to check out the orphan and stray references, and see if these are likely to change the h-index (as well as the other bibliometric measures) when they are combined with the correctly matching citations to recalculate the metrics. In many cases, they would and in some cases quite significantly. Unfortunately, this is a really arduous process, more so in Wos than in Scopus. Still, for certain purposes this must be done to avoid discrimination against certain researchers. From the practical rather than the high-brow academic aspect, those most vulnerable to massively underreported, automatically calculated h-indexes may have: easy to misspell names (long ones with many consonants and few vowels, such as Brzezinski and many other Polish names); more than one, inconsistently used middle initial in the references (such as Michael E.D. Koenig); an accented character or unusual diphthong in their names, which asks for trouble and spelling errors (such as Jacsó); written for journals whose title is abbreviated in dozens of different ways (such as the *Journal of the American Society for Information Science and Technology*); published many, often cited books (such as Wilfrid F. Lancaster) and conference papers (such as Mike Thelwall or Lokman Meho). In addition, one can always count on a large percentage of misspelled journal titles and incorrect volume, issue and page numbers of even the simplest data elements by us authors (including – embarrassingly enough – myself). Some of these factors are related to the characteristics of the database content or software, others are generally true.

In spite of its gross deficiency, GS is the least discriminating against authors of books that are cited, by virtue of using millions of books from the Google Books project. PsycBooks is good too, but tiny even for the single field of psychology.

Special conventions for punctuation of the database content creator may play a role too. Those with a hyphenated last name, such as Judit Bar-Ilan, really do not have a chance to get a fair h-index in WoS without merging in references that cannot find the

master record because of the ancient convention of ISI to eliminate the hyphen and apostrophe, as well as the okina used in many Hawaiian names, and it is still a “curse of the mummy”. One must also search for “Barilan J” to find some of her best cited papers. The sample information on the search template warns users to search for “O’Brian” as “OBrian” also (both with and without the apostrophe), but in the twenty-first century this is something that should typically be done by the software automatically.

The importance of using the Cited Reference search option cannot be overestimated. Orphan and stray references are ignored in calculating the h-index and the other bibliographic measures. In my estimation, these represent at least 10-12 percent of the matching references in WoS.

When I recalculated the h-index of Wilfrid F. Lancaster (Jacsó, 2008e) for the contribution to the Festschrift issue of *Library Trends* for his 75th birthday, it doubled his original h-index from 14 to 28. Except for one journal article that did not have a master record in WoS, but had many orphan references, the greatest contributors to his recalculated h-index were the orphan references to his books. Stray references that did not match the master records in WoS also contributed to the rise of his h-index to a reasonable level (More details are provided about the process of manually recalculating the h-index using the Cited Reference search mode in Jacsó (2008e).

Scopus handles orphan and stray references in a much more user-friendly way with the recent introduction of a new feature. There is now an additional tab displayed on the main results list. Clicking on, it shows the stray and orphan references. As of early July, there were more than 55 million such references. The number is not known for WoS but must be in the same league because although it may have fewer orphan records for journal articles, it has more orphan records for conference papers. The ratio of stray records is assumed to be the same, but the absolute number of them is higher in WoS because it has cited references for almost twice as many records as Scopus.

It adds to the second-class citizen status of orphan and stray records in WoS that the results list shows a maximum of 50 entries per screen (Scopus’s limit is 200 per page) and limits the entries to 500 (Scopus’s limit is 2,000). In addition, WoS does not allow on-screen sorting of the results sets, which would be very important for getting a feel for the most cited papers, whose orphan and stray citations are not given credit in the automatic calculation of the h-index. One can get a good first impression in Scopus when sorting by the times cited data element, focusing on the values just below the automatically computed h-index value.

This is not enough in itself because a good dozen singleton orphan or stray references for a number of papers or books or conference papers added together may increase the automatically computed h-index value of, say, 11 for a researcher by 30-40 percent.

WoS must provide better tools for handling the orphan and cited references to make better use of its huge set of such records for calculating the h-index, as well as the several other informative bibliometric measures.

It could also introduce its own innovation by allowing the loosening of the citation matching algorithm by showing a list of stray references automatically that are close to matching references except for the wrong starting page numbers, volume numbers and/or missing issue numbers. Marking the non-matching parts in red in a longer and

dynamically sortable list and offering check-boxes next to the item could make the process far more efficient.

