Measuring the usage of individual research articles

Based on a presentation given at the UKSG seminar 'Mandating and the scholarly journal article: attracting interest on deposits?', London, 29 October 2008

Publisher and Institutional Repository Usage Statistics (PIRUS) is a COUNTER-led project tasked with examining the viability of developing usage reports for individual articles by combining data gathered from publishers, aggregators and institutional repositories (IRs). The project studied current practices for storing and labelling individual articles in both publisher and repository environments, successfully tested a mechanism for automatically gathering usage data from IRs, and proposed a simple XML-based report for displaying article usage statistics that can, in principle, be implemented by any entity that hosts and provides online access to articles. PIRUS is optimistic that further development of the work so far undertaken could lead to a roll-out of a facility allowing publishers to track more use of their content than they currently do and to provide enhanced information to their authors, while IRs would gain a more demonstrable, quantitative measure of return on resources invested.

RICHARD GEDYE
Research Director
Oxford University Press

It is seven years now since COUNTER was formed with a mission to establish and maintain a set of universally accepted 'best practices' for the production and distribution of usage statistics reports for online information products by vendors of those products. Since its formation in 2002 COUNTER has indeed developed a number of codes of practice, initially for journals and databases and latterly for books. The code of practice for journals and databases is about to enter its third version, and COUNTER-compliant usage statistics over the last five years have become an increasingly standard component of all online products which want to be taken seriously in the market. Over 100 publishers, aggregators and hosting platforms now provide these statistics, representing more than 15,000 full-text online journals, as well as a growing number of online books.

At the outset, COUNTER's aim was to encourage a better usage statistics service for one particular stakeholder in the world of online information dissemination, that is, the institutional library purchaser. However, for academic and scientific journals, there are at least five sets of stakeholders for whom usage statistics are of some interest: firstly, the library community, who will use them for a variety of purposes including collection development, marketing efficacy assessment, and demonstration of value for money at budget submission time; secondly, publishers, owners, and editors of journals, who will be interested less in usage statistics at the individual subscribing institution level and more in the overall statistics for their journal. Meanwhile, a third group with a potential interest in usage statistics is the author community, and whereas editors, journal owners and libraries are going to be mainly interested in journal-level statistics, authors understandably are going to have a particular interest in usage of their own articles. This interest is likely to be shared both by any grant-giving bodies who have funded the authors' research and the institutions within which that research has been carried out.

Indeed, at COUNTER's inception, there was much discussion about whether we should provide article-level usage statistics to the library community. The general consensus seven years ago was that this would be too much information. It was not what our key library constituency required, and in general over the first few years of COUNTER's life, the demand for usage statistics at the article level was pretty low.

However, interest in measuring individual article usage has recently started to grow. A number of factors have contributed towards this. With some publishers already offering authors information about the usage of their articles, authors and funding agencies are becoming increasingly aware of the potential for developing a reliable and global overview of the usage of particular articles. So the issue is starting to move up a number of agendas. We are beginning to see examples of the idea of online usage becoming an accepted alternative measure of article (and journal) value. For example, usage-based metrics have been considered as a tool for use in the UK Research Excellence Framework, while the recent JISC Usage Statistics Review Project is aimed at 'formulating a fundamental scheme for repository log files and at proposing a standard for their aggregation to provide meaningful and comparable item-level usage statistics for electronic documents'[1]. On an international scale, the European Knowledge Exchange (KE) multinational consortium issued a report last year that specifically recommended developing standards for usage reporting at the individual article level. In fact the report of the KE's Institutional Repositories Workshop Strand on Usage statistics[2] was quite blunt in its recommendation to 'lobby COUNTER to add article level stats.'

The provenance of this particular KE report (an Institutional Repositories Workshop) is quite telling, because the extent to which authors are increasingly becoming mandated by their funders or their institutions to deposit copies of their accepted articles into subject-based or institutional repositories is, I believe, a further factor in an increasing realization that there would be benefits in measuring usage at the article level and doing so in a uniform and co-ordinated way. The benefits become all the more obvious when you consider that in today's world an article may be available from a growing number of different sources – from the main journal website where it has been formally published, but also from commercial aggregators like Ovid, ProQuest, EBSCO, Gale, LexisNexis, Westlaw, etc, as well as from subject-based repositories like PubMed Central, UK PubMed Central, the physics ArXiv, and from any number of authors' local institutional repositories.

So if we want to assess an article's impact by counting its usage, we will sometimes want to measure more than simply the usage that is coming from the main publisher's website. It was with this thought in mind that COUNTER started to consider what procedures might be developed so that publishers could, by maximizing the amount of usage that they capture and report, enhance the service which they provide as part of their relationship with authors.

Meanwhile, on the technical side, measuring article usage is now becoming potentially more of a practical proposition, with the emergence of a number of developments, in particular the implementation of XML-based usage reports by COUNTER in the new Release 3 of its Journals and Databases Code of Practice.[3] This makes more granular reporting of usage more practical, especially as it is coupled with the implementation by COUNTER of the SUSHI protocol, which facilitates the automatic harvesting of usage statistics from lots of sources by machine and then their consolidation.

So the seeds were sown of a new COUNTER-led project which was eventually launched under the name of PIRUS (Publisher and Institutional Repository Usage Statistics) and whose mission we have distilled as 'To develop a global standard for collecting and distributing article usage data, whatever the source'.

The project team included representatives of institutional repositories (Cranfield University, Oxford University, Southampton University), UK PubMed Central, CrossRef and Oxford University Press, as well as COUNTER.

In order to go about achieving our mission, we set ourselves three aims: firstly, to develop COUNTER-compliant usage reports at the individual article level; secondly, to create guidelines that would enable any entity that hosts online journal articles to produce such reports; and thirdly, to propose ways in which those reports could be consolidated at a global level in a standard way. As the six-month project progressed, our mission remained unchanged, but we modified somewhat the ways in which we felt the mission could be most realistically achieved, as will shortly become clear.

