NEW METRICS FOR RESEARCH OUTPUTS

OVERVIEW OF THE MAIN ISSUES

A. Background

The only metric that has been in widespread use hitherto is the Journal Impact Factor (JIF), developed by the Institute for Scientific Information, ISI (now part of the Thomson Reuters group). The metric was originally developed as a measure of the impact of individual journals and was intended as a tool for publishers. It is calculated annually for all the journals covered by Thomson's Web of Science (WoS) database. The JIF prevailed pretty well unrivalled over two decades because no other provider had the breadth and depth of content to create alternative, meaningful metrics for measuring anything else from the research literature.

Quick facts

Web of Science:
10,000 journals
110,000 conference proceedings

Scopus:
16,000 journals
520 conference proceedings
315 book series
36 million items in total

Open Access journals (data from the DOAJ):
3750 journals
222,400 articles

***Open access repositories* (data from OpenDOAR):**
1280 repositories (article total unknown)

This has changed now with the launch in 2004 of Elsevier's Scopus, a database that contains journals and other source items. Elsevier is well-placed to develop new impact measures using the content of Scopus and indeed is actively courting the research community with this in mind.

A third front has also opened up in the form of Open Access literature. This resides in the open article databases of commercial open access publishers, such as PLoS and BioMed Central (BMC), in the open article databases produced by the not-for-profit open access publishers[1], as a portion of the content of subscription-based publishers (the portion that has been opened up as a result of so-called 'hybrid, author pays' options) and in institutional or subject-based repositories such as arXiv and PubMed Central[2].

None of these sources alone, though, approximates to the whole journal literature and the one that gets closest, Scopus, is proprietary. It should be noted in that respect that Scopus collects its content by means of the donation of journal data by other publishers. This is significant in the context of anticipated publisher attitudes towards participation in any scheme to develop a new corpus of open literature on which to develop novel metrics for research measurement.

B. Issues

There are a number of issues pertinent to the development of new ways of measuring research and the main ones are as follows:

[1] The outputs of both commercial and not-for-profit open access publishers are collated by the Directory of Open Access Journals: www.doaj.org

[2] OpenDOAR (www.opendoar.org) or ROAR (http://roar.eprints.org)

i. Inappropriateness of a single metric

There has been over-reliance on the JIF as a research measure. In fact, it has never measured research. It has measured something to do with research journals. This metric was intended to be a comparative tool for publishers but it has been used – incorrectly – to assess the performance of individuals, departments and even institutions. Whether or not it measures something about a researcher or research group, it is clear that use of a single metric is not an appropriate way to conduct comparative analysis of such a complex set of activities that research represents (unless that single metric is the product of multivariate regression).

ii. Primary bases for measuring research

In the broadest sense there are two main types of metric:

- User (reader)-generated metrics = usage-based analysis (e.g. downloads, 'reads', abstract views, etc)
- Author-generated metrics = citation-based analysis (formal citations, acknowledgments, links, etc)

The focus for the publisher discussion group has been the latter – citation-based metrics. It is possible to develop many relevant and incisive metrics given a substantial Open Access literature, as well as to develop new techniques for tracking and monitoring research developments using the research literature and other forms of output. The primary need will be the development of metrics that fit the needs of particular research communities, acknowledging that few measures are comfortably and sensibly applicable across the board. Research outputs differ markedly across disciplines and measures that fit well the products of humanities research will be different to those that are appropriate for the social sciences, which in turn may not suit the requirements of the natural and physical sciences.

iii. Publisher compliance

Until Open Access is the strongly-predominant format for research outputs, which may take another three-to-five years, some degree of Closed Access publisher compliance will be needed to build the database upon which new metrics can be developed. For metrics based on citation analysis, publishers will need to supply, at a minimum, citation data from the articles they publish. For metrics based on usage, publishers and repositories will need to supply usage data for collation into a single database. For other kinds of metric, such as those measuring certain types of trend, for example, article full-texts will be needed. Preliminary work is going on in all these areas using the currently available Open Access corpus and this work can be accelerated as soon as access to the full-text of more of the research literature increases.

iv. Collation of the data for analysis

The database(s) that will provide the raw material for the development of new metrics must be collated from various sources (from different publishers and, probably, from repositories). The question of which party should be entrusted to collect data from publishers and make them available as a single database is still undecided.

Discussions within the 'pioneer' publisher discussion group have identified a number of possible candidates and major contributors: amongst the non-commercial candidates the front-runners are CrossRef and the NLM (PubMed), with a preference for the former (see section D).

Whatever the final solution in this regard, there will be considerable technical work ahead to bring data in different formats together into one, internally-consistent database that can be queried effectively.

v. Metrics development and service provision

The overall aim is to provide an Open Access database available to all-comers, some of whom may wish to invest in the development of new metrics and/or new services based upon the use of these metrics. Some commercial database providers are likely to pursue this path (as some already are) and it is also likely that many non-commercial players will be interested in developing new metrics and possibly services based on these. There are already examples of such enterprises.

University-ranking services and the metric(s) they use

THE-QS World University Rankings
Uses academic peer review, employer survey, faculty/student ratio, citations, international faculty numbers and international student numbers in a weighted algorithm to produce overall placings
http://www.topuniversities.com/worlduniversityrankings/

Academic Ranking of World Universities (produced by Shanghai Jiao Tong University)
Uses weighted system comprising four factors: quality of faculty, quality of education, research output and size of institution
http://ed.sjtu.edu.cn/ranking.htm

For instance, there exist several university-ranking services, each based on a different metric, provided on a free-to-use Web access model (see sidebox, left, for the main two services). Given access to a reasonably comprehensive citation database, such players – and others – may rise to the challenge of developing appropriate metrics and rankings, plus other types of serve based on the metrics, for the world's research community. Some of these players will be existing commercial companies, perhaps some of them scholarly publishers themselves: others will be within the research community itself or new start-ups, and a variety of business models can be expected.

