The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1468-4527.htm

SAVVY SEARCHING

The pros and cons of computing the h-index using Google Scholar

Péter Jacsó
University of Hawaii, Honolulu, Hawaii, USA

Abstract

Purpose – A previous paper by the present author described the pros and cons of using the three largest cited reference enhanced multidisciplinary databases and discussed and illustrated in general how the theoretically sound idea of the h-index may become distorted depending on the software and the content of the database(s) used, and the searchers' skill and knowledge of the database features. The aim of this paper is to focus on Google Scholar (GS), from the perspective of calculating the h-index for individuals and journals.

Design/methodology/approach – A desk-based approach to data collection is used and critical commentary is added.

Findings – The paper shows that effective corroboration of the h-index and its two component indicators can be done only on persons and journals with which a researcher is intimately familiar. Corroborative tests must be done in every database for important research.

Originality/value – The paper highlights the very time-consuming process of corroborating data, tracing and counting valid citations and points out GS's unscholarly and irresponsible handling of data.

Keywords Databases, Information retrieval, Search engines, Referencing

Paper type Viewpoint

The introductory part (Jacsó, 2008) to this series of columns about the pros and cons of using the three largest cited reference enhanced multidisciplinary databases discussed and illustrated in general how the theoretically sufficiently sound idea of the h-index (Hirsch, 2005) may become distorted depending on the software and the content of the database(s) used, and the searchers' skill and knowledge of the database features.

In this column, Google Scholar (GS) is under the microscope from the perspective of calculating the h-index for individuals and journals. An enhanced version of this paper with annotated screenshots is posted at www.jacso.info/h-gs, as GS results – for various reasons – are often irreproducible, which is not conducive to genuine scholarly research. The examples for hit counts and citation counts misrepresented by GS and used by third-party utility programs to calculate the h-index are mostly for *Online Information Review*, and for the author of this very paper. This is not merely for myopia and egotism, but for the fact that the very time-consuming process of corroborating the data, tracing purportedly citing papers, counting valid citations and pointing out GS's handling of data can be the most directly demonstrated through this tiny microcosm for readers of *Online Information Review*. It is also important to realise that effective corroboration of the h-index and its two component indicators can be done only on persons and journals with which the researcher is intimately familiar. Corroborative tests must be done in every database for important research whose results may affect people, just as canaries were used to signal dangers in the coal mines.

Online Information Review
Vol. 32 No. 3, 2008
pp. 437-452

1468-4527
DOI 10.1108/14684520810889718

Dead canaries

The deficiencies in the GS software from bibliometric and scientometric perspectives, dwarf the content limitations. The consequences are present in the entire GS universe for the simple reason that most of the problems are caused by the GS software: by the damaged parsing and slapdash citation-matching algorithms. The problems are caused not merely by typos and other inaccuracies in the source data, nor by missing one or two highly cited articles and a dozen lowly cited papers well below the reasonably calculated h-index. Vanclay (2007) convincingly explained and illustrated the stability and robustness of the h-index.

The most serious software deficiencies across the board, even though not visible in every search, do not bother casual searchers who are hunting for a few good papers, but they influence, and may distort, the h-index computed by third-party utilities which inevitably show Garbage In/Garbage Out symptoms. These utilities cannot help but base their calculations on the first 1,000 records (at best) of the often much higher number of questionable hits and citations often reported by GS, especially when computing the h-index for very productive and/or very highly cited periodicals such as *Science*, or *The Lancet*. The ratios of substantial errors may be different in the test microcosm and the GS universe: some have fewer, others have more.

The context of the search

GS's popularity is well-deserved for situations when finding a few good papers (or at least their bibliographic records as pointers) is the primary purpose. The main appeal of GS is that it almost always can lead the users to a few good open access papers, or documents that are not open access but – from the perspective of the end-users – are "freely" available through subscription-based databases in libraries to which the searcher has access. GS also deserves credit for making the information retrieval process smooth and simple (especially if the library has a link resolver) without the need to:

- identify the best candidates from the variety of databases available through the library; then
- learn the particular software used by the databases; and
- run and refine searches in the different systems.

In this context, GS provides instant gratification, and certainly satisfies the overwhelming majority of users, as long as they need only a couple of good papers. As GS is itself free, and can remarkably improve resource discovery and document delivery, it is no wonder that the acceptance of GS by academic librarians has significantly increased since its debut (Neuhaus *et al.*, 2008).

Beyond the instant gratification, the most important virtue of GS is that, in addition to the tens of millions digital journal articles and conference papers available for free searching (even if not for free viewing) – courtesy of the major scholarly publishers, it also covers all kinds of literature that were either print-born or born digital. These include the content of millions of books passed on to GS from the Google Print project with brotherly love, and millions of preprints and reprints courtesy of research and educational institutions, and patents courtesy of the taxpayers.

Apart from journal articles, the other materials are poorly covered (if at all) by most of the subscription-based academic databases. However, GS also has millions of items

which are not in the same league as the materials mentioned above, certainly not from a citation indexing perspective such as assignments posted on the web by students in undergraduate or even graduate courses that must have a bibliography, and entries from blogs and discussion lists. They are there by virtue of being digital, not by virtue of their scholarly value. (I am not a great fan of blogs but there are some good ones, just as there are some master's degree theses which are as good as many papers published in scholarly journals, but they are the exceptions.) This is a relatively minor concern compared with the software problems to be discussed. The content base is certainly there for calculating the h-index, although with some reservations regarding, for example, papers published more than 15 years ago; but content reservations are applicable to all of the alternatives.

The flip side of this much-improved access to digital materials is that papers unavailable digitally remain barely known to GS, as its content is created entirely automatically, just as it is in Yahoo, Ask, Exalead, and GigaBlast. The difference is that Google created GS purportedly to accommodate scholarly literature, while there is no Yahoo Scholar, Ask Scholar or ExaLead Scholar.

