

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

The introductory part (Jacso, 2008b) to the series of papers 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 get distorted depending on the software and the content of the database(s) used, and the searchers' skill and knowledge about the features of the database.

In this issue Google Scholar 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 Google Scholar results –for various reasons- are often irreproducible, which is not conducive to real scholarly research. The examples for hit counts and citation counts misrepresented by Google Scholar 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, re-counting valid citations and pointing out Google Scholar's unscholarly and irresponsible handling of data can be the most directly demonstrated through this tiny microcosm for readers of *Online Information Review*. It is also important to realize that effective corroboration of the h-index and its two component indicators, can be done well only for persons and journals that the researcher is intimately familiar with. Corroborative tests must be done in every database for important research whose results may effect people, just as canaries had to be used to signal dangers in the coal mines.

Dead canaries

The lethal deficiencies in the software of Google Scholar -from the bibliometric and scientometric perspectives- dwarf the content limitations. The consequences are present in the entire Google Scholar universe for the simple reason that most of the problems are caused by the software of Google Scholar, by the brain damaged parsing and slapdash citation matching algorithms. The problems are caused not merely by typos and other inaccuracies in the source data, neither by missing one or two highly cited articles and a dozen lowly cited papers well below the reasonably calculated h-index. These do not influence the h-index which is quite stable and robust – as Vanclay (2007) convincingly explained and illustrated.

The most serious software deficiencies which are present across the board even though not visible in every search, don't bother the casual searchers hunting for a few good papers, but they influence and may distort the h-index computed by the third-party utilities which inevitably show the GIGO (Garbage In/Garbage Out) symptoms. They 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 frequently reported by Google Scholar, especially when computing the h-index for very productive and/or very highly cited periodicals, such as Science, or Lancet. The ratios of substantial errors may be

different in the test microcosm and the Google Scholar Universe, but they are like the dead canaries in the coal mines, some have fewer, others have more.

The context of the search

Google Scholar'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 Google Scholar 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 the searcher is associated with. Google Scholar also deserves credit to make the information retrieval process smooth and simple (especially if the library has a link resolver) without the need to a) identify the best candidates from the variety of databases available through the library, then b) learn the particular software used by the databases, and c) run and refine searches in the different systems.

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

Beyond the instant gratification, the most important virtue of Google Scholar 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 born or borne digital. These include the content of millions of books passed on to Google Scholar 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, Google Scholar 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 because of the blogorrhea, but there are some good ones, just as there are some master 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 to the software problems to be discussed. The content base is certainly there for calculating the h-index, although with some reservations regarding, for example, for papers published more than 15 years ago, but content reservations are applicable to all of the alternatives.

The flop side of this much improved access to digital materials is that papers which are not available digitally, remain barely known to Google Scholar (and to the users) as its content is created entirely automatically just as it is in Yahoo, Ask, Exalead, and GigaBlast. The difference is that Google created Google Scholar –purportedly to accommodate scholarly literature- while there is no Yahoo Scholar, Ask Scholar, or ExaLead Scholar. (There is a Windows Live Academic [WLA] database from Microsoft

which is separate from the Windows Live database and theoretically analogous to Google Scholar, but it is more like Windows Live Elementary (public) - even after the introduction of the citation count feature just as I was working on this paper.)

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. The former indicates the productivity of authors (a traditionally essential criterion in the academic world) and journals, while the latter may serve as an indicator for 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 Jorge E Hirsch, 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 (Hirsch, 2007) of professors, researchers, journals, and institutions in scholarly publishing.

In spite of 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 getting used more frequently than it may appear through just the scholarly and professional publications. The h-index is now shown in many resumes, applications for jobs, grants, sabbaticals, etc. Even in scholarly publications the majority of the authors take the "cited by" values as reported by Google Scholar at face value, and rush to conclusions in comparing these counts with those of Web of Science and Scopus.

Although Google Scholar 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 realize the incredible credibility gap between the hit counts and the citation counts as reported by Google Scholar and the reality.

