THE DILEMMAS OF PERFORMANCE INDICATORS OF INDIVIDUAL RESEARCHERS

AN URGENT DEBATE IN BIBLIOMETRICS

PAUL WOUTERS

WOLFGANG GLÄNZEL

JOCHEN GLÄSER

ISMAEL RAFOLS

INTRODUCTION

In the last quarter of the 20th century, bibliometrics evolved from a sub-discipline of library and information science to an instrument for evaluation and benchmarking (Glänzel, InScit, 2006; Wouters 2013). As a consequence of this shift in perspective, new fields of applications and challenges opened to bibliometrics, although many tools were still designed for use in the context of scientific information, information retrieval and libraries. In other words, these became used in a context for which they were not designed (e.g., the Journal Impact Factor (JIF)).

This development has been joined by an intensified interest in the evaluation of individual researchers. The publication of the Hirsch Index in 2005 (Hirsch, 2005) and its popularisation by the journal *Nature* (Ball, 2005) has given this a strong stimulus. According to Hirsch, his index seemed the perfect indicator to assess the scientific performance of an individual author because "it is transparent, unbiased and very hard to rig". The h-index combines publication activity with citation impact. For example, an author with a h-index of 14 has created 14 publications that each have been cited at least 14 times each. So neither authors with a long list of mediocre publications, nor an author with a one won-

der hit are rewarded by this indicator. Nevertheless, the h-index turned out to have several severe disadvantages to be wearing the crown of "the perfect indicator" (cf. Glänzel, 2006). As Hirsch acknowledged himself, it cannot be used for cross-disciplinary comparison. A field in which many citations are exchanged among authors will produce a much higher average Hirsch index than a field with much less citations and references per publication. Moreover, the older one gets, the higher ones h-index will be. Furthermore, confidence intervals of empirical h-indexes are huge (Glänzel, 2010) such that this indicator is not suited for ranking individuals or research units and rankings based on the h-index may also be influenced in rather counter-intuitive ways (Waltman & Eck, 2012). Although many variants of the h-index have been published, none of them has turned out to be the perfect indicator. In fact, we cannot expect any indicator to be the perfect one. Nevertheless, there is ample evidence that the use of the h-index and other bibliometric indicators (such as the JIF) has become pervasive in the estimation of the scientific and even scholarly impact of a body of work by an individual scientific author. For example, many biomedical researchers mention the value of their h-index on their CV. In publications lists, one can regularly see the value of the JIF mentioned after the journal's name. In some countries, for example Turkey and China, one's salary can be determined by the value of either the h-index or the journal's Impact Factor one has published in.

This situation is clearly not desirable. If researchers are being evaluated, they should be aware of the criteria used and these criteria should be justified for the purpose at hand. This requires that users of performance indicators should have clear guidelines. It seems rather obvious that the bibliometric community has an important responsibility to inform and provide such guidelines. However, at the moment, there is no consensus yet about such guidelines. Individual bibliometric centres do indeed inform their clients about the use and limitations of their indicators. Moreover, all bibliometric centres have the habit of publishing their work in the scientific literature, often including technical details of their indicators. However, this published work is not easily accessible to non-expert users such as deans of faculties and research directors. The literature is too technical and distributed over too many journals and books. It needs synthesizing and translation into plain language which is easily understandable.

So how should the community of scientometricians relate to this development? What should the responsibility be of scientometric and bibliometric experts in the process of research evaluation? Should science and technology indicators be used at this level? If so, how should their limitations be interpreted? In what sense are we in need of a heigthened ethical awareness in the field of scientometrics, informetrics and bibliometrics? We are fully aware that these questions are not new. In fact, they have been raised several times at scientometric and bibliometric conferences, almost from the very start of the field. But social relationships are always dynamic and this certainly holds for the scientific and scholarly system. The increased role of indicators in general and of scientometric performance indicators in particular makes it necessary to address these questions again and in the context of the evolving practices of research evaluations and assessments. This was our motivation to propose two debates at the subsequent scientometric conferences this year. At the 14th ISSI Conference 15-19 July in Vienna, a special plenary session was organized with a joint presentation by Wolfgang Glänzel and Paul Wouters, followed by responses by Henk Moed and Gunnar Sivertsen. At the STI2013 conference, "Translational twists and turns: science as a socio-economic endeavour" 4-6 September in Berlin, a full plenary was devoted to bibliometrics of individual researchers, chaired by Ben Martin (SPRU), with presentations by Wolfgang Glänzel, Paul Wouters, Marc Luwel, and Jochen Gläser. In this short report, we wish to give an impression of this discussion with the aim to further stimulate this exchange of ideas, experiences and, of course, technical knowledge.

