BioOne COMPLETE

The Poverty of Citation Databases: Data Mining is Crucial for Fair Metrical Evaluation of Research Performance

Author: Krell, Frank-Thorsten

Source: BioScience, 59(1) : 6-7

Published By: American Institute of Biological Sciences

URL: https://doi.org/10.1525/bio.2009.59.1.2

BioOne Complete (complete.BioOne.org) is a full-text database of 200 subscribed and open-access titles in the biological, ecological, and environmental sciences published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Complete website, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/terms-of-use.

Usage of BioOne Complete content is strictly limited to personal, educational, and non - commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

BioOne sees sustainable scholarly publishing as an inherently collaborative enterprise connecting authors, nonprofit publishers, academic institutions, research libraries, and research funders in the common goal of maximizing access to critical research.

Downloaded From: https://bioone.org/journals/BioScience on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

The Poverty of Citation Databases: Data Mining Is Crucial for Fair Metrical Evaluation of Research Performance

FRANK-THORSTEN KRELL

For a long time, the journal impact factor has been used to evaluate the scientific performance of authors. It is increasingly recognized, however, that judging an author's scientific performance should take into account that author's scientific output, and not the output of other authors publishing in the same journal—the citation rates of papers in one journal can vary enormously, and the journal impact factor fails to that consider that variance.

The number of citations an author attracts is a reliable measure of the attention the author receives from the scientific community, or, in other words, of the scientific impact of an author. (Attention is a lame arbiter of scientific quality, but that is a problem that cannot be solved by any simple metrics.) In 2005, Jorge E. Hirsch proposed a simple, elegant measure of an author's impact: the *h* index, which is the number of an author's papers (*h*) with at least *h* citations. Other author-based indexes have been proposed, such as the *g* index, which, given a set of papers ranked in decreasing order of the number of citations received, is the largest number such that the top *g* articles received together at least g^2 citations (Egghe 2006). The *g* index better takes into account the citation scores of top articles.

Guillaume Chapron and Aurélie Husté claim in the July 2006 issue of *BioScience* that the "*h* index...can very easily be computed from most literature databases." It can, but is the resulting index representative of the author's impact? Bar-Ilan (2008) asks rightly, "Which *h*-index?" after calculating the *h* index for Israeli researchers using the Web of Science (WoS), Scopus, and Google Scholar. Depending on the data source, *h* indexes of the same author can vary by a factor of up to eight (31 vs. 4), often by a factor of two. Any of the three sources might provide the highest *h* index, depending on the individual author, indicating that all three sources are randomly incomplete. However, the contents of these three databases have never been compared with a complete data set. Admittedly, a complete data set is difficult to obtain if all available databases are incomplete. Here I perform such a comparison for the first time.

The data I use relate to my own publications—I am the only author for whom I have such data available. The complete data set contains all citations of my papers from WoS ($n = 181$), Scopus ($n = 101$), and Google Scholar ($n = 172$), in addition to citations I found in the literature during the last 20 years: the total, as of May 2008, is 704 citations. The citation databases contain only a small portion of the citations of my papers: WoS, 25.7 percent; Scopus, 14.3 percent; and Google Scholar, 24.4 percent. This poor coverage dramatically affects my *h* and *g* indexes. From the comprehensive count, 14 papers were cited at least 14 times, and my *g* index is 20. WoS would give me $h = 7$ and $g = 10$; Scopus $h = 6$ and $g = 9$; and Google Scholar, although not having the highest coverage, $h = 8$ and $g = 11$. The poor coverage of citation databases cuts my performance indicators by half, and my case is not an isolated one.

Why do citation databases miss three-quarters of my citations? Is it just me, or is everybody affected in the same way? The coverage of my field, organismic entomology and taxonomy, is particularly deficient in all available citation databases. For example, WoS covers 69 percent of "Biological sciences—animals and plants," according to Moed (2005, p. 125), who takes into account both the coverage of journal literature by WoS and the importance of journals (measured as the percentage of references to documents published in a journal relative to total references). However, considering the covered entomological journals in relation to the existing journals, the coverage of entomologic taxonomical journals by WoS is at most 3 percent (27 out of about 900 entomological journals with taxonomical content that are held by the library of the Natural History Museum in London).

