



Project
MUSE[®]
Scholarly journals online

Testing the Calculation of a Realistic h -index in Google Scholar, Scopus, and Web of Science for F. W. Lancaster

PETER JACSO

ABSTRACT

This paper focuses on the practical limitations in the content and software of the databases that are used to calculate the h -index for assessing the publishing productivity and impact of researchers. To celebrate F. W. Lancaster's biological age of seventy-five, and "scientific age" of forty-five, this paper discusses the related features of Google Scholar, Scopus, and Web of Science (WoS), and demonstrates in the latter how a much more realistic and fair h -index can be computed for F. W. Lancaster than the one produced automatically. Browsing and searching the cited reference index of the 1945–2007 edition of WoS, which in my estimate has over a hundred million "orphan references" that have no counterpart master records to be attached to, and "stray references" that cite papers which do have master records but cannot be identified by the matching algorithm because of errors of omission and commission in the references of the citing works, can bring up hundreds of additional cited references given to works of an accomplished author but are ignored in the automatic process of calculating the h -index. The partially manual process doubled the h -index value for F. W. Lancaster from 13 to 26, which is a much more realistic value for an information scientist and professor of his stature.

INTRODUCTION

The h -index was developed by Professor Jorge E. Hirsch of the Department of Physics at the University of San Diego. It was published in the prestigious *Proceedings of the National Academies of Science* (Hirsch, 2005) soon after its preprint appeared in arXiv, the excellent and widely used preprint repository focusing primarily on physics (<http://arxiv.org/pdf/>

LIBRARY TRENDS, Vol. 56, No. 4, Spring 2008 ("The Evaluation and Transformation of Information Systems: Essays Honoring the Legacy of F. W. Lancaster," edited by Lorraine J. Haricombe and Keith Russell), pp. 784–815

(c) 2008 The Board of Trustees, University of Illinois

physics/0508025). It was welcomed much more widely and quickly than any other bibliometric and scientometric indicators received before (Lancaster, 1991).

Hirsch summarized the essence in a terse abstract: "I propose the index h , defined as the number of papers with citation number $\geq h$, as a useful index to characterize the scientific output of a researcher." He then explains that "A scientist has index h if h is his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each." This means that an author with $h=16$ has 16 publications each of which received 16 or more citations. The h -index varies widely from discipline to discipline and even within disciplines and research areas. In library and information science, for example, a h -index of 16 is a high value, but in, say astronomy and retrovirology, it is considered to be a relatively low value.

SHORT LITERATURE OVERVIEW

Immediately after publication there was already a flurry of formal and informal comments and reactions by researchers from various disciplines with only a few dismissive and skeptical comments (Purvis, 2006; Ashkanasy, 2007; Berger, 2007), and plenty of supporting ones, in serious news sources, listserv fora and blog sites, beyond the many academic journals. It was cited by more than sixty papers by the end of August 2007. The most telling sign of the importance and appreciation of the h -index was that editors of *Scientometrics* found a way to squeeze in a paper about the h -index in its December 2005 issue (Bornmann and Daniel, 2005), then dedicated its April 2006 issue to the topic, with several substantial articles by some of the most respected scientometricians followed by three more in May, June, and July, then two more in 2007 in that journal alone. The papers approached the topic from a variety of theoretical (Egghe and Rousseau, 2006; Liang, 2006; Egghe, 2006, 2007a; Schubert, 2007; Glänzel, 2006; and practical angles (Costas & Bordons, 2007; Imperial & Rodriguez-Navarro, 2007; Vanclay, 2007).

There are several case studies that present the h -index for a variety of target groups. These include the prominent scholars, educators, and researchers in a specific field (Kelly & Jennions, 2006; Saad, 2006; Cronin & Meho, 2006 Oppenheim, 2007), lesser known researchers in the broad field of physics (Schreiber, 2007a), institutions within a country (Prathap, 2006), researchers of a discipline within a country (Salgado and Páez, 2007), researchers within a country in different fields (Imperial & Rodriguez-Navarro, 2007; Packer & Meneghini, 2006, Meneghini & Packer, 2006), across countries in a field of specialization (Oelrich, Peters, and Jung 2007), and in the highly select group of scientometrics, the winners of the award commemorating John Derek de Solla Price (Bar-Ilan, 2006a).

Some of the best papers about the *h*-index voiced reservations about the details of the proposed model, but they indicated their support of the theory of Hirsch by suggesting variant and derivative indexes built on the idea of Hirsch (Batista, et al, 2006; Egghe, 2006, Vanclay, 2006; Barendse, 2007; Jin et al, 2007). Several papers compared the *h*-index with other, traditional measures (van Raan, 2006; Barendse, 2007; Costas & Bordons, 2007).

The *h*-index was begging to be applied to journals, to complement the controversial Journal Impact Factor, and several papers confirmed and applied this extension (although not for a lifetime measure given the volume of papers in many journals) (Braun, Glänzel, & Schubert, 2006; Schubert & Glänzel, 2007; Olden, 2007).

Although it is not about the *h*-index, an excellent article by Butler and Visser (2006) about the need for extending citation analysis to nonsource materials (i.e., to material types, document genres, specific journals not covered by a database) is essential for understanding the context of the *h*-index. I will come back to their well-designed, nationwide research later, as their conclusions are likely to apply not only to researchers in Australia but around the world. The warnings of Cronin, Snyder, and Atkins (1997), and earlier by Line (1979), should be heeded by everyone who evaluates the research performance of scholars in the sciences, the social sciences and especially in arts and humanities for the preference of non-journal sources in the research area.

Two aspects of the concept of Hirsch received special interest: the bias of the *h*-index for extensively self-citing authors (Schreiber, 2007b; Vinkler, 2007), and its robustness and relative insensitivity to missing records for highly cited papers (Vanclay, 2007; Rousseau, 2007). There are several other relevant papers cited in the sections on Google Scholar, Scopus, and Web of Science, which provide a broader background for these three systems, beyond the perspective of the *h*-index itself, often comparing the alternatives. These latter two papers offer a good transition to the focus of my research for this Festschrift, which illustrates through the example of Wilf Lancaster. One has to be careful with searching by author name(s). In WoS, the last name must come first, followed by the first and—if applicable—middle initial(s) and no punctuation at all. In Scopus, the order does not matter but initial(s) must be followed by a dot. If the last name is entered first, it must be followed by a comma. The template suggests that the comma must be followed by a space, but it is actually not needed, and the software removes it if it is entered, echoing back an odd-looking format, which is different from what the user entered, and from the way it looks in the record. In Google Scholar, punctuation is ignored so “FW Lancaster” and “F.W. Lancaster” bring up the same number of hits. On the other hand, you should put the name in between quotes (which had no effect until about mid-1996, but it is important because otherwise Google Scholar picks up records for articles authored by, say “M Lancaster,” “F

Smith,” and “W Black” because its software does not handle repeatable fields, such as authors, appropriately). Purely software-generated *h*-indexes, which ignore the “orphan references” are actually sensitive to even just a few missing records for publications, which are highly cited, but are ignored in the process of automatically generating the *h*-index.

OUTLINE

First, the features of Google Scholar and Scopus are discussed from the *h*-index perspective, followed by a more detailed analysis of the pros and cons of Web of Science (WoS) from the perspective of generating the *h*-index in general, and for F. W. Lancaster in particular. These three systems have the broadest disciplinary coverage among the databases, which are fully or partially enhanced by well-tagged cited references, which is one of the pre-requisites for counting and keeping track of the citations given and received (Jacso, 2008b). It is another question as to why the developers of Google Scholar apparently did not make use of any of the metadata which are available in tens of million records, which the developers had access to (Jacso, 2008a).

Given the space limitation of this Festschrift, I can provide only limited coverage of the issues, but a multipart series about the content and software advantages and disadvantages of using Google Scholar, Scopus, and WoS and for calculating a rational *h*-index (Jacso, 2008c, 2008d, 2008e) are to be published in 2008.

Here it is demonstrated how the existing features of WoS can be exploited to arrive at a much more credible, and traceable *h*-index, which is at least twice as high as the automatically generated *h*-index in WoS, and 4–8 times higher than the *h*-index produced by Scopus. The ratio depends on which of the two automatic *h*-index generation options of Scopus is chosen by the user (as discussed later). I am reluctant to provide any comparative score with Google Scholar, simply because its hit counts and citation counts remain as untraceable and inflated (Jacso, 2006a, 2006b) as they were at the launch of the beta version in 2004. All three systems have limitations (and so do all the citation-enhanced indexing and abstracting databases (Jacso, 2004b), but the deficiencies in Google Scholar are so voluminous, unscholarly, and often so hidden that its hit counts and citation counts should not be accepted even as a starting point for evaluating the research output of real scholars. In an interesting twist, Google Scholar can help in revealing the shortcomings of WoS and Scopus by showing information about publications that are not covered by the automatic *h*-index generators of either.

THE OEUVRE OF LANCASTER

It is not for sheer snobbery that I use the French loan word. As a single word it refers to “a substantial body of work constituting the lifework of

a writer” according to the Merriam-Webster Collegiate Dictionary (and many other good dictionaries). It is important—especially in a Festschrift (which is a borrowed German term and a general material designation in AACR2)—because Hirsch meant the *h*-index to estimate the lifetime cumulative impact of a researcher, not just the combination of his or her productivity and citedness in journals and other publications in the past decade.

For the examples used here, it is essential to know that F. W. Lancaster published in English fifteen books (in twenty editions), about thirty-five other monographic works, including technical reports, forty-five chapters in edited books and conference proceedings, close to forty book reviews and two hundred articles in periodicals. Lancaster also edited about a dozen books.

Several of his works were translated into Portuguese, Spanish, French, Spanish, Chinese, Japanese, and Arabic. These statistics are important for gauging the variety of document genres and publication types, especially because WoS and Scopus cover as source publications almost exclusively only his publications in journals and in a few other periodical publications, such as the *Annual Review of Information Science and Technology*.