However, the situation is entirely different when the purpose of the search is to assist in decisions on such matters as hiring, promotion, tenure, granting of research awards, allocating funds, ranking of research activity, renewal of journals, cancellation of standing orders, etc. In such cases searches are done in order to determine how many articles, books, book chapters, conference papers and other scholarly publications were written by an author (or group of authors), or how many papers were published in a journal, and how often were these cited. Number of papers published indicates the productivity of authors (traditionally an essential criterion in the academic world) and journals, while citations may serve as an indicator of the impact of the authors and journals. These indicators and their ratio have been the major benchmark for teaching and research faculty, and for collection evaluation, for decades.

With the development of the h-index by Hirsch (2007), there is a fairly new yardstick which combines the productivity and citedness indicators in an innovative way to evaluate the past performance, and even predict the future potential of professors, researchers, journals, and institutions in scholarly publishing.

Despite its appeal and simplicity, the h-index must not be accepted as an almighty single indicator for performance. The issue is important because the h-index, as a combined indicator of researchers' publishing productivity and citedness, is used more frequently than it may appear through just the scholarly and professional publications. The h-index is now shown in many resumés, and applications for jobs, grants, sabbaticals, etc. Even in scholarly publications a majority of authors take the "cited by" values as reported by GS at face value, and rush to conclusions in comparing these counts with those of Web of Science (WoS) and Scopus.

Although GS does not present the results in any logical order, verifying the validity of the reported hit counts (for productivity measure) is relatively easy using one of the utilities, or scraping and converting the result list into a spreadsheet, sorting it by title to discover and remove the many duplicates, triplicates and quadruplicates. Verifying the "citation counts" (for the citedness measure), however, is an extremely time-consuming process, but in real scholarly research this is not unusual.

Researchers at least should take random samples to corroborate the "cited by" counts, and pay close attention to the plausibility of "citation counts" to realise the

significant credibility gap between reality and hit counts and citation counts as reported by GS.

From the launch of the service, it has been hopeless to derive any factual information from Google, Inc. regarding the dimension of the content of the database, its size, girth (width, length, and depth combined), or the sources included. It is surprising that, despite this secrecy GS has been so widely embraced by researchers and librarians for scientometrics purposes without reservation. Medical librarian Dean Giustini (2008) at the University of British Columbia dedicated a blog to GS (http://weblogs.elearning.ubc.ca/googlescholar/); Wentz (2004), Manager of Imperial College Library in London, claimed on the MEDLIB-L discussion list that the "cited by" facility of GS is spectacular (later he withdrew that conclusion publicly, and the original document is no longer available). Goble (2006) of Manchester University referred to Google as the "Lord's gift" (she meant GS) in an aside of her otherwise impressive presentation. I presented a very different view of GS at the closing plenary session (Jacsó, 2006) for reasons which are still valid today.

There have been efforts to calculate the h-index and/or gauge the extent of GS's coverage of documents in various disciplines (Neuhaus *et al.*, 2006), or by groups of individuals (Cronin and Meho, 2006; Norris and Oppenheim, 2007; Oppenheim, 2007; Bar-Ilan, 2008, Sanderson, 2008), or by journals (Vanclay, 2008). These require an arduous process, especially in GS.

The note in Lokman and Kiduk's informative comparison of WoS, Scopus and GS on citation counts and ranking of 25 library and information science faculty members is sobering: "WoS data took about 100 hours of collecting and processing time, Scopus consumed 200 hours, and GS a grueling 3,000 hours". I am not surprised.

GS dispenses utterly unreliable indicators through its hot counts and citation counts and makes it inconvenient and discouraging to trace the purportedly citing items even if one has access to most of the digital journals of the discipline. Third party utility programs cannot help in this.

Lokman and Kiduk, however, know the ins-and-outs of responsible citation analysis. They are aware of the serious limitations of GS's document parsing and citation matching algorithms which are not so good in identifying authors and matching citations. That is why Lokman and Kiduk's team spent 3,000 hours verifying and correcting GS's hit counts and citation counts.

None of the reference-enhanced databases are perfect, but Scopus and WoS have a reasonable transparency about their database content, as well as about their record creation and citation-matching processes. They have master records with cited references, and they show the bibliographic and reference details of the citing records. GS does not show the cited references it extracted from the records, and it does not provide a link to records which appear with the (citation) prefix. For records with links one must go to the primary documents (most of them are available only for subscribers), find the link in the cited reference list that purportedly cites the target article.

This is a tedious process when there are hundreds of cited references in the citing documents, especially when they are not in alphabetical order, but in citation order. If the cited references do not have the title of the paper (typical in the citation style of many science journals), the process is gruelling.

GS lumps into a single result list the regular records and the ones prefixed with the label (citation) for which its software could not find a master record. These orphan and stray references can significantly increase hit counts and citation counts. In WoS and Scopus these are stored in separate files and require additional processes. Few searchers know about this feature, but Cronin and Meho (2006) refer to it. Even fewer are willing to combine the hit counts and the citation counts in the separate result list produced from the master records with sufficiently matching citations and the result list of the orphan/stray references because it is a tedious process, especially in WoS. In the next two issues this topic will be further explained to illustrate how much the combination of these result list can improve the h-index of certain types of researchers.

There are others who know the parsing and citation-matching ability of GS, but lack the time to verify its reported citation counts. Bar-Ilan (2008), the mathematician who can handle citation analysis issues equally well from practical and theoretical perspectives, tells it as it is. In her paper on comparing the h-index of 40 of the most highly cited Israeli scientists, she warns that "one has to take into account that the sources and the validity of the citations in GS were not examined in this study. Examining the citing items for GS was beyond the scope of the current study." I cannot blame her and others who accept the citation counts as reported by GS, but this was likely to have been calculated in the development of GS's citation-matching algorithm, and may be one of the reasons for the secrecy about details of the system.

GS very often regales users with worthy content for free, but it very often shortchanges the users with its numbers at every step of the search process by claiming more than it delivers. True, GS does not calculate the h-index, nor does it rank the hit list by citedness, but it offers hit counts and citation counts for the source items that appear in the result list. The problem is that they are often dead wrong because of the inferior parsing and citation-matching software elements.

What is in a name?