From the launch of the service, it has been hopeless to get 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 in spite of this secrecy of North Korean dimensions, Google Scholar was so widely embraced by researchers and librarians for scientometrics purposes. Colin Cunniff, Dean of the University of British Columbia dedicated a blog to Google Scholar (<http://weblogs.elearning.ubc.ca/googlescholar/>), Reinhard Wentz, manager of the Imperial College Library in London, claimed on the MEDLIB-L discussion list that the "cited by" facility of Google Scholar is spectacular (later he withdrew that conclusion publicly, and the original document is not available anymore). Carol Goble (2006) of

Manchester University referred to Google as the “Lord’s gift” (she meant Google Scholar) in an aside of her otherwise impressive presentation at the UKSG 2006 conference. I presented a very different view about Google Scholar at the closing plenary session of this conference (Jacso, 2006) for reasons which are still valid today.

There have been efforts to calculate the h-index and/or gauge the extent of Google Scholar’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 Google Scholar.

The note in Meho and Yang’s (2007) informative comparison of Web of Science (Wos), Scopus and Google Scholar (GS) on the citation counts and ranking of 25 library and information science faculty members is very 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.

Google Scholar dispenses utterly unreliable indicators through its hit 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. The third party utility programs cannot help in this.

These authors, however, know the ins-and-outs of responsible citation analysis. They are aware of the serious limitations of Google Scholar’s document parsing and citation matching algorithms which are as good in identifying authors and matching citations, as Alzheimer patients are at the pre-dementia stage in recognizing relatives, and matching their names and relationship. That’s why their team has spent 3,000 hours on verifying and correcting Google Scholar’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. Google Scholar 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 the records with links, one must go to the primary documents (most of them are available only for subscribers), find in the cited reference list the one 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 extremely grueling.

Google Scholar 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 the hit counts and the 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

the fact how much the combination of these result lists can improve the h-index of certain types of researchers.

There are others who know the parsing and citation matching debility of Google Scholar, but don't have the time needed to verify its reported citation counts. Judit 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 about 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 can't blame her, and others who accept the citation counts as reported by Google Scholar, but it is like changing money into local currency in, say, Iraq, and not counting and checking banknote by banknote the money received. This attitude will encourage the back alley money changers to go bolder and bolder. This attitude was likely to have been calculated in the development of Google Scholar's citation matching algorithm, and may be one of the reasons for the secrecy about all the details of the system.

Google Scholar very often regales the 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. Even if it is just natural unintelligence that cripples the citation matching it is like having an undoubtedly useful paramedic to provide first aid, take you to the hospital but the patient may not want the paramedic to do the lab works, the X-rays, interpret the results, do the diagnosis and engage in a cranial surgery. True, Google Scholar does not calculate the h-index, does not even rank anymore the hit list by citedness, but it does the equivalent of the lab test, offering the 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 issue. Still, you never know in Google Scholar for sure what is in a name. There is no option to browse in Google Scholar at all, so you just search blindfolded. Neither is there any software feature to distinguish authors with the same name and first initial, as there are in WoS and Scopus. The only chance to distinguish JE Hirsch the physicist from JE Hirsch the audiologist is to limit the search to the closest broad subject category. This is quite risky because only a small segment of the database has such codes assigned. His 2005 article about the h-index is assigned to the physics category. His 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 following odd family names, 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 Google Scholar) 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 who allegedly is the author of 17,300 papers in the Emerald collection, 12,400 attributed by Google Scholar to “M Profile”. *In Online Information Review* and in 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 Google Scholar universe where “F Password” is far the most productive author (102,000 hits reported by Google Scholar), 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 Google Scholar. As mentioned before, F Password is the most prominent member with 102,000 papers attributed to him/her by Google Scholar. Obviously, “Forgot password” is a much more common element on the menus than the “My profile” option, and those other authors reported by Google Scholar are as dead souls as Chichikov’s serfs. The important thing for a researcher is to recognize the symptom of the dead canaries in the coal mine, and proceed accordingly in interpreting the hit counts and citation counts.