The research for this Festschrift showed that in WoS alone there are more than 2,000 references to his nonjournal publications, and to inconsistently or incorrectly cited/recorded journal publications, in addition to the 650 references that were automatically matched with and attributed to the 131 master records for his publications by WoS. The rest of the references became orphan because there were no master records at all for certain publications, or ones that matched exactly, letter by letter, the content of the cited references, either because of errors by citing authors (like myself) or by data entry operators making errors of omission or commission (Jacso, 1997).

As will be demonstrated later, many—but not all—of these extra two thousand references are of crucial importance. They are important, not for bibliographic control or obsession, but for determining his real *h*-index even if currently this cannot be done using the automatic *h*-index generator functions of WoS (or Scopus).

Lancaster’s career age or scientific age (to use Hirsch’s term) is forty-five years in my estimation, as his first cited article was published in *The New Scientist* in 1964.

GOOGLE SCHOLAR

Google Scholar is an excellent tool for finding information about documents that may not be represented in traditional bibliographic databases. It is even more valuable in leading users to open access versions of primary documents, but using it for bibliometric or scientometric purposes, such as for determining the *h*-index of a person or a journal, is another question.

Google Scholar may have spoiled the users by virtue of being fast and free, but it does so by playing fast and loose with hit counts and citation counts. I included in this paper several screenshot illustrations for the major problems not only for facilitating the understanding of serious problems, which may have a wide-ranging affect, but also because the master records may be deleted as has been the case after my articles or PowerPoint presentations discussing the deficiencies were published or were posted on the Web. This is illustrated in a screenshot gallery about Google Scholar, the mis-matchmaker at <http://www2.hawaii.edu/~jacso/extra/> inspired by the disappearance of a record for an important chapter in the *Annual Review of Medicine*. For fairness, on some occasions the software errors in Google Scholar producing voluminous sets of false hit counts and citation counts were later fixed—which is the appropriate reaction to the criticism.

The No-Brainers

The extent and volume of inflated hit counts and citation counts cannot be fully determined, nor can they be traced and corroborated systematically, but even my chance encounters with absurd hit and citation counts, followed up by test searches for obviously implausible and often clearly nonsense hit counts and citation counts indicate the severity of the problems. For example, it is discouraging to see that Google Scholar still cannot handle even elementary search operations correctly that are not a problem for the mainstream Google engine.

The intentionally very broad search term “Lancaster” returns 442,000 hits, and then the query “Lancaster OR Lancaster” meant to broaden the search with a possible misspelled variant yields almost 100,000 fewer “hits” than the logic of the Boolean operation would dictate (see Figure 1). It is not the nonsense variety of the usually helpful “Did You Mean” recommendation, neither the absolute number that is bothersome here but the meltdown of the Boolean logic, the tenet of search operations and commonsense. I have not seen any professional information service that would behave in such a senseless way. I find it interesting that the Boolean “OR” operator sometimes works correctly both on small and very large sets. This reminds the user that Google Scholar rarely returns the same hit counts for the very same query even a few minutes later—not a good sign for scholarly research.

Google Scholar keeps playing fast and loose with the hit counts also when the search is to be limited to year ranges. One does not need even a high school diploma to realize the oddity that there are more results for the previous search for the shorter time span than for the longer one. (See Figure 2.)

Google Scholar does not have the essential feature of sorting the results by decreasing citedness count (or any other user-selectable sort

ETA [Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

lancaster Search [Advanced Scholar Search](#) [Scholar Preferences](#) [Scholar Help](#)

Articles - Recent articles Results 1 - 10 of about 442,000 for [lancaster](#) [[definition](#)].

[\[book\] The econometric analysis of transition data - all 2 versions »](#)
[T Lancaster - 1990 - cambridge.org](#)

ETA [Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

lancaster OR lancaster Search [Advanced Scholar Search](#) [Scholar Preferences](#) [Scholar Help](#)

Articles - Recent articles Results 1 - 10 of about 348,000 for [lancaster OR lancaster](#).

Did you mean: lancaster OR lancaster

[\[book\] The econometric analysis of transition data - all 2 versions »](#)
[T Lancaster - 1990 - cambridge.org](#)

Figure 1. Wrong Boolean OR operation in Google Scholar

[http://scholar.google.com/scholar?hl=en&lr=&q=lancaster&as_ylo=1907&as_yhi=2007](#) ? ▾

[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

lancaster 1907 - 2007 Search [Advanced Scholar Search](#) [Scholar Preferences](#) [Scholar Help](#)

Articles - Recent articles Results 1 - 10 of about 116,000 for [lancaster](#)

[http://scholar.google.com/scholar?hl=en&lr=&q=lancaster&as_ylo=1917&as_yhi=2007](#) ? ▾

[Web](#) [Images](#) [Video](#) [News](#) [Maps](#) [more »](#)

lancaster 1917 - 2007 Search [Advanced Scholar Search](#) [Scholar Preferences](#) [Scholar Help](#)

Articles - Recent articles Results 1 - 10 of about 118,000 for [lancaster](#)

Figure 2. More hits in Google Scholar for the shorter time span than for the longer one

elements). It used to have this feature in its early days, but it was abandoned or unpredictable. Third party software programs can help in this regard, but they may hide the warning signs obvious from the raw hit list, and those programs may disappear or become dysfunctional when a change is made in the output format of Google Scholar, or when Google blocks the service as it happened with the first *h*-index generator (<http://www.brics.dk/~mis/hnumber.html>).

Phantom Links

Many other essential features of Google Scholar are confusing. Looking up a “master record” may lead the user on a wild goose chase. This is the case with the master record for Lancaster’s most cited book (which is listed as the tenth item for the search of “FW Lancaster” as an author). It takes the user to a conference paper by a group of Cuban researchers related to the economic feasibility of a drainage project in Venezuela. Its relationship to Lancaster’s book is not evident, and no further details about a possible connection are available unless the document is purchased. (See Figure 3.)

Phantom Master Records

The result list often has items and citation counts which are grossly misleading. For example, the second item in Figure 4 suggests that Lancaster wrote in 2005 an article in the *Journal of the Medical Library Association*

[Information retrieval systems; characteristics, testing and evaluation.](#)

FW Lancaster - New York - [bases.bireme.br](#)

Base de dados : REPIDISCA. Pesquisa : 55503 [Identificador único]. Referências encontradas : 1 [refinar]. Mostrando: 1 .. 1 no formato [Detalhado]. página 1 de 1, 1 / 1, REPIDISCA, seleciona. para imprimir. Id: 55503. ...

[Cited by 203](#) - [Related Articles](#) - [Cached](#) - [Web Search](#) - [Library Search](#)

REPIDISCA
Id: 55503
Autor: Guillama Rodríguez, José Luis; Alfonso Fleites, Fausto Eduardo; Sampedo Delgado, Germán; Artiles, Rafael.
Título: Estudio de factibilidad económica : drenaje pluvial de la ciudad de Coro, Estado de Falcón, República de Venezuela
Fonte: In: Asociación Cubana de Ingeniería Sanitaria y Ambiental. Memoria : Il congreso AIDIS de Norteamérica y el Caribe : IV congreso de la Asociación Cubana de Ingeniería Sanitaria y Ambiental. Santiago, AIDIS, 1995. p.597-605, Tab.
Idioma: Es; Es.
Conferência: Apresentado em: Congreso AIDIS de Norteamérica y El Caribe, 2 Congreso de la Asociación Cubana de Ingeniería Sanitaria y Ambiental, 4, Santiago de Cuba, 5-9 jun. 1995.

Figure 3. Google Scholar’s “master record” for one of the most cited books of Lancaster takes the user to this unrelated record for further detail

All articles - [Recent articles](#) Results 1 - 10 of about 365 for author:"fw lancaster". (0.04)

[Vocabulary Control For Information Retrieval](#) - all 3 versions »
FW Lancaster - 1986 - [eric.ed.gov](#)
 ED075999 - Vocabulary Control for Information Retrieval.
[Cited by 136](#) - [Related Articles](#) - [Cached](#) - [Web Search](#) - [Library Search](#)

[Indexing and Abstracting in Theory and Practice](#) - all 3 versions »
FW Lancaster - J Med Libr Assoc, 2005 - [pubmedcentral.nih.gov](#)
 This third edition of what has become a classic among textbooks in schools of library and information science (and related programs) has been thoroughly updated to reflect the evolving technological advancements in the field. ...
[Cited by 125](#) - [Related Articles](#) - [Web Search](#) - [Library Search](#) - [BL Direct](#)

[book] [Toward Paperless Information Systems](#) - all 2 versions »
FW Lancaster - 1978 - Academic Press, Inc. Orlando, FL, USA
 Google, Inc. Subscribe (Full Service), Register (Limited Service, Free), Login.
 Search: The ACM Digital Library The Guide. Feedback Report a problem Satisfaction survey. Toward Paperless Information Systems. Purchase this Book Purchase this ...
[Cited by 83](#) - [Related Articles](#) - [Web Search](#) - [Library Search](#)

[book] [Information Retrieval Today](#)
FW Lancaster, AJ Warner - 1993 - Information Resources Press Arlington, VA, USA
[Cited by 64](#) - [Related Articles](#) - [Web Search](#) - [Library Search](#)

[book] [The measurement and evaluation of library services](#)
 SL Baker, **FW Lancaster** - 1991 - Information Resources Press
[Cited by 47](#) - [Related Articles](#) - [Web Search](#) - [Library Search](#)

Figure 4. Confusing result list in Google Scholar for the query “FW Lancaster” as author

about indexing and abstracting, one of his major topics which was already cited 125 times, making it the most cited item of *JMLA*.

Actually, it is a review of the third edition of one of Lancaster’s classic books in 2003. The review itself was not cited, book reviews rarely are, which would be a consolation for the real author of the review, Virginia Lingle, who is not recognized, let alone acknowledged by Google as the author of the review.

Phantom Authors

There are many instances when the mix-up in Google Scholar is more enigmatic. In most of the cases when the name of the real author is replaced by parts of the digital text that Google Scholar fancies to be the authors, it is bad news for any genuine authors.