Hirsch developed his index for evaluating the scholarly research output of individuals, so it is obvious that name searching is of the highest priority. Still, you never know in GS for certain what is in a name. There is no option to browse in GS, so you just search blind. Neither is there any software feature to distinguish authors with the same name and first initial, as there is in WoS and Scopus. The only chance to distinguish J.E. Hirsch the Physicist from J.E. Hirsch the audiologist is to limit the search to the closest broad subject category. However, this is quite risky, because only a small segment of the database has such codes assigned. Hirsch's 2005 article about the h-index is assigned to the physics category. Hirsch's 2007 article is not assigned to any of the predefined categories. You may qualify the search by keywords, but you are left on your own which keywords to use, and how many of them.

Sooner or later your search will produce strange names. Although you will not search for the odd family names noted below, they will show up as co-authors; or if you search by journal or keyword, they also appear as single author, and on a bad day when you search by the title of your own work you may find it under any of the names that I used in the "canary test".

For example, the most prolific author in the Emerald journals (according to GS) is "F Password", who purportedly authored 13,800 papers for journals of this publisher. (The archive of Emerald is not aware of such an author.) If the search is extended to the

entire family (i.e. not using first initial), the most productive author would be the person with the last name "Profile", allegedly the author of 17,300 papers in the Emerald collection, 12,400 attributed by GS to "M Profile". In *Online Information Review* and its two previous titles, "M Profile" (76 publications) is just a notch ahead of "F Password" (74 publications). But this is not true for the GS universe, where "F Password" is by far the most productive author (102,000 hits reported by GS), which attributes merely 12,800 works to "M Profile". System-wide the most prolific authors are members of the "Password" family, with 910,000 publications attributed to it by GS. "F Password" is the most prominent family member, with 102,000 papers attributed to him/her by GS. Obviously, "Forgot password" is a much more common element on the menus than the "My profile" option, and those authors reported by Google are as dead souls as Chichikov's serfs. The important for a researcher to proceed accordingly in interpreting the hit counts and citation counts (Figure 1).

No wonder that authors, journals and the numerical-chronological designations (publication year, volume, issue and starting page numbers) are mis-identified for millions of documents. As a consequence, the citation-matching algorithm of GS is equally unreliable, often yielding excessive and obviously absurd numbers of false positives and false negatives. GS plays fast and loose with the numbers, the hit counts and the citation counts. The software module which presents the results has stopped ranking the result list by citation counts, and now uses a new ranking algorithm. Its explanation (http://scholar.google.com/intl/en/scholar/about.html) does not true. It promises that GS aims to sort articles the way researchers do, weighing the full text of each article, the author, the publication in which the article appears, and how often the piece has been cited in other scholarly literature. Considering the absurd author names mentioned above and their frequency as reported by GS, one may have doubts. Further examples will shed more lights onto the name problems. This simple example below shows what an idle claim is the one about ranking.

Figure 1.
The ultra-prolific researcher F Password

GS does not assign a rank number but the Publish or Persih (PoP) utility (www.harzing.com) does show what was the rank order number of the items in the result list. Here, is a duplicate pair, each with four citations. These are from the same journal; they have same per year citation frequency; they have the same full text, same authors, same publication year (if currency is a ranking factor), so there is no distinction between them, and thus they should have the same rank, should they not? Well, they do not. One is ranked as the 102nd, and the other as the 402nd item. This is quite a rank difference, especially in a population of 432 records for papers published in *Online Information Review*. Actually, there is a difference, as there is a typo at the end of the name of the fourth author, "Weekes" instead of "Weeks" in one of them, which also uses e-prints in the last word of the title, instead of the e-preprints. So was it penalised for the lower ranking? No, that got the much better ranking (Figure 2).

You can see more oddities from the tiny sample below that the parser has managed to convert "Julie M Still" to "Julie M" from the Emerald archive, and "Martin Myhill" to "M Martin" from Ingenta. There are many others in this small sample, such as "S Carol" for "Carol S Bond", "G David" for "David Green", or "Peter J" for this author – all correct in the sources, but GS's parser must have used the first letter of the last name for first initial, and spelling out the first name in full – rather unfortunate both for the productivity and for the citedness statistics of the individuals.

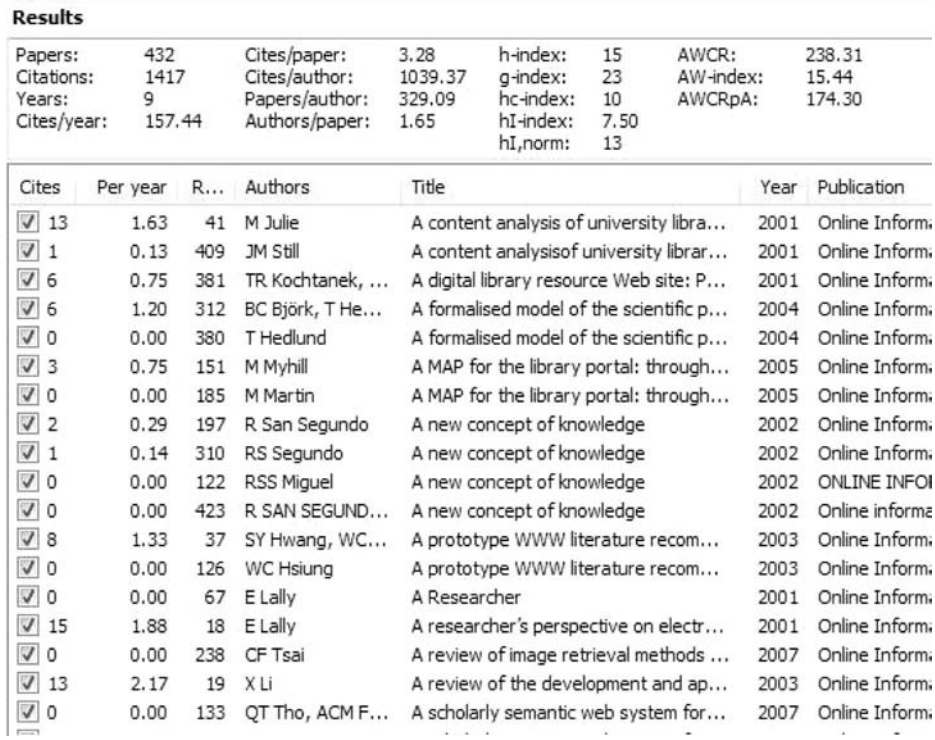
"Julie M Still" is particularly hard hit, because 13 of the references to her article are attributed to "M Julie", so if the searcher looks up her name in the correct format, as "J M Still", there will be only a single article citing her, and she loses the 13 others. You can also see the odd quadruplicate case for "Rosa San Segundo Miguel", who may now regret having a four-element name, just as I regret having insisted for too long that the accents on my first and last names be used. Of course, my family name even without the accent make most of my citers misspell it as Jasco, and there go my citation counts (Figure 3).

