SAVVY SEARCHING

The plausibility of computing the h-index of scholarly productivity and impact using reference-enhanced databases

Péter Jacsó
University of Hawaii, Hawaii, USA

Abstract

Purpose – This paper aims to provide a general overview, to be followed by a series of papers focusing on the analysis of pros and cons of the three largest, cited-reference-enhanced, multidisciplinary databases (Google Scholar, Scopus, and Web of Science) for determining the h-index.

Design/methodology/approach – The paper focuses on the analysis of pros and cons of the three largest, cited-reference-enhanced, multidisciplinary databases (Google Scholar, Scopus and Web of Science).

Findings – The h-index, developed by Jorge E. Hirsch to quantify the scientific output of researchers, has immediately received well-deserved attention in academia. The theoretical part of his idea was widely embraced, and even enhanced, by several researchers. Many of them also recommended derivative metrics based on Hirsch's idea to compensate for potential distortion factors, such as high self-citation rates. The practical aspects of determining the h-index also need scrutiny, because some content and software characteristics of reference-enhanced databases can strongly influence the h-index values.

Originality/value – The paper focuses on the analysis of pros and cons of the three largest, cited-reference-enhanced, multidisciplinary databases.

Keywords Databases, Indexing

Paper type General review

The h-index, developed by Jorge E. Hirsch to quantify the scientific output of researchers, has immediately received well-deserved attention in academia. The theoretical part of his idea was widely embraced, and even enhanced, by several researchers. Many of them also recommended derivative metrics based on Hirsch's idea to compensate for potential distortion factors, such as high self-citation rates. The practical aspects of determining the h-index also need scrutiny, because some content and software characteristics of reference-enhanced databases can strongly influence the h-index values. This paper provides a general overview of these issues, to be followed by a series of papers focusing on the analysis of pros and cons of the three largest, cited-reference-enhanced, multidisciplinary databases (Google Scholar, Scopus and Web of Science) for determining the h-index.

Bibliometric, informetric and scientometric measures are being increasingly employed as a means of evaluating scientists, researchers, universities, research institutions and their departments at the national and international levels. These measures are employed in making decisions for awarding tenure, grants and allocation

Online Information Review
Vol. 32 No. 2, 2008
pp. 266-283

1468-4527
DOI 10.1108/14684520810879872

of research funds. Such metrics form an integral part of the research assessment and evaluation in the UK, Australia, New Zealand and several other countries. These have been almost exclusively based on the citation indexes of the Institute for Scientific Information (now Thomson Scientific).

The recently introduced new bibliometric measure, the h-index (Hirsch, 2005), was developed to characterise, by means of a single number, both the productivity (number of publications) and the impact or influence (citedness of publications) of scholars. It immediately generated immense interest among researchers and academic administrators. This interest was well deserved because of the convincing and transparent theoretical foundation and the practical simplicity of creating h-index lists from simple database searches for individuals, and for research groups, institutions and journals.

Beyond the simplicity and convenience, however, consideration must be given also to the prevailing software and content limitations of the databases, and to the peculiarities of the h-index-generating utility programs that are used in the process. The lack of necessary knowledge about the content and software, and the search skills of the persons who determine the h-index, may aggravate the situation.

The h-index is generated from a set of results produced through queries, performed by searchers with very different search skills. They use a variety of databases that can report the citedness of the works of an author represented by master records in the result lists. This process is different from the calculation of the controversial but widely used Journal Impact Factor, which is also a single-number bibliometric measure based on citation analysis, prepared centrally from a partially normalised dataset by highly qualified specialists and reported annually in the journal citation reports of Thomson Scientific.

As long as the software can report at least the citation counts for each item, and preferably can sort the result by citation count, lists of h-indexes may be created using several different databases and database editions (Web of Science and some EBSCO databases), choosing different options for generating the h-index (Scopus), or different third-party programs (Google Scholar).

While there are several databases which may be ideal not only for discovering and identifying relevant documents for a research topic or by an author, journal or institution, and also searching the full-text and delivering the source documents, they may be entirely inappropriate for computing the h-index. This is obvious when a database does not offer citation counts for its bibliographic master records, as is the case with the excellent psychology and behavioural sciences databases of EBSCO, which has full text records for 85 per cent of its nearly 700,000-item database, but does not provide information about the citedness of items or offer citation searching.

In other cases, however, when the result list does include such information and the software offers citation-searching options, it may not be obvious that the database is still inappropriate for h-index purposes, as is the case with many of the otherwise worthy databases of EBSCO, where bibliographic records are enhanced with cited references very selectively, and only for the past few years.

This introductory paper explains the essential practical requirements for creating h-index lists, and the considerations in interpreting the results, as they may be highly distorted due to shortcomings in the breadth (retrospectivity, depth), and consistency of coverage in the databases, and potential deficiencies in the citation-matching algorithm and in the handling of the quasi-matches and non-matches by the software.

A series of subsequent papers will present the specific pros and cons of the three largest, multidisciplinary databases enhanced by or based on cited references: Google Scholar, Scopus and Web of Science for determining and interpreting the h-index, and the reasons for considerable differences for the same researchers. Judit Bar-Ilan (2007) sums up the most succinctly the problem in the title ("Which h-index?") of her recent paper about the h-indexes produced by Web of Science, Scopus and Google Scholar for a group of prominent Israeli scientists.

The concept and caveats of the h-index

Hirsch (2005) summarised the h-index as "the number of papers with citation number $\geq$ h, as a useful index to characterize the scientific output of a researcher". For example, J.E. Hirsch has an h-index of 50, if he has at least 50 papers that were cited at least 50 times.

Hirsch demonstrates his new measure showing the h-index of 21 most prominent physicists, from E. Witten ($h = 112$) to S.W. Hawking ($h = 62$). (The latter ties with M.S. Dresselhaus, who was reported in Hirsch's paper by a significantly lower $h = 62$ index than she would have deserved according to my calculations. As of early February 2008, Hawking's h-index is still 62, while Dresselhaus has $h = 76$, and it is not caused by a sharp surge in the citations received by her works in the past two years).

