Jointly published by Akadémiai Kiadó, Budapest and Springer, Dordrecht

Scientometrics, Vol. 62, No. 1 (2005) 117–131

Impact of bibliometrics upon the science system: Inadvertent consequences?

PETER WEINGART

Institute for Science & Technology Studies, University of Bielefeld, Bielefeld (Germany)

The introduction of bibliometric (and other) ranking is an answer to legitimation pressures on the higher education and research system. After years of hesitation by scientists, science administrators and even politicians in many of the industrialized countries, the implementation of bibliometrics based (and other types of) rankings for institutions of higher education and research is now being introduced on a full scale. What used to be an irritation to the parties concerned has suddenly become a fad. In contrast to this rather sudden enthusiasm, there is very little reflection on the impacts of this practice on the system itself. So far empirical data on the impact of bibliometric rankings seem to be available only for two cases: Australia and the British research assessment exercise (RAE). Thus, the actual steering effects of bibliometric rankings, the reactions of the system are largely unknown. Rankings are in urgent demand by politics. The intended effect is to create competition among institutions of higher learning and research and thereby to increase their efficiency. The rankings are supposed to identify excellence in these institutions and among researchers. Unintended effects may be 'oversteering', either by forcing less competitive institutions to be closed down or by creating oligopolies whose once achieved position of supremacy cannot be challenged anymore by competitors. On the individual level the emergence of a kind of 'chart' of highly cited stars in science can already be observed (ISI HighlyCited.com). With the spread of rankings the business administration paradigm and culture is diffused through the academic system. The commercialization of ranking is most pronounced in the dependence of the entire practice on commercial providers of the pertinent data. As products like ISI's Essential Science Indicators become available, their use in the context of evaluation tasks is increasing rapidly. The future of the higher education and research system rests on two pillars: traditional peer review and ranking. The goal must be to have a system of informed peer review which combines the two. However, the politicized use of numbers (citations, impact factors, funding etc.) appears unavoidable.

The evaluation craze out of control?

When the first bibliometric based evaluations of research institutions were carried out – by Martin and Irvine in 1983 in the UK – the reaction of the scientists concerned was predictable. They challenged the possibility of the enterprise on methodological grounds, and they threatened to take the analysts to court because they feared that the

Received August 17, 2004

Address for correspondence:
PETER WEINGART
Institute for Science & Technology Studies, University of Bielefeld
P. O. Box 100131, D–33501, Bielefeld, Germany
E-mail: weingart@uni-bielefeld.de

0138–9130/2005/US $ 20.00

results would have adverse effects (WEINGART, 2001: p. 316). The reaction was predictable because first of all the very attempt to measure research performance by 'outsiders', i.e. non-experts in the field under study conflicted with the firmly established wisdom that only the experts themselves were in the position to judge the quality and relevance of research and that the appropriate mechanism to achieve that, namely peer review, was functioning adequately. The second reason for scepticism if not outright rejection was the methodology employed. Bibliometric measures, although quantitative and therefore seemingly objective, appeared to be theoretically unfounded, empirically crude, and dependent on data that were known to be imprecise. The rejection of bibliometric indicators on the part of the scientific community was supported by policy makers and government administrators, although mostly because of disinterest.

Since then times have changed in several respects. As budgets for research have leveled off and priority decisions re-distribute rather than add funds the pressure to legitimate such decisions has focused interest on measures that do not involve policy-makers in experts' arguments that they are unable to engage in. First the focus, at least in the German higher education system, was on the regulation of teaching loads and student flows by numerical formula, implemented in the 1970s. That will not be of concern here but as a historical example is indicative because for the first time it demonstrated that the seemingly complex world of teaching with its different subjects, types of instruction and levels of qualification could be regulated by the application of a few crude numbers. Of course, here the matching of student numbers and teaching capacities and thus ultimately the control over the number of staff was the objective. Although indicators of research began to be developed in the 1970s as well, they were not implemented until a little further down the line when the assessment of departments, of individual researchers, and the ranking of universities became an important instrument for the competitive allocation of funds replacing the supposedly more costly system of block grants.