Figure 1. The ultra-prolific researcher F Password, and other important “authors” suggested by Google Scholar

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 Google Scholar is equally unreliable, often yielding excessive, and obviously absurd number of false positives and false negatives. Google Scholar plays fast and loose with the numbers, the hit counts and the citation counts. The software module which presents

the results, stopped ranking the result list by citation counts, and uses a new ranking algorithm like a population commitment in an election campaign, and is as much true.

It promises that Google Scholar 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 Google Scholar, 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.

Google Scholar does not assign a rank number but the Publish or Perish (PoP) utility (<http://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 4 citations. They 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 not they? Well, they don't, One is ranked as the 102nd the other as the 402nd item. 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 penalized for the lower ranking? No, that got the much better ranking.

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

You can see more oddities from the tiny sample below that the parser has managed to extract Julie M Still's name as Julie M from the Emerald archive, and Martin Myhill's name as 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. These are all correct in the sources, but Google Scholar's parser must have been on something to use 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 JM 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 to have a name of four elements just as I regret to have insisted for too long to use the accents on my first and last name and as everyone else who has accented

characters in their names. Of course my family name even without the accent makes most of the citers misspell it as Jasco, and there go my citation counts.

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

As we saw earlier, Google Scholar tends to attribute citations to authors and journals that don't deserve it. The worst type of such attributions is when a pseudo author created by Google Scholar takes away the citation from the legitimate author. The most notorious pseudo author is 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 coauthor). The Emerald archive shows correctly the data, but Google Scholar attributes these to the author M Profile, and deprives the legitimate authors from 10 and 4 citations.

Icon Key: Requires login or subscription Backfiles

[Select all](#) | Add to the marked list:

- [Online IR system evaluation: online databases versus Web search engines](#)
Author(s): Hong (Iris) Xie
Online Information Review; Volume: 28 Issue: 3; 2004 Research paper
[View HTML](#) | [View PDF](#) (84 KB) | [Reprints & Permissions](#)
- [Ease of use versus user control: an evaluation of Web and non-Web interfaces of online databases](#)
Author(s): Hong (Iris) Xie, Colleen Cool
Online Information Review; Volume: 24 Issue: 2; 2000 General review
[View HTML](#) | [View PDF](#) (114 KB) | [Reprints & Permissions](#)

[Ease of use versus user control: an evaluation of Web and non-Web interfaces of online databases - all 3 versions »](#)

M Profile - Online Information Review, 2000 - emeraldinsight.com

... Author(s): Hong (Iris) Xie, Colleen Cool Journal: **Online Information Review** ISSN: 1468 ... Reference Links: 0 Article URL: <http://www.emeraldinsight.com/10.1108> ...

[Cited by 10](#) - [Related Articles](#) - [Web Search](#) - [Check 1cate!](#)

[Online IR system evaluation: online databases versus Web search engines - all 3 versions »](#)

M Profile - Online Information Review, 2004 - emeraldinsight.com

... Web search engines Author(s): Hong (Iris) Xie Journal: **Online Information Review** ISSN: 1468 ... Reference Links: 0 Article URL: <http://www.emeraldinsight.com/10.1108> ...

[Cited by 4](#) - [Related Articles](#) - [Web Search](#) - [Check 1cate!](#)

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

Senior researchers without empathy and with high h-index or just for blind love of Google Scholar may bagatellize such unintended identity and citation theft, but they may have been hit already (without knowing) or will likely to be hit in the future. Google Scholar will take away the identity and citation 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 Google Scholar and has had some good catch also to improve the h-index. In this case below, two authors are robbed from 110 citations and of the recognition of their authorship if you search by their name. In some European countries omitting the author name from the publication is infringement of the moral component of authors' copyright, an unknown concept in the U.S. copyright law.

The screenshot shows a Google search interface. At the top left is the 'gle BETA' logo. To its right are links for 'Web', 'Images', 'Video', 'News', 'Maps', and 'more »'. Below these is a search box containing the text 'author:"I Introduction"' and a 'Search' button. To the right of the search box are links for 'Advanced Scholar Search', 'Scholar Preferences', and 'Scholar Help'. Below the search box is a green bar with the text 'All articles - [Recent articles](#) Results 1 - 100 of about 5,990 for author:"I Introduction'.