When this happens with authors straight in a row with a number of their publications, their *h*-index will not benefit from their well-cited publications. The group of authors in the following example, who have published their findings in prominent journals and received a substantial number of citations, will not appreciate that Google Scholar has removed their names and replaced them with those of other researchers, such as “I AntiCancer,” “C San Diego,” “S Clinic,” and “C La Jolla.” (See Figure 5.)

[Genistein Inhibits the Growth of Human-Patient BPH and Prostate Cancer](#)
 | [AntiCancer](#), C San Diego, S Clinic, C La Jolla - [The Prostate](#), 1998 - [doi.wiley.com](#)
 ... 1 [AntiCancer](#), Inc., San Diego, California 2 Department of Surgery, University of
 California School of Medicine, San Diego, California 3 Mercy Hospital and ...
[Cited by 64](#) - [Related Articles](#) - [Web Search](#)

Original Article

Genistein inhibits the growth of human-patient BPH and prostate car

Jack Geller ^{1 2 *}, Lida Sionit ², Christine Partido ², Lingna Li ¹, Xiuying Tan ¹, Tyler Youngkin ²,
 Robert M. Hoffman ^{1 4}

¹AntiCancer, Inc., San Diego, California

²Mercy Hospital and Medical Center, San Diego, California

³Scripps Clinic, La Jolla, California

⁴Department of Surgery, University of California School of Medicine, San Diego, California

* Correspondence to Jack Geller, AntiCancer, Inc., 7917 Ostrow Street, San Diego, CA 92111

Figure 5. Phantom authors in a Google Scholar example, and real authors appearing in the source documents but deprived of their authorship by Google Scholar

Admittedly, the efforts taken by Google Scholar to create last names and first initials from the affiliations data element (such as “C San Diego” from the Mercy Hospital and Medical Center in San Diego) is remarkable. Funding agencies where administrators may run Google Scholar searches for free to verify the productivity and citedness clout of these researchers will not find those papers where the authors’ names were replaced. In this case, at least six of their articles are not accessible through their names.

The author names are not replaced in all the records when such phantom names are added, so this syndrome may affect only a few million authors. However, it is a powerful action because a single phantom name can wipe out several real authors in an author group. This happens when Google Scholar designates “V. Cart” to be the author for more than 85,000 articles. Some of these may have been written by Victor Cart or Veronica Cart, but eyeballing the result list and the source page of the articles clearly indicates that most “V. Cart” entries as author names were created by the software fancying the View Cart menu option to be an author name for unknown reasons. The View Shopping Cart menu option also qualifies even when the search term is enclosed between double quotes. Unfortunately, “V. Cart” alone deprives all the real authors, who are made invisible and unassociated with the articles by the action of the software.

Phantom Citations

Then comes the problem of phantom citations. In cases when the searcher has access to the full documents, it is often found that the purportedly citing papers do not have matching references. This further undermines the “scholarly” status of Google Scholar when it comes to counting citations

received, and using them to rank scholars. There is no possibility to trace this master record itself for one of Lancaster's books, but subscribers to the *Journal of Documentation* may access the article purportedly citing the document by "F Wilfrid" and "F Wilfrid" (as understood by the software of Google Scholar). (See Figure 6.) The lucky ones will find that Maria Pinto does indeed cite Lancaster's other book and the article (not shown here) that she coauthored for *Library Trends* with Lancaster, but not the book claimed by Google Scholar. The reason for this is the utterly loose citation matching algorithm of Google Scholar.

While it definitely could be an advantage to learn the citedness of some of Lancaster's books through Google Scholar, the software is not ready to handle correct book records, many of which were apparently borrowed from Google Books. It adds to the confusion that the query template still uses the label "Return articles written by" even if there are many non-article type items in Google Scholar.

Very often there are nonclickable entries in the results list, such as ones that were extracted very poorly from journals and are marked with the (citation) tag, and most of the records for books (also the results of extraction, except for the ones linking to the very rewarding <http://books.google.com> site, which does not show the shoddiness of Google Scholar). These nonclickable entries prevent traceability, which is quite an essential function, especially given Google Scholar's many serious deficiencies.

I am less concerned about the duplicate, triplicate, and quadruplicate records that dilute the result lists, and increase the hit counts, although there were more than eighty such records in the test results for the search on "FW Lancaster." These are variants of other references and, as I suggest later, these should be normalized by human intervention (assisted by an intelligent software), rather than just scattering them around and diluting the result list.

Phantom Publication Years

It is bothersome, too, that many records Google Scholar produces are purportedly about papers to be published in the next year or even in the next decade, and these records show how many hundred or even thousand times the future articles have already been cited. Actually, these are not publication years but page numbers of the articles, the number of patients in the survey described, or a variety of other four digit numbers that Google mistook for publication years, and still makes even two-digit volume and issue numbers four-digits long to pass as publication years (Jacso, 2008a).

The ones for the year 2009 and after are easy to spot; my concern is for how many million records Google Scholar creates fancy false publication years that further debilitate its mentally handicapped matching algorithm. Luckily, Google Scholar does not generate (yet) an *h*-index, but

[BOOK] **If you want to evaluate your library--**

F Wilfrid, F Wilfrid - University of Illinois, Graduate School of Library and
[Cited by 1](#) - [Related Articles](#) - [Web Search](#)

Lalmas, M. and Ruthven, I. (1998), "Representing and retrieving structured documents using the Dempster-Shafer theory of evidence: modelling and evaluation", *Journal of Documentation*, Vol. 54 No. 5, pp. 529-65.

Abstracting in
digital
environments

Lancaster, F.W. (1998), *Indexing and Abstracting in Theory and Practice*, 2nd ed., Graduate School Library and Information Science, Urbana-Champaign, IL.

Liddy, E. (1991), "The discourse-level structure of empirical abstracts: an exploratory study", *Information Processing & Management*, Vol. 27 No. 1, pp. 55-81.

607

Luhn, H.P. (1958), "The automatic creation of literature abstract", *IBM Journal of Research and Development*, Vol. 2, pp. 159-65.

Figure 6. The purportedly citing reference in Google

there are third party utilities that do, along with other measures that take into consideration the publication year. It is not always the fault of these programs, but once again we encounter the Garbage In, Garbage Out principle. Plausibility tests to be built in the *h*-index generating programs could certainly help in getting a more rational *h*-index. I could list here the *h*-index of F.W. Lancaster as reported by Google Scholar, but it would be irresponsible, as every number reported by Google Scholar needs an IRS audit before accepting it.

The variety of items in Google Scholar is great, although simple catalog entries and items in course reading lists perhaps should not be counted as citations from academic journals, books, and conference proceedings. In addition, Google Scholar also has much wider coverage of scholarly foreign language materials than either Scopus or WoS. However, its pathetic software has a long way to go to make use, at a scholarly level, of the unprecedented access that hundreds of scholarly publishers offered to Google, Inc. exclusively—to at least tens of million articles, conference papers, and books in their digital collections.

There is mass adulation in the media, and even in academia, for Google Scholar. Fortunately, academic and research librarians did not start to cancel their traditional databases when Google Scholar was launched according to the survey by Mullen and Hartman (2006). It is equally fortunate, that there are many competent critics from the library and information science and technology discipline who publish about the issue. Most of them give a usually well-balanced view of the pros and cons of this service in the most respected and/or most widely read academic and professional LIS journals to keep the information professionals informed. The most useful papers focusing on Google Scholar include Mayr & Walter, 2007; Kousha & Thelwall, 2007; Robinson & Wusteman, 2007; Pomerantz, 2006; White, 2006; Callicott & Vaughn, 2005; Neuhaus et al., 2006; and Mullen

& Hartman, 2006. Particularly useful are those articles that put the emphasis on comparing Google Scholar with Web of Science and/or Scopus, such as Meho & Yang, 2007; Schroeder, 2007; Norris & Oppenheim, 2007. Bar Ilan's (2006b) paper within the domain of computer science extends the comparison of Google Scholar to CiteSeer (which in my opinion has a far better citation matching algorithm than Google Scholar and could serve as a model for open access citation-based search system). The paper of Bar-Ilan, Levene, and Lin (2007) in the premier issue of the *Journal of Informetrics* outlines the measures that could be used for comparing the degree of similarity of the ranking of results retrieved from Web of Science, Scopus, and Google Scholar and sets the scene for the next step in database evaluation: the comparative-competitive citation analysis.

The authors of these papers often have a much higher opinion of the scholarly capabilities of Google Scholar than I have, but they have convincing evidence within the sphere of their surveys and comparison, and thus provide a healthy dose of rational criticism of my stance.

SCOPUS

Scopus was born in 2004 with the best software module for presenting result lists. When the *h*-index was introduced, users could very efficiently scroll down in chunks of maximum two hundred items per page in the result list in Scopus, sorted by decreasing order of citation counts in order to eyeball the *h*-index value, the point where the number of citations received by a publication is equal to or larger than its rank order number. WoS beat Scopus in coming up with an automatic *h*-index generator in late 2006, but by mid-2007 Scopus came out with two automated *h*-index generation options. One is good, but the other one is very unfair to accomplished scholars, because it excludes from the *h*-index calculation any paper published before 1996—even when they have been cited extensively since 1996.

It is the tail wagging the dog syndrome, and the making of virtue out of necessity principle (in this case a false, backfiring virtue) that is unfair to researchers whose writings before 1996 are totally ignored even when they were cited after 1995 and remain cited.

The necessity comes from the fact that in Scopus, only records for works published after 1995 are enhanced by cited references. There are about fifteen million such records in Scopus (Jacso, 2008d), plus another set of about seven thousand records for pre-1996 publications. They provide excellent links to the cited references, which often splendidly overcome the limitations of controlled vocabulary searches. Their implementation in Scopus is technically superb, elegant, and effective (Jacso, 2008d).

It is another question that from the point of view of citation analysis in general, and the *h*-index in particular, the composition of the database creates a situation where the *h*-index of Scopus disses senior researchers who are beyond their teens measured in scientific age. I do not hold the

punch line, but let the reader know right here that the *h*-index value of “FW Lancaster” is three by the good-looking but ill-conceived automatic alternative, and six by the better one, which produces the same result, as the manual look-up process, but with more details.