Indicators of research quality are not yet generally accepted. The US government, despite its bent on performance indicators for the rationalization of budgetary decisions, does not use bibliometric measures of research (ROESSNER, 2002; FELLER, 2002).* In the EU the situation is very mixed with various degrees of institutionalization of bibliometric indicators. The extreme is probably represented by Finland, "the only country in which the journal impact factor has been canonized in the law of the land," implying that the publication of just one paper in a higher impact journal can boost the budget of a university hospital by about US$ 7000 (ADAM, 2002: p. 727). But the lure of quantitative measures appears to be increasingly attractive to other governments as well. By way of a mix of copying other examples, outside pressure

* Private assessment by S. Cozzens. Roessner's and Feller's articles give an overview of performance indicators for the evaluation of S&T programs in the US in general.

from a spreading accountability culture and mutual observation of actors one can now witness internationally a dramatic shift away from the well founded scepticism to an uncritical embrace of bibliometric numbers. This change of mind is not limited to policy makers and administrators but has taken hold of deans, department chairmen, university presidents and officials in funding agencies and research councils as well, i.e. of representatives of the scientific community that were most strongly opposed to external evaluation of research with any means.

This new demand for numbers unlocking the secrets of the world of research and internal allocation of prestige and rewards, allowing outsiders a direct look at the international standing or provincial isolation of their local scientists, thus giving them the power to dismantle unfounded claims to fame, has brought many players into a rapidly growing market of research evaluation and bibliometric analyses in particular. Several countries have set up their own institutions to collect and process data on the performance of their own research installations. Others use any one of the independent and either university-based or commercial institutes or research groups specializing in bibliometric studies to do particular or routine evaluations for them. In the US the NSF/NSB Science Indicators Report, published since the 1970s, was the first to contain bibliometric output indicators. France has set up its 'Observatoire des Sciences et des Techniques' (OST), and so have the Netherlands (NOWT). Both the Swiss and the German Science Councils make use of bibliometric indicators in their reports. The focus of these and other agencies' reporting is primarily the national science systems.

All of them are up to now and for the foreseeable future dependent on one single provider of data, the Institute of Scientific Information (ISI), the producer of the only multidisciplinary databank of scientific literature that contains citation data and thus allows the compilation of citation counts and impact factors of journals as well as the development of more sophisticated measures such as co-citation maps. While originally conceived as a literature databank designed to identify uses of knowledge and networks of researchers, ISI's database soon proved its value as a tool for sociology and history of science research as well as for the evaluation of research institutions and even of individual researchers.

After many years of somewhat reluctant response to this sideline use of their products, ISI has now recognized the growing importance of the demand for bibliometric indicators and has begun producing tailor-made evaluation tools such as ISI Essential Science Indicators and ISI Highly Cited Com. These are powerful tools that allow anyone with an Internet access to a university library to identify the highly cited scientists of their local university, the relative impact of that university compared to others in the country or internationally, or the rank of that university in a particular field and so on. These tools are now actively marketed, and a growing demand contributes to their rising price.

The new owner of ISI, Thomson Company of Toronto, aggressively commercializes them thereby promoting direct use by anyone willing to pay the fees. This policy has at least two far reaching effects. First, the intermediary research groups that hitherto cleaned the crude ISI data, prepared the data for specified purposes and developed the skills to interpret them are being squeezed out of that market. Once that will have happened the competence of these groups, pertaining to knowledge of regional or national institutions, of language and, thus, of names will be lost. So will be their skills in constantly refining the indicators by doing research on their applications in evaluation procedures. Above all, this will convey the image that the data are correct and do not need any costly cleaning. Second, the ready availability of such seemingly exact indicators suggests that any layperson can evaluate researchers and their products. In fact, however, their methodological and operational origins are concealed from the end user who is not able to reflect upon the theoretical assumptions implied in their construction. This has led to a growing number of incidents in which administrators in government science policy and higher education agencies refer to these data when negotiating budget decisions, or when department chairs use them for recruiting and salary decisions. As the discourse on the accountability of science and the evaluation of research institutions picks up momentum demand for ISI's 'off the shelf' indicator packages is growing. The healthy scepticism of years ago, albeit often for the wrong reasons, appears to have given way to an uncritical embrace of bibliometric measures and to an irresponsible use.

The implications of this development are disquieting, at least. The evaluation process that was hitherto internal to science, i.e. peer review, has been 'externalized', i.e. made accessible to the lay public by proxy, namely numbers reflecting the quantitative aspects of the communication process in science. These numbers become the basis of budgetary decisions directly affecting the research process as well as the operation of universities, of clinics and other research institutions dependent on public funds. The production of these numbers is in the hands of a commercial company that presently holds a virtual world monopoly on them and, whether conscious of it or not, structures political decisions affecting research systems all over the world by the profile and the quality of the data it provides to its customers. The evaluation of research, and the budgeting of university departments based on it, to the extent that they depend on bibliometric data, have effectively been handed over to a private company with commercial interests.[*] This makes the critical examination of the validity and reliability

[*] There are presently no serious efforts anywhere to challenge ISI's position. Cf. *Nature*, 415, 14. Feb., 2002, 728. One implication of Thomson's status as a private company may be that the pressure to market data prematurely increases and the quality of the data decreases. Cf. for such a comment S. Müller, Das Monopol, *Deutsche Universitätszeitschrift (DUZ)*, 21 (2003) 10–11.

of ISI's data as well as of the uses made of them especially by governments and of the unintended steering effects of their use a task of paramount importance – in the interest of both governments and the scientific community.