THE ISSI CONFERENCE IN VIENNA

To initiate a process of a more professional guidance for the application of bibliometric indicators in the evaluation of individual researchers, we asked the organizers of the ISSI conference to devote a plenary to this problem, which they kindly agreed to. At the plenary, Wolfgang Glänzel and Paul Wouters presented "The dos and don'ts in individual level bibliometrics". We do not think this is a final list, more a good start with ten dos and don'ts. A start for reflection, experiments and the exchange of experiences. In the following, we sketch our proposals for applying bibliometrics on individual researchers as well as the ensuing debate at the conference.

TEN THINGS YOU MUST NOT DO:

1. Don't reduce individual performance to a single number
2. Don't use journal impact factors as measures of quality of individual researchers
3. Don't apply hidden "bibliometric filters" for selection
4. Don't apply arbitrary weights to co-authorships
5. Don't rank scientists according to one indicator
6. Don't merge incommensurable measures
7. Don't use flawed statistics
8. Don't blindly trust one-hit wonders
9. Don't compare apples and oranges
10. Don't allow deadlines and workload to compel you to drop good practices.

TEN THINGS YOU CAN DO:

1. Basic measures such as numbers of publications and citations are still relevant statistical measures
2. Analyze collaboration patterns of researchers
3. Always combine quantitative and qualitative methods
4. Use citation context analysis
5. Analyze subject profiles of individual researchers
6. Make an explicit choice between the analysis of the full oeuvre or comparative analysis using a citation window
7. Combine bibliometrics with career analysis
8. Clean bibliometric data carefully and use external sources
9. Don't take this list of dos and don'ts too absolutely: even some don'ts can be used given the right context
10. Help users to interpret and use your results.

Of course, the complex business of research assessments cannot be reduced to a simple list of "commandments". In other words, we do not want to initiate a bibliometric police with absolute rules. The context of the evaluation should always determine which indicators and methods to use. Therefore, some don'ts in our list may sometimes be perfectly useable, such as the application of bibliometric indicators to make a first selection among a large number of candidates. Also, in informed peer review of large institutions it may be inevitable to use bibliometric filters to zoom in on the most relevant work for closer inspection. In all those cases, these filters should however be made explicit. After all, the researchers who are subjected to assessment should be able to provide counter-arguments if they think the filters have been used inappropriately.

Our presentation was commented on by Henk Moed (Elsevier) with a presentation on "Author Level Bibliometrics" and by Gunnar Sivertsen (NIFU, Oslo University) with comments on the basis of his extensive experiences in research evaluation. Henk Moed built on the concept of the multi-dimensional research matrix which was published by the European Expert Group on the Assessment of University Based Research in 2010, of which he was a member (*Assessing Europe's University-Based Research - Expert Group on Assess-*

ment of University-Based Research, 2010). This matrix aims to give global guidance to the use of indicators at various levels of the university organization. The first row of the matrix discusses goals, output dimensions and bibliometric and other indicators at the individual level. The matrix as a whole does not focus on the problem of how to evaluate individual researchers. Still, the matrix is surely a valuable contribution to the development of more professional standards in the application of performance indicators. In his presentation, Moed discussed a number of case studies which clearly showed that no absolute rules can be expected. It all depends on the goal of the assessment as well as on the state of affairs in the research area involved. Moed argued that the data should be verified by the researchers themselves (already a standing practice in most if not all of the main bibliometric centres). A key problem he identified is the attribution of scientific performance to an individual when in reality most research is based on collaborative work within and between teams. One of the three cases he presented involved a country in which science policy suspects that their researchers are not oriented enough toward international networks. In this case, a policy measure could be to stimulate and reward publication in top international journals. For this, the number of publications in those type of journals could be an appropriate bibliometric indicator.