The first search result is a PDF document titled '[PDF] Reactions of Transition Metal Complexes with Fullerenes (C 60, C 70, etc.) and Related Materials - Full-Text @ My Library - all 5 versions »'. Below the title is the text 'I Introduction - Chem. Rev, 1998 - dns.ntu-ccms.ntu.edu.tw'. The next line is 'Page 1. Reactions of Transition Metal Complexes with Fullerenes (C 60 , C 70 , etc.) and Related Materials Alan L. Balch* and Marilyn M. Olmstead The Department of Chemistry, University of California, Davis, California 95616 ...'. At the bottom of the result are links for 'Cited by 110 - Related Articles - View as HTML - Web Search'.

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

I don't know with how many papers and authors and citations this misappropriation of identity and citations has happened, but many of the canaries stopped singing in my test. I do know that some still sing because Google Scholar did not unseat the real author(s) just added the interloper. It even goes one step further and gives citations to researchers who had nothing to do with authoring the paper. Google Scholar 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 pre-print server as gathered and parsed by Google Scholar, he finds himself in strange company – thanks to Google Scholar. 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 Google Scholar's parser think that three of the listed physicists are co-authors? Why the others in the list were not promoted to c-author status? How often are people who are mentioned in a paper designated by Google Scholar as co-authors? How would this effect the h-index if fractional points are to be used in proportion to the number of co-authors? I demonstrated earlier that Google Scholar happily makes up author names from menu options and chapter headings, as well as publication years from page numbers, and practically from any number that appear on a page and Google Scholar fancies that it could be a good enough publication year. These are more than signs of blissful ignorance, they are signs of a brain-damaged software. It is worth to think about this before popping the next question which seeks answer to what is in a number, a hit count, and a citation count.

[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 impact 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:

For scientists that earn a Nobel prize, the impact and relevance of their research work is unquestioning the rest of us, how does one quantify the impact and relevance of an individual's scientific 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 Google Scholar (top)

What is in a number?

In Google Scholar you never know, and you should never trust what it reports. The basic rule for black market money changing and Google Scholar-based h -index calculation is just the opposite of the one for casino gambling: always count and verify your money, hits, and citations while standing in the back alley, or sitting at your PC. Unfortunately, it can be done only 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 Google Scholar's omnipresent bluffs. When Google Scholar reports that it has 513 records for papers published in *Online Information Review* from 2000 (when the journal got this title), it should not be taken at face value. It is just like the initial asking price in the bazaar, a warm-up for driving a bargain.



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

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 Google Scholar, it is obvious that what you get is not 432 records for 432 articles, reviews, and editorials.

Not as if 432 were too many hits at first glance, but because there are almost always duplicates in the result sets of Google Scholar. This is like when the back alley money changer folds a banknote in half, so the harried buyer who just thumbs through the banknotes on the longer side of the stack (as we usually do), does not realize that he is short changed.

The duplicates are there because Google Scholar hoards records from many sources, like the hostess who is afraid that the food plates for the party would not look big enough, fetches what she can from the garage, the kids' rooms and backpacks, and slaps cheap macaroni, stale potato salad, a hill of nachos on top of the gourmet plate.



Figure 8. Consecutive steps to make Google Scholar its last offer

Google Scholar does not offer any sort option, and the duplicates are not queuing up like passengers at a bus stop in London. Luckily, the Publish or Perish (PoP) utility developed by Tarma Software Research Pty does it, and that 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 there is one quadruplicate, so the total number of unique records is close to 360, i.e. 70% of the initial promise of Google Scholar, and 83% of its last offer. It is not a

good deal, but Google Scholar often has much worse rate of duplicates and triplicates. One of the reasons of 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 Google Scholar's parsing disability (to be discussed later).

Google Scholar 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 who have 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. I think that project gave the idea to create Google Scholar, not what the PR department of Google promotes for syndicated sweet stories of childhood deprivation of current scientific materials in the rural schools of future developers in their native countries.