Validity and reliability of bibliometric indicators in the evaluation of institutions and individuals

Since their inception questions of validity and reliability of bibliometric indicators have been the concern of researchers engaged in the development of such indicators. These formerly academic questions become an urgent issue of policy making as indicators are being implemented and tied to budgetary decisions, i.e. that so called evaluation based funding (EBF) is expanded. In addition to these traditional concerns linked to the construction of any policy relevant indicator arises another one: the unintended and/or the adaptive effects of the actual application of these indicators. Especially the former concerns are raised in relation to peer review. However, part of the reason for the increased popularity of quantitative bibliometric indicators among public officials is the growing scepticism and disenchantment with peer review. Initial doubts about its openness (reaching back to the 1970s in the US) triggering allegations of 'old boy networks' have been seconded by a number of fraud scandals reaching high up into the elite layers of the biomedical and physics establishments. Due to both public critique of the scientific community's self control and to political pressures to re-distribute research funds on the basis of large scale evaluations, the availability and practicality of numerical indicators that promise greater transparency and objectivity become very attractive to policymakers. In effect the trust lost by the peer review mechanism has been shifted to the use of numerical indicators. This is, of course, tantamount to the loss of autonomy for the scientific community and to greater involvement of the political public in the direction of its affairs.

The peer review process, especially the reliability and consistency of peer evaluations, have been the target of many empirical analyses. The most active disciplines in terms of the concern about the functioning of their own peer review are the medical sciences and psychology. Recently the physicists have joined them.[*] The findings were, indeed, not encouraging. Different approaches to testing and measuring the reliability of the judgments of peers both in decisions about research proposals to funding agencies and about articles to be published always reveal the same results: Peer evaluations diverge, they contradict each other, and they do not remain consistent over time. Cichetti in a review of a multitude of studies concludes that the reviewers of

[*] Four international conferences on peer review in the medical sciences were organized by *JAMA* in the 1990s. Cf. for an overview of similar activities and analyses of peer review HIRSCHAUER (2002). As an aside: it is an interesting question why these disciplines are especially concerned about their peer review mechanism.

research proposals have more agreement about which proposals not to fund than about which ones to support. In the review of articles for journals it is the other way around: Reviewers agree more about acceptance than about rejection of articles (CICHETTI, 1991; BAKANIC et al., 1989). Others have pointed to the considerable role of extraneous factors such as sheer luck or being well integrated into the right networks, or belonging to the right institution (COLE et al., 1981).

Upon closer inspection, however, these findings are not surprising given the nature of the scientific communication process of which peer review is an integral part. The process is open, controversial and ongoing. Differences of opinion are essential for the productivity and innovativeness of the process and for preventing the undue dominance of just one opinion. Unanimity would be the exception and, consequently, is very rare until a particular research question is settled and the researchers' attention is directed elsewhere. The expectation of unanimous evaluations stems from a 'disappointed scientism' coupled to issues of justice (HIRSCHAUER, 2002). As the basis for the various critiques of peer review in general and for justifying the use of bibliometric indicators in particular, this expectation creates the wrong benchmark.

Why is this relevant to evaluations based on bibliometric measures? First of all, peer review "remains the backdrop against which all other types of research evaluation appear, and often the standard against which their validity is judged" (ROESSNER, 2002: p. 86). To the extent that the introduction of these measures was and still is based on the distrust towards peer review, it is mistaken for two reasons:

1. It assumes that these measures are independent of the peer review process.
2. It assumes that they are more exact than peer review because, being quantitative, they appear to be more objective.

In fact, publication and citation measures are representations of the communication process as it unfolds in journal publications. Thus, they also embody the peer review evaluations that have led to actual publications. For that very reason they cannot be more exact or objective than peer review judgments.

The actual advantages of evaluations based on bibliometric measures over peer review are on a different level. 1. The measures on which the evaluations are based are 'non-reactive', i.e. the results are based on a large number of 'incidents' (publications and citations involving the decisions of reviewers). These decisions are not motivated by their being counted for purposes of evaluation. 2. The measures are usually based on a much larger number of such 'incidents' than a limited review process is. Therefore they allow for a broader perspective most likely eliminating personal biases due to limitations of personal knowledge. It is these properties that justify the use of bibliometric indicators.