The decibel of dissing grows hand-in-hand with their career age. It is a bad sign that the search for “FW Lancaster” as author yields only twenty-six hits, one-fifth of what WoS has master records for. From the perspective of the *h*-index the picture becomes somewhat better when book reviews (which are rarely cited) are excluded from the hit counts, well covered by WoS, but to a lesser extent by Scopus.

However, it is a worse sign that even the higher *h*-index value of Lancaster in Scopus is less than half of what WoS reports through its automatic *h*-index generation feature, which is not good enough in the LIS discipline for a person of Lancaster’s accomplishments and eminence.

The reason for the low hit number (see Figure 7) is that Scopus does not have comprehensive coverage of prominent LIS journals in which Lancaster primarily published, even if Scopus goes back to the mid-1960s with bibliographic data. The low number of master records for Lancaster is a problem in itself, especially when there are no master records, which would be pegs to allow the software to hang its hat of cited references on for well over a hundred of Lancaster’s articles, many of which have been cited much more than six times, nor for his books, many of which have been cited more than a hundred times. (WoS does not have master records for Lancaster’s books, either.)

Even when there are master records to attribute citations to, the Scopus citation counts are significantly lower than the citation counts in WoS, because twenty-two of the twenty-six records in Scopus are for articles published before 1996, and articles get most of their citations in their second and third years. (See Figure 8.)

Earlier publications do get cited after 1996, of course, but there is another twist on top of the above with these publications. One of the two options of Scopus for automatic *h*-index generation and display is the *h*-index on the Author Details page. (See Figure 9.) It further reduces Lancaster’s *h*-index to three because of the ill-conceived idea and policy to ignore in the *h*-index calculation all the publications of the authors published before 1996, even if many of them may have been cited by the authors from 1996 onward. There are three Author Details pages for “FW Lancaster,” because if there is any difference in the subject categories assigned to the author, or in the author affiliation, Scopus creates a new Author Details page. Out of these four other articles, only one was published since 1996, and it was not cited (according to Scopus), so it has no influence on this ultra-low *h*-index.

This policy does not affect those whose career age is less than twelve years, and they are best taken care of by Scopus because for the 1996–

Your query: AUTHOR-NAME(lancaster,f w) [Edit](#) [Save](#) [Save as Alert](#) [RSS](#) [Search History](#)

Refine Results [Open](#)

Results: 26 Search within results [Go](#)

[Print](#) [Output](#) [Citation tracker](#) [Add to list](#) Select: All Page 1 to 20 [Next](#)

	Document (sort by relevance)	Author(s)	Date	Source Title	Cited By
1.	<input type="checkbox"/> Types and levels of collaboration in interdisciplinary research in the sciences Abstract + Refs View at Publisher Show Abstract	Qin, J., Lancaster, F.W., Allen, B.	1997	<i>Journal of the American Society for Information Science</i> 48 (10), pp. 893-916	27
2.	<input type="checkbox"/> BIBLIOMETRIC TECHNIQUES APPLIED TO ISSUES MANAGEMENT: A CASE STUDY. Abstract + Refs View at Publisher Show Abstract	Lancaster, F.W., Lee, Ja-Lih	1985	<i>Journal of the American Society for Information Science</i> 36 (6), pp. 389-397	15
3.	<input type="checkbox"/> Acquired immunodeficiency syndrome (AIDS) and the epidemic growth of its literature Abstract + Refs View at Publisher Show Abstract	Self, Ph.C., Filardo, Th.W., Lancaster, F.W.	1989	<i>Scientometrics</i> 17 (1-2), pp. 49-60	12
4.	<input type="checkbox"/> Persuasive communities: a longitudinal analysis of references in the Philosophical Transactions of the Royal Society, 1665-1990. Abstract + Refs View at Publisher	Allen, B., Qin, J., Lancaster, F.W.	1994	<i>Social Studies of Science</i> 24 (2), pp. 279-310	11
5.	<input type="checkbox"/> CORRELATION BETWEEN PERTINENCE AND RATE OF CITATION DUPLICATION IN MULTIDATABASE SEARCHES. Abstract + Refs View at Publisher Show Abstract	Neway, Julie M., Lancaster, F.W.	1983	<i>Journal of the American Society for Information Science</i> 34 (4), pp. 292-293	8
6.	<input type="checkbox"/> Abstracts and abstracting in knowledge discovery Abstract + Refs Show Abstract	Pinto, M., Lancaster, F.W.	1999	<i>Library Trends</i> 48 (1), pp. 234-248	7
7.	<input type="checkbox"/> Evaluation of interactive knowledge-based systems: Overview and design for empirical	Lancaster, F.W., Ullvila, I.W., Humnhrev, S.M.	1996	<i>Journal of the American Society for Information</i>	4

Figure 7. Implausibly low hit counts and *h*-index value for FW Lancaster in Scopus

2007 time period it has almost the same number of records enhanced by cited references, about 12.2 million records, that WoS has for 1996–2007. In WoS there is no search option to determine the number of records enhanced by cited references. However, in the Dialog version of the three ISI citation databases, the simple search for such records “SELECT PY=1996:2007 AND NR > 0” provides this information. As of mid-October, 2007, when the test was done, the number of such records was 12.4 million. My earlier tests of Dialog have indicated that about 6–7 percent of records appear in two or three of the separate citation indexes because some journals and their articles fit the Science, Social Science and/or the Arts & Humanities categories, so the grand total of the above hits from the three databases must be reduced. Hence, the net number of records enhanced by cited references was just a notch below 12 million in the Dialog version of the three citation databases together. In WoS such duplicates are automatically removed. Scopus allows searching (indirectly) by the words in the title of about 44 million orphaned or stray cited references. This is in addition to the titles of 11.9 million master records that have one or more cited reference(s)—to documents not covered by Scopus or that did not match the master records. For example, the search for the term *h*-index in the title provided 30 additional hits, citing more than 20 papers or notes in journals, bulletins, and newsletters, which indeed are not covered by Scopus.

Year	Document title	WoS	Scopus
		227	97
1997	Types and levels of collaboration in interdisciplinary research in the sc	25	27
1985	BIBLIOMETRIC TECHNIQUES APPLIED TO ISSUES MANAGEMEN	27	15
1989	ACQUIRED IMMUNODEFICIENCY SYNDROME (AIDS) AND THE E	16	12
1994	PERSUASIVE COMMUNITIES - A LONGITUDINAL ANALYSIS OF R	14	11
1983	THE CORRELATION BETWEEN PERTINENCE AND RATE OF CITA	8	8
1999	Abstracts and abstracting in knowledge discovery	6	7
1996	Evaluation of interactive knowledge-based systems: Overview and de	5	4
1986	FACTORS INFLUENCING SOURCES CITED BY SCIENTISTS - A C	8	3
1972	EVALUATING EFFECTIVENESS OF AN ON-LINE, NATURAL LANG	30	2
1992	USE OF LITERATURE BY EAST EUROPEAN SCIENTISTS - WHAT	5	2
1969	EVALUATING PERFORMANCE OF A LARGE COMPUTERIZED INF	13	1
1971	EVALUATION OF PUBLISHED INDEXES AND ABSTRACT JOURNA	12	1
1986	QUALITATIVE ASPECTS OF THE BRADFORD DISTRIBUTION	5	1
1990	DOES PLACE OF PUBLICATION INFLUENCE CITATION BEHAVIO	9	1
1991	THE CONTRIBUTION OF SCIENTISTS TO THE POPULAR LITERA	4	1
1998	Redundancy and uniqueness of subject access points in online catalo	0	1
1965	A CASE STUDY IN APPLICATION OF CRANFIELD SYSTEM EVALL	4	0
1966	EVALUATING SMALL INFORMATION RETRIEVAL SYSTEM	6	0
1972	CRITICAL EVALUATION OF A COMPUTER-BASED MEDICAL LITE	2	0
1972	SELECTIVE DISSEMINATION	3	0
1978	ASSESSING BENEFITS AND PROMISE OF AN INTERNATIONAL IN	6	0
1980	ONLINE SYSTEMS IN THE COMMUNICATION PROCESS - PROJE	7	0
1982	THE FUTURE OF INDEXING AND ABSTRACTING SERVICES	11	0
1987	COMPARING THE SCATTER OF CITING AND CITED LITERATURE	1	0
1996	Article quality	0	0

Figure 8. Comparison of WoS and Scopus cited reference counts for the 26 master records in Scopus for FW Lancaster

WEB OF SCIENCE




Thomson ISI was the first to integrate a software tool to automate the calculation of the *h*-index (Jacso, 2007a). It was part of the new Citation Reports feature (not to be confused with the Journal Citation Reports), which not only makes more convenient the calculation of the *h*-index of

Lancaster, F. W.

[Find unmatched authors](#)  Feedback
Personal

Name	Lancaster, F. W.	
Author ID	7005143503	
Affiliation	University of Illinois, Grad. Sch. of Lib. and Info. Science	Champaign-Urb.

Research

Documents	22	 Add to list	
Cited By	81	 Citation tracker	
<i>h</i> Index	3	 <i>h</i> -graph	<small>The <i>h</i> Index considers Scopus articles published after 1995.</small>
Co-authors	28		
Web Search	3		
Subject Area	Computer Science Medicine Social Sciences More...		

 [Find unmatched authors](#)
History


Publication range	1965-1999	
Source history	Library Resources and Technical Services Journal of the American Society for Information Science Science More...	 documents  documents  documents

Figure 9. Author Details page in Scopus for F.W. Lancaster

authors, but also provides additional bibliometric data and citation analysis options for the result set from which the *h*-index was computed. This was a timely and very well-designed move, including the compact visual representation of the yearly distribution of publications and citations of the author. (See Figure 10.)