Our project was divided into three phases.

Phase 1: survey of current storage and labelling practice

Phase 1, which we started in August 2008 and ran for two months, was a survey of both publishers and institutional repositories to look, firstly, at current practices in the application of individual article identifiers and other metadata associated with stored articles (because clearly if you cannot easily identify these things, you are never going to get very far in attempting to measure their usage), and, secondly, at how different *versions* of individual articles are identified.

Our e-mail/telephone survey of publishers was conducted by PIRUS project leader Peter Shepherd of COUNTER who polled a sample of 15 journal publishers, aggregators and hosts: American Chemical Society, American Institute of Physics, Atypon, BioMed Central, EBSCO, Elsevier, Informa, Ingenta, Institute of Physics Publishing, Nature, OUP, Ovid, SAGE, Springer and Wiley-Blackwell. While there was general enthusiasm for the concept, one publisher had misgivings about the principle of providing usage statistics at the individual article level, while more were concerned about the practicalities of providing such reports. It was, however, reassuring to discover that all the publishers who responded use digital object identifiers (DOIs) to identify all versions of a single published work. The responses revealed a couple of items of concern, however, that made us realize that we probably had not communicated our mission quite as well as we had intended. There was some concern that providing usage reports to *institutional customers* on all articles published in all journals they subscribe to is one of our goals, and it is not. Clearly any resulting reports would be unmanageably large, not to mention of limited use. Our initially more modest ambition is to develop procedures and define a template that would provide reports on the usage of specific articles to specific authors or research funders. At a later stage, if work currently being carried out elsewhere on standards for definitive institutional and author identification bears fruit, it might be possible to develop reports that summarize usage of the research output of a single institution (which could become a useful tool in research-assessment contexts) or the usage of all articles by one author.

Our survey of institutional repositories was carried out by Paul Needham of Cranfield University, who made a point of ensuring that, as a minimum, he covered sites using the four systems (DSpace, Eprints, Fedora and Digital Commons) whose software accounts for over 80% of all digital repository implementations worldwide.

An encouraging aspect of the survey response was that the overwhelming majority of IR respondents add digital object identifiers to their records, where they are available. This is a promising finding since it indicates a potentially workable way of mapping articles stored in institutional repositories to their equivalent 'versions of record' on publishers' websites. However, significant challenges remain to be addressed. For example, there is currently no standard or automated process whereby DOIs are sought, retrieved and allocated to IR items, and there is a wide variation in which field in the repository article record the relevant DOI is stored (although Eprints sets a good example here by usually storing it in a field specifically designated for the DOI).

Phase 2: collecting and collating article usage data from multiple sources

Phase 2 of the PIRUS Project is formally described as 'the development of draft usage reports and protocols for the recording and reporting of individual article usage from a number of sources, and testing these with data from specific publishers and institutional repositories'.

To achieve this, we identified two specific goals that we needed to aim for: firstly, to define and identify relevant source items, and secondly, to look at how to collect data about usage of these items and then collate and display it in an appropriate form. The first of these two goals set us an interesting challenge – how to consistently identify peer-reviewed research articles, wherever they are hosted. Our current thinking is to use the following criteria: firstly, only items with a DOI will qualify to be counted; secondly, the 'resource type' element needs to be clearly identified so that we can exclude any agreed exceptions to the above. Ideally, agreement would be reached on the use of a term such as 'journal article' as a 'resource type' descriptor.

Meanwhile, achieving the second of the two goals of Phase 2 – resolving how we might collect and collate the usage data from multiple sources – set us what turned out to be an even more interesting challenge. After some discussion within the project team, we identified three usage collection strategies. The first one, which was closest to our original vision, was that all institutional repositories and aggregators should compile and make available a COUNTER-approved report for every article hosted. A draft XML-structured template for this has been developed and is published in our project's final report.

One concern, however, was that this might not scale well. So we started to explore alternative methodologies. In particular we decided to explore a technique similar to that which is currently used with great success by the Google Analytics service – that is, the concept of small tracker files embedded in each downloadable item. These files comprise short scripts which send a brief identifying message to a central server every time a request for an item results in a successful download. We have written scripts and conducted a number of tests of this methodology in which messages transmitted as a result of item download requests from both DSpace and Eprints repositories at Cranfield and Southampton respectively have been successfully received and resolved by an independent third party

Our third usage collection strategy is to recommend that locally held institutional repository usage data be structured and exposed in such a way that it can be harvested by any agreed-upon independent third party.

Phase 3: bringing it all together

The final phase of the PIRUS Project is to bring together all the different components of our research, summarize our findings and proposals, and submit that as a report both to JISC, who have funded our research via the PALS Metadata and Interoperability Programme (a collaborative venture with the Publishers Association and ALPSP) and to COUNTER for consideration as the launching pad for the development of a new article-based code of practice. By the time this paper is published, our report will be available on the COUNTER website.[4]

References

1. http://ie-repository.jisc.ac.uk/250/1/Usage_Statistics_Review_Final_report.pdf (Accessed 2 February 2009)
2. http://www.knowledge-exchange.info/Default.aspx?ID=164 (Accessed 2 February 2009)
3. http://www.projectcounter.org/r3/Release3D9.pdf (Accessed 2 February 2009)
4. www.projectcounter.org (Accessed 2 February 2009)

■ **Richard Gedye**
Research Director
Journals Division
Oxford University Press
Great Clarendon Street
Oxford OX2 6DP, UK
Tel: +44-(0)1865-353785
E-mail: richard.gedye@oxfordjournals.org
Web: www.oxfordjournals.org