Hirsch had a novel idea and has presented it simply and convincingly in *PNAS* (*Proceedings of the National Academy of Sciences of the USA*), a most influential academic journal (quite tellingly, this short paper, which probably has the most succinct abstract, and must be one of the few papers whose author dares to use first person singular in *PNAS*, has been cited over 100 times within two years).

There are a number of papers that suggest refinements, and enhancements of the original idea (Egghe, 2006; Jin *et al.*, 2007; Vanclay, 2006), or extend its application beyond individuals (Banks, 2006; Braun *et al.*, 2006; Prathap, 2006), and only a very few that have negative reactions to it (Ashkanasy, 2007; Purvis, 2006). There are numerous papers about applying and testing the idea of Hirsch in various disciplines (Bar-Ilan, 2007; Cronin and Meho, 2006; Harzing and van der Wal, 2008; Meho and Yang, 2007; Oppenheim, 2007; Schreiber, 2007a; Saad, 2006).

Hirsch mentions some of the caveats that must be applied in interpreting the h-index, and he duly notes that it should be used together with other assessment criteria. Among his caveats he warns that "there will be differences in typical h values in different fields, determined in part by the average number of references in a paper in the field, the average number of papers produced by each scientists in the field, and the size (the number of scientists) of the field". Hirsch aptly demonstrates this by showing the h-index of ten of the most prominent researchers in the biosciences, from S.H. Snyder ($h = 191$) to A. Ullrich ($h = 120$). The lowest h-index in the biosciences is higher than the highest one in physics, and almost twice as high as the lowest one in physics. The median h among the top 10 physicists and bioscientists is 75 and 147 respectively. The difference would be larger if the calculation of the h-index for the group of bioscientists had not been restricted to 1983-2005, but calculated from 1955, as it was done for the physicists.

Hirsch elaborates on the issue, warning that "scientists working in non-mainstream areas will not achieve the same very high h values as the top echelon of those working in highly topical areas". In my experience there are significant differences even within a discipline or sub-discipline, such as cataloguing digital resources versus cataloguing Eastern European incunabula. Finally, Hirsch raises the issue of self-citation, which was the topic of several of the articles triggered by the introduction of the h-index (Schreiber, 2007b; Vinkler, 2007).

Hirsch admits that self-citations can increase an individual's h-index, but he argues that self-citations would have far less impact on the h-index than on the measure of total number of citations received. He concludes by saying that "I have proposed an easily computable index, h, which gives an estimate of the importance, significance, and broad impact of a scientist's cumulative research contributions".

There is also a warning – although only in a tiny eight-point footnote – that "of course, the database used must be complete enough to cover the full period spanned by the individual's publications". Completeness is just one of my main concerns when putting the theory into practice, and not only for the reason that completeness of the database is a peripheral issue for most users, but also because it is not easy to realise the other content and software limitations of the databases used to determine the h-index. These non-obvious limitations may grossly distort the results.

It is natural to lead in this discussion with the question, what is the h-index of J.E. Hirsch? The answer depends on which databases one uses, and how. Table I shows the huge differences in the h-indexes computed from eight databases, ranging from the impressively high h-index in WoS for a researcher with 30 years of publishing activity to the absurdly low h-index in the very large multidisciplinary databases of EBSCO.

In Web of Science (WoS) Hirsh's h-index is = 51, based on 9,288 references to 179 publications by "our" J.E. Hirsch in periodicals covered by WoS. It is to be noted that six records for papers by J.E. Hirsch published before 1980 are not included in the 1980-2008 edition of WoS used for this test, but their exclusion does not change the h-index, even if it had been calculated from the WoS editions starting their coverage from 1945 or 1955. Hirsch started publishing in the mid-1970s, and none of his first six papers have been cited more than 51 times. The term "edition of WoS" is just my practical shorthand, not an official term, and it depends on the libraries' decision how many years of coverage the library chose when it licensed the WoS database (Figure 1).

In Scopus his h-index is 33, based on 3,938 references to 168 publications by "our" J.E. Hirsch in sources covered by Scopus. In WoS, Scopus and Google Scholar the searches were limited to physics and other disciplinary areas for filtering the set for Hirsch's source papers produced by the name and initials combination. This eliminates records for papers by authors in other disciplines with the same name and initials

Database	h-index	Master records	Total citations	Avg citations
WoS (Web of Science)	51	179	9,288	51.888
ADS (Astrophysics Data System)	41	181	6,633	36.646
Google Scholar	38	219	6,191	28.269
PROLA (Physical Review Online Archive)	36	127	4,912	38.677
Scopus	33	168	3,938	23.440
arXiv	14	49	719	14.673
Ebsco MegaFILE	1	10	4	0.400
Academic Search Premier	0	16	0	0

Table I. The h-indexes for physicist Jorge E. Hirsch from eight databases

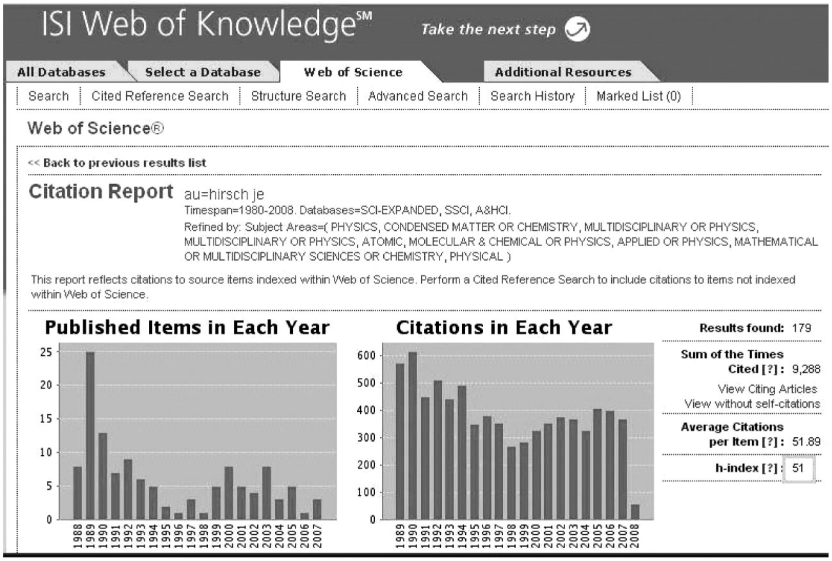

Figure 1.
The Citation Report feature of WoS shows – among other informative bibliometric measures – the h-index

combination, the ones who are not "our" J.E. Hirsch. In the astronomy and physics databases this is implied. In cases where two researchers have the same family name and middle initial(s), and work in the same subject field, this problem must be handled in a tedious way. That is why in many h-index lists 10-15 per cent of the originally considered potential researchers' h-indexes are eliminated or not calculated (Figure 2).