It must be mentioned that the same informative bibliometric chart and index data are generated for any type of query by topical terms, journal name, author affiliations at the country, and/or the institutional level—as long as the set is not larger than 10,000 records. As pointed out earlier (Jacso, 2007a), the limit should be increased well above 10,000 items for the simple reason that these types of queries may exceed that limit. For author searches the limit is not a problem because WoS does not create phantom authors with 85,000 papers as Google Scholar does for “V Cart.” The limit is a problem when the *h*-index of journals or institutions is to be computed, as the number of hits may be well over 10,000. Although the queries can be restricted to shorter time frames in consecutive rounds to

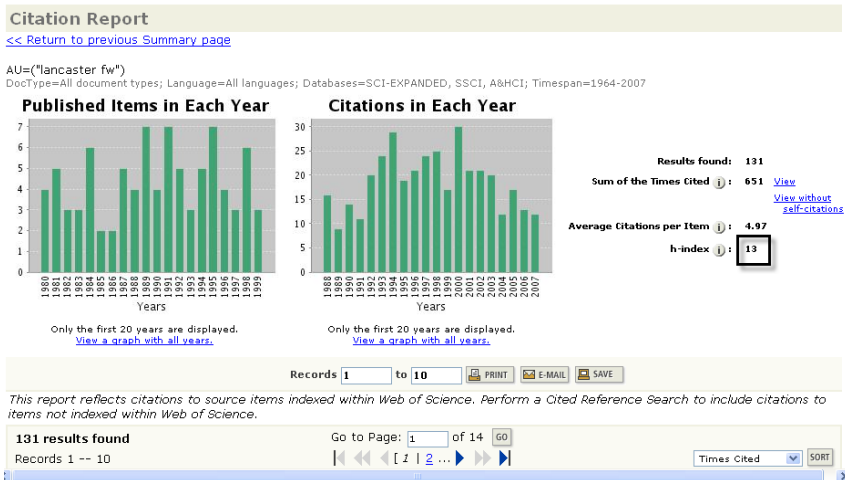


Figure 10. Profile of the result set for the simplest query about FW Lancaster as author in Web of Science

make the sets smaller than 10,000 items, it makes the process staggered and more cumbersome. This would certainly make the wait time to generate the chart and the related indicators longer, and it still would be a very reasonable response time in exchange for the results.

Database Size

Hirsch mentions in a footnote of his paper that “of course the database used must be complete enough to cover the full period spanned by the individual’s publications.” (The footnote appears with a dagger on p. 16569.) This is a crucial issue as in Hirsch’s closing statement: The index “gives an estimate of the importance, significance, and broad impact of a scientist’s *cumulative* (emphasis added) research contributions” (p. 16572). And here comes the rub.

I not only agree with but have been advocating (Jacso, 1997) the importance of the completeness of the database and its retrospective coverage (along with other essential content criteria). But there is an additional criterion. What matters really is not merely the completeness of retrospective coverage of the databases used for generating the h -index (or at least a result list of bibliographic references in decreasing order of citedness). The very few databases that calculate and report the h -index or at least the citedness of publications that appear in the result lists of a search, should meet several other criteria in an optimum scenario (Jacso, 2004a). These databases should cover all the appropriate publication types. The software should be

able to recognize with a high level of accuracy the match between the citing references and the source items, and should calculate the hit numbers correctly. The interested users must be able to formulate their queries in a way to accommodate spelling and abbreviation variants of the authors' names and to distinguish authors with identical last names, first and middle initials.

There are problems and potential pitfalls in all these regards. It aggravates the problem that the dozens of articles that include *h*-index lists for researchers in various disciplines use only the automatically generated *h*-index values. These can much more distort the realistic *h*-index of researchers as envisioned by Hirsch than the built-in or perceived real and potential bias factors. It is not merely the completeness of the database but the completeness of the citation-enhanced part of the database (Jasco, 2007b) that matters, and WoS stands above the other databases since it was created by Eugene Garfield specifically for citation-based searching and citation analysis purposes (Garfield, 1955). This is by far the most important distinction that sets apart Scopus and WoS. In Scopus somewhat less than half of the records are enhanced by their cited references. In WoS all qualifying records (whose articles have references) are enhanced with those cited references.

I focus in the rest of this essay on WoS, since users can determine the *h*-index on their own, depending on the WoS edition licensed, and the person's willingness to engage in a series of steps to discover hundreds or—in the case of Lancaster—thousands of orphan or stray references available. Currently, WoS is the only database that has enough systematic and comprehensive coverage to evaluate accomplished researchers through citation enhanced records of at least journal articles as source documents, and several hundred million orphan and stray cited references, which have not been/could not be attributed to any of the nearly forty million master records in the 1945–2007 edition of WoS. (After I submitted my manuscript, Scopus introduced a new feature which alerts the users that there are more items for the search, and lists these under a separate tab, labeled as “More.” These are short records extracted from cited references that do not have a perfectly matching master record pair to add the citation to. These are exactly the orphan and stray records discussed in this essay, which can make a huge difference in computing a realistic *h*-index. WoS has had a somewhat similar index for a long time but it is less visible, convenient, and sophisticated as I discuss below.)

It is no accident that out of the dozens of articles that I have read and the dozens of *h*-lists that I have seen published until the end of October 2007, only a single *h*-list was based on Scopus. There certainly will be *h*-lists generated from Google Scholar, which has the great advantage of including records for books, a significant number of foreign language articles, conference papers, and useful gray literature sources. It also has the huge disadvantage of the lethal maladies of its grossly undereducated software.

WoS Editions

WoS is available in many different editions across the thousands of libraries that subscribe to it. The libraries may specify how far back they want to go. A library may have the 1996–2007 edition (with 15 million records), or the 1980–2007 edition (with 25.4 million records), or the 1945–2007 edition (with 38.7 million records), or the Century (of Science) edition, which goes back to 1900, and includes almost 40 million master records for items in scholarly, academic, and professional publications. In my estimation it also includes several hundred million cited references in a compact format (including many references to the same items in a huge number of spelling and morphological variants. This is due to the hundreds of bibliographic styles for various reference formats, and to the fallacies of us authors, editors, indexers, and data entry operators).

There is an additional twist: the Social Science component can go back only to 1956, and the Arts and Humanities component only to 1975, even if the library chooses the 1945–2007 WoS edition (totaling 38.6 million master records for the three citation index components). It would be essential to report for any *h*-index list exactly what edition was used because of the above variations. Hirsch clearly states that he used the 1955–2005 WoS edition (which at the time of his research must have had about 35 million master records). He is the exception rather than the rule among the *h*-list creators in stating the edition of the database used. I argue also for reporting the exact search query in order to recreate the query by an interested party to corroborate the results, and the filtering applied. For example, “FW Lancaster” as author would also retrieve records for two articles by a chemist with the same initials, but a good searcher should spot those and exclude them from the search (or at least the one that was cited fifty-two times). In this case spotting the items to be excluded is easy as the other “FW Lancaster” published in the 1950s, much earlier than *our* “FW Lancaster” started his career in information science, so those records stand out. (It is noted emphatically here, that the syntax of the queries are different across databases, and across host software programs. In WoS, authors are searchable with or without punctuation marks as long as the last name is entered first and followed by the first and middle initials in the right order, such as “Lancaster FW,” “Lancaster F W,” “Lancaster F.W.,” “Lancaster F. W.,” “Lancaster, FW,” etc.).

Then again, searching also for “Lancaster WF” is a good defensive search strategy (in the cited reference search mode, for picking up references for his papers erroneously cited with WF as the first and middle initials. It gets more complex when searching for “Lancaster F,” as it produces a much larger set to eyeball for selecting the right one(s). The new Author Identification module of WoS helps in this matter, but it is not a panacea.

The Software Generated h-index

In many cases the difference of database size is significant when using the edition going back to 1980 for all three citation index components versus the edition going back to 1945 in Science, 1956 in Social Science, and 1975 in Arts and Humanities. The software generated *h*-indexes reflect the publications of the researchers only from 1980 onward, and only those that were covered in the database from 1980.

In the case of Lancaster, the former edition has only 88 master records for Lancaster's publications with a total of 374 citations received, and an *h*-index of 11. For the 1945–2007 edition, the figures are: 131 master records (having removed 2 master records for another "FW Lancaster") with a total of 651 cited references attributed to the correct master records, and an *h*-index of 13. Manually adding the only master record where Lancaster's name appears with only his first initial increases his *h*-index to 14 because that paper was cited 27 times. There were no master records for "W. Lancaster," nor for misspelled variants, but these do appear in the cited references. With more unusual, more difficult last names and less distinguishing first and middle initials this stage of identifying the variants for the right author in the master records may be much more time consuming, such as for this very author whose name is too often misspelled in master records. Recently I started to drop the accent from the last character of my name, because it made me lose citations that used only the base character, and the last character deleted, or transformed it into some special character, and thus did not match the accented character in the master records. I could help this, but not the many misspellings caused by the common character-pair mix-up when Jacso appears as Jasco in the citing references.

In case of coauthorships more attention is needed even at this stage and especially at the look-up stage of the cited author index (discussed below). Identifying and bringing together the master records for an author in a set for having the software to generate the *h*-index is the relatively easy step, especially in WoS and Scopus, both of which recently introduced a software module for author disambiguation (Jacso, 2007c).

Searching and Browsing the Cited Reference Index

The really hard part is the next process of locating orphan and stray citation records in the Cited Reference Index, which can be searched by the combination of author last name, middle and first initial, the cited work and cited year, and a matching list can be browsed. Truncation can also be used in the process.

A database coverage that seemingly fits the time span may still not guarantee appropriate completeness, because the coverage of a particular journal (where the author published a paper in, say, 1981) may have started only in 1983 in WoS, or the journal may not have been covered at all, or its coverage was stopped for various reasons. This is the case

with Lancaster's article about "Evaluating Collections by Their Use" in *Collection Management*, a journal not covered by WoS as a source journal, but that appears in the cited reference index of WoS twenty-four times as an orphan entry. This is important, since the software generated *h*-index does not reflect this citedness score. There are dozens of other sources where Lancaster published his papers but the source was not covered, or not at the time when he published the paper. For this reason, there are no master records for such papers. Human error may also cause the absence of master records, when an article was skipped in processing the issue of a journal.