Unfortunately, the developers of Google Scholar believed that their parsing software would be smarter in automatically extracting metadata from the full-text archives than the process of creating metadata by librarians. What Google misses the most is not another masseuse, chef, veterinarian and psychiatrist resident in GooglePlex but an experienced, no-nonsense librarian. In lack 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 cataloging (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. It does a better job parsing old OCR-ed manuscripts on brittle paper from the Ottoman era than Google Scholar's utterly unintelligent parser does of digital files. The same can be said about its citation matching software. Google Scholar 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. That's where the PoP program can pop in which calculates the h-index and many of its 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 Google Scholar 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 skeptical, just read my recent evaluation of the basic search features of Google Scholar (Jacso, 2006a). Whenever you use the PoP software, which is far the most sophisticated and far the most resistant to blocking by Google, keep in mind that if it gets garbage from Google Scholar 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

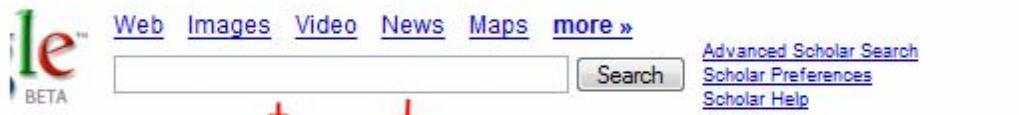
If it is a citation count reported by Google Scholar it is almost always less than what it looks like. Take as an example the citation count reported for my paper entitled Google Scholar: the pros and the cons, published in *Online Information Review* in 2005. It is reported to be cited 57 times – good news for the author and also for the publisher. Bad news for both (although not new for this author), that the number is just not true. Right at the beginning when asking Google Scholar to show the money, it tells that actually there are 55 citing references, and it can show 53. As usual, it can't tell the truth even when the numbers are very small, and when there is no reason to use the ballpark estimation which is –according to cliché claim “for users' convenience” . Some of the purportedly citing scholarly documents were intellectually not accessible for me as they are in Chinese. I didn't have physical access to 4 source documents 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, the folded bank notes in the stack, if you recall.

There are four that don't cite me, let alone the specific paper. An additional one is easy to spot as it obviously could not cite any Google Scholar paper for a simple reason: it was written several month before Google Scholar was launched, and a year before I wrote my purportedly cited paper. It is a LIS master thesis by a person called DC Field according to Google Scholar. Actually DC Field is created by Google Scholar from the Dublin Core Field label for the metadata section. The author is Meghan Lafferty in reality, but the wrong name is a lesser problem from my perspective. Not surprisingly, my paper is not mentioned in the thesis. It is a real enigma what made Google Scholar 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, once I submit this long overdue paper.



311 articles - [Recent articles](#) Results 1 - 30 of 30 for author:jacso

[Google Scholar: the pros and the cons - all 5 versions »](#)
P Jacsó - [Online Information Review, 2005 - emeraldinsight.com](#)
Abstract Purpose – To identify the pros and the cons of Google Scholar.
Design/methodology/approach – Chronicles the recent history of the Google Scholar search engine from its inception in November 2004 and critiques it ...
Cited by 57 - [Related Articles](#) - [Web Search](#) - [Check 1citate!](#)



Results 1 - 53 of about 55 citing [Jacsó: Google Scholar: the pros and the cons](#).

[\[PDF\] SILS Electronic Theses and Dissertations - all 2 versions »](#)
DC Field - [etd.ils.unc.edu](#)
Meghan Lafferty. A Comparison of Subscribed and Non-subscribed Titles in the Springer Link Electronic Journal Package. A Master's Paper for the MS in LS degree. April, 2004. 74 pages. Advisor: Robert M. Losee
[Related Articles](#) - [View as HTML](#) - [Web Search](#)

[Metadata and semantics research - all 4 versions »](#)
MA Sicilia - [Online Information Review, 2006 - emeraldinsight.com](#)
Abstract Purpose – The purpose of this Guest Editorial is to introduce the papers in this special issue. Design/methodology/approach – A brief summary of the main contributions of the papers included in this issue is provided. ...
[Cited by 1](#) - [Related Articles](#) - [Web Search](#) - [BL Direct](#) - [Check 1citate!](#)