However, these undoubted advantages are not absolute. They are compromised by the large scale implementation of evaluation schemes linked to bibliometric indicators

and collective reactions to them (cf. next section). In addition to these fundamental considerations, policy-makers and science administrators must be aware of the methodological and technical problems that are attached to the use of bibliometric indicators. A few commonly known ones can illustrate this.

Since the indicators are often (and preferably) based on a large volume of accumulated data, they contain data processing mistakes, and since they are selective for certain journals as well they only represent selective parts of the whole communication process. Depending on the data base these measures may entail biases towards countries, disciplines, and journals (BRAUN et al., 2000; ZITT et al., 2003).

Another problem is that of definitions of fields. In certain cases publications are excluded because the definition of a field in the data bank based on a particular journal set is incomplete or overlaps with other definitions. In particular, interdisciplinary fields present a problem to proper categorization. A seemingly clearly defined research field like 'high temperature superconductivity' has connections to 'low temperature' and 'solid state physics', 'physical chemistry', 'materials science' and 'thin film preparation' that make a clear cut delineation of the field impossible. Thus, such problems of delineation of disciplines may ultimately lead to mistaken citation counts.[*]

Furthermore, and generally speaking, too little is known about the use of citations in the scientific communication process, positive, negative, or perfunctory (CASE & HIGGINS, 2000; CRONIN, 2000). For the time being the application of citation indicators has to be based on the conviction emerging from a number of studies that, given sufficiently large numbers, different motives for citing an article neutralize each other. What remains is the attention to the piece cited. We also know that different disciplines have developed very different customs of citing. Articles in basic biomedical research are being cited six times more often than articles in mathematics. Such regularities have to be taken into consideration when comparisons between institutions across disciplinary lines are undertaken. An accepted theory of citation decisions, however, on which the better informed use of citation indicators could be based is lacking and may never be achieved (VAN RAAN, 1998; SMALL, 1998).

Finally, an additional problem arises from a statistical viewpoint. In many evaluations based on citation counts, especially those of individuals or small institutions, the numbers are small. Single digit differences of citations may be due to the time window chosen, they may depend on the particular position of papers in the communication under way, on the amount of time an article has had a chance to be cited, and thus they may change rapidly. In institutional evaluations and rankings the relatively small number of citations involved can lead to 'extreme' cases such as that

[*] This refers to an incident when a ranking was published in a popular science journal and drew criticism from scientists who could make the case of their having been ranked unjustly. The subsequent attempt to have a methodological note published in the same journal was rejected as 'not being of interest to the readers' (WEINGART, 1993).

one highly cited publication may decide the relative position of a respective institution regardless of the 'quality distribution' throughout its staff compared to others. Needless to say that the author of that paper may have left the institution a long time ago while its rank is still on record. Small differences or differences based on small numbers cannot justify budgetary or salary decisions because they do not reliably indicate meaningful differences of competitive effort, of productivity and even less so of quality of institutions or individuals.

The general conclusion to be drawn from these insights is common knowledge among researchers and evaluators who are experienced in using bibliometric measures: they can only be applied on a high level of aggregation, they must be carefully constructed with respect to precise questions, and they must be interpreted with great care with the technical and methodological problems in mind.

How does this relate to peer review? The use of bibliometrics can have a beneficial effect on the peer review process in several respects. Precisely because bibliometric measures are based on mass data they reveal macro-patterns in the communication process that cannot be seen from the highly limited and selective perspective of the individual researcher. Bibliometrics can 'inform', for example, about the unsuspected connection between research fields that are not yet institutionally connected. The unique contribution of bibliometrics to the collective communication process in science and their greatest value to the scientific community itself as well as to policy makers and the public is in providing this 'greater picture'. However, bibliometric analysis and evaluation do not replace peer review for an obvious reason: The interpretation of these patterns, of unexpected contradictions to the common wisdom of the community or other irregularities must be left to the experts in the respective fields or at least assisted by them. Peer judgment must complement bibliometric analyses wherever necessary.

Another highly important function of bibliometrics in peer review may be to 'control' and thereby 'strengthen' peer review. The rapid decline of attention for a research field that had been prominent before and whose institutional dominance tends to protract past relevance may be likely to escape the review process because of its inherent selectivity and/or vested interests involved. Peer review judgments (especially in policy related evaluative contexts) that are counter-checked by bibliometric studies are better protected against the operation of 'old boy networks' which, in turn, will strengthen the outside credibility of the mechanism.