A comprehensive and current bibliography is an essential tool for discovering and using such orphan and stray references in recalculating the *h*-index for journal articles. For this paper Lancaster's curriculum vitae (which is an appendix to this Festschrift) was used, although the bibliography therein does not include book reviews, letters to the editors, etc. These are rarely cited, so are not relevant from the perspective of the *h*-index.

Creating Pseudo Master Records for Nonsource References

There are several items in the bibliography for which there are no master records in WoS, but they appear with higher and much higher citedness frequency than the software generated *h*-index value, and these must be accounted for in recalculating a realistic *h*-index. This requires a rather arduous job of searching and browsing the cited reference index to collect and organize the data.

Although I looked up all the items in the bibliography in any reasonable variations of F. W. Lancaster and his coauthor names, this is not required. Except for such a festive occasion as Lancaster's seventy-fifth birthday, there is no reason for trying to "keep track of each fallen robin." In the index there were 712 entries under the cited author name format "Lancaster FW" alone, with a total citation count of 2,607 (including the 131 entries with a 651 total citation count, and including less than 1 percent for other cited authors named "Lancaster FW"). There were 19 entries with the "Lancaster W.F." name format and 103 citations (all for "our" author), and 51 "Lancaster F" name format with 446 citations (most for other than "our" author).

Searching under coauthor names and their variant and anticipated misspellings also brought up dozens of works and hundreds of orphan and stray citations (such as Harricombe in addition to the correct Haricombe) because they did not have a master record, or differed in any of the following data elements from the master record: cited author last name, first and middle initial; publication year, volume and issue numbers; starting page numbers; and the cited work's title. In the case of journal articles this is always the name of the journal, not the title of the article. The title of the work is limited to twenty characters, which imposes

ISI Web of KnowledgeSM Access the new version! Web of Science

FINISH SEARCH >> View the articles that cite the selected references.
The completed search will be added to the search history.

(Limit by language and document type)

CITED REFERENCE INDEX Go to Page: 1 of 1 Go

References 1 -- 13

SELECT PAGE SELECT ALL* CLEAR ALL or select specific references from the list.
When desired references have been selected from all pages, click FINISH SEAR

Search Select	Times Cited**	Cited Author	Cited Work	[SHOW EXPANDED TITLES]	Year	Volume	Page
<input type="checkbox"/>	1	HARICOMBE LJ	COMMUNICATION SPR		1994		
<input type="checkbox"/>	2	HARICOMBE LJ	CREATING AGILE LIB M		1998		
<input type="checkbox"/>	1	HARICOMBE LJ	CREATING AGILE LIBRA		1998		
<input type="checkbox"/>	2	HARICOMBE LJ	LIBR QUART		1993	63	508
<input type="checkbox"/>	1	HARICOMBE LJ	LIBR TRENDS		1998	47	1
<input checked="" type="checkbox"/>	1	HARICOMBE LJ	OUT COLD ACAD BOYCOT		1995		
<input checked="" type="checkbox"/>	3	HARICOMBE LJ	OUT COLD ACAD BOYCOT		1995		
<input checked="" type="checkbox"/>	1	HARICOMBE LJ	OUT COLD ACAD BOYCOT		1995		1
<input checked="" type="checkbox"/>	1	HARICOMBE LJ	OUT COLD ACAD BOYCOT		1995		112
<input checked="" type="checkbox"/>	1	HARICOMBE LJ	OUT COLD ACADEMIC BO				158
<input checked="" type="checkbox"/>	7	HARICOMBE LJ	OUT COLD ACADEMIC BO		1995		
<input checked="" type="checkbox"/>	1	HARICOMBE LJ	THESIS U ILLINOIS UR		1992		
<input checked="" type="checkbox"/>	1	HARRICOMBE LJ	COLD ACAD BOYCOTTS I		1995		

* "Select All" adds the first 500 matches to your cited reference search, not all matches.

Figure 11. Orphan references in Web of Science for a Haricombe & Lancaster book with 16 citations in 8 reference variants

strange omissions and abbreviations beyond the regular and predictable ones. In case of book citations, cited page numbers or number range appear in lieu of the chronological/numerical designations.

Even the slightest difference of a space, or a different page number, makes a citation record an orphan or stray as shown in Figure 11 on the excerpt of the index for the book, *Out in the Cold: Academic Boycotts and the Isolation of South Africa*—rightly—under the first author's name. (The Haricombe misspelled variant came up because, contemplating a possible spelling error, my defensive search strategy was "Harricomb* OR Haricomb*.")

I believe that these often overshrunk reference records were the results of the Procrustean bed of the 80-column Hollerith cards. The inconsistent references in the source documents and the inconsistent omissions and abbreviations all contribute to the wild varieties and enigmatic formats.

Even seemingly distinct titles may be time consuming to locate exhaustively. I thought that finding references to his book titled *If You Want to Evaluate Your Library* would be a walk in the park. Looking it up under the word "want*" and the name "Lancaster" immediately brought up 2 entries with a total citation count of 2. This was obviously inappropriate. I just thought the *If* and *You* would be dropped. When the name and the word "If*" brought up 14 entry variants with a citation count of 48, I

ISI Web of KnowledgeSM Access the new version! Web of Science

CITED REFERENCE INDEX
References 1 -- 17

Go to Page: 1 of 1 GO

SELECT PAGE SELECT ALL CLEAR ALL or select specific references from the list.
When desired references have been selected from all pages, click FIN

to complete your search.

Select	Times Cited**	Cited Author	Cited Work	[SHOW EXPANDED TITLES]	Year	Volume	Page
<input type="checkbox"/>	1	BRADFORD SC	YOU WANT EVALUATE YO		1988		
<input type="checkbox"/>	1	LANCASTER F	YOU WANT EVALUATE YO		1988		
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALAUTE YO		1988		
<input type="checkbox"/>	3	LANCASTER FW	YOU WANT EVALUATE LI		1993		
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE LI		1988		
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE LI		1988		145
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE YO				
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE YO		1998		
<input type="checkbox"/>	13	LANCASTER FW	YOU WANT EVALUATE YO		1993		
<input type="checkbox"/>	18	LANCASTER FW	YOU WANT EVALUATE YO		1988		
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE YO		1988		3
<input type="checkbox"/>	2	LANCASTER FW	YOU WANT EVALUATE YO		1988		5
<input type="checkbox"/>	2	LANCASTER FW	YOU WANT EVALUATE YO		1988		39
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE YO		1988		123
<input type="checkbox"/>	1	LANCASTER FW	YOU WANT EVALUATE YO		1988		168
<input type="checkbox"/>	1	*U ILL GRAD SCH LI	YOU WANT EV YOUR LIB		1988		8
<input type="checkbox"/>	1	WILFRID LF	YOU WANT EVALUATE YO		1988		

* "Select All" adds the first 500 matches to your cited reference search, not all matches.
** Times Cited counts are for all databases and all years, not just for your current database and

Figure 12. Variations of a title and edition in the Cited Reference Index in Web of Science

thought “I am done,” but tried under the name combined with “You*,” and it returned an additional 13 entries with 49 citations, a real win. Of course there may be some additional items with his name absent or misspelled, or the word “want*” misspelled, but because it was already so high for his original *h*-index, it simply was not necessary. You have to know when to stop. (See Figure 12.)

The most difficult are those citations where the lead-in part of the cited works are the same for hundreds of entries for different items whose entries scatter widely from “inf retrieval” to “information retr” to “information retrieval” and only the publication year may give a hint (in those cases when it is correctly cited) if it is the first or second edition of Lancaster’s book on information retrieval systems. Or is it a reference to one of his other books whose title starts with the same character string of “information retrieva” truncated in the index, such as *Information Retrieval Today*, or *Information Retrieval On-Line*? (See Figure 13.)

The result list was created by the defensive strategy of Lancaster alone as cited author and “INF* RETR*” as cited words. The continuation of the 89-item list in Figure 14 shows that a defensive approach is a must. The

Select	Times Cited**	Cited Author	Cited Work	[SHOW EXPANDED TITLES]	Year	Volume	Page
<input type="checkbox"/>	1	LANCASTER	INFORMATION RETRIEVA				
<input type="checkbox"/>	1	LANCASTER	INFORMATION RETRIEVA		1973		
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVE		1997		
<input type="checkbox"/>	1	LANCASTER F	INFORM RETRIEVAL TOD		1993		
<input type="checkbox"/>	6	LANCASTER F	INFORMATION RETRIEVA		1993		
<input type="checkbox"/>	4	LANCASTER F	INFORMATION RETRIEVA		1979		
<input type="checkbox"/>	1	LANCASTER F	INFORMATION RETRIEVA		1979		140
<input type="checkbox"/>	5	LANCASTER F	INFORMATION RETRIEVA		1973		
<input type="checkbox"/>	2	LANCASTER F	INFORMATION RETRIEVA		1973		206
<input type="checkbox"/>	2	LANCASTER F	INFORMATION RETRIEVA		1972		
<input type="checkbox"/>	4	LANCASTER F	INFORMATION RETRIEVA		1968		
<input type="checkbox"/>	1	LANCASTER F	INFORMATION RETRIEVA		1968		126
<input type="checkbox"/>	2	LANCASTER F	INFORMATION RETRIEVA		1968		183
<input type="checkbox"/>	1	LANCASTER FW	INF RETR ONL		1973		
<input type="checkbox"/>	2	LANCASTER FW	INFORM RETRIEVAL ON		1973		
<input type="checkbox"/>	1	LANCASTER FW	INFORM RETRIEVAL ONL		1973		
<input type="checkbox"/>	2	LANCASTER FW	INFORM RETRIEVAL SYS		1979		
<input type="checkbox"/>	1	LANCASTER FW	INFORM RETRIEVAL SYS		1968		
<input type="checkbox"/>	2	LANCASTER FW	INFORM RETRIEVAL TOD		1993		
<input type="checkbox"/>	1	LANCASTER FW	INFORMAITON RETRIEVA		1968		
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETREIVA		1979		
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIECA		1968		
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEFA		1979		
<input type="checkbox"/>	3	LANCASTER FW	INFORMATION RETRIEVA				
<input type="checkbox"/>	26	LANCASTER FW	INFORMATION RETRIEVA		1993		
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA		1982		
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA		1981		
<input type="checkbox"/>	2	LANCASTER FW	INFORMATION RETRIEVA		1981		105

Figure 13. Enigmatic, inaccurate, and inconsistent entries for different editions of the same work and for different works in Web of Science

correctness of none of the data elements can be taken for granted, not even in the case of the fairly easy last name of our author, and his well-justified insistence on using only his first and middle initial (since when fully spelled out, both could lead to far more variant entries).