[Take Me Back: Validating the Wayback Machine](#)
J Murphy, NH Hashim, PO'Connor - [jcmc.indiana.edu](#)
Although fields such as e-commerce, information systems, and computer-mediated

Figure 9. Phantom citations from papers are like counterfeit notes

Google Scholar's citation matching algorithm does not check that all the elements are in a single entry in the bibliography and delivers fool's money through the false citation counts. Even competent researchers, familiar with citation indexing may overlook this. For example, Vanclay (2008) in a manuscript posted at various preprint servers, asserts that WoS excludes a number of articles from the journal of *Forest Ecology & Management* (FEM), which are highly cited in Google Scholar. His top example is a journal article purportedly cited 114 times according to Google Scholar. I checked the first 11 citing items (one I did not have access to). 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 Google Scholar recognizes many more articles from the journal *Ecology & Management*. Google Scholar 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 Google Scholar's citation matching algorithm. It is also a warning about how loose the criteria may be which apparently ignores 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 make fool's money, and embarrass authors who may proudly but wrongly claim that their paper in *Forestry Ecology & Management* has been cited well over 100 times.

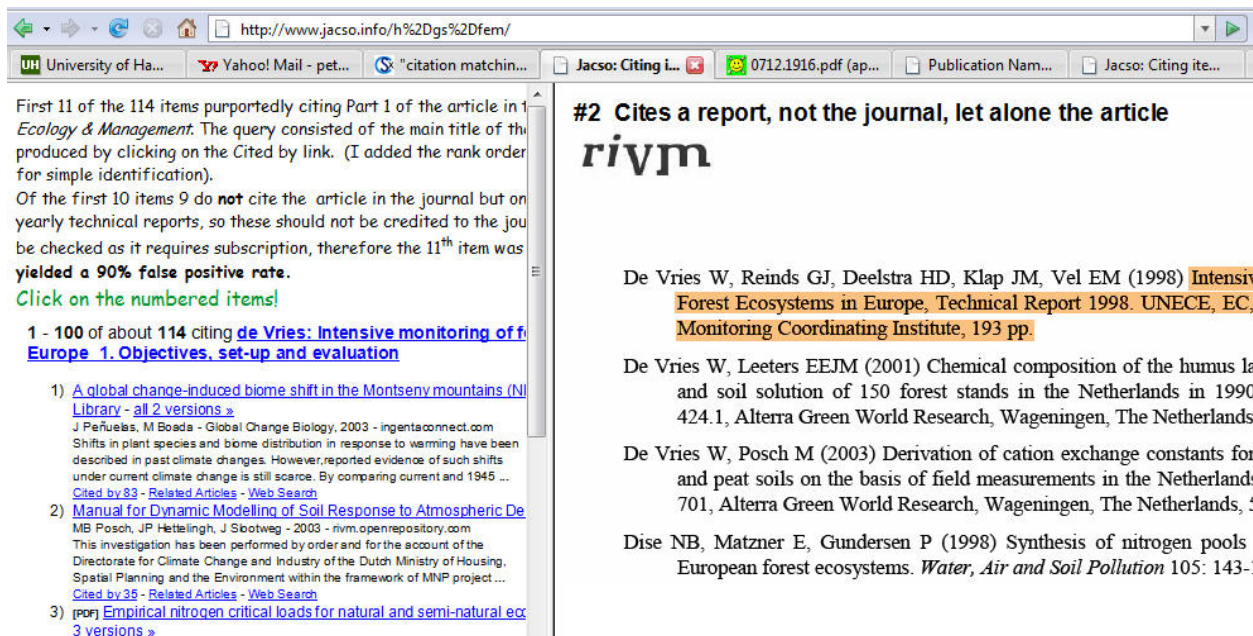


Figure 10. False positives – references to different items

But the four papers that are listed by Google Scholar as citing my paper about Google Scholar are not just false positives but phantom citations where the author's name does not appear at all in the bibliography. It is like getting counterfeit money – quite a bad transaction.