In Google Scholar his h-index is 38 (based on 6,191 references to 219 records for papers attributed to J.E. Hirsch by Google Scholar). It is to be noted that in Google Scholar the master records and the orphan reference records (to be discussed later) are merged in the result lists. The hit counts and the citation counts of Google Scholar keep changing dramatically. If they were increasing, it could be chalked up to adding new records, but often these counts decrease because of deleting records from the database (see the example of disappearing records after I demonstrated an extreme mismatch at www2.hawaii.edu/ ~ jacso/extra/

In the Astrophysics Data System (ADS) the h-index of J.E. Hirsch is 41, based on 6,642 references to 181 master records for papers of Hirsch available in this database. The Physics Review Online Archive (PROLA) of the American Physical Society, which recently introduced the feature of sorting search results by papers' citedness, yields an h-index of 36 based on 4,912 references to 127 records for papers by Hirsch in APS journals. There are two reasons for this high but realistic h-index from the a publisher's digital archive. One is that Hirsch wrote more than 90 per cent of his papers for one of the APS publications. The other reason is that that the number of citing references in PROLA is not limited to sources published by APS but includes citations from articles in journals of other publishers who participate in the CrossRef project, created to facilitate inter-publisher linking to cited and citing references. This is an important development, because (as of early January 2008) more than 2,500 publishers participate in CrossRef, with nearly 20,000 journals and 31 million articles. Some other publishers, such as Springer, also started to use CrossRef to report the citedness of papers both

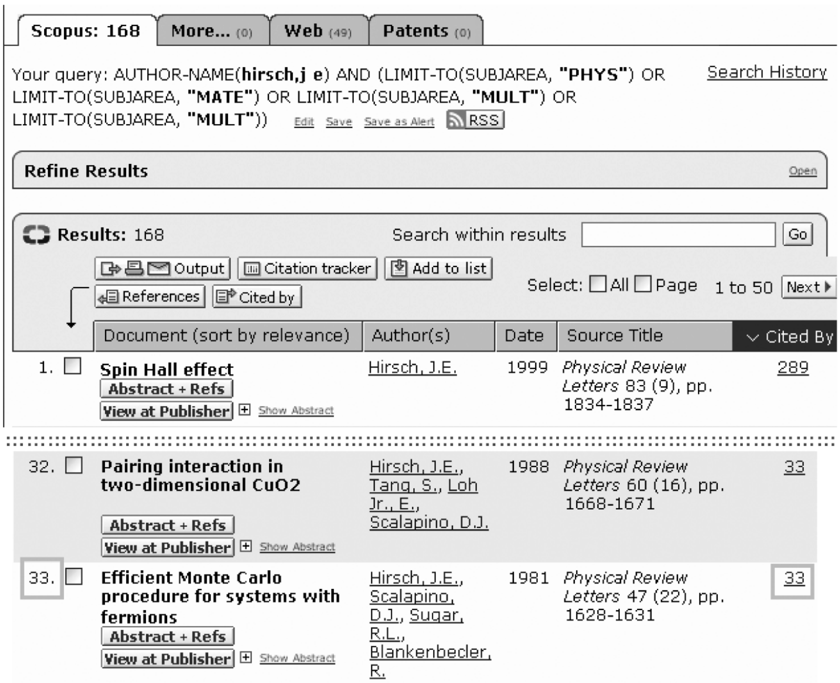

Figure 2.
Scrolling-down in the result list sorted by citedness – is one of several options for determining the h-index in Scopus

from their own journals and also from journals of other publishers that are CrossRef members. Unfortunately, CrossRef is not searchable directly by author name for end-users, nor does it report the citedness of the papers. The same is true for the new interface of the digital archive of the Institute of Physics, which seems to be a step backward. The American Institute of Physics' Scitation system covers several publishers' journals, and it offers searching by author name (the taken-for-granted option) but does not report the citedness of articles. From the perspective of the h-index this is a limitation of several other online information services, such as Scitopia, and WorldWide Science, which would have a broader base of coverage by combining records for papers published in journals of several large publishers. Currently, none of these services and databases can be used for calculating the h-index (Figures 3 and 4).

At the other end of the spectrum are the two largest EBSCO databases. They report citation counts in the master records but do not offer sort options. It would be possible to sort the results manually, but it is not worth the effort because so few records show citation counts. In the EBSCO MegaFILE database the h-index is 1, based on four references to the master records for 10 papers authored by J. E. Hirsch. In EBSCO Academic Search Premier the h-index is 0 because none of the 16 master records for papers of J.E. Hirsch show citation counts, as if none of them would have been cited by articles covered in these databases. This is not so, because coming through the back door, and using the citation searching option, reveals such articles, but the citation matching algorithm does not see the match even when the references are accurate and would match master records (Figure 5).

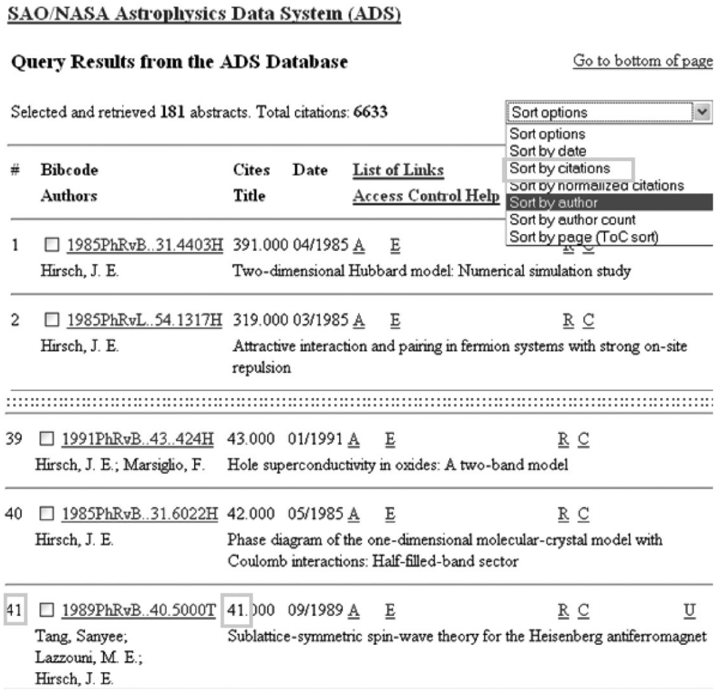

Figure 3.
The scroll-down spotting of the h-index in the result list sorted by citedness in ADS