SUMMARY OF FINDINGS

The entries of the orphan and stray citation records are cumbersome to identify, collect, unify, and aggregate, but identify, unify, collect, and aggregate one must. A responsible searcher should make every reasonable attempt to do so, in order to arrive at a reasonable *h*-index. In spite of the time-consuming process it may be worth it, especially when the process would significantly increase the *h*-index of the author. In Lancaster's case

<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1971	
<input type="checkbox"/>	81	LANCASTER FW	INFORMATION RETRIEVA	1968	
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	11
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	32
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	34
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	58
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	74
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	82
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	130
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	181
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1968	209
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1965	124
<input type="checkbox"/>	1	LANCASTER FW	INFORMATION RETRIEVA	1963	
<input type="checkbox"/>	1	LANCASTER FW	INFORMATIONAL RETRIE	1968	
<input type="checkbox"/>	1	LANCASTER H	INFORMATION RETRIEVA	1978	193
<input type="checkbox"/>	1	LANCASTER L	INFORMATION RETRIEVA	1968	
<input type="checkbox"/>	1	LANCASTER RW	INFORMATION RETRIEVA	1979	
<input type="checkbox"/>	1	LANCASTER W	INFORMATION RETRIEVA	1993	
<input type="checkbox"/>	1	LANCASTER W	INFORMATION RETRIEVA	1968	
<input type="checkbox"/>	2	LANCASTER WF	INFORMATION RETRIEVA	1993	
<input type="checkbox"/>	2	LANCASTER WF	INFORMATION RETRIEVA	1979	
<input type="checkbox"/>	2	LANCASTER WF	INFORMATION RETRIEVA	1968	
<input type="checkbox"/>	1	LANCASTER WR	INFORMATION RETRIEVA	1979	154

Figure 14. Continuation of the browsable index entries in Web of Science

the process doubled his *h*-index value from the software generated original value of 13 to 26 even though I was unable to identify and aggregate the citation counts of more than a hundred orphan references because of their very skimpy or erroneous content of the cited references.

I preferred to leave them alone rather than assign them to the most likely master record or most likely pseudo master record that I created for orphan and stray references for the same work/same edition. Leaving no stone unturned would have increased F. W. Lancaster's *h*-index by an additional 3–4 points.

Of the top cited 25 works collected from the Cited Reference Index only 10 had master records in WoS—all of them for journal articles. Except for one article, the top 12 cited works are all books or monographic reports without master records. That one article does have a master record—with 52 citing references. Looking it up in the cited reference index added 11 orphan records, which mostly had slightly different pagination from the pagination in the master record. This, of course, did not increase the *h*-index because even the original number of correctly matched citing references was twice as high as the new and improved *h*-index; it was a gratifying find at the end of the citation digging process, going out with a bang.

CONCLUSIONS

Google Scholar has no option for elementary functions, such as numbering the result lists, marking records, truncating, or sorting the results of a search by decreasing citation counts, nor does it have a built-in, automatic *h*-index generator. Its result list on the screen can be scrapped and converted into a spreadsheet. There are third-party utilities that can do this, but they cannot help with the underlying problems—the unreliable, grossly inflated citation counts, the phantom author names by the hundreds of thousand, which often replace the name of the real authors, and the phantom citations caused by its primitive citation matching algorithm.

Scopus has outstanding software features, and it can compute a reasonable *h*-index but only for researchers with a maximum twelve years of scholarly publishing activity in journals and to a lesser extent in conference proceedings in the various fields of the sciences. WoS has the longest retrospective coverage for most of the sciences, many of the social sciences, and a few of the arts and humanities disciplines, if one has access to a WoS edition with coverage from at least the mid-1980s for covering the living and active researchers. Not covering and thus not creating master records for books is the serious limitation of WoS, because in the career of many scholars in several disciplines, books are the most cited works of scholars.

Linda Butler and Martin Visser argued convincingly about the importance of extending the citation analysis and evaluation of researchers to books and other materials, which are not covered by Web of Science as source documents (Butler & Visser, 2006). I could not agree more with them, especially since they make their case based on the analysis of more than thirty thousand publications by Australian researchers, comparing the results with the coverage of the sources of their publications in Web of Science. They could not refer to the *h*-index, of course, since it was not yet developed when they wrote their paper. Their results, however, strongly reinforce the need, in the evaluation of researchers, to include coverage of sources that are not covered by Web of Science (nor by Scopus, either), in order to be fair to researchers beyond the realm of chemistry, physics and biomedicine, where 80–90 percent of the publication outlets used by researchers in these disciplines are covered by WoS (and even to a broader extent by Scopus, although for a much shorter time span). This is exactly the motivation for my methodology presented here.

The directly searchable Cited Reference Index compensates to some extent for the lack of coverage of potentially important journals, books, and technical reports. Enhancement of its software by a software module, which would analyze and compare the entries in the Cited Reference Index with the quality, accuracy, and intuition of the CiteSeer system would take out the most arduous part of the process. It could show po-

[CiteSeer.IST Home](#) **Check:** The following citations are predicted to all refer to the same paper. [Details](#)

- F. Lancaster. *Vocabulary Control for Information Retrieval*, Second Edition. Information Resources, Arlington, VA, 1986.
- Lancaster, F. (1986). *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington VA.
- F. W. Lancaster. *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, VA, 1986.
- Lancaster F.W., *Vocabulary Control For Information Retrieval* (2nd ed), Arlington, Virginia, Information Resources Press (1986).
- Lancaster, F.W. *Vocabulary control for information retrieval*. Washington, D.C.: Information Resources Press, 1972.
- Lancaster, F. W. 1986. *Vocabulary Control for Information Retrieval*. Arlington, Va.: Information Resources Press.
- Lancaster, F. (1986). *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington VA.
- F. W. Lancaster, *Vocabulary control for information retrieval*, Information Resources Press, Arlington, VA, 1986.
- Lancaster, F. , *Vocabulary Control for Information Retrieval*, Information Resources Press, 1972.
- F. W. Lancaster. *Vocabulary Control for Information Retrieval*, (2nd. edition). Information Resources Press, 1986.
- F. Lancaster, "Vocabulary Control for Information Retrieval", Information Resource Press, 1972. - 499--
- F. W. Lancaster. *Vocabulary Control for Information Retrieval*. Information Resources Press, Arlington, VA, second edition, 1986.
- CiteSeer.IST - Copyright [Penn State](#) and [NEC](#)

Figure 15. Intelligent name recognition and advice for the user in CiteSeer

tentially matching orphan/stray and master records with adjustable precision scales on the data elements, such as issue number and starting page number (and with a check-box to let the user select/deselect entries as needed). The software approach would cost far less than retrospective enhancement of records, and would draw the extra benefit from its existing and precious, but currently much underused, content. Figure 15 is an example of a record from CiteSeer to illustrate how smartly it recognizes the many variant formats of not only the author's name and the title, but also imprint details.

In his very short reminiscences in the last chapter of the Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems (dedicated to the pioneers of science information systems) (Bowden, Hahn, & Williams, 1999), Lancaster (1999) started with this: "My biggest moment in the field of information science occurred in 1968 when I learned that my first book had been accepted for publication by John Wiley." Very tellingly just half a page later, he ends his comments saying: "There have been many notable events in my career ... but getting my first book published was definitely the highlight."

I could have written this contribution about many issues where I share the interests of Wilf Lancaster and learned a lot from him, such as indexing and abstracting, citation analysis, the evaluation of information systems, or designing and building databases. But I thought that the most appropriate contribution I could make for this Festschrift would be to esti-

mate his much more realistic *h*-index, this exciting new metric of scholarly impact, by demonstrating the practical limitations of a potentially great, theoretically sound, new performance indicator by recalculating the software generated *h*-index and giving appropriate credit to Lancaster's obvious labor of love, the books that he wrote and that got cited the most in his oeuvre.

NOTE

1. Beyond the bibliographic citation, this work of Lancaster about bibliometric methods in assessing productivity and impact of research deserves more than a passing mention, because—beyond the obvious reason of being of high relevance to my topic—it is little known, and very little cited by researchers. The reason for this is that the book was published in India and is scarcely available in academic libraries; only three OCLC member libraries are identified as holding it in the subscription-based WorldCat. This thin book is thick in content, and is a perfect companion for putting the *h*-index in perspective, especially for those who will use the *h*-index for evaluating others without knowing much about quantifiable measures of research performance in general. I was lucky during my recent lecture tour in India to spot the book at one of the university libraries and could order it through an agent in India just when I started working on this project.