As we saw earlier, GS tends to attribute citations to authors and journals that do not deserve it. The worst type of such attributions is when a pseudo-author created by GS takes away the citation from the legitimate author. The most notorious pseudo authors are "F Password" and – for records extracted from the Emerald Collection – "M Profile". Obviously they are dead souls, while the authors deprived of their citations are living, working researchers. Take as an example two articles that Hong Iris Xie published in *Online Information Review* (one with Colleen Cool as co-author). The Emerald archive shows correctly the data, but GS attributes these to the author "M Profile" and deprives the legitimate authors of ten and four citations, respectively, (Figure 4).

A senior researcher without empathy and with a high h-index, or for blind love of GS, may downplay such unintended identity and citation theft, but they may be hit

Cites	Per y...	Rank	Authors	Title	Year
☑ 4	0.57	102	WG Town, BA Vickery, J Kuras...	Chemical e-journals, chemical e-pr...	2002
☑ 4	0.57	402	WG Town, BA Vickery, J Kuras...	Chemical e-journals, chemical e-pr...	2002
☑ 4	0.57	366	ACM Fong, SC Hui, HL Vu	Effective techniques for automati...	2002
☑ 4	0.57	263	C Chen, H Chen, K Chen, J Hsi...	The design of metadata for the Di...	2002
☑ 4	0.50	53	A Díaz, P Gervás, A García, I C...	Sections, categories and keyword...	2001

Figure 2.
Odd ranking of duplicate pair with same citation count



Results

Papers:	432	Cites/paper:	3.28	h-index:	15	AWCR:	238.31
Citations:	1417	Cites/author:	1039.37	g-index:	23	AW-index:	15.44
Years:	9	Papers/author:	329.09	hc-index:	10	AWCRpA:	174.30
Cites/year:	157.44	Authors/paper:	1.65	hI-index:	7.50		
				hI,norm:	13		

Cites	Per year	R...	Authors	Title	Year	Publication
13	1.63	41	M Julie	A content analysis of university libra...	2001	Online Inform
1	0.13	409	JM Still	A content analysisof university librar...	2001	Online Inform
6	0.75	381	TR Kochtanek, ...	A digital library resource Web site: P...	2001	Online Inform
6	1.20	312	BC Björk, T He...	A formalised model of the scientific p...	2004	Online Inform
0	0.00	380	T Hedlund	A formalised model of the scientific p...	2004	Online Inform
3	0.75	151	M Myhill	A MAP for the library portal: through...	2005	Online Inform
0	0.00	185	M Martin	A MAP for the library portal: through...	2005	Online Inform
2	0.29	197	R San Segundo	A new concept of knowledge	2002	Online Inform
1	0.14	310	RS Segundo	A new concept of knowledge	2002	Online Inform
0	0.00	122	RSS Miguel	A new concept of knowledge	2002	ONLINE INFOI
0	0.00	423	R SAN SEGUND...	A new concept of knowledge	2002	Online informa
8	1.33	37	SY Hwang, WC...	A prototype WWW literature recom...	2003	Online Inform
0	0.00	126	WC Hsiung	A prototype WWW literature recom...	2003	Online Inform
0	0.00	67	E Lally	A Researcher	2001	Online Inform
15	1.88	18	E Lally	A researcher's perspective on electr...	2001	Online Inform
0	0.00	238	CF Tsai	A review of image retrieval methods ...	2007	Online Inform
13	2.17	19	X Li	A review of the development and ap...	2003	Online Inform
0	0.00	133	QT Tho, ACM F...	A scholarly semantic web system for...	2007	Online Inform

Figure 3.
Some names with initialized last name and spelled out first names among the duplicates and quadruplicates

already (without knowing) or will likely to be hit in the future. GS will take away the identity and citations of authors for much higher cited works as well. My long-time favorite author, "I Introduction" that some deny to exist, has nearly 6,000 papers reported by GS and has had some good catch to improve the h-index. In this case two authors are robbed of 110 citations and of the recognition of their authorship. In some European countries omitting the author name from the publication is infringement of the moral component of copyright, an unknown concept in US copyright law (Figure 5).

I do not know for how many papers, authors and citations this misappropriation of identity and citations has happened because GS did not unseat the real author(s), but rather just added the interloper. It even goes one step further and gives citations to researchers who had nothing to do with authoring the paper. GS is quite inventive in adding co-authors.

For example, Hirsch wrote his seminal paper alone about the h-index, but in the long list of versions in mirror sites of the arXiv prep-print server gathered by GS, he finds himself in strange company – due to GS. What should make one really pause is that his "co-authors" are the physicists whose h-index he calculated, and included in an enumerative list. What made GS's parser think that three of the listed physicists were co-authors? Why were others in the list not promoted? How often are people mentioned in a paper designated by GS as co-authors? How would this affect the h-index if fractional points are to be used in proportion to the number of co-authors? I demonstrated earlier that GS happily makes up author names from menu options and