Intended and unintended steering effects of bibliometric measures

All concerns about the validity and reliability of bibliometric measures are academic as long as they remain research tools, but they are of high political significance once these measures are implemented as indicators on which distributive decisions are based. One crucial problem then is if they attain their objectives as tools of policy making. Do

individuals and institutions react in the way intended by the application of bibliometric (and other) measures or do they in some ways evade or circumvent the intended goals?

Bibliometric indicators, when applied in conjunction with budgetary decisions and other types of sanctions, inadvertently become so-called reactive measures. That means when they affect people, these react to the implementation of such measures by altering their behavior. Behavior change is intended. For example, the link of citation measures to the allocation of funds is supposed to induce researchers to engage in more competitive publication routines in order to increase their publication activity and publish their papers in high impact factor journals. In many cases funding formulas are linked to more than just one indicator combining, for example, bibliometric measures with received external grants as indicators. The latter is intended to induce researchers to apply for research grants. A further indicator sometimes entered into funding formulas supposed to measure research quality, the number of doctoral students supervised, is intended to achieve a greater output of PhDs. Sir Gareth Roberts, president of Wolfson College, Oxford, sees the reform of the British Research Assessment Exercise having to go exactly in this direction: "Figures such as the number of doctorates produced, external research income and number of papers produced could be used as proxies for research quality to work out how much research funding a university should receive" (ROBERTS, 2003).

Each of these indicators assumes a one-dimensional mode of reaction or an incentive compatibility, but that assumption is illusionary. Researchers can and are known to increase their publication count by dividing their articles to a 'least publishable unit', they can propose relatively conservative but safe research projects, and they can lower their standards for their PhD candidates. These are just examples how individuals can manipulate indicators or evade their intended steering effects. As a commentary in *Nature* notes: "Scientists are increasingly desperate to publish in a few top journals and are wasting time and energy manipulating their manuscripts and courting editors. As a result, the objective presentation of work, the accessibility of articles and the quality of research itself are being compromised" (LAWRENCE, 2003a: p. 259). What is true for individuals is also true for institutions, they can do the same. Obviously, the effectiveness of research policy employing evaluative indicators depends entirely on the sound theoretical base of the indicators and on the requisite knowledge about the reactions they trigger among the individuals and organizations whose behavior they are supposed to change.

So far only very few studies have been undertaken to identify the effectiveness and unintended reactions of this kind to bibliometric measures as well as secondary consequences for the university or the communication process in science as a whole. Sociology of science and ethnographic studies show that scientists do, indeed, react to non-epistemic influences (GLÄSER et al., 2002: p. 16). An Australian study showed that upon the implementation of formula based funding, i.e. in that case, linking the number

of publications in peer reviewed journals to funding, the number of publications, indeed, went up, but the quality of the papers had not increased as measured by citations. "With no attempt made to differentiate between quality, visibility or impact of the different journals when funding is allocated, there is little incentive to strive for publication in a prestigious journal" (BUTLER, 2003: p. 41). The obviously one-dimensional incentive set by policy led to foreseeable counter-productive reactions.

The Spanish National Commission for the Evaluation of Research Activity (CNEAI) rewards individual researchers with salary bonuses for publishing in prestigious journals. A study suggests at least the plausible conclusion that the researchers have responded by increasing their research output (JIMÉNEZ-CONTRERAS, 2003: pp. 133, 138). Comparing the Australian with the Spanish experience Butler states that "in the Spanish case CNEAI achieved its stated aims, which were to increase productivity and the internationalisation of Spanish research. In contrast, the Australian funding formulas were designed to reward quality, but in fact reward quantity" (BUTLER, 2003: p. 44). Worse yet, Australia fell even behind nearly all OECD countries.

A comparison between two Australian universities (Queensland and Western Australia) "provides further support to the assumption that the coupling of increasing quantity and decreasing quality is due to the introduction of quantity-based funding formulas" (GLÄSER et al., 2002: p. 14). UWA introduced a quantity of research output based funding formula while UQ sought to improve its status with a recruitment drive for bright young and international researchers. While the UWA status in terms of its relative citation impact (RCI) declined UQ could even increase its RCI significantly (GLÄSER et al., 2002: p. 14).

Another study on changes in universities suggests that there is now a bias in favor of research quantity rather than quality, that there is a bias towards short-term performance, not long-term research capacity, and that there is a bias in favor of conventional approaches (MARGINSON & CONSIDINE, 2000: p. 17 cited in GLÄSER et al., 2002: p. 12). This reflects that under a régime of evaluation-based funding scientists have been found to publish more but less riskful, mainstream rather than borderline papers, and try to place them in lower quality journals as long as they are in the ISI journal index. Under such circumstances publishing has become an end to boost publication counts and to obtain funds, a legitimate but unintended reaction as, e.g. in the Australian case, price tags can be attached to publications: A$ 3000 for an article in a peer reviewed journal, A$ 15000 for a book (BUTLER, 2003: p. 40).