Automatically launching a search and reporting the total – as is done in Scopus – could have the same effect and would alert the users to the importance of looking up stray and orphan references. Providing an option to recalculate the h-index incorporating the orphan and stray records marked by the user would reap the benefit of tens of millions of ignored references and calculate a more reasonable h-index.

There are many additional issues related to computing a reasonable h-index from cited reference enhanced databases, such as the handling of self-citations and multiple authorships. There are several highly relevant papers that address one or more of these issues and other controversial aspects of the h-index (Bar-Ilan, 2008; Meho and Yang, 2007a, b; Schreiber, 2007; Vanclay, 2007). Clearly, there is no consensus on this matter. But the variety of derivatives proposed by the most experienced scientometricians (Bornmann *et al.*, 2008; Egghe, 2006a, b, c; Harzing, 2008; Liang, 2006; Jin *et al.*, 2007; Vanclay, 2006; van Raan, 2005) also clearly shows that the seminal work of Jorge E. Hirsch triggered further enhancements based on expert testing, and there is great interest in developing alternatives and rational variants that can co-exist in order to help in the measurement and evaluation of the publishing productivity and impact of researchers.

References

- Bar-Ilan, J. (2006), "H-index for price medalists revisited", *ISSI Newsletter*, Vol. 2 No. 1, pp. 3-5.
- Bar-Ilan, J. (2008), "Which h-index? A comparison of WoS, Scopus and Google Scholar", *Scientometrics*, Vol. 74 No. 2, pp. 257-71.
- Bar-Ilan, J., Levene, M. and Lim, A. (2007), "Some measures for comparing citation databases", *Journal of Informetrics*, Vol. 1 No. 1, pp. 26-34.
- Bauer, K. and Bakalbasi, N. (2005), "An examination of citation counts in a new scholarly communication environment", *D-Lib Magazine*, Vol. 11 No. 9, available at: www.dlib.org/dlib/september05/bauer/09bauer.html
- Bornmann, L. and Daniel, H. (2005), "Does the h-index for ranking of scientists really work?", *Scientometrics*, Vol. 65 No. 3, pp. 391-2.
- Bornmann, L., Mutz, R. and Daniel, H. (2008), "Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine", *Journal of the American Society for Information Science and Technology*, Vol. 59 No. 5, pp. 830-7.
- Cronin, B. and Meho, L. (2006), "Using the h-index to rank influential information scientists", *Journal of the American Society for Information Science and Technology*, Vol. 57 No. 9, pp. 1275-8.
- de Moya-Anegón, F. *et al.*, (2007), "Coverage analysis of Scopus: a journal metric approach", *Scientometrics*, Vol. 73 No. 1, pp. 53-78.
- Dess, H.M. (2006), "Scopus", *Issues in Science and Technology Librarianship*, Winter 2006, available at: www.istl.org/06-winter/databases4.html
- Egghe, L. (2006a), "Theory and practise of the g-index", *Scientometrics*, Vol. 69 No. 1, pp. 131-52.
- Egghe, L. (2006b), "How to improve the h-index", *The Scientist*, Vol. 20 No. 3, pp. 315-21.
- Egghe, L. (2006c), "An improvement of the h-index: the g-index", *ISSI Newsletter*, Vol. 2 No. 1, pp. 8-9.

