Science & Society

EMBO reports

Scientometrics in a changing research landscape

Bibliometrics has become an integral part of research quality evaluation and has been changing the practice of research

Lutz Bornmann[1] & Loet Leydesdorff[2]

Quality assessments permeate the entire scientific enterprise, from funding applications to promotions, prizes and tenure. Their remit can encompass the scientific output of individual scientists, whole departments or institutes, or even entire countries. Peer review has traditionally been the major method used to determine the quality of scientific work, either to arbitrate if the work should be published in a certain journal, or to assess the quality of a scientist's or institution's total research output. Since the 1990s, quantitative assessment measures in the form of indicator-supported procedures, such as bibliometrics, have gained increasing importance, especially in budgetary decisions where numbers are more easily compared than peer opinion, and are usually faster to produce. In particular, quantitative procedures can provide important information for quality assessment when it comes to comparing a large number of units, such as several research groups or universities, as individual experts are not capable of handling so much information in a single evaluation procedure. Thus, for example, the new UK Research Excellence Framework (REF) puts more emphasis on bibliometric data and less on peer review than did its predecessor.

Even though bibliometrics and peer review are often thought of as alternative methods of evaluation, their combination in what is known as informed peer review can lead to more accurate assessments: peer reviewers can enhance their qualitative assessment on the basis of bibliometric and other indicator-supported empirical results. This reduces the risk of distortions and mistakes as discrepancies between the peers' judgements and the bibliometric evaluation become more transparent. Although this combination of peer review and bibliometrics is regarded as the ideal method for research evaluation, the weighting of both can differ. The German Research Foundation (DFG), for example, encourages applicants to submit only their five most relevant publications, which is a manageable number for the reviewers. On the other side, the Australian Research Council (ARC) and the UK REF focus on bibliometric instruments for national evaluations to the detriment of peer review. The weighting of the two instruments can also change over time: the new REF weights bibliometrics higher than the former Research Assessment Exercise.

Bibliometrics has various advantages that make it suitable for the evaluation of research. The most important one is that bibliometrics analyses data, which concerns the essence of scientific work. In virtually all research disciplines, publishing relevant research results is crucial; results that are not published are usually of no importance. Furthermore, authors of scientific publications have to discuss the context and implications of their research with reference to the state of the art and appropriately cite the methods, data sets and so on that they have used. Citations are embedded in the reputation system of research, as researchers express their recognition and the influence of others' work.

Another advantage of using bibliometrics in research evaluation is that the bibliometric data can be easily found and assessed for a broad spectrum of disciplines using appropriate databases: for example, Web of Science (WoS) or Scopus. The productivity and impact even of large research units can therefore be measured with reasonable effort. Finally, the results of bibliometrics correlate well with other indicators of research quality, including external funding or scientific prizes [1,2]. Since there is now hardly any evaluation that does not count publications and citations, bibliometrics seems to have established itself as a reliable tool in the general assessment of research. Indeed, it would not last long if reputations and awards based on bibliometric analyses were arbitrary or undeserved.

However, bibliometrics also has a number of disadvantages. These, though, do not relate to its general applicability in research evaluation—this is no longer doubted—but relate to whether such an analysis is done professionally according to standards [3], which are often known only to experts.

> *"... bibliometrics can only be applied to disciplines where the literature and its citations are available from appropriate databases."*

First, bibliometrics can only be applied to disciplines where the literature and its citations are available from appropriate

1 Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Munich, Germany. E-mail: bornmann@gv.mpg.de
2 Amsterdam School of Communication Research (ASCoR), University of Amsterdam, Amsterdam, The Netherlands. E-mail: loet@leydesdorff.net
DOI 10.15252/embr.201439608 | Published online 11 November 2014

databases. While the natural sciences are well-represented in such databases, the literature of the technical sciences, the social sciences, and the humanities (TSH) are only partly included. Bibliometrics can therefore only yield limited results for these disciplines. Google Scholar is often seen as a solution, but it is not clear what Google Scholar considers as a citation; the validity of the data is therefore not guaranteed [4].