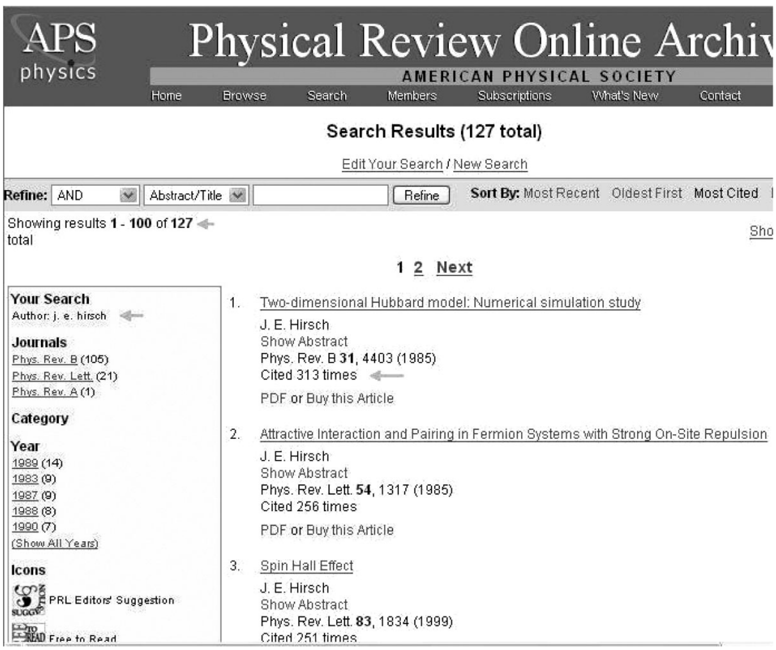

Figure 4.
PROLA allows scroll-down spotting of the h-index in the result list that can be sorted by several criteria, including citation counts

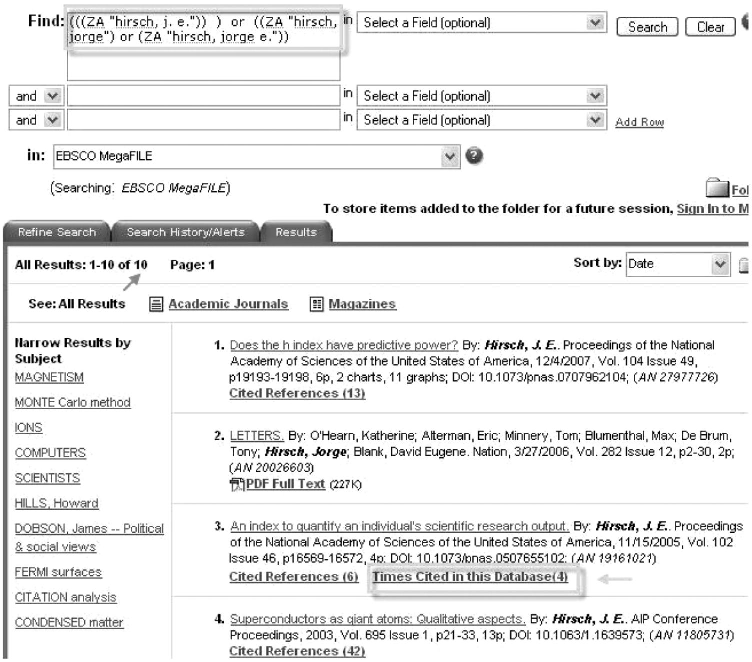

Figure 5.
The h-index in EBSCO MegaFILE based on a single article cited four times in the database out of the ten records for papers authored/co-authored by J.E. Hirsch

These variable h-index values, and especially the absurdly low h-index values for J.E. Hirsch in the two EBSCO databases, offer the perfect example to illustrate the reasons for the significant differences, and the limitations -from the h-index perspective- of otherwise good quality, very large databases with cited reference enhanced records.

Of course, databases are not licensed with the purpose of determining the h-index of researchers. EBSCO keeps a low profile about the enhancement of many of its databases with cited references and emphasises the enhancement of indexing/abstracting records with the full text of the source documents; this may be much more important for the majority of end-users. Having said that, it is quite likely that h-indexes will be generated from several of the EBSCO databases, which is why they are prominently discussed in this paper. Google Scholar, Scopus and Web of Science will have their own in-depth evaluation from the perspective of the h-index in the forthcoming issues, so they are discussed here only briefly and to provide context.

The completeness of a database referred to by Hirsch is a far more complex issue than the footnote in his seminal paper may suggest, i.e. that "the database used must be complete enough to cover the full period spanned by the individual's publications". The prerequisites are much broader and include scope, size, retrospectivity, composition, breadth and consistency of source coverage. These are relevant even for databases without cited references for comprehensive searches (Jacso, 1997), but for h-index calculation these requirements extend to the subset of the cited reference enhanced subset of the database. Not even the most accomplished scientists (except for information professionals specialising in bibliometric studies) may be fully aware of these database traits, especially when it comes to the subset enhanced with cited references. The dimensions of that subset are usually the most enigmatic features of the databases.

Size and scope

The size of the database is usually readily available to searchers, but the size of the reference-enhanced subset rarely is. For example, there is a variety of editions of the CINAHL database, and the catalogue of EBSCO provides useful information about some of the content features, but not with regard to the cited reference enhanced subset of the different editions. All of them have such subsets, usually between 18-20 per cent of the entire database – except for CINAHL Select where 90 per cent of the articles are available in full text, but merely 3.2 per cent of the records are enhanced by cited references. This information is not readily available for the various CINAHL editions, and not even librarians may know how to determine the number of records enhanced by cited references as a means of understanding the comprehensiveness of the cited reference enhanced subset.