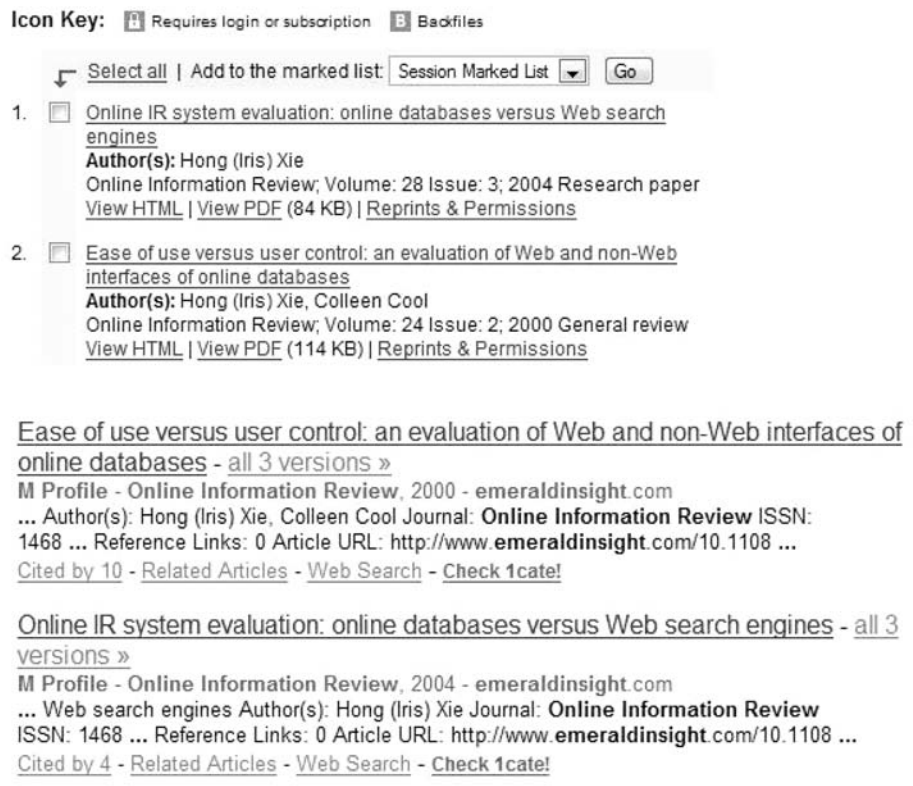


Figure 4.
Two records as they appear in the Emerald archive and in GS

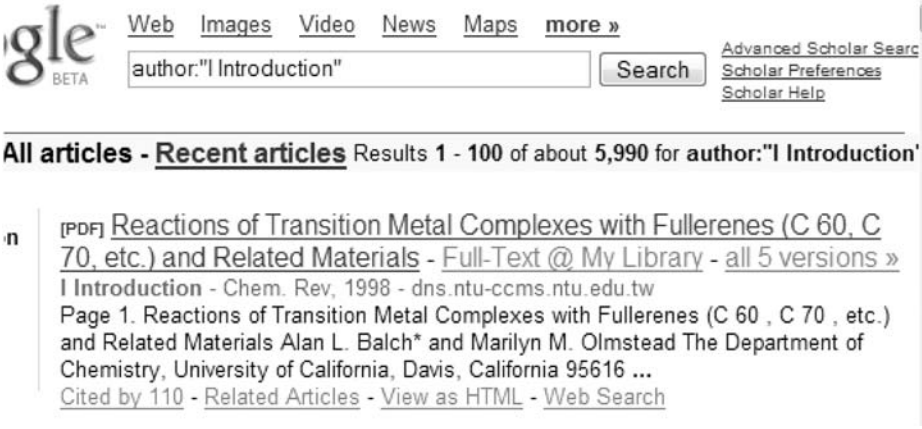


Figure 5.
Identity and citation misappropriation as intellectual property lawyers would say

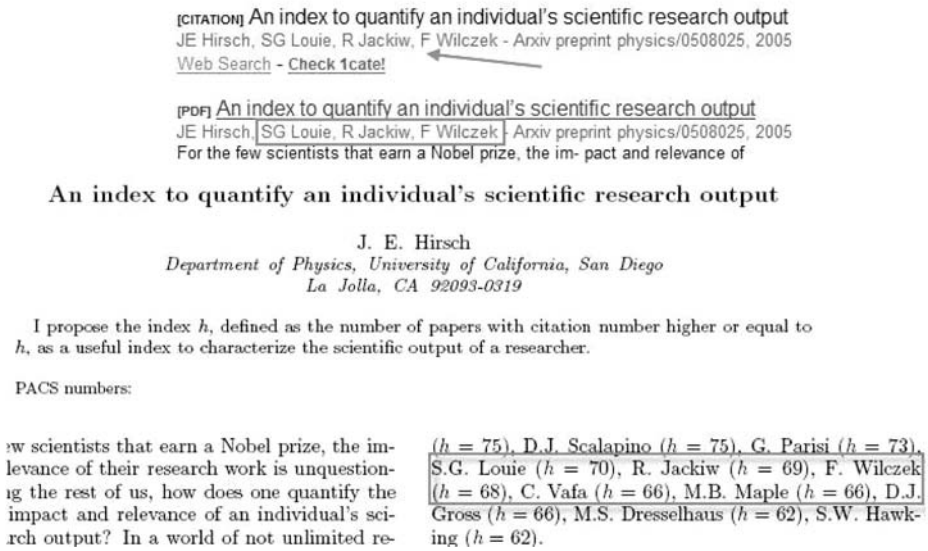
chapter headings, as well as publication years from page numbers, and practically from any number that appears on a page. These are signs of damaged software. It is worth thinking about this before popping the next question, which seeks answers to what is in a number, a hit count, and a citation count (Figure 6).

What is in a number?

In GS you never know, and you should never trust what it reports. The basic rule for GS-based h-index calculation is: always count and verify your hits, and citations. Unfortunately, it can be done only for up to 1,000 hits and citations. At least within this limit it can be quickly done by progressing in increments of 100 items (or just jumping to the last page of the result list) to call GS's bluff. When GS reports that it has 513 records for papers published in *Online Information Review* from 2000 (when the journal received this new title), it should not be taken at face value (Figure 7).

Proceeding to the second round (displaying the result list from 101 to 200) shows a lower number (490). Then it keeps decreasing, and the last offer is 432 records. Having dealt with GS, it is obvious that what you get is not 432 records for 432 articles, reviews and editorials.

It is not as if 432 were too many hits at first glance, but rather because there are almost always duplicates in the result sets of GS.