Second, bibliometric data are numerical data with highly skewed distributions. Their evaluation therefore requires appropriate statistical methods. For example, the arithmetic mean is relatively inappropriate for citation analysis, since it is strongly influenced by highly cited publications. Thus, Göttingen University in Germany achieved a good place in the current Leiden ranking, which uses a mean-based indicator, because it could boast one extremely highly cited publication in recent years. The Journal Impact Factor—the best known indicator for the importance of journals—is similarly affected by this problem: since it gives the average number of citations for the papers in a journal during the preceding 2 years, it may be determined by a few highly cited papers and hardly at all by the mass of papers, which are cited very little or not at all.

The h-index—a bibliometric indicator which is now similarly well known as the Journal Impact Factor—is unaffected by this problem, as it is not based on the mean. Rather, it measures the publications in a set with a specific minimum of citations (namely h) so that the few highly cited publications play only a small role in its calculations. The h-index, however, has other weaknesses that make its use in research evaluation questionable; the

arbitrary limit for the selection of the significant publications with at least h citations is criticised; it could just as well be h^2 citations.

Third, citations need time to accumulate. Research evaluation on the basis of bibliometrics can therefore say nothing about more recent publications. It has now become standard practice in bibliometrics to allow at least 3 years for a reliable measurement of the impact of publications. This disadvantage of bibliometrics is chiefly a problem with the evaluation of institutions where the research performance of recent years is generally assessed, about which bibliometrics—the measurement of impact based on citations—can say little. In the assessment of recent years, one can only use bibliometric instruments to evaluate the productivity of the researchers of an institution and their success in publishing their manuscripts in respected journals.

Here, the most important question is how long the citation window should be to achieve reliable and valid impact measurement. There are many examples where the importance of research results has become apparent only decades after publication [5]. For example, the "Shockley–Queisser limit" describes the limited efficiency of solar cells on the basis of absorption and reemission processes. The original reception of the paper was rather timid, but today, it has become one of the relatively few highly cited papers in a field that has developed relatively synchronously with rapidly growing solar-cell and photovoltaic research.

"There are many examples where the importance of research results have become apparent only decades after publication"

Although such papers constitute probably one in every 10,000 papers [5], the standard practice of using a citation window of only 3 years nevertheless seems to be too small. In one study, of the 10% of highest cited papers identified using a 30-year window, more than 40% are excluded from this elite collection when a 3-year window is used [6]. When a 20-year window is used, 92% are still included, and a 10-year window yields 82% of the 30-year highest cited papers. Based on his results, Wang recommends that researchers should report "the potential errors in their evaluations when using short-time windows, providing a paragraph such as: 'Although a citation window of 5 years is used here, note that the Spearman correlation between these citation counts and long-term (31 years) citation counts will be about 0.87. Furthermore, the potential error of using a 5-year time window will be higher for highly cited papers because papers in the top 10% most cited papers in year 5 have a 32% chance of not being in the top 10% in year 31'" [5].

This tendency to focus on the citations of papers published during the last 2 or 3 years assumes a rapid research front, as in the biomedical sciences. However, disciplines differ in terms of the existence and speed of research fronts and their historical developments. A recent study has distinguished between "transitory knowledge claims" in research papers at the research front and "sticky knowledge claims" that may accumulate citations during ten or even more years [7].

As bibliometrics has developed into a standard procedure in research evaluation, with both advantages and disadvantages, a further question is now whether bibliometric measurement and assessment is likely to change scientific practice, as fixing on particular indicators for measuring research performance generally leads to an adaptation of researchers' behaviour. This may well be intentional: one reason for research evaluation is to increase research performance, namely productivity. However, there are also unintended effects. For example, in order to achieve a desired increase in publication volume, some researchers choose a publication strategy known as salami slicing: The results of a research project are published in many small parts, although they could also be published in a few large papers or a single one. This behaviour is not generally considered to help the progress of research, but it may improve bibliometric scores.