Meanwhile, various projects are already attempting to study stakeholder requirements and develop metrics that can be applied in different situations. For example, the EU-funded EERQI project (European Education Research Quality Indicators)[3] has recently begun, aiming to develop markers and indicators of quality for education research. The British Academy published a report last year on peer review in the humanities and social sciences that contains a chapter addressing the issue of appropriate metrics in these disciplines[4] and has followed up with a consultation exercise this year. These are just examples of many activities currently underway on this front as new research assessment frameworks are developed in the UK[5] and Australia[6].

[3] http://www.eerqi.eu/

[4] http://www.britac.ac.uk/reports/peer-review/index.cfm

[5] http://www.hero.ac.uk/uk/research/research_quality_and_evaluation/research_excellence_framework__ref_.cfm

[6] http://www.arc.gov.au/era/default.htm

A number of tools are already built that can deliver citation-based rankings and other metrics based on citation analysis of Open Access content. The main ones are Citebase[7] (currently working on arXiv), Citeseer[8] (currently working on the Open Access computer science literature) and CitEc[9] (working on the Repec Open Access economics database).

vi. Author identification and other technical issues

There are a number of technical issues pertaining to the development of metrics for measuring individual author performance. Not least of these is being able to identify individual authors satisfactorily. At present, the only really significant development in this area has been that in the Netherlands, a country that has had the foresight to establish a national research activity database (METIS) and to assign a unique personal number (digital author identifier, DAI) to every researcher employed by or affiliated to a Dutch university. In every other situation author disambiguation remains a problem. It is something with which Thomson ISI (and, presumably, Scopus) has wrestled but no satisfactory solution has emanated so far from commercial database producers.

Individual institutional repositories can disambiguate authors within the institution, usually by means of a system based upon unique identifiers referenced against an institutional personnel database.

vii. Usage indicators

The possibility of complementing citation analysis with a usage statistics service is a future option. COUNTER has established codes of practice with respect to how vendors report usage, standardising this reporting so that libraries and publishers can make full use of the data for comparative analysis.

Many institutional repositories are also collecting data on usage, reporting on downloads and other factors such as where usage is coming from and referral patterns.

The promise, therefore, is of a future aggregation of publisher and repository reporting services to provide a single point of reference. This is already being investigated on a project basis through a JISC-funded project involving COUNTER and Cranfield University as project partners. The findings from that project may pave the way for developments in this area.

C. Aims for the source database

The aims in compiling a database of citation data for the purpose of developing new metrics are that it should:

- be an Open Access citation database: the intent is to produce a database that any player can use to develop metrics or rankings, or to carry out bibliometric research
- minimally, provide citation data

[7] http://www.citebase.org
[8] http://citeseer.ist.psu.edu/citeseer.html
[9] http://citec.repec.org/

- optimally, provide full-text content so that indicators using data contained in the text may also be developed and bibliometricians can exploit the data to study research trends and other characteristics
- be as comprehensive as possible, combining data from all relevant sources: the sources that must be included are:
  - data from Closed Access publishers
  - data from all Open Access sources – subject repositories, institutional repositories and Open Access publisher databases
- be current; that is, provide data immediately without an embargo period

D. Possible solutions

Two players with large collections of current citation data have so far been identified as possible solutions from discussions in the publisher group:

i. CrossRef builds the citation database

CrossRef currently has around 120 member publishers and covers approximately 5000 journals. Not all member publishers provide citation data to the service and publishers may opt into or out of allowing specific third parties to access and index their data via the CrossRef database. Some member publishers are, though, willing to release data to third parties and content from these publishers would form the core of the database. Optimally, all member publishers would permit CrossRef to build a citation database that is freely accessible by all-comers who would wish to develop research metrics and services based upon them. It is unlikely, however, that Elsevier would cooperate in such a venture since it would undermine Scopus's offering. A database formed from the rest of CrossRef member publishers' content would represent around 105 of the total journal literature.

ii. The US National Library of Medicine

The NLM holds huge volumes of publication data including the PubMed Central database which represents the biggest single source of such data in the life sciences. One option is for the NLM to negotiate with publishers with the aim of collecting citation data from as many as possible in order to augment the biomedical content of the NLM with data from other disciplines

iii. The NLM and CrossRef work together

Under this option the NLM might collect citation data provided by CrossRef (at least, the data that CrossRef member publishers opt into providing) and aggregate those with the citation data that the NLM already collects, providing a broad-scope database that would be made available to would-be users via the NLM.

iv. Other solutions

The options above are not the only possibilities but they do represent the simplest currently available 'big' solutions. There are two other main routes to the goal:

- Establishment of a new database from scratch. Scopus covers 16,000 journals and, with the exception of its own 2350 titles, the data from all of them are contributed free by other publishers keen promote their content through this service. Setting up another database and asking those same publishers to contribute data to that is an option. If they all complied, the database would represent about two-thirds of the journal literature.

- As funder and institutional Open Access mandates increase there will be a rapid growth in source content available for harvesting from institutional and subject-based Open Access repositories. This material will be available in most cases after a short embargo and in many cases prior to publication in journals. Aggregated with the content from Open Access journal publishers, along with the Open Access articles from 'hybrid' journals, this will quite shortly constitute a sizeable database on which to develop new metrics.

Alma Swan
22 November 2008