REFERENCES

- Ashkanasy, N. M. (2007). Playing the citations game. *Journal of Organizational Behavior*, 28(6), 643–645.
- Banks, M. G. (2006). An extension of the Hirsch index: Indexing scientific topics and compounds. *Scientometrics*, 69(1), 161–168.
- Bar-Ilan, J. (2006a). H-index for Price medalists revisited. *ISSI Newsletter*, 2(1), 3–5.
- Bar-Ilan, J. (2006b). An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes. *Information Processing and Management*, 42(6), 1553–1566.
- Bar-Ilan, J., Levene, M., & Lin, A. (2007). Some measures for comparing citation databases. *Journal of Informetrics*, 1(1), 26–34.
- Barendse, W. (2007). The strike rate index: A new index for journal quality based on journal size and the *h*-index of citations. *Biomedical Digital Libraries*, Article #4.
- Batista, P. D., Campiteli, M. G., Kinouchi, O., & Martínez, A. S. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1), 179–189.
- Berger, M. (2007). The problematic ratings game in modern science. *South African Journal of Science*, 103(1–2), 2–3.
- Bornmann, L., & Daniel, H-D. (2005). Does the *h*-index for ranking of scientists really work? *Scientometrics*, 65(3), 391–392.
- Bowden, M. E., Hahn, T. B., & Williams, R. W. (1998). Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems, American Society for Information Science and Technology and the Chemical Heritage Foundation.
- Braun, T., Glänzel, W., & Schubert, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69(1), 169–173.
- Butler, L., & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics*, 66(2), 327–343.
- Callicott, B., & Vaughn, D. (2005). Google Scholar vs. library scholar: Testing the performance of Schoogle. *Internet Reference Services Quarterly*, 10(3–4), 2005.
- Costas, R., & Bordons, M. (2007). The *h*-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics*, 1(3), 193–203.
- Cronin, B., & Meho, L. (2006). Using the *h*-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57(9), 1275–1278.
- Cronin, B., Snyder, H., & Atkins, H. (1997). Comparative citation rankings of authors in monographic and journal literature: A study of sociology. *Journal of Documentation*, 53(3), 263–273.
- Egghe, L. (2006). Theory and practise of the *g*-index. *Scientometrics*, 69(1), 131–152.

- Egghe, L. (2007a). Dynamic h-index: The Hirsch index in function of time. *Journal of the American Society for Information Science and Technology*, 58(3), 452–454.
- Egghe, L. (2007b). Item-time-dependent Lotkaian informetrics and applications to the calculation of the time-dependent h-index and g-index. *Mathematical and Computer Modelling*, 45(7–8), 864–872.
- Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, 69(1), 121–129.
- Garfield, E. (1955). Citation indexes for science: A new dimension in documentation through association of ideas. *Science*, 122 (3159), 108–111.
- Glänzel, W. (2006). On the h-index - A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67(2), 315–321.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academies of Science*, 102(46), 16569–16572.
- Imperial, J., & Rodríguez-Navarro, A. (2007). Usefulness of Hirsch's h-index to evaluate scientific research in Spain. *Scientometrics*, 71(2), 271–282.
- Jacso, P. (1997). Content evaluation of databases. *Annual Review of Information Science and Technology*, 32, 231–267.
- Jacso, P. (2004a). Citedness scores for filtering information and ranking search results. *Online Information Review*, 28(5), 371–376.
- Jacso, P. (2004b). Citation-enhanced indexing/abstracting databases. *Online Information Review*, 28(3), 235–238.
- Jacso, P. (2006a). Deflated, inflated and phantom citation counts. *Online Information Review*, 30(3), 297–309.
- Jacso, P. (2006b). Dubious hit counts and cuckoo's eggs. *Online Information Review*, 30(2), 188–193.
- Jacso, P. (2007a). *Web of Science, Peter's Digital Reference Shelf*. Retrieved May 9, 2008, from <http://www.galegroup.com/reference/peter/200701/wos.htm>.
- Jacso, P. (2007b). The dimensions of cited reference enhanced database subsets. *Online Information Review*, 31(5), 694–705.
- Jacso, P. (2007c). Software issues related to cited references. *Online Information Review*, 31(6), 892–905.
- Jacso, P. (2008a). Google Scholar Revisited. *Online Information Review*, 32(1), 102–114.
- Jacso, P. (2008b). The plausibility of calculating the h-index in cited reference enhanced databases. *Online Information Review*, 32 (2), 266–282
- Jacso, P. (2008c). The pros and cons of calculating the h-index from Google Scholar. *Online Information Review*, 32(3), in press.
- Jacso, P. (2008d). The pros and cons of calculating the h-index from Scopus. *Online Information Review*, 32(4). Forthcoming.
- Jacso, P. (2008e). The pros and cons of calculating the h-index from Web of Science. *Online Information Review*, 32(5). Forthcoming.
- Jeang, K. (2007). *Impact factor, H-index, peer comparisons, and retrovirology: Is it time to individualize citation metrics? Retrovirology, Article 4*. Retrieved May 9, 2008, from <http://www.retrovirology.com/content/4/1/42>.
- Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52(6), 855–863.
- Kelly, C. D., & Jennions, M. D. (2006). The h-index and career assessment by numbers. *Trends in Ecology and Evolution*, 21(4), 167–170.
- Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google/Web citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58(7), 1055–1065.
- Lancaster, F. W. (1991). *Bibliometric methods in assessing productivity and impact of research*. Bangalore: Sarada Ranganathan Endowment for Library Science.
- Lancaster, F. W. (1999). Getting published. In M. E. Bowden, T. B. Hahn, & R. V. Williams, (1999). *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*. Medford, NJ: Information Today.
- Liang, L. (2006). H-index sequence and h-index matrix: Constructions and applications. *Scientometrics*, 69(1), 153–159.
- Line, M.B. (1979). The influence of sources used on the results of citation analyses. *Journal of Documentation*, 35(4), 265–284.

- Line, M.B. and Brittain, J. M. (1973). Sources of citations and references for analysis purposes: A comparative assessment. *Journal of Documentation*, 29(1), 72-86.
- Mayr, P., & Walter, A. (2007). An exploratory study of Google Scholar. *Online Information Review*, 31(6), 814-830.
- Meho, L. I., & Yang, K. (2007). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, 58(13), 2105-2125.
- Meneghini, R., & Packer, A. L. (2006). Articles with authors affiliated to Brazilian institutions published from 1994 to 2003 with 100 or more citations: II—identification of thematic nuclei of excellence in Brazilian science. *Anais da Academia Brasileira de Ciências*, 78(4), 855-883.
- Mullen, L. B., & Hartman, K. A. (2006). Google Scholar and the library Web site: The early response by ARL libraries. *College and Research Libraries*, 67(2), 106-122.
- Neuhaus, C., Neuhaus, E., Asher, A., & Wrede, C. (2006) The depth and breadth of Google Scholar: An empirical study. *Portal*, 6 (2), 133-140.
- Norris, M., & Oppenheim, C. (2007). Comparing alternatives to the Web of Science for coverage of the social sciences literature. *Journal of Informetrics*, 1(2), 161-169.
- Oelrich, B., Peters, R., & Jung, K. (2007). A bibliometric evaluation of publications in urological journals among European Union countries between 2000-2005. *European Urology*, 52(4), 1238-1248.
- Olden, J. D. (2007). How do ecological journals stack-up? Ranking of scientific quality according to the h-index. *Ecoscience*, 14(3), 370-376.
- 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*, 58(21), 297-301.
- Packer, A. L., & Meneghini, R. (2006). Articles with authors affiliated to Brazilian institutions published from 1994 to 2003 with 100 or more citations: I—the weight of international collaboration and the role of the networks. *Anais da Academia Brasileira de Ciências*, 78(4), 841-853.
- Pomerantz, J. (2006). Google Scholar and 100% availability of information. *Information Technology and Libraries*, 25(1), 52-56.
- Prathap, G. (2006). Hirsch-type indices for ranking institutions' scientific research output. *Current Science*, 91(11), 1439.
- Purvis, A. (2006). The h-index: Playing the numbers game. *Trends in Ecology and Evolution*, 21(8), 422.
- Robinson, M. L., & Wusterman, J. (2007). Putting Google Scholar to test: A preliminary study. *Program*, 41(1), 71-80.
- Rousseau, R. (2007). The influence of missing publications on the Hirsch index. *Journal of Informetrics*, 1(1), 2-7.
- Saad, G. (2006). Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively. *Scientometrics*, 69(1), 117-120.
- Salgado, J. F., & Páez, D. (2007). Scientific productivity and Hirsch's h-index of Spanish social psychology: Convergence between productivity indexes and comparison with other areas. *Psicothema*, 19(2), 179-189.
- Schreiber, M. (2007a). A case study of the Hirsch index for 26 non-prominent physicists. *Annalen der Physik (Leipzig)*, 16(9), 640-652.
- Schreiber, M. (2007b). Self-citation corrections for the Hirsch index. *Europhysics Letters*, 78(3).
- Schroeder, R. (2007). Pointing users toward citation searching: using Google Scholar and Web of Science. *Portal*, 7(4), 243-248.
- Schubert, A. (2007). Successive h-indices. *Scientometrics*, 70(1), 201-205.
- Schubert, A., & Glänzel, W. (2007). A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1(3), 179-184.
- van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups *Scientometrics*, 67(3), 491-502.
- Vanclay, J. K. (2006). Refining the h-index [5]. *Scientist*, 20(7), 14-15.

- Vanclay, J. K. (2007). On the robustness of the *h*-index. *Journal of the American Society for Information Science and Technology*, 58(10), 1547–1550.
- Vinkler, P. (2007). Eminence of scientists in the light of the *h*-index and other scientometric indicators. *Journal of Information Science*, 33(4), 481–491.
- White, B. (2006). Examining the claims of Google Scholar as a serious information source. *The New Zealand Library and Information Management Journal*, 50(1), 11–24.

Peter Jacso is a professor at the Department of Information and Computer Sciences at the University of Hawaii. For his teaching he received the Pratt-Severn Faculty Innovation Award in Library and Information Studies from the Association of Library and Information Science Educators, the Outstanding Information Science Teacher Award of the American Society for Information Science & Technology, and the Institute for Scientific Information. He authored several books, and conference papers. His papers were published in various research publications such as the *Annual Review of Information Science & Technology*, *Current Science*, *Cortex*, *Library Software Review*, and *Library & Information Science Research*. His columns, editorials, and database reviews appeared in *Online*, *Database*, *Online Information Review*, *Computers in Libraries*, as well as in his Web-born review column hosted by the Gale Group. His writings earned him awards of the American Library Association, Learned Information, Ltd., Emerald, and Oryx Press. He has been very active on the conference circuit in the United States, Europe, and Southeast Asia.