- Fingerman, S. (2005), "Scopus: profusion and confusion", *Online*, Vol. 29 No. 2, pp. 36-8.
- Fingerman, S. (2006), "Web of Science and Scopus: current features and capabilities", *Issues in Science and Technology Librarianship*, Fall, available at: www.istl.org/06-fall/electronic2.html
- Gavel, Y. and Iselid, L. (2008), "Web of science and Scopus: a journal title overlap study", *Online Information Review*, Vol. 32 No. 1, pp. 8-21.
- Glänzel, W. and Persson, O. (2005), "H-index for prize medalists", *ISSI Newsletter*, Vol. 1 No. 4, pp. 15-18.
- Goodman, A. (2004), "Google Scholar vs. real scholarship", available at: www.traffick.com/2004/11/google-scholar-vs-real-scholarship.asp
- Goodman, D. and Deis, L. (2007), "Update on Scopus and Web of Science", *The Charleston Advisor*, Vol. 8 No. 3, pp. 15-18.
- Harzing, A-W. (2008), "Publish or perish – metrics", available at: www.harzing.com/pop.htm#metrics
- Harzing, A-W. and van der Wal, R. (2008), "Google Scholar – a new data source for citation analysis", available at: www.harzing.com/pop_gs.htm
- Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academy of Sciences*, Vol. 102 No. 46, pp. 16569-72.
- Iglesias, J.E. and Pecharroman, C. (2006), "Scaling the h-index for different scientific ISI fields", Preprint available at: <http://arxiv.org/ftp/physics/papers/0607/0607224.pdf>
- Jacso, P. (1998), "Analyzing the journal coverage of abstracting/indexing databases at variable aggregate and analytic levels", *Library & Information Science Research*, Vol. 20 No. 2, pp. 133-51, available at: www.jacso.info/PDFs/jacso-analyzing.pdf
- Jacso, P. (2005a), "Google Scholar: the pros and the cons", *Online Information Review*, Vol. 29 No. 2, pp. 208-14, available at: www.jacso.info/PDFs/jacso-google-scholar-pros-and-cons.pdf
- Jacso, P. (2005b), "As we may search – comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases", *Current Science*, Vol. 89 No. 9, pp. 1537-47, available at: www.ias.ac.in/currsci/nov102005/1537.pdf
- Jacso, P. (2006), "Deflated, inflated and phantom citation counts", *Online Information Review*, Vol. 30 No. 3, pp. 297-309, available at: www.jacso.info/PDFs/jacso-deflated-inflated.pdf
- Jacso, P. (2007a), "Scopus", available at: www.gale.cengage.com/reference/peter/200711/scopus.htm
- Jacso, P. (2007b), "The dimensions of cited reference enhanced database subsets", *Online Information Review*, Vol. 31 No. 5, pp. 694-705, available at: www.jacso.info/PDFs/jacso-dimensions-of-cited-refs.pdf
- Jacso, P. (2007c), "Web of Science", available at: www.gale.cengage.com/reference/peter/200701/wos.htm
- Jacso, P. (2008a), "Google Scholar revisited", *Online Information Review*, Vol. 32 No. 1, pp. 102-14, available at: www.jacso.info/PDFs/jacso-GS-revisited-OIR-2008-32-1.pdf
- Jacso, P. (2008b), "The plausibility of computing the h-index of scholarly productivity and impact using reference enhanced databases", *Online Information Review*, Vol. 32 No. 2, pp. 262-83, available at: www.jacso.info/PDFs/jacso-h-index-plausibility-OIR-2008-32-2.pdf
- Jacso, P. (2008c), "The pros and cons of computing the h-index using Google Scholar", *Online Information Review*, Vol. 32 No. 3, pp. 437-51, available at: www.jacso.info/PDFs/jacso-pros-and-cons-of-computing-the-h-index.pdf
- Jacso, P. (2008d), "The pros and cons of computing the h-index using Scopus", *Online Information Review*, Vol. 32 No. 4, p. 5.