[CITATION] An index to quantify an individual's scientific research output
JE Hirsch, SG Louie, R Jackiw, F Wilczek - Arxiv preprint physics/0508025, 2005
Web Search - Check 1cate!

[PDF] An index to quantify an individual's scientific research output
JE Hirsch, SG Louie, R Jackiw, F Wilczek - Arxiv preprint physics/0508025, 2005
For the few scientists that earn a Nobel prize, the im- pact and relevance of

An index to quantify an individual's scientific research output

J. E. Hirsch
Department of Physics, University of California, San Diego
La Jolla, CA 92093-0319

I propose the index h, defined as the number of papers with citation number higher or equal to h, as a useful index to characterize the scientific output of a researcher.

PACS numbers:

...w scientists that earn a Nobel prize, the im- ...levance of their research work is unquestion- ...g the rest of us, how does one quantify the impact and relevance of an individual's sci- ...rch output? In a world of not unlimited re-

($h = 75$), D.J. Scalapino ($h = 75$), G. Parisi ($h = 73$), S.G. Louie ($h = 70$), R. Jackiw ($h = 69$), F. Wilczek ($h = 68$), C. Vafa ($h = 66$), M.B. Maple ($h = 66$), D.J. Gross ($h = 66$), M.S. Dresselhaus ($h = 62$), S.W. Hawking ($h = 62$).

Figure 6.
Persons listed in the article as subjects of a test (bottom), are promoted to co-authors by GS (top)

Figure 7.
The first hit count reported is like the asking price in the bazaar

The duplicates are there because GS hoards records from many sources (Figure 8).

GS does not offer any sort option, and the duplicates are not queuing like passengers at a bus stop. Luckily, the PoP utility developed by Tarma Software Research Pty does it, and this makes it easier to herd the scattered records from the result list, and count how many net records are there. In our example there are 318 non-duplicate records; the rest are duplicates, triplicates and one quadruplicate, so the total number of unique records is close to 360, or 70 per cent of the initial promise of GS, and 83 per cent of its last offer. It is not a good deal, but GS has much worse rates of duplicates and triplicates. One of the reasons for this is the hoarding of records from so many secondary sources, primarily from indexing/abstracting databases such as ERIC and PASCAL, which do not use the same title and/or the same name format as the publishers' collection. The other reason is GS's parsing disability (to be discussed later).

GS would have done much better to focus on the digital collections of the hundreds of scholarly publishers who are members of the CrossRef association (www.crossref.org), which is the DOI link registration agency for scholarly and professional publications.

These publishers are the ones with well-tagged, huge, full-text digital archives of more than 30 million articles and other publications. After all, the whole idea came from the fact that Google, Inc. was commissioned to create the CrossRef database many years ago.

Unfortunately, the developers of GS believed that their parsing software would be smarter in automatically extracting metadata from the full-text archives than the

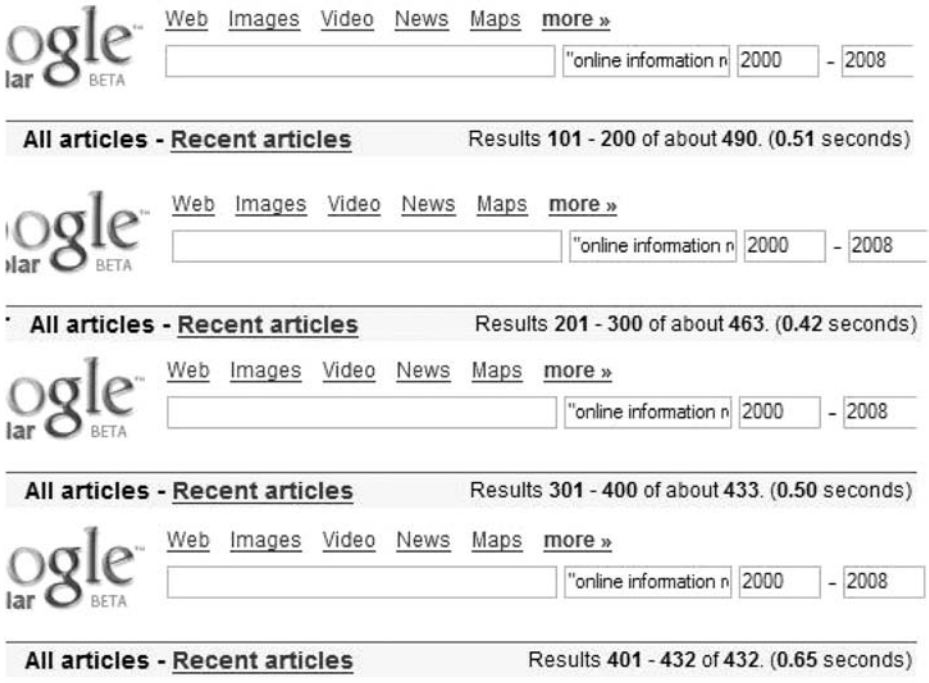


Figure 8.
Consecutive steps to make GS its last offer

process of creating metadata by librarians. What Google misses the most is an experienced, no-nonsense librarian. In the absence of such a person, the developer chose not to use the existing metadata which identify and tag the title, author, journal name, publication year and other traditional data elements of descriptive and subject cataloguing (pardon the expression).

There are good parsers and bad parsers, and some are superbly trained by developers. Such is the one used for the Astrophysics Data Systems (ADS) project. ADS does a better job of parsing old OCR-ed manuscripts on brittle paper from the Ottoman era than GS's does of digital files. The same can be said about the citation-matching software. GS has no such essential output options as marking selected records, sorting a set, exporting a subset. It does not even number the elements in the set, and it does not calculate the h-index. This is where the PoP program can pop into calculate the h-index and many of it variants.