Since detailed citation studies are costly and time consuming, many evaluating bodies have taken a short cut. They "look at scientists' publication records and evaluate the quality of their output in terms of the impact factors of the journals in which their papers appear – figures that are readily available" (ADAM, 2002: p. 727). Impact factors of journals are "the poor man's citation analysis" (van Raan). They are problematic as indicators of research quality when compared between fields because of different

citation practices. They are also unreliable because of the highly uneven distribution of citations in a given journal which means that a certain paper may be published in a high impact journal but receive fewer citations than papers in a less renowned journal. Per Seglen notes that "there is a general correlation between article citation counts and journal impact, but this is a one-way relationship. The journal does not help the article; it is the other way around" (ADAM, 2002: p. 727). Impact factors in their undifferentiated form are outdated and should not be used as measures in any evaluative context at all. Yet, they are probably the most popular bibliometric measure of all. So much so that the journal *Nature* carries promotional flyers with its newest impact factor in great letters and slogans such as 'No *Nature*, no impact'.

On an anecdotal level the editorial policy of journals is accused of being influenced by impact factor considerations. The increase of medically related papers in top biology journals is attributed to "their beneficial effects on the impact factor, rather than for their scientific quality" as is the publication of review articles in specialized journals, since they are "cited more often than research papers" (LAWRENCE, 2003b: p. 836). Thus, it is not surprising that publishers of scientific journals are eager to use favorable impact factors for the promotion of their products. This has led a well known journal in critical care medicine (*Shock*) to an attempted manipulation of the communication process that borders on the absurd. Upon the provisional acceptance of an article the associate editor added that the journal "presently requests that several references to *Shock* are incorporated in the reference list." When the manuscript was sent back with the required revisions the editor insisted that before sending the manuscript to the publisher it would be greatly appreciated "if you could incorporate 4-6 references of appropriate articles that have been published in *Shock* in your revised manuscript urgently. This would be of tremendous help…to the journal". This goes even further: upon publication of the article the author is requested to send copies to colleagues and urge them to cite it.[*] *Shock* is by no means the only journal attempting to boost its impact factor by putting some gentle pressure on its authors. The journal *Leukemia* was even accused of 'manipulation' for the same editorial policy. Eugene Garfield came to its defence, perhaps because he fears for the future of the impact factor. But his argument, that the editors are "justified in asking authors to cite equivalent references from the same journal" as a means to counter the 'Matthew effect' is hardly convincing (SMITH, 1997; GARFIELD, 1997).

Whether successful or not, and however far spread at this time, this kind of practice demonstrates that not only the behavior of individuals but that of organizations may be affected by bibliometric measures in ways that are clearly unintended. Long before they assume the magnitude of structural effects they are warning signs. In the case of the impact factor it is more clearly the fact than with other indicators: "It has evolved to become an end in itself – the driving force for scientists to improve their reputation or

[*] Copies of letters are in the possession of the author.

get a position, and causes damaging competition between journals" (LAWRENCE, 2003b: p. 836). Together with the growing realization of unintended adaptation effects of the British Research Assessment Exercise through a few studies, they are urgent reason to do more thorough research on adaptation processes in reaction to evaluation-based funding schemes in general and the use of bibliometric measures in particular. What effects do they have on the content of knowledge, on the questions asked, the methodologies used, the reliability of results? What effects do they have on the communication process in science, on the mechanisms of organized scepticism, on the attribution of excellence and reputation? In some of the recent cases of fraud or premature speed into publication, the use of bibliometric measures and the resulting pressure to publish have been identified as causing that behavior. If proven to be true this link would be the ultimate evidence that the rush into EBF does more harm than good. It would amount to the fact that the academic culture in which knowledge production thrived on a unique combination of competition, mutual trust and collegial critique is being destroyed. Whether what will emerge in its place will be easier to direct and less costly to sustain is an entirely open question.

The publics of bibliometric indicators

Part of the future culture of knowledge production becomes already visible on the borderlines between the scientific world and the world of information and data production as well as the publishing business and the media. To understand what is happening one has to realize that the evaluation industry that has been created is serving several different publics. One of them are policy makers who have brought this industry into existence and are responsible for its growth by using it as a tool to exert control over the operation of research institutions in the name of public interest. Their motives are legitimated by reference to the public interest that tax funds are spent efficiently and parsimoniously on research serving the needs and interests of that public. Another public are the media that, in turn, refer to the public interest that the operation of research institutions, their relative status and quality, be made transparent to the lay public.