-
- Jacsó, P. (2008e), "Testing the calculation of a realistic h-index in Google Scholar, Scopus and Web of Science for F.W. Lancaster", (to appear in the special issue of Library Trends, Summer 2008), available at: www.jacso.info/PDFs/lancaster.pdf
- Jeang, K-T. (2007), "Impact factor, h index, peer comparisons, and retrovirology: is it time to individualize citation metrics?", *Retrovirology*, Vol. 4 No. 42.
- Jin, B., Liang, L., Rousseau, R. and Egghe, L. (2007), "The R- and AR-indices: complementing the h-index", *Chinese Science Bulletin*, Vol. 52 No. 6, pp. 855-63.
- Liang, L. (2006), "H-index sequence and h-index matrix: constructions and applications", *Scientometrics*, Vol. 69 No. 1, pp. 163-9.
- Mayr, P. and Walter, A. (2007), "An exploratory study of Google Scholar", *Online Information Review*, Vol. 31 No. 6, pp. 814-30.
- Meho, L.I. and Yang, K. (2007a), "Fusion approach to citation-based quality assessment", paper presented at 11th International Conference of the International Society for Scientometrics and Informetrics, 25-27 June, Madrid, available at: www.slis.indiana.edu/faculty/meho-fusion-approach.pdf
- Meho, L.I. and Yang, K. (2007b), "Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 13, pp. 2105-25.
- Myhill, M. (2005), "Google Scholar", *The Charleston Advisor*, Vol. 6 No. 4.
- Neuhaus, C. and Daniel, H. (2008), "Data sources for performing citation analysis: an overview", *Journal of Documentation*, Vol. 64 No. 2, pp. 193-210.
- Neuhaus, C., Neuhaus, E. and Asher, A. (2008), "Google Scholar does to school: the presence of Google Scholar on college and university web sites", *Journal of Academic Librarianship*, Vol. 34 No. 1, pp. 39-51.
- Neuhaus, C., Neuhaus, E., Asher, A. and Wrede, C. (2006), "The depth and breadth of Google Scholar: an empirical study", *Portal: Libraries and the Academy*, Vol. 6 No. 2, pp. 127-41.
- Norris, M. and Oppenheim, C. (2007), "Comparing alternatives to the Web of Science for coverage of the social sciences' literature", *Journal of Informetrics*, Vol. 1 No. 2, pp. 161-9.
- Noruzi, A. (2005), "Google Scholar: the new generation of citation indexes", *Libri*, Vol. 55 No. 4, pp. 170-80.
- Oppenheim, C. (2007), "Using the h-index to rank influential British researchers in information science and librarianship", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 21, pp. 297-301.
- Robinson, M.L. and Wusteman, J. (2007), "Putting Google Scholar to the test: a preliminary study", *Program*, Vol. 41 No. 1, pp. 71-80.
- Sanderson, M. (2008), "Revisiting h measured on UK LIS and IR academics", *Journal of the American Society for Information Science and Technology*, early view edition published in advance of the print edition on 18 March 2008, available at: <http://dx.doi.org/10.1002/asi20771>
- Schreiber, M. (2007), "Self-citation corrections for the Hirsch index", *Europhysics Letters*, Vol. 78 No. 3, available at: <http://dx.doi.org/10.1029/0295-5075/78/30002>
- Scopus (2008), "Scopus in detail: facts & figures", available at: www.info.scopus.com/detail/facts/ (accessed July 2008).
- Tenopir, C. (2005), "Google in the academic library: undergraduates may find all they want on Google Scholar", *Library Journal*, Vol. 130 No. 2, p. 32, available at: www.libraryjournal.com/article/CA498868.html

-
- Vanclay, J.K. (2006), "Refining the h-index", *Scientist*, Vol. 20 No. 7, pp. 14-15.
- Vanclay, J.K. (2007), "On the robustness of the h-index", *Journal of the American Society for Information Science and Technology*, Vol. 58 No. 10, pp. 1547-50.
- van Raan, A.F.J. (2005), "Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups", *Scientometrics*, Vol. 69 No. 1, pp. 117-20.
- Vine, R. (2006), "Google Scholar", *Journal of Medical Library Association*, Vol. 94 No. 1, pp. 97-9, available at: www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1324783
- Vinkler, P. (2007), "Eminence of scientists in the light of the h-index and other scientometric indicators", *Journal of Information Science*, Vol. 33 No. 4, pp. 481-91.
- Walters, W.H. (2007), "Google Scholar coverage of a multidisciplinary field", *Information Processing & Management*, Vol. 43 No. 4, pp. 1121-32.
- White, B. (2006), "Examining the claims of Google Scholar as a serious information source", *New Zealand Library & Information Management Journal*, Vol. 50 No. 1, pp. 11-24, available at: <http://eprints.rclis.org/7657>
- Wleklinski, J.M. (2005), "Studying Google Scholar: wall to wall coverage?", *Online*, Vol. 29 No. 3, pp. 22-6.
- Yang, K. and Meho, L.I. (2006), "Citation analysis: a comparison of Google Scholar, Scopus, and Web of Science", *Proceedings 69th Annual Meeting of the American Society for Information Science and Technology (ASIST)*, Vol. 43, available at: http://eprints.rclis.org/archive/00008121/01/Yang_citation.pdf

Further reading

- Jacsó, P. (1997), "Content evaluation of databases", *Annual Review of Information Science and Technology*, Vol. 32, pp. 231-67, available at: www.jacso.info/PDFs/jacso-content-arist.pdf
- Jacsó, P. (2005c), "Comparison and analysis of the citedness scores in Web of Science and Google Scholar", *Lecture Notes in Computer Science*, Vol. 3815, pp. 360-9, available at: www.jacso.info/PDFs/jacso-comparison-analysis-of-citedness.pdf

Corresponding author

Péter Jacsó can be contacted at: jacso@hawaii.edu