It also produces pretty statistics which could be informative, but with the duplicates and triplicates, the frequent omissions of the second, third, etc. authors, the number of papers published, the authors/paper, and papers per author indicators are of little use. The natural unintelligence of the GS parser has serious implications also for citation matching, citation counts and the h-index; therefore I am not lacking the self-citation adjusted indicators, because the citation matcher would do a frightening self-citation analysis that would yield higher numbers than the one which does not remove the self-citation. If you wonder why am I so sceptical, just read my recent evaluation of the basic search features of GS (Jacsó, 2006). Whenever you use PoP software, which is far the most sophisticated and most resistant to blocking by Google, keep in mind that, if it receives garbage from GS, it cannot make gold of it. I am most concerned about the inflated citation counts, even if it makes everyone look better.

Fool's money and counterfeit money

If it is a citation count reported by GS, it is almost always less than it appears. Take as an example the citation count reported for my paper entitled "Google Scholar: the pros and the cons", published in *Online Information Review* in 2005. It is reportedly cited 57 times – good news for the author and also for the publisher. But it is bad news for both (although not new for this author) that the number is just not true. Right at the beginning when asking GS to "show the money", it tells that actually there are 55 citing references, and it can show 53. As usual, it cannot tell the truth even when the numbers are very small, and when there is no reason to use the ballpark estimation for the "users' convenience" cliché. Some of the purportedly citing scholarly documents were as inaccessible for me as they are in Chinese. I did not have physical access to four source documents in order to judge them. One was a blog reference which would not likely contribute to promoting me to professor emeritus when I retire. Six of the items are duplicates.

There are four that do not cite me, let alone the specific paper. An additional one is easy to spot, as it obviously could not cite any GS paper for a simple reason: it was written several month before GS was launched, and a year before I wrote my purportedly cited paper. It is an LIS master's thesis by a person called D C Field, according to GS. Actually D C Field is created by GS from the Dublin Core Field label for the metadata section. The author is actually Meghan Lafferty, but the wrong name is a lesser problem from my perspective. Not surprisingly, my paper is not mentioned in the thesis. It is an enigma as to what made GS claim that it cites my paper. The same

is true for the other non-citing papers. These are more unnerving than the usual false positives. These leave me in the dark and will make me check the validity of all the citations (Figure 9).

GS's citation matching algorithm does not check that all the elements are in a single entry in the bibliography and delivers false citation counts. Even competent researchers familiar with citation indexing may overlook this. For example, Vanclay (2008), in a manuscript posted on various preprint servers, asserts that WoS excludes a number of articles from the *Journal of Forestry Ecology & Management* (*FEM*) which are highly cited in GS. His top example is a journal article purportedly cited 114 times

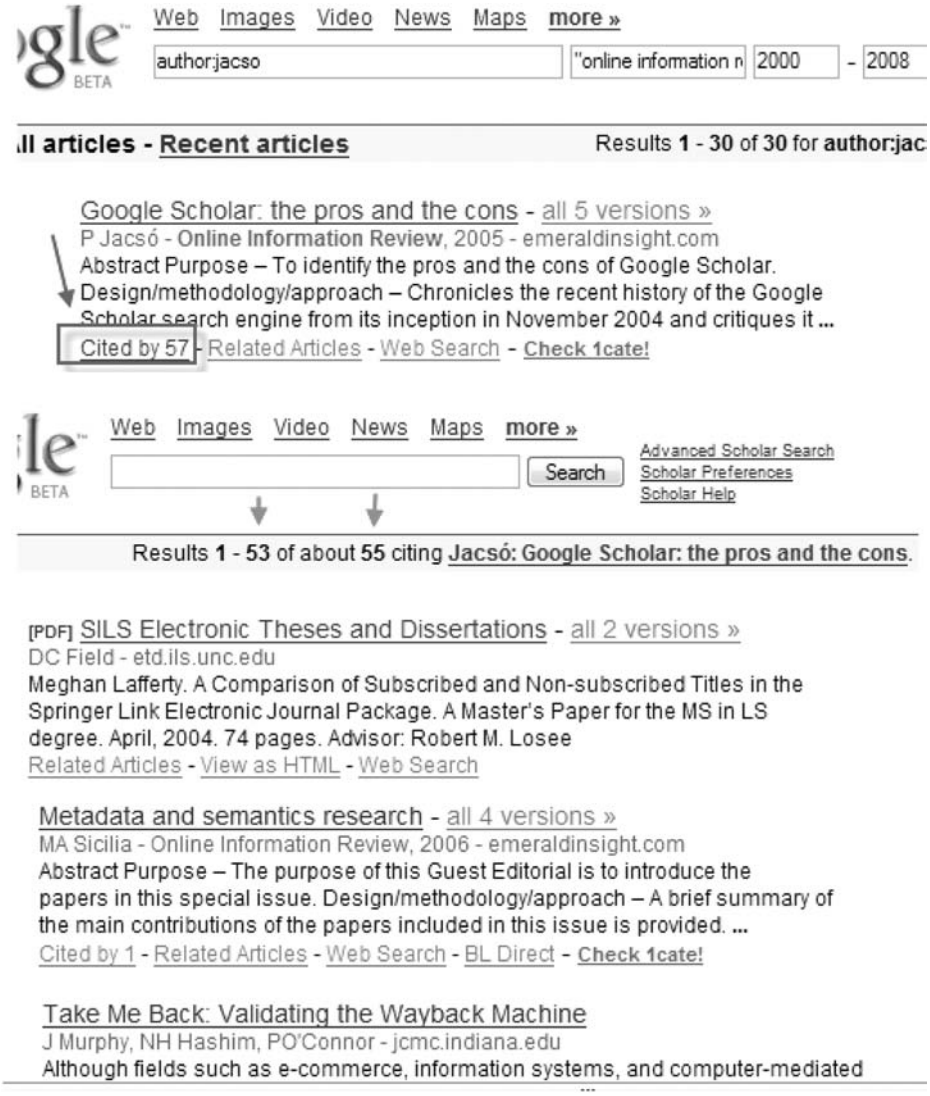


Figure 9.
Phantom citations from papers are like counterfeit notes

according to GS – I checked the first 11 citing items (one I did not have access to), and there was only a single item that cited the article in the journal Vanclay refers to; all the other references were to one of the several yearly updated technical reports that had part of the same title as the journal article. Vanclay's whole article focuses on journals, and this example adds nothing to support his argument that GS recognises many more articles from the journal than WoS. GS lumps together a series of technical reports and a journal article, awarding the citations to the journal. This is a typical mis-recognition and mis-attribution scenario in GS's citation-matching algorithm, and a warning about how loose the criteria may be, apparently ignoring the source and the publication year in the matching process. I posted at http://jacso.info/h-gs-fem a file which shows the relevant reference excerpts in the documents purportedly citing the journal article. Such references embarrass authors who may proudly but wrongly claim that their paper in *FEM* has been cited more than 100 times (Figure 10).