The legitimating power of these publics is best demonstrated by the rhetoric of the public representation of evaluation data by both the producers of these data and the media.

Undue simplification is only one of the problems which arise when, for example, cumulative data of publication counts and of grants appear in the media without any weighting by appropriate factors such as size of institution. Resulting conclusions such as rankings are meaningless and misleading but evidently seem to serve the media's needs to dramatize. Likewise, ISI offers rankings of universities on *Science Watch* under such titles as "Harvard runs high in latest 'Top Ten' Research Roundup". This

ranking is based on citations-per-paper (impact) score for each university in 21 fields, based on papers published and cited between 1997–2001. That figure was compared to world baseline figures representing the impact for the field during the same period. This produces relative impact scores expressed as percentages. Sometimes rankings are based on the hundredth of a decimal point. The exactness suggested by such measures may be a promotional gimmick for ISI's products. Policy makers are confronted with one dimensional rankings which are, in fact, multidimensional. Any superficial attempt at interpreting such rankings without assistance from experts who know how these numbers are created in the first place, and what they represent, is in the context of policy decisions misleading, meaningless and irresponsible. One could even go further and say that it is unethical given the unhealthy combination of the unavoidable limitations of competence on the part of policy makers and media exposure.

On another level ISI employs the language of media hype. The company conveys an image of individual popularity contests by presenting highly cited scientists in its *Science Watch*. Their January/February 2003 headline reads: "Astrophysicist Andrew Fabian on Rocketing to Prominence", this evaluation being based on more than 6000 citations over the last decade. The language of marketing and sensationalized competition has penetrated the hitherto self-contained discourse of peer review. This is not to claim that the scientific community did not know competition before the days of bibliometric indicators, quite the contrary.[*] But it rarely ever had an outside audience nor commentators employing the language of sports events.

One may speculate about the repercussions of this development. It seems highly likely that the orientation to media prominence that is already visible in other contexts will be strengthened. Short term successes such as high positions in rankings that will be watched and commented like the national soccer league, and that may trigger favorable decisions from science councils and university administrations are likely to gain prevalence over more sustained strategies. A metaphor too far fetched? The magazine *Science* commented already in 1997, that "the tactics of soccer managers have taken over the world of higher education". According to the journal's assessment the results of that year's RAE in the UK "revealed how soccer style transfers of researchers and other tactics aimed at improving department's rating are now part of British academic life" (WILLIAMS, 1997: p. 18). This loss of control over its own system-specific time scale and mode of evaluation will probably have a profound long term impact on knowledge production. Unfortunately this will never be known in detail as there will be no possibility to compare.

[*] The most prominent example has been the Watson & Crick story about the discovery of the double helix as told by Watson.

Conclusions

As is often the case technologies may be used wisely or irresponsibly. Bibliometric indicators are a research based social technology, and because they convey knowledge that unlocks an otherwise hidden process to policy makers and the media it is prone to being instrumentalized for all kinds of interests involved in science policy. The evaluation hype that has taken hold of the research and higher education sector has moved the indicators from the niches of academia into a strategic position in policy making. This means that questions of validity and reliability, theoretical foundation and quality of data, assume a political role. Bibliometricians as well as policy makers and administrators have a responsibility for the quality of their instruments and the quality of their uses. The tendencies described above are indications of another reality. The warning, therefore, is against the commercialized marketing of generalized products whose quality is questionable, against the uncritical use of bibliometric measures independent of the peer review process, and against their use without regard for the consequences both individually and institutionally.

At least the following principles should be observed. Bibliometric (and similar) indicators:

1) have to be applied by professional people, trained to deal with the raw data. (ISI's data are not cleaned, and are not fit to perform in depth analyses);

2) should only be used in accordance with the established principles of 'best practice' of professional bibliometrics (VAN RAAN, 1996);

3) should only be applied in connection with qualitative peer review, preferably of the people and institutions being evaluated. The principle is that bibliometric indicators support peer review and can possibly correct it where individual evaluations are confronted with aggregated data and patterns. On the other hand, the bibliometric measures may be corrected by peer review judgment where formal algorithms fail. This conjunction of bibliometric measures with traditional peer review, i.e. so-called 'informed peer review', can serve the legitimate needs of transparency of the general public, and at the same time it retains the expert nature of the judgments that have to be passed. Bibliometric indicators have become such a powerful tool in the context of science policy making and budgetary decisions that their potentially misleading and even destructive use must be acknowledged. By virtue of their potency the application of these indicators warrants a professional code of ethics.

*

This paper is an edited version of the author's contribution to the conference "Bibliometric Analysis in Science and Research" held in Jülich (Germany), on 5-7 November, 2003.