For many libraries availability and searchability of full-text may be more important than the availability of records enhanced by cited references, and rightly so, even though the cited and citing references can significantly enrich the search experience and improve the results. After all, access to the full text is the meat for most users, and, if one does not eat it with the most appropriate silverware, it is less of a concern than the other way around, being equipped with Martha Stewart-endorsed utensils, but no meat to eat.

It is quite obvious from the results for the h-index test that no matter how large is the EBSCO MegaFILE (33 million master records) or the Academic Search Premier database (14 million master records), they are not yet ready for calculating the h-index.

The size of the MegaFile is practically the same as in Scopus (33.3 million records), and 10 per cent larger then the 1980-2008 edition of Web of Science (30 million records). It is much smaller than the 1945-2008 edition of WOS, which has 39 million master records, but that is not the reason for its inadequacy for h-index calculation. The reason for that is that the size of the cited reference enhanced subset of about 735,000 records represents barely more the 2 per cent of the database. In the smaller Academic Search Premier database the cited reference subset of 577,500 records represents 4 per cent of the entire database. It is to be mentioned again, that in case of Web of Science, the library defines the time span (and thus the edition) to be licensed, in other cases the editions are predefined by the content providers, or it comes in just a single edition.

From the perspective of citation-based searching and the computing of the h-index the first main issue is the size of the cited reference enhanced subset of the databases not just the size of the entire database.

By comparison, in Scopus nearly 12.5 million records (37 per cent of the full database of 33.3 million records) are enhanced by cited references. In the 1980-2008 edition of WoS this ratio is about 80 per cent, yielding more than 23.6 million records enhanced with cited references. The 1945-2008 edition of WoS has 31 million cited reference enhanced records (80 per cent) out of its 39 million records. In the edition of WoS covering 1996-2008, the number of records enhanced by cited references is about 12 million out of the nearly 15 million records, somewhat smaller than that subset in Scopus. Obviously, EBSCO is not in the same league as Scopus or WoS, when it comes to cited reference enhanced subset, and that's why EBSCO does not advertise this feature widely.

One of the other reasons for the poor performance of MegaFile and Academic Search Premier is the poor coverage or non-coverage of many of the top-ranking physics journals. It is a "double whammy" for Hirsch, who published most of his papers in journals not covered by these two EBSCO databases, and his papers received the majority of their citations from papers published in these very journals not covered by these two databases of EBSCO.

The scope of the database is another factor to consider. When h-indexes are to be determined for researchers in a specific discipline, smaller databases with a focus on that discipline might serve the purpose very well, even though it may not be obvious from the name of the database. The open access Astrophysics Data System (ADS) (http://adswww.harvard.edu) yielded good results for my h-index tests for some of the physicists listed as most prominent in Hirsch's paper, even when their work had nothing to do with astronomy. The reason for this is that the database is as much for physics as for astronomy. Actually, ADS covers all physics areas and not just astrophysics, and Hirsch is mostly cited by articles in physics, except for his paper about the h-index which is much more widely cited in other science and social science journals than in physics journals (ADS has more than four million records for physics papers, more than 1.5 million for astronomy, and 0.5 million are harvested from the arXiv repository).

If the scope of the cited reference enhanced database is very narrow, but in that area its coverage is comprehensive and consistent, it can be used for determining the h-index for the researchers publishing in that discipline or sub-discipline. A good example for this is the legendary SPIRES-HEP bibliographic database at Stanford, which was the first publicly accessible web server in the US set up to offer open access to a bibliographic database. Its design clearly shows the fingerprints of a librarian, Louise Addis, who was the highly interested and capable librarian at the Stanford Linear Accelerator Center (SLAC) when the SPIRES project was launched.

However, for the test of h-index of some of the most prominent physicists on Hirsch's lists it could provide a realistic h-index only for Stephen Hawking. Actually, SPIRES-HEP is the database where Hawking receives his highest h-index ($h = 6$) – even above what the 1945-2008 WoS edition produces. The physicists on Hirsch list who work(ed) and published in other specialty areas of physics had a very low h-index in SPIRES (Hirsch had $h = 2$, and some Nobel laureates in physics had the same or $h = 1$), because the scope of the database is limited to high energy and particle physics, and it does a very good job at that.

Retrospectivity

In discussing one of his tests for determining the h-index for winners of the Nobel Prize in physics, Hirsch mentions that he used (for a calculation) the "first published paper year or 1955, the first year in the ISI database". Actually it is the first year of coverage in the ISI database that his university licenses, but another university may have a larger edition of WoS, where the first year of coverage is 1945 or 1901.

Even the decade older edition of WoS can make a considerable difference in determining the h-index for mature, well-established researchers. The database edition used by Hirsch would not be fair for calculating the h-index for those Nobel laureates in physics who received their award in the early decades of the of the twentieth-century. No database can be fair to them, because the papers of those Nobel laureates, and the thousands that cited them, were written around the turn of the century, well before the start of coverage of most databases. These are absent from the calculation of the h-index, except for the Century of Science edition of WoS, but even that edition covers only a small portion of their papers. The Chemical Abstracts, and PsycINFO databases also go back to the beginning of the twentieth-century, but those records are not enhanced by cited references.

Consider Wilhelm Conrad Röntgen, the first Nobel Prize winner in physics in 1901. His most important papers (including the landmark paper, "On a new kind of rays in science") were written in the 1890s and received the majority of their citations before or around the turn of the century. Röntgen, Max Planck, Fermi and several other famous physicists would have an unrealistically low h-index from any databases (even from the Century of Science where their h-indexes are 2, 12 and 18, and Einstein achieves just 43). The shorter the time span of the retrospective enhancement of records by cited references, the lower the h-index – even for some of the most prominent researchers.

This limitation in retrospective coverage is not navel gazing pondering about scientists who lived more than a century ago. In the mini case study by Hirsch of the Nobel laureates in physics for the past two decades the symptom of low h-index is there: the range in the 1955-2005 WoS edition is from 22 to 79, the average is 41, and the median is 35 – not an index value in physics that would indicate their Nobel-worthy research. It is not their fault but is due to the limitations of the databases used.