But the four papers that are listed by GS as citing my paper about GS are not just false positives but phantom citations, where the author's name does not appear at all in the bibliography.

It may be worth it to pause, suspend the examination of GS, and gingerly ask its developers publicly about the implications of also this. If we can believe in statistics, TechCrunch reported a 32 per cent decrease in GS's usage in the past year (www.techcrunch.com/2007/12/22/2007-in-numbers-igoogle-googles-homegrown-star-performer-this-year/). Notess (2008) posted a note in March about this, noting that "I have found general Web searches often more effective than GS searches for at least some scholarly documents."

Some of the early fans of GS changed their minds. Reinhard Wetz posted a blog on MEDLIB-L withdrawing his enthusiastic praise for GS. Actually, he went much further, writing that:

> Google Scholar's ability to identify citations is at best dodgy, but more likely misleading and based on very spurious use of algorithms establishing similarity and relationships between references [. . .] Google Scholar should withdraw the "cited by" feature from its Beta version and probably not offer it in the final version.

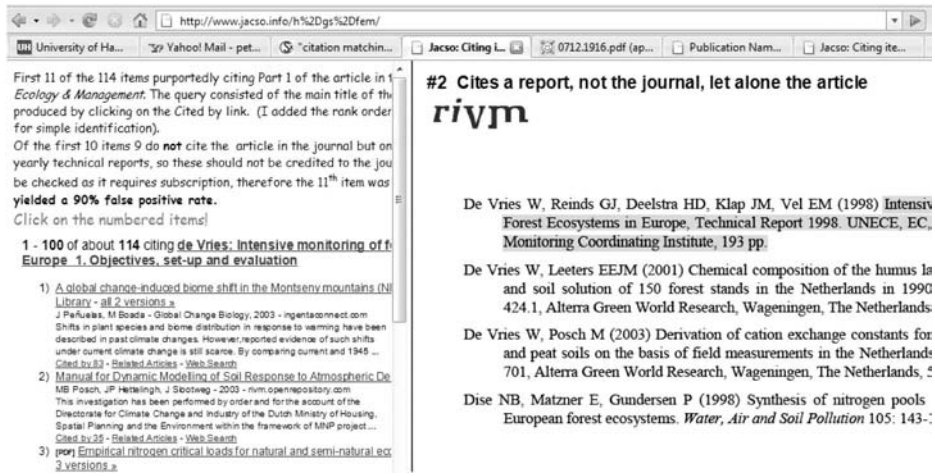


Figure 10.
False positives – references to different items

Dean Giustini also lost his enthusiasm and patience, when he wrote in early January:

> Scholar is not as useful as promised, and many web searchers are now moving back to regular Google for indiscriminate *scholarly* trawling of the web [...] Unless it changes its course, GS well go the way of the dodo bird eventually.

My suggestions:

- Keep using GS for resource discovery and as a metasearch engine.
- Do not cancel your WoS or Scopus subscription.
- Think twice before using GS to calculate h-indexes without a massive corroboration of the raw data reported by GS.

References

Bar-Ilan, J. (2008), "Which h-index? A comparison of WoS, Scopus and Google Scholar", *Scientometrics*, Vol. 74 No. 2, pp. 257-71.

Cronin, B. and Meho, L.I. (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.

Giustini, D (2008), "Google's growth rates", available at: http://weblogs.elearning.ubc.ca/googlescholar/archives/044168.html

Goble, C. (2006), "Science, workflows and collections", paper presented at the UKSG Conference at Warwick University, Coventry, 3-5 April, available at: www.uksg.org/sites/uksg.org/files/imported/presentations8/goble.ppt

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.

Hirsch, J.E. (2007), "Does the h-index have predictive power?", available at: http://arxiv.org/PS_cache/arxiv/pdf/0708/0708.0646v2.pdf

Jacsó, P. (2006), "Puppy love versus reality: the illiteracy, innumeracy, phantom hit counts and citation counts of Google Scholar", Plenary closing session presentation at the UKSG Conference at Warwick University, 3-5 April, available at: www2.hawaii.edu/ ~ jacso/conferences/UKSG-GS-ppt-innumeracy-illiteracy.ppt

Jacsó, P. (2008), "The plausibility of computing the h-index of scholarly productivity and impact using reference enhanced databases", *Online Information Review*, Vol. 32 No. 2, pp. 266-83, available at: www.jacso.info/PDFs/jacso-h-index-plausibility-OIR-2008-32-2.pdf

Neuhaus, C., Neuhaus, E. and Asher, A. (2008), "Google Scholar goes 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.

Notess, G. (2008), "Scholar down, books up", available at: www.searchengineshowdown.com/blog/2008/01/scholar_down_books_up.shtml

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. 2, pp. 297-301.

Sanderson, M. (2008), "Revisiting h measured on UK LIS and IR academics", *Journal of the American Society for Information Science and Technology*, available at: http://dx.doi.org/10.1002/asi20771 (accessed 18 March).

Vanclay, J. (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.

Vanclay, J. (2008), "Ranking forestry journals using the h-index", available at: http://arxiv.org/abs/0712.1916 (accessed 17 March 2008).

Wentz, R. (2004), "WoS versus Goggle Scholar: cited by correction", available at: http://listserv.acsu.buffalo.edu/cgi-bin/
wa?A2 = ind0412B&L = medlib-l&P = R5842&I = -3&m = 95812

Further reading

Jacso, P. (2008), "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

Meho, L.I. and Yang, K. (2007a), "Fusion approach to citation-based quality assessment", paper presented at the 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, 25-27 June, 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 vs 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/

Corresponding author

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