Bibliometrics, assessment and UK research

Based on a paper presented at the UKSG seminar 'Measure for Measure, or Much Ado About Nothing? Measuring the quality and value of online journals', London, Thursday 14 June 2007

The Research Assessment Exercise (RAE), based on peer review, has enabled the UK to improve its comparative international research performance. But the RAE is changing after 2008 to a system based on metrics, within which a bibliometric quality indicator will be a critical component. Bibliometrics, using publication and citation counts, have many advantages and have been extensively developed over the last 20 years. But they also have serious challenges, some real and some apparent, and researchers will need expert advice to develop and work with an effective and supportive metrics system. This paper discusses the background to the change in metrics, identifies a range of possible problems that need to be tackled, and confirms the critical role of those with prior expertise in guiding the establishment and then the development of a sensible system.

JONATHAN ADAMS
Director
Evidence Ltd

Introduction

People who work in universities need informed and expert guidance about serials and their significance over the next few years because their research assessment system is changing to a completely new mechanism using bibliometrics. Forecasts suggest that some serious trip-wires – possibly even pitfalls – will face researchers and managers who react to these changes without proper information.

The UK has had a cyclical Research Assessment Exercise (RAE) since 1986. In the early 1980s, both government and university saw that research had a problem: it was becoming more expensive and there were diversifying opportunities. Resources had become too thinly distributed, so the system for their allocation had to change to become more selective. That would inevitably bring about some concentration of those resources. The first research selectivity exercise was introduced in 1986, becoming the RAE in 1989. It was run under a slightly different system in 1992, when the former polytechnics became universities. It ran again in 1996 and 2001. Universities are getting geared up for its next run in 2008.

The RAE takes submissions across about 70 subject areas ('Units of Assessment' or UoAs) grouped, in 2008, under super-panels to take account of interdisciplinarity. The RAE system is based on peer review. Each peer panel receives information from the universities, with data relating to staff, funding, post-graduates and published outputs for that UoA. There is a statement about the strategy and track record of each unit. The outcome of the peer review is a grade from 1 to 5, or even 5*. That grade affects both core funding and reputations, so these are critical issues for researchers.

Changing the assessment process

It is agreed that the most critical data for peer assessors are usually the outputs: the publications of the staff. We have analysed these, which may be journal articles, books, chapters, conference proceedings or even art and video. Where outputs come from journals we can track them back to the serial in which they were published and collate their citation counts. We can group the outputs by UoA and by institution within each UoA. We can attach to that analysis the grade awarded by the peer panel, so we can explore many aspects of publication behaviour and quality assessment.

There has been a progressive shift to serials as the preferred source for submissions. We think that that process is likely to go further in 2008. In broad-based science areas, the total number of outputs submitted in 2001 dropped slightly, but journal articles grew from 90% of submitted material to something close to 96%. In areas such as clinical medicine it was near 100%. Engineering is of particular interest. In 1996, the engineers told us that journal articles just were not an important channel for them: key conference series were where they had direct communication with industrial users. But, by 2001, we saw a step up from 57% to 78% submitted outputs from serials.

In social sciences and in humanities we again see a shift to journals. People in the humanities say that they do not use journals: the key channel is still the monograph. But a recent European humanities exercise has found many different specialists in the humanities producing long lists of key journals for their area, so evidently it has an increasing importance there as well.

By 2001, the RAE had become a problem. It had come to be seen by universities as expensive and burdensome, and by government as suspicious (because it was based on peer review, which they do not wholly trust). Government would like to see something a little harder, a little more quantitative, and a little more transparent. So, in early 2006, the Treasury announced that RAE2008 would be the last of its type: assessment was to shift to a metrics-based process.

After consultation, Alan Johnson, Minister for Higher Education, announced that there would be three key metrics. These metrics would initially apply to what are called STEM subjects: science, technology, engineering and medicine. Publication and citation data were central, because they were the preferred quality measure. As well as outputs, the other metrics were to be an input measure, probably funding, and an 'infrastructure' measure, represented by post-graduate training. That corresponds roughly to the structure of previous submission data: the RA4 section (grants), the RA3 (post-graduate research students) and the RA2 (outputs).

Problems in using bibliometrics

This sounds reasonable, so why might we be wary about bibliometrics and research assessment? Serials form a key part of what people think represents their work, so they should be fine as a basis for a metric system. When researchers use citation data as a guide to quality, however, they do so in a very informed and expert way. That is very different from an algorithm that blindly determines how much money a university is going to get.

Individuals can pick up warning signals that formulae cannot. For example, citations may index quality, but not all citations are collated. Thomson covers 8,500 or so journals. In science subjects, 75% of the cited material is in those Thomson journals, so 25% is not. In some subjects the deficit is higher. Citation characteristics also vary between fields, so the context for 'normalizing' citation counts needs to be very carefully set. The government proposal is to use just six main STEM subject areas – a very coarse-grained approach.