To see the impact of just a ten-year difference in retrospectivity on the h-index, I used the 1945-2008 and the 1955-2008 edition of WoS (Hirsch used the latter), and chose two physicists from the middle point of the two decades, Bertram N. Brockhouse and Clifford G. Shull, who shared the 1994 Nobel prize.

For Brockhouse the difference was just one point ($h = 33$ in the former, and $h = 32$ in the latter) between the two editions of WoS. However, in Shull's case the h-index from the 1945-2008 edition was 35, and from the 1955-2008 was only 26. The reason for this is that the latter did not include records for the source items and many of the citing items from 1946-1954, when Shull worked at what is known now as the Oak Ridge National Laboratory.

Further analysis shows that he wrote more than 40 per cent of all of his papers during his years at Oak Ridge. These earned more than 50 per cent of the total citations his works received (or more precisely that could be attributed to him by the software). Using the WoS edition with a ten-year-shorter time span does not include in the h-index generation process these cited and citing papers, and that is the reason for the significantly lower h-index in Hirsch's list for Shull.

Then again, the h-index computed from the 1955-2005 edition of WoS is still considerably higher than the h-index computed from other databases that can be used for h-index calculation in physics. Scopus, Google Scholar, ADS, and PROLA the h-index for Shull were 12, 19, 18, and 19. Brockhouse's h-index in Scopus (9) clearly shows the influence of its much shorter time span of records enhanced by cited references *vis-à-vis* Google Scholar (21), ADS (21) and PROLA (19). Depending on the target group or the individual whose "importance, significance, and broad impact" is to be measured, the span of the cited reference enhanced subset of the database is very important for fairness.

I discussed the dimensions of the cited reference enhanced subset of databases in a recent column (Jacso, 2007), but my focus was different there: the huge advantage of easily discovering items which cite or were cited by a few highly relevant papers on the user's topic, irrespective of the terminological and morphological differences of the subject words. No library would license a cited reference enhanced database just for the sake of allowing a few users to determine the h-index of some researchers. However, it would be increasingly important to inform customers of this aspect of the database.

Thomson Scientific does not need to do this, because whatever edition a library licenses, all the records include the cited references that appeared in the source documents. On the other hand, in most other databases this type of enhancement of records with cited references goes back only a few years.

The producers of the PsycINFO database deserve credit for providing details of this aspect clearly, emphasising that cited references have been added comprehensively only from 2001, and in the years before the enhancement is very selective. About 690,000 of the 2.4 million records are enhanced in PsycINFO, and 86 per cent of them are records for papers published after 2000. It may be useful for relatively young researchers in psychology in determining their h-index, but for established researchers the h-index produced from PsycINFO would not be fair. Scopus also goes out of its way to alert users that cited referenced were added to records only from 1996 (actually there are some thousands of records from earlier years that were also enhanced). In most of the cited reference enhanced databases this information is not mentioned. For example, the EBSCO databases give no hint about the retrospectivity of the cited reference subset, but their small proportion (less than 5 per cent) in the various editions (Premier, Complete, Alumni) of the Academic Search database might give a clue to skilled users that such records are available only for the past few years, making their databases inappropriate for even gauging the "cumulative contributions" of scientists, as Hirsch meant to do.

The public relation message of chemical abstracts seems to provide an easy-to-misunderstand statement, when its content description mentions that it has "links to cited references dating back to the beginning of the twentieth-century". Technically and legally it is true, but what the user really needs to know is that cited references have been added only to records entered from 1997. Given its yearly addition of more than a million records, chemical abstracts undoubtedly offers the largest collection of reference enhanced records in chemistry from the past 11 years, but for determining the h-index of chemists' whose contributions in the past 20-30 years have had the most influence, it is not the preferred resource.

Source base

This criterion refers to the types of source documents covered by the database and enhanced by cited references. The PsycINFO database stands out from the crowd of traditional indexing/abstracting databases by covering a wide variety of primary sources beyond journals, such as monographs, collective works, book chapters, reference books, textbooks, conference proceedings and dissertations. About 25 per cent of the materials belong to one of these categories. From the reference enhanced subset perspective, the coverage is more limited because records for conference papers and dissertations have not been enhanced by cited references, while 63,000 books and book chapters are already enhanced by cited references. This is true primarily for recent (twenty-first century) books and book chapters. Records for classic psychology books have not yet been enhanced in spite of their importance in this discipline.

In physics and in most of the hard sciences books play a much smaller role in scientific communication than journal articles and conference papers. Still, books should not be ignored in the calculation of the h-index for those scientists who have put considerable energy into writing books, which subsequently became highly cited.

The most important move by Google Scholar since its launch was the autonomous indexing of cited references in books (harvested mostly from the Google Books service) and conference proceedings. These significantly improved the quality of the source base, and in some cases the h-index of researchers, as I discuss below in the treatment of stray and orphaned references.

Journal base

For decades journals (and some other serial publications such as annual reviews) have been the almost exclusive source documents processed by Thomson ISI for citation-based searching, and for calculating journal impact factors and other bibliometric measures. There is a well-defined set of scholarly journals in every discipline that can serve as a benchmark in judging the scope of journals and the breadth of their coverage by cited reference enhanced databases needed to compute a reasonably fair h-index. Their coverage must be also consistently comprehensive. This is usually not a problem in publishers' digital archives, but there are exceptions. In our professions most ALA journals have limited digital availability, and the same is true for the library science journals published by the Haworth Press. In information science Wiley does a very poor job in digitising back issues of the *Annual Review of Information Science & Technology* (*ARIST*), as well as of the *Bulletin* and the *Proceedings of the American Society for Information Science* (the earlier name of the society).

The gaps in digitisation of ASIST publications is just one of the reasons why this publisher's digital archive is totally inappropriate for the h-index calculation of information scientists. Of the 42 volumes of *ARIST*, only five (2002 to 2006) are available digitally. For the *Bulletin* of ASIST, Wiley's Interscience database does not have in digital format the issues of its first 21 volumes, and the *Proceedings* of the Society are included only for 2002-2006, missing 38 volumes. It needs further investigation to determine how incomplete the digital versions of ASIST journals are in other disciplines. This lack of digital availability from the publisher is different from the situation when the library decides to license only a subset of the digital archives of journals.