It is also desirable for researchers to publish in respected journals. Yet since these journals only publish newsworthy results or results with a possible high impact, a stronger focus on respected journals in research evaluation raises the risk of scientific malpractice when results are manipulated or falsified to satisfy this requirement. The risk of this behaviour should not be unreasonably increased by research evaluation processes, in which, for example, scientists in China are sometimes financially rewarded according to the Impact Factors of the journals in which they publish their papers [8].

In national scientific systems, in which research evaluation or bibliometrics plays a major role, indicators are often used without sufficient knowledge of the subject. Since the demand for such figures is high and the numbers are often required speedily or inexpensively, they are sometimes produced by analysts with little understanding of bibliometrics. For example, such amateur bibliometricians may be inclined to use the h-index because it is a popular and modern indicator that is readily available and easy to calculate. Yet, these assessments often do not take into account that the h-index is unsuitable for comparing researchers from different subject areas and with different academic ages. Amateur bibliometricians also often wrongly use the Journal Impact Factor to measure the impact of single pieces of work, although the Journal Impact Factor only provides information about the performance of a journal.

"..., a further question is now whether bibliometric measurement and assessment is likely to change scientific practise..."

There is a community of professional experts in bibliometrics who develop advanced indicators for productivity and citation impact measurements. Only experts from this community should undertake a bibliometric study that involves comparisons across fields of science. These centres of professional expertise have generated analytical versions of the databases and can be found, for example, at the Centre for Science and Technology Studies (CWTS, Leiden) or the Centre for Research & Development Monitoring (ECOOM, Leuven).

Fourth, a range of suppliers of bibliometric data, such as Elsevier or Thomson Reuters, have developed research evaluation systems that allow decision-makers to produce results about any given

research unit at the press of a button. This "desktop bibliometrics" also increases the risk that such analyses are applied without sufficient knowledge of the subject. Furthermore, these systems often present themselves as a black box: the user does not know how the results are calculated; but even simple indicators such as the h-index can be calculated in different ways. This is why the results of bibliometric analyses do not always correspond to the current standards in bibliometrics.

> *"The state no longer has faith that excellent research alone is automatically best for society."*

Fifth, bibliometrics can be applied well in the natural sciences, but its application to TSH is limited. Even if research in these disciplines is published, these publications and their citations are only poorly represented in the literature databases that can be used for bibliometrics. The differing citation culture—in particular the different average number of references per paper and thereby the different probability of being cited—is widely regarded as the cause of this variation. Based on an analysis of all WoS records published in 1990, 1995, 2000, 2005 and 2010, however, a study found that almost all disciplines show similar numbers of references in the reference lists [9]. This suggests that the comparatively low citation rates in the humanities are not so much the result of a lower average number of references per paper, but caused by the low fraction of references that are published in the core set of journals covered by WoS.

Furthermore, the research output in TSH is not only publications, but other products such as software and patents. These products and their citations are hardly reflected in the literature databases. Thus, for example, a large part of the publications and other research products from the TSH area are missing from the Leiden University Ranking, which is based on data in WoS. Even the indicator report of the German Competence Centre for Bibliometrics (KB), which assesses German research based on bibliometric data from WoS, underrepresents publications from the TSH areas.

So far, scientometric research has developed no satisfactory solution to evaluate TSH in the same sophisticated way that is used for the natural sciences. Various initiatives have therefore tried to develop alternative quality criteria. For example, the cooperative project "Developing and Testing Research Quality Criteria in the Humanities, with an emphasis on Literature Studies and Art History" of the Universities of Zurich and Basel, supplies Swiss universities with instruments to measure research performance and compare research performance internationally.

Until the 1990s, politicians had faith that pushing the quality of science to the highest levels would automatically generate returns for society. Quality controls in research were primarily concerned with the use of research for research. Triggered by the financial crisis and by growing competition between nations, the direct societal benefits of research have moved increasingly into the foreground of quality assessments. The state no longer has faith that excellent research alone is automatically best for society. Basic research in particular has become subject to scrutiny, since it is more difficult to show a link between its results and beneficial applications. Recent years have therefore seen a tendency to implement evaluation procedures that attempt to provide information on the societal impacts of research. For example, applicants to the US National Science Foundation have to state what benefits their research would bring beyond science. As part of the UK REF, British institutions also have to provide information about the societal impacts of their research.