Gunnar Sivertsen strongly agreed with the main thrust of the discussion. Moreover, he made clear that the discussion should not be restricted to the bibliometric community itself. On the contrary, the main audience of bibliometric guidelines and standards should be the researchers themselves and administrators in universities and funding agencies.

The ensuing debate led to a large number of suggestions. A few speakers emphasized that this debate is indeed not new and was already addressed at the ISSI conference in Chicago in 1995. A key point was the issue of responsibility: it is clear that researchers themselves and the evaluating bodies should carry the main responsibility for the use of performance indicators. However, they should be able to rely on clear guidance from the technical experts. How must this balance be struck? Should bibliometricians refuse to deliver indicators when they think their application would be unjustified? Should the association of scientometricians publicly comment on misapplications? Or should this be left to the judgment of the universities themselves? Several calls were made to publicly criticize applications of bibliometric measures that are deemed harmful. At the same time, it is not yet clear in whose name these statements should be made. The plenary did not solve these issues yet. However, a consensus seemed to be emerging that more explicit guidance by bibliometricians is required (building on the work from the past) and that researchers should have a clear address to which they can turn to with questions about the application of performance indicators either by themselves or by their evaluators.

THE STI2013 CONFERENCE

The plenary at the STI2013 conference started with an introduction by Ben Martin. He reported on a dramatic case of the rise and fall of a young professor in economics in Germany. This researcher had conducted one research project for his PhD and was able to generate a rather impressive number of publications on the basis of this dataset. Because he was so productive, he was able to attract more external research funding. Subsequently, he became an attractive target for headhunting by universities looking for high profile researchers who might help them increase their ranking position in various league tables and in getting grants for the Excellence Initiative program in Germany. And indeed, a university was quite eager to hire him for precisely this purpose. It was only then that a few members of the selection committee decided to actually read the publications by this young economics researcher. It struck them that these publications were quite like each other. Not very surprising given that

the research was all based on a single dataset from his PhD project. It turned out that he had published a large number of variations of the same article in different journals without anyone noticing these duplications. This discovery was the beginning of the end of the career of this formerly promising economist. A number of journals began retracting these publications, although not with the cooperation of the researcher. This process is still ongoing. A sobering tale, according to Martin, and one which shows that the abuse of bibliometrics is now a serious concern for all parties involved in the management and development of scientific research.

Wolfgang Glänzel then informed the audience about the discussion at the ISSI conference. He emphasized especially the need for more information and guidelines among researchers, managers and policy makers. The bibliometrics community should play a role in providing at least the core of these guidelines. He also emphasized that the ISSI conference had made clear that we need some organization to which questions about the proper use of bibliometric indicators can be addressed. ISSI as an organization may not be in the best position to play this role, given its scientific rather than professional role. This may therefore be a role that the main bibliometric centres should take up, in a somewhat coordinated way.

In his presentation, Paul Wouters argued in favour of a portfolio approach and showed how CWTS has been developing bibliometric profiles at the level of the individual researcher. He also presented the philosophy behind the 7th Framework project ACUMEN which aims to enable individual researchers to enrich their Curriculum Vitae with properly calculated and relevant bibliometric indicators as well as qualitative evidence. The portfolio approach has also been proposed by other researchers such as Bornmann (2013).

Marc Luwel focused on the tensions in the concrete practice of science policy making. There is an increasing need for justification of the budgets devoted to research. This needs to be based on verifiable empirical evidence, hence the need for performance indicators among which also bibliometric indicators. Nevertheless, Luwel stressed, quantitative indicators cannot be used as the sole basis. Luwel: "Beware of the lone librarian cooking a toxic cocktail of publication and citation data!"