Consistency of coverage

The coverage of journals in cited reference enhanced databases can be surprisingly uneven. Missing issues and volumes in indexing and abstracting records present a double problem, because in addition to not counting the articles for the productivity measure and their references for the citedness measure, the references of the papers citing articles in the missing issues do not have the master records as pegs on which to hang the citations.

This is a problem with several journals in EBSCO Academic Search Premier, which has only a few years of coverage for such journals as *Library and Information Science*

Research, *Program*, *Aslib Proceedings*, *Journal of Internet Cataloging, Science & Technology Libraries*, *Online Information Review*, and has missing issues for many of them, such as *Knowledge Quest*, *Music Reference Service Quarterly*, *International Journal of Information Security*. It is to be noted that acquiring information about the consistency of the completeness of journal coverage is difficult and time consuming even for a skilled information professional, and omissions can greatly influence the h-index of the authors of omitted papers, and of the journals themselves.

Software issues

Currently there are only a few databases (including the publishers' digital archives) that have the essential tools for determining the h-index, let alone doing so efficiently. Disappointingly, the Dialog Information System, which was among the few to enhance its system for the handling of cited references in millions of records of ISI citation indexes, has fallen behind the pack. A few years ago, when PsycINFO records were partially enhanced by cited references, there were no efforts by Dialog to implement the options for searching by cited author, cited journal and cited year. They are searchable as free text, such as the abstracts, but not as distinct elements as cited author, cited title word, cited year or cited journal. This is akin to eating soup with a fork.

No wonder there is no automatically generated information about the citedness of papers, let alone a sort option by this criterion, or the ability to produce h-indexes in this implementation of PsycINFO. Neither was the software upgraded for the ISI citation databases to allow automatic reporting of papers' citedness counts, let alone a facility to permit sorting the result lists by citedness count, or calculating the h-index.

The implementation of PsycINFO on Ovid and OCLC is far from ideal, and inadequate for computing the h-index. EBSCO provides the citedness count on PsycINFO (and in many of its own databases) but does not offer the facility to sort the result set by that information. CSA comes closest to facilitating the calculation of the h-index by allowing the downloading of records directly to the RefWorks software, including the times-cited count. The set can be converted into a spreadsheet with the press of a button, and sorted by the citedness count in a swift process (EBSCO also allows downloading records to RefWorks, but it does not send the citedness count data along with the bibliographic data).

Beyond Web of Science and Scopus, only SPIRES, ADS and PROLA offer adequate software tools for calculating the h-index. The valuable arXiv preprint collection has many implementations, and it is one of the three components of ADS, which is the only host to provide citation counts and to allow sorting of the sets by citedness.

Google Scholar no longer ranks the result list by citedness count systematically, which previously was the one and only sort possibility; neither does it offer the downloading of records. At least there is a number of third party utilities which can scrape the screen and even calculate the h-index. It is another issue that these are at the mercy of Google Inc., which can and does prevent access by these utilities to Google Scholar.

Citation matching

The most critical question is the correct citation matching by the software. The variety of reference styles required by journals makes this process very error-prone. Instead of trying to describe the difficulties, I rather present an example – on the topic of citation matching from CiteSeer, one of the smartest autonomous citation matching systems. What is shown here is just the tip of the iceberg. Papers that are cited hundreds and thousands of times show much more variety then these 14 references that cite the same book in many different ways even in this simple case.

There are more problematic documents from a referencing aspect, such as government documents, conference papers and large author groups. In case of journal articles the various abbreviations and punctuations in journal titles, and in volume, issue and page numbers, raise many additional problems. Author names with accented characters are a special challenge. Röntgen may appear also as Rontgen, Roentgen or Rntgen when the software swallows the accented character. German, Dutch, Italian and Spanish names with prefixes represent another challenge. Add to this double middle initials with or without punctuation and you have a picture of the name difficulties even in simple Latin scripts. Author names with many adjacent consonants are more often misspelled than most of the Japanese names with the consonant-vowel pattern, such as Tanaka or Morimoto. Japanese, Chinese and Koran names, however, are very problematic when they are transliterated into the same Latin character-string. The American Physical Society just announced an experiment to include the names of Chinese, Japanese and Korean authors also in their own script to help distinguish their identity (Sprouse, 2008) (Figure 6).

Only SPIRES and Web of Science offer some help by allowing direct browsing of author names, and the former can also search automatically for name variants. Scopus has an indirect way to browse author and journal name variants; this will be discussed in its forthcoming review. The Institute of Physics does not even offer a search field for authors in its new interface. The new advanced search template of Windows Live Academic, shows a separate cell for four first and middle initials but allows only one character, so there are 1,350 matches for Hirsch J. Hirsch J.E. was not among the first

CiteSeer.IST Home **Check:** The following citations are predicted to all refer to the same paper. Details

Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2002.

Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2003.

Hanna Pasula, Bhaskara Marthi, Brian Milch, Stuart Russell, and Ilya Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing Systems 15. MIT Press. 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In NIPS 15. MIT Press, Cambridge, MA, 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity Uncertainty and Citation Matching*. In S. Becker, S. Thrun, and K. Obermayer, editors, Advances in Neural Information Processing Systems 15. MIT Press, 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In NIPS, 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Info. Proc. Systems, 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Proc. Advances in Neural Information Processing. MIT Press, 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Processing Systems 15, Vancouver, British Columbia, 2002. MIT Press.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*, 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2003.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In Advances in Neural Information Processing (NIPS), 2002.

H. Pasula, B. Marthi, B. Milch, S. Russell, and I. Shpitser. *Identity uncertainty and citation matching*. In NIPS, 2002.

Figure 6.
Excerpt from the variety of references to a book with simple author names, no middle initials, prefixes and accented characters

100, not an encouraging start even for a "plain vanilla" search, let alone for determining a researcher's h-index.