It may be worth to pause, suspend the adulation of Google Scholar, and gingerly ask its developers publicly about the implications of this malpractice. If we can believe in statistics, TechCrunch reported a 32% decrease in Google Scholar's usage in the past year << <http://www.techcrunch.com/2007/12/22/2007-in-numbers-igoogle-goggles-homegrown-star-performer-this-year/>>>. Maybe searchers realized that Google Scholar may have a diploma from a diploma mill Gregg Notess posted a note in March, about this, and wrote that "I have found general Web searches often more effective than Google Scholar searches for at least some scholarly documents".

Some of the early fans of Google Scholars changed their minds. Reinhard Wetz posted a blog on MEDLIB-L withdrawing its enthusiastic praise from Google scholar. 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”. 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 engines. Unless it changes for its users, Google Scholar is willing to be the dodo bird eventually.

My suggestion: keep using Google Scholar for resource discovery and a metasearch engine. Don’t cancel your WoS or Scopus subscription. Think twice before using Google Scholar to calculate h-indexes without a massive corroboration of the raw data reported by Google Scholar.

References

- Bar-Ilan, J. (2008). “Which h-index? – A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, Vol 74, No.2 pp. 257-271.
- Cronin, B., & 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. 1275-1278.
- 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. Presentation at the UKSG Conference at Warwick University April 3-5, 2006. available at <http://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. 16569-16572.
- Hirsch, J.E. (2007). “Does the h-index have predictive power?” available at http://arxiv.org/PS_cache/arxiv/pdf/0708/0708.0646v2.pdf
- Jacso, 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 April 3-5, 2006. available at <http://www2.hawaii.edu/~jacso/conferences/UKSG-GS-ppt-innumeracy-illiteracy.ppt>
- Jacso, P (2008a). “Google Scholar Revisited”. *Online Information Review* Vol 32. No. 1 pp. 102-114. available at <http://www.jacso.info/PDFs/jacso-GS-revisited-OIR-2008-32-1.pdf>
- Jacso, P (2008b). “The Plausibility of Computing the h-index of Scholarly Productivity and Impact Using Reference Enhanced Databases. *Online Information Review* Vol 32 No. 2, pp. 262-283 available at <http://www.jacso.info/PDFs/jacso-h-index-plausibility-OIR-2008-32-2.pdf>
- Meho, L. I., & Yang, K. (2007). “Fusion approach to citation-based quality assessment. In: 11th International Conference of the International Society for Scientometrics and Informetrics, Madrid, Spain, June 25-27, 2007. available at www.slis.indiana.edu/faculty/meho-fusion-approach.pdf
- Meho, L.I., Yang, K. (2007) “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 (JASIST)* Vol. 58. No. 13, pp 2105-2125; open access preprint available at <http://dlist.sir.arizona.edu/1733/>
- Neuhaus, C. et al (2006). “The Depth and Breadth of Google Scholar: An empirical study”, *portal: Libraries and the Academy*, Vol. 6 No. 2, pp. 127-141.
- Neuhaus, C., Neuhaus, E., & Asher, A. (2008). “Google Scholar Goes to School: The Presence of Google Scholar on College and University Web Sites”. *The Journal of Academic Librarianship*, 34(1), 39-51.

Notess, G. (2008). "Scholar Down, Books Up" available at http://www.searchengineshowdown.com/blog/2008/01/scholar_down_books_up.shtml

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. 161-169.

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, Mark (2008). "Revisiting h measured on UK LIS and IR academics". *Journal of the American Society for Information Science and Technology* (early view edition published in advance of the print edition on March 18, 2008; available at <http://dx.doi.org/10.1002/asi20771>

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 - 1550

Vanclay, J. (2008). "Ranking forestry journals using the h-index". Revised manuscript deposited March 17, 2008 at <http://arxiv.org/abs/0712.1916>

Wentz, R. (2004). "WoS versus Google Scholar: Cited by...: Correction"
<http://listserv.acsu.buffalo.edu/cgi-bin/wa?A2=ind0412B&L=medlib-l&P=R5842&I=-3&m=95812>