I thank Grit Laudel and Jochen Gläser for their comments on that previous version and above all Matthias Winterhager for his assistance in preparing this paper without which it would not have come about.

References

ADAM, D. (2002), The counting house. *Nature*, 415 (6873) : 726–729.

BAKANIC, V., MCPHAIL, C., SIMON, R. J. (1989), Mixed messages – referees comments on the mauscripts they review. *Sociological Quaterly,* 30 (4) : 639–654.

BRAUN T., GLÄNZEL W., SCHUBERT, A. (2000), How balanced is the Science Citation Index's journal coverage? – A preliminary overview of macrolevel statistical data. In: CRONIN, B., ATKINS, H. B. (Eds), *The Web of Knowledge: a Festschrift in Honor of Eugene Garfield*. Medford, NJ: Information Today Inc. & The American Society for Information Science, pp. 251–277.

BUTLER, L. (2003), Modifying publication practices in response to funding formulas. *Research Evaluation*, 17 (1) : 39–46.

CASE, D. O., HIGGINS, G. M. (2000), How can we investigate citation behavior? A study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, 51 (7) : 635–645.

CICHETTI, D. V. (1991), The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences,* 14 : 119–135. Discussion: 135–186.

COLE, S., COLE, J. R., SIMON, G. A. (1981), Chance and consensus in peer review. *Science*, 214 : 881–886.

CRONIN, B. (2000), Semiotics and evaluative bibliometrics. *Journal of Documentation*, 56 (3) : 440–453.

FELLER, I. (2002), The good, the indifferent, and the less than attractive: II. *Research Evaluation,* 11 (2) : 95–99.

GARFIELD, E. (1997), Editors are justified in asking authors to cite equivalent references from same journal. *British Medical Journal,* 314 : 1765.

GLÄSER, J., LAUDEL, G., HINZE, S., BUTLER, L. (2002), *Impact of Evaluation-Based Funding on the Production of Scientific Knowledge: What to Worry About, and How to Find Out*. Expertise for the German Ministry for Education and Research.

HIRSCHAUER, S. (2002), *Expertise zum Thema "Die Innenwelt des Peer review. Qualitätszuschreibung und informelle Wissenschaftskommunikation in Fachzeitschriften."* Expertise for the German Ministry for Education and Research.

JIMÉNEZ-CONTRERAS, E., DE MOYA ANEGÓN, F., LÓPEZ-CÓZAR, E. D. (2003), The evolution of research activity in Spain: The impact of the National Commission for the Evaluation of Research Activity (CNEAI). *Research Policy*, 32 (1) : 123–142.

LAWRENCE, P. A. (2003a), The politics of publication, *Nature*, 422 (20 March) : 259–261.

LAWRENCE, P.A. (2003b), Rank injustice, *Nature*, 415 (21 February) : 835–836.

MARGINSON, S., CONSIDINE, M. (2000), *The Enterprise University: Power, Governance and Reinvention in Australia*. Cambridge, UK: Cambridge University Press.

MARTIN, B. R., IRVINE, J. (1983), Assessing basic research – some partial indicators of scientific progress in radio astronomy. *Research Policy*, 12 (2) : 61–90.

MÜLLER, S. (2003), Das Monopol, *Deutsche Universitätszeitschrift (DUZ)*, 21 : 10–11.

ROBERTS, G. (2003), *The Guardian,* 7 January.

ROESSNER, J. D. (2002), Outcome measurement in the USA: state of the art. *Research Evaluation*, 11 (2) : 85–93.

SMALL, H. G. (1998), Citations and cinsilience in science. *Scientometrics*, 43 (1) : 143–148.

SMITH, R. (1997), Journal accused of manipulating impact factor. *British Medical Journal*, 314 : 463.

VAN RAAN, A. F. J. (1996), Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises, *Scientometrics,* 36 : 423–435.

VAN RAAN, A. F. J. (1998), In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, 43 (1) : 129–139.

WEINGART, P. (1993), Der Forschungsindex. *Bild der Wissenschaft*, 5 : 32–39.

WEINGART, P. (2001), *Die Stunde der Wahrheit. Zum Verhältnis der Wissenschaft zu Politik, Wirtschaft und Medien in der Wissensgesellschaft*. Weilerswist: Velbrück Wissenschaft.

WILLIAMS N. (1997), UK universities: Jostling for rank. *Science*, 275 (5296) : 18–19.

ZITT M., RAMANANA-RAHARY, S., BASSECOULARD, E. (2003), Correcting glasses help fair comparisons in international science landscape: Country indicators as a function of ISI database delineation. *Scientometrics*, 56 (2) : 259–282.