The presentation by Jochen Gläser linked to Luwel's talk by taking one more step in the area of applied ethics and the dilemmas in the application and use of bibliometric indicators. Gläser made clear that he does not see himself as a bibliometrician pur sang, but more as a sociologist who is interested in combining bibliometric methods with other social science methods, such as surveys and interviews. Moreover, he is not himself involved in applying bibliometrics for research assessments and can therefore take a step back and present reflections that may be useful to the community. He presented a couple of scenarios in which bibliometric reports are carefully crafted but basically ignored by the decision makers or only cherry-picked in order to justify decisions that were going to be taken anyway. And indeed, often it is not clear to what extent and how the bibliometric reports that we produce are actually useful for quality decision making. The reports may very well start to live a life of their own. Gläser discussed to what extent the current available literature on applied ethics is useful for the bibliometric community (not very much) and in what sense we can learn from the communities that have been producing guidelines and standards for their profession (more promising perhaps). His presentation made clear that ethics is indeed inevitable and highly relevant, albeit not in all respects completely new. Some of the bibliometric centres have been following implicit and explicit ethical rules without always formalizing them. Gläser also mentioned the proposal by Wouters in a parallel session at STI2013 on bibliometric standards to initiate a bibliometric "Ombudsman Office" to which researchers that have been evaluated can

turn in case they feel misjudged or harmed by the performance analysis and these conflicting interests cannot be solved with the bibliometric experts involved. He saw this proposal as a long-term goal. In the shorter term, he proposed that the next few STI conferences should all devote more attention to these ethical and political issues, for example by organizing special session devoted to them.

As authors of this short report, we think this last proposal is an excellent idea. It might for example be possible to not only discuss formal research papers in these sessions, but we might want to focus more on exchanging experiences. This could take the form of submitting problems or case studies rather than the normal papers. The discussion could also be organized in more engaging and discussion oriented way.

The ensuing discussion at the STI2013 conference again made clear that these issues have become quite urgent for many practitioners in the field of science & technology indicators. The issue of responsibility rose to the top again of most urgent issues, although it is also clear that it will also remain a rather complex web of problems. This complexity was illustrated by a number of participants who addressed the limitation of the present discussion to the European and Anglo-Saxon context. The role of indicators in China, Turkey, Iran, and South East Asia is clearly different and perhaps even more dominating. As a result, many researchers have had to set explicit indicator based targets in their career development. This means that the current debate should be linked to these practices.

In the discussion an additional fundamental problem was raised: to what extent is the ethical problem in reality caused by a fundamental problem in the state of our knowledge in bibliometrics? Often, we actually do not know what exactly is represented by the patterns we see by the indicators. Therefore, it was argued, we should first of all try to generate more robust knowledge, using also a much larger variety of databases in addition to the traditional citation indexes.

Perhaps this questions might be a good agenda for the next series of STI and ISSI conferences? In parallel, we will be organizing a workshop on these issues targeted at the users of bibliometric indicators, including the scientific communities in the first half of 2014. Stay tuned!

REFERENCES

Assessing Europe's University-Based Research - Expert Group on Assessment of University-Based Research. (2010). Research Policy. European Commission. doi:10.2777/80193

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature,* 436(7053), 900. Retrieved from http://dx.doi.org/10.1038/4369000a

Bornmann, L. (2013). A better alternative to the h index. *Journal of Informetrics,* 7(1), 100. doi:10.1016/j.joi.2012.09.004

Glänzel, W., On the opportunities and limitations of the h-index (in Chinese). *Science Focus,* 1 (1), 2006, 10–11. English version accessible at: http://eprints.rclis.org/9378/

Glänzel, W., On Reliability and Robustness of Scientometrics Indicators Based on Stochastic Models. An evidence-based opinion paper. *Journal of Informetrics,* 2010, 4 (3), 313-319

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America,* 102(46), 16569–72. doi:10.1073/pnas.0507655102

Waltman, L., & Eck, N. J. Van. (2012). The Inconsistency of the h-index. *Journal of the American Society for Information Science and Technology,* 63(2007), 406–415. doi:10.1002/asi