Syntactically matching names do not necessarily give a green light for proceeding with the process. The set retrieved may include researchers with the name J.E. Hirsch, who are not the physicist, or are physicists, but not the one who worked at various campuses of the University of California and started to publish in the mid-1970s. Their records must be excluded from the search. Obviously, the searcher must be familiar with the publication activity of the scientists whose h-indexes are to be determined, and must have good search skills to look up variant versions of the names, such as Jorge Hirsch, Jorge E. Hirsch, and do the limiting/filtering operations to exclude from the results set the hits for J.E. Hirsch the audiologist, and J.E. Hirsch the surgeon, and decide if J.E. Hirsch of the National University of Buenos Aires should be retained in the set or not. Some software can help in this process.

Orphaned and stray references

With all these oddities, another crucial issue is what happens with the references that cannot be matched. In some systems you may be able browse or search them, and manually add them to the master records that collect the references. This is an arduous process (Jacso, 2008). No wonder that, excepting the studies by Cronin and Meho (2006), Meho and Yang (2007), Oppenheim (2007), and a few others, most of the h-lists compilations known to me, relied only on the h-index values as reported automatically in the master records. But these do not include the stray references that do not match sufficiently for the software to add them to the master record. These may be in the thousands for some authors and are ignored in calculating their h-index or spotting it by scrolling down in the result list sorted by citedness. In case of Google Scholar the problem is often the opposite. Its software finds a match even when there is not even a reference to the author, let alone to a specific work of the author in the purportedly citing document. These phantom citations can grossly inflate the hit counts in Google Scholar (Jacso, 2006). Even though the h-index is rather robust (Vanclay, 2007), and a few omissions for papers cited less often than the h-index would not cause significant changes, there is an additional important reference type whose handling may result in a considerable lower h-index than the researcher would deserve. I call these the orphan references, even though it is not an inherent feature, because they may be orphaned only in some cited reference enhanced databases, but not in others.

Using the master records with the citation counts listed is always very convenient, but also often very unfair to the researchers whose h-index is to be determined. In addition to the stray cited references which cannot be matched against a master record to hang the citations on, there are the orphan references to books, chapters, conference papers and articles in journals not covered by the database as source documents, or are in issues omitted by the database producer for reason or another. This is not strictly a matching issue because there are no master records for these items, so the references have nothing to be matched against. Inclusion of master records for books is probably the greatest asset of Google Scholar from a citation matching and h-index calculation perspective.

How the software handles the stray and orphaned records is an important issue, because these may significantly change the h-index of individuals who published highly cited books, as did S.W. Hawking in Hirsch's lists. In my tests I found that

Hawking would have a 10-12 per cent higher h-index if the citations received by his books from the articles in journals covered in WoS and Scopus could have been attached by the software to a master record, and thus taken into account in calculating his h-index. But books are not source documents in WoS, and in Scopus they have become very selective source documents only recently. References given to books become orphans, because they cannot be associated with, and accrued through, a master record, and thus are ignored in the automatic calculation of the h-index.

I realised what a difference this makes when preparing my contribution to a festschrift for the 75th birthday of F.W. Lancaster (Jacso, 2008), which was a noble opportunity to engage in the extra work needed to determine a reasonable h-index for such an important personality in the information industry. As that paper will be published later than this one, I do not want to reveal the extent of the increase of his h-index, but I can say that it was very much worth the arduous manual matching.

Future columns in *Online Information Review* will review how the content and software features of Google Scholar, Scopus and Web of Science can meet the criteria for calculating a reasonable, if not perfect, h-index to assess scholarly productivity and impact.

References

Ashkanasy, N.M. (2007), "Playing the citations game", *Journal of Organizational Behavior*, Vol. 28 No. 6, pp. 643-5.

Banks, M.G. (2006), "An extension of the Hirsch index: indexing scientific topics and compounds", *Scientometrics*, Vol. 69 No. 1, pp. 161-8.

Bar-Ilan, J. (2007), "Which h-index? A comparison of WoS, Scopus and Google Scholar", *Scientometrics*, Vol. 74 No. 2, pp. 257-71.

Braun, T., Glänzel, W. and Schubert, A. (2006), "A Hirsch-type index for journals", *Scientometrics*, Vol. 69 No. 1, pp. 169-73.

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.

Egghe, L. (2006), "Theory and practice of the g-index", *Scientometrics*, Vol. 69 No. 1, pp. 131-52.

Harzing, A. and van der Wal, R. (2008), "A Google Scholar h-index for journals: a better metric to measure journal impact in economics and business?", paper presetned at the 2008 Academy of Management Annual Meeting, 8-13 August 2008.

Hirsch, J.E. (2005), "An index to quantify an individual's scientific research output", *Proceedings of the National Academies of Science*, Vol. 102 No. 46, pp. 16569-72.

Jacsó, P. (1997), "Content evaluation of databases", *Annual Review of Information Science and Technology*, Vol. 32, pp. 231-67.

Jacsó, P. (2006), "Deflated, inflated and phantom citation counts", *Online Information Review*, Vol. 30 No. 3, pp. 297-309.

Jacsó, P. (2007), "The dimensions of cited reference enhanced database subsets", *Online Information Review*, Vol. 31 No. 5, pp. 694-705.

Jacsó, P. (2008), "Testing the calculation of a realistic h-index in Google Scholar, Scopus and Web of Science for F.W. Lancaster", *Library Trends*, Summer (in press).

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.

Meho, L.I. and Yang, K. (2007), "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, available at: http://dlist.sir.arizona.edu/1733/

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.

Prathap, G. (2006), "Hirsch-type indices for ranking institutions' scientific research output", *Current Science*, Vol. 91 No. 11, p. 1439.

Purvis, A. (2006), "The h-index: playing the numbers game", *Trends in Ecology and Evolution*, Vol. 21 No. 8, p. 422.

Saad, G. (2006), "Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively", *Scientometrics*, Vol. 69 No. 1, pp. 117-20.

Schreiber, M. (2007a), "A case study of the Hirsch index for 26 non-prominent physicists", *Annalen der Physik (Leipzig)*, Vol. 16 No. 9, pp. 640-52.

Schreiber, M. (2007b), "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

Sprouse, G.D. (2008), "Which Wei Wang?", *Reviews of Modern Physics*, Vol. 80, p. 1.

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.

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.

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