> *"... productivity no longer only means publication output, and the impact of publications can no longer be equated simply with citations"*

Evaluating the societal impacts of research does not stop at the traditional products of research, such as prizes or publications, but includes other elements such as software, patents or data sets. The impact itself is also measured more broadly to include effects on society and not just on research. However, there are still no accepted standard procedures that yield reliable and valid information. Often, a case study is carried out in which an institution describes one or several examples of the societal impacts of its research. The problem is that the results of case studies cannot be generalised and compared owing to a lack of standardisation.

So-called altmetrics—the number of page views, downloads, shares, saves, recommendations, and comments from social media platforms, such as Twitter, Mendeley and Facebook—could provide a possible alternative to bibliometric data. A perceived advantage of altmetrics is the ability to provide recent data, whereas citations need time to accumulate. Another perceived advantage is that alternative metrics can also measure the impact of research in other sectors of society, as social media platforms are used by individuals and institutions from many parts of society.

However, it is not clear to what extent these advantages—speed and breadth of impact—really matter. The study of altmetrics began only a few years ago and is now in a state similar to that of research into traditional metrics in the 1970s. Before alternative metrics can be applied to research evaluation—with possible effects on funding decisions or promotions—there are a number of open questions. What kind of impact do the metrics measure, and with what category of persons? How reliable are the data obtained from social media platforms? How can the manipulation of social media data by users be counteracted or prevented? Finally, metrics need to be validated by correlating them with other indicators: is there, for example, a connection between alternative metrics and the judgment of experts as to the societal relevance of publications?

This new challenge of measuring the broad impact of research on society has triggered a scientific revolution in scientometrics. This assertion is based on a fundamental change in the taxonomy of scientometrics: productivity no longer only means publication output, and the impact of publications can no longer be equated simply with citations. Scientometrics should therefore soon enter a phase of normal science to find answers to the questions mentioned above. Such corresponding alternative indicators should be applied in research evaluation only after altmetrics has been thoroughly scrutinised in further studies.

It is clear that scientometrics has become an integral part of research evaluation and

plays a crucial role in making decisions about national research policies, funding, promotions, job offers and so on, and thereby on the careers of scientists. Scientometrics therefore has demonstrated that it provides reliable, transparent and relevant results, which it largely achieves with citation-based data if it is done correctly. The next challenge will be to develop altmetrics to the same standards.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Diekmann A, Naf M, Schubiger M (2012) The impact of (Thyssen)-awarded articles in the scientific community. *Kölner Z Soz Sozialpsychol* 64: 563 – 581
2. Luhmann N (1992) *Die Wissenschaft der Gesellschaft*. Frankfurt am Main, Germany: Suhrkamp
3. Bornmann L, Marx W (2014) How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics* 98: 487 – 509
4. Bornmann L, Marx W, Schier H, Rahm E, Thor A, Daniel HD (2009) Convergent validity of bibliometric Google Scholar data in the field of chemistry. Citation counts for papers that were accepted by Angewandte Chemie International Edition or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts. *J Inform* 3: 27 – 35
5. van Raan AFJ (2004) Sleeping beauties in science. *Scientometrics* 59: 467 – 472
6. Wang J (2013) Citation time window choice for research impact evaluation. *Scientometrics* 94: 851 – 872
7. Baumgartner SE, Leydesdorff L (2014) Group-based trajectory modeling (GBTM) of citations in scholarly literature: dynamic qualities of "transient" and "sticky knowledge claims". *J Assoc Inf Sci Technol* 65: 797 – 811
8. Shao J, Shen H (2011) The outflow of academic papers from China: why is it happening and can it be stemmed? *Learned Publishing* 24: 95 – 97
9. Marx W, Bornmann L (2014) On the causes of subject-specific citation rates in Web of Science. *Scientometrics* (in press)