Would bibliometrics correspond to the peer review used in the past? How much change would we see? To evaluate this, we analysed the impact of RAE2001 outputs from the HEIs that submitted in Unit of Assessment 14 (UoA14) biology. We found a broad relationship between peer-reviewed RAE grade and average bibliometric impact, but we also found a lot of residual variance. It could not be determined from this data whether a unit with bibliometric impact around world average was a 3a, 4, 5, or even 5* grade department. So there is a concurrence, but more information is built into peer review than we can build into a metric algorithm.

Citation growth characteristics vary between fields, and larger fields tend to have higher citation rates. Molecular biology is a field of roughly the same size as animal sciences but it has very different citation growth characteristics. In the coarse-grained approach proposed, these two would be aggregated, creating a potential conflict as to how universities best present information.

Within animal sciences, we have very different citation rates for journals. Is that going to put pressure on academics to publish in particular places? We already know that the journal impact factor is being used by research managers – who do not understand what such factors mean – to put pressure on researchers to publish in particular serials. If we look closely at the dynamics of environment, ecology and molecular biology journals, we can track the long-term accumulation of citations relative to 1985. At first, the growth rate for environment and ecology lags behind molecular biology. Over a longer time, environment and ecology picks up above molecular biology. The subtleties of normalizing the data are going to make some difference to the outcomes of the metrical analysis.

How are journals submitted to a different extent for the assessment exercise? We can see what people identified as the material that best represented their research. For example, we can look at UoA18 chemistry and compare journal five-year impact with the percentage of articles selected for the RAE. We found two journals of high impact with high rates of submission for assessment, but they were followed by a cloud of mainstream journals. There is no evidence of any overall relationship between impact and submission rates. The quality of the article and of the journal require separate human judgements, but a metric system may ignore that.

This comes back to the issue of normalizing the data. We looked at UoA13 psychology data across 4, 5 and 5* units. We normalized the citation data relative to the field average and relative to the journal average. Each normalization produced a different pattern in relation to grade, so choosing the frame of reference will affect the outcomes.

Other factors worrying researchers are issues like self-citation. Self-citation is a normal part of the research process. Every publication builds on the work of its predecessors. Is it self-citation when somebody in your laboratory cites your paper? Is it self-citation when somebody in your university cites the paper? Who is going to go through all the journals, articles and citations to check which ones are the self-citations? If you take out self-citations then you may also change the behaviour of researchers, and we do not know what the outcomes of that might be.

'May-flies' and 'Sleeping Beauties' have been proposed as problems. Do some papers peak quickly and then die, and would they be affected by changing the citation window that might be used? We have done some work for OSI on that, and we do not think it is an issue. Leiden University has looked at whether there are papers that do not get cited for many years and then suddenly take off. There are such papers, but no more than you would expect by chance in a distribution informed by the rest of the material.

Are negative citations an issue? Eugene Garfield said citations were impact, not necessarily excellence. A classic paper on 'cold fusion' received many citations because it was wrong. But it had a big influence on other people's work, so it had significant impact. Wrong and trivial papers get no citations.

What sort of metric is going to be produced? 'Point' metrics are part of the problem. We all use things like 'average citation counts'. That is a metric, but it is also just one point in a distribution. Research activity is highly skewed, so research metrics are skewed. A lot of people do not get much research money and a few people get lots. The same thing happens with publication data. Some people publish few papers and others publish a lot; many papers get few citations and a few get exceptional attention.

Where does significance lie in these distributions? Do we really want to look only at an average, or at something a little more informed? *Evidence* has been building profiles to move away from averages. To illustrate the approach, we can create an impact profile for two UK universities. We look at citation counts rebased to each article's world average for the field and year. Then, we pull out uncited papers – about 15% of UK publications are uncited. For the rest we lump the papers by impact into bins structured by their relationship to world average (half, quarter, twice, four times, etc.). We get something that looks like a normal curve, which is an easy image to hold and to compare between institution, country and so on.

This takes us away from averages to something more informed. When we think about distributions it is not just the average, but the volume at different impact levels. We can pick out the articles with exceptional impact that have the greatest influence on the field and perhaps on subsequent products and processes with economic impact.

Conclusions

What will happen next? The nature of the UK assessment system is up for debate. Publication patterns are changing: it is clear that UK social science researchers are tending more towards an American paradigm. People are beginning to think about their research and outputs in different ways. The new metrics will affect the sciences and engineering first, but it will also affect other disciplines in due course. All will need information about the general models that will influence the way things are managed and presented in their institution, and the way research managers change their expectations of researchers.

The system will be dynamic. I have identified some possible changes of behaviour, not all of which are entirely positive. The specific metrics will therefore need to be negotiated, because there are many interested parties. The algorithm that is built up for the outputs component has to link with funding inputs and the information on research students. Weighting factors, balances, data quality, data sources, and validation of the data all need to be carefully built in.

The outcome remains, for the present, very unclear. Researchers will need a lot of relevant, sound information to enable them to make proper judgements.

■ **Jonathan Adams**
Director
Evidence Ltd
103 Clarendon Road
Leeds LS2 9DF, UK
E-mail: enquiries@evidence.co.uk
Tel: +44 (0) 113 384 5680
Fax: +44 (0) 113 384 5874
http://www.evidence.co.uk