The coverage in other taxonomic disciplines is not much better. For new descriptions of marine species, a data set from 2002–2003 shows that only 36 percent were published in journals with an impact factor—that is, covered by WoS (Bouchet 2006). Brown and colleagues (2008) found that none of the established databases and search engines covers references on selected fossil amphibians anywhere near completeness. Compared with a comprehensive library-based search, the coverage was between 4 and 23 percent, with Google Scholar in the lead (Scopus and WoS were not studied). Other scientific disciplines, such as molecular biology and biochemistry (biological sciences related to humans, chemistry, or clinical medicine) are covered to a much larger extent (84 to 92 percent in WoS; Moed 2005). The different coverage of different disciplines makes performance indicators relying on citation databases impossible to compare among fields.

doi:10.1525/bio.2009.59.1.2

Downloaded From: https://bioone.org/journals/BioScience on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use

Not only do popular databases miss a significant proportion of citations, depending on the scientific discipline, but they also include a significant amount of irrelevant citations, such as self-citations. Self-citations might boost the visibility of an author's work and in the long term increase this author's citation rate, but self-citations represent neither the attention an author receives nor any sort of impact on the field—they indicate only that an author is aware of and builds on his or her own papers. Thus, self-citations are not appropriate to be considered in assessments of research performance and should be eliminated from any data set used to calculate performance indicators. In my own comprehensive citation count, self-citations amount to 27 percent. Other authors have found self-citation rates from 11 to 67 percent (e.g., Schreiber 2008).

Self-citations increase the *h* factor, often by a rather small amount of one or two units, but sometimes by up to six units (in my case, by four units, from 10 to 14). The *g* index seems to be even more affected by self-citations (Schreiber 2008), although in my case the difference is just three units (17 vs. 20). Using data sets from citation databases, the effect of self-citations on my *h* and *g* indexes is reduced to one unit, as their coverage of my self-citations (originating mainly from my taxonomic papers) is even poorer than the overall coverage. With better coverage, the influence of self-citations increases.

When data from available databases are cleared of self-citations and complemented by any additional citations, indexes such as *h* and *g* might give an honest and fair indication of the attention an author receives from the scientific community. To distinguish such indexes from uncritically obtained figures, I propose to call them honest indexes. The honest *h* index (h_h) is defined as the *h* index considering all available citations, except for self-citations by the author for whom the index is determined.

Conclusions

Research evaluation by means of simple metrics is widely considered the quick and easy alternative to assessing the actual quality of scientific output. It seems to be objective and does not require any expertise. The quickest metrical assessment—using the raw data sets provided by WoS, Scopus, Google Scholar, or any other professional citation database—suffers from unacceptable data deficiency. Metric research evaluation should be done only after serious data mining and clean up. Completing citation data sets and removing errors and self-citations would lead to honest citation indexes. Any scientist in risk of metric evaluation might consider building up a personal database of citations to calculate one's personal h/h_h and g/g_h indexes. This is particularly advisable in fields with low coverage in available bibliographic databases or with a significant portion of nonjournal literature, such as organismic zoology and botany, geosciences, social sciences, and humanities and arts (see Moed 2005, p. 126). It would be an additional burden on the shoulders of scientists, but auditors and administrators are unlikely to assume the burden of performing such fair and honest evaluation. If the data are on the table, however, nobody can deny them.

References cited

Bar-Ilan J. 2008. Which *h*-index? A comparison of WoS, Scopus and Google Scholar. Scientometrics 74: 257–271.

Bouchet P. 2006. The magnitude of marine biodiversity. Pages 31–62 in Duarte CM, ed. The Exploration of Marine Biodiversity: Scientific and Technological Challenges. Bilbao (Spain): Fundación BBVA.

Brown LE, Dubois A, Shepard DB. 2008. Inefficiency and bias of search engines in retrieving references containing scientific names of fossil amphibians. Bulletin of Science, Technology and Society 28: 279–288.

Chapron G, Husté A. 2006. Open, fair, and free journal ranking for researchers. BioScience 56: 558–559.

Egghe L. 2006. Theory and practice of the *g* index. Scientometrics 69: 131–152.

Moed HF. 2005. Citation Analysis in Research Evaluation. Dordrecht (Netherlands): Springer.

Schreiber M. 2008. The influence of self-citation corrections on Egghe's *g* index. Scientometrics 76: 187–200.

Frank-Thorsten Krell (e-mail: frank.krell@dmns.org) is Curator of Entomology at the Department of Zoology, Denver Museum of Nature & Science, in Colorado.

Downloaded From: https://bioone.org/journals/BioScience on 29 Mar 2024
Terms of Use: https://bioone.org/terms-of-use