# Measuring up to expectations?: Usage data and electronic journals

*Terry Hulbert*

An Overview of the UKSG seminar at the Scientific Societies Lecture Theatre, London,
Monday 22 October, 2001

*Terry Hulbert, Business Development Manager, Institute of Physics Publishing, Dirac House, 6 Temple Back, Bristol BS1 6BE*
*E-mail: terry.hulbert@iop.org*

**Introduction**

It seems only a couple of months ago that publishers ran around in a frenzy of activity, keen to sign bilateral reference linking agreements with each other. Negotiating these and subsequently becoming involved in organising my company's involvement in the CrossRef initiative certainly kept me very busy. How times have changed. Only a year or so later nobody really remarks about reference linking any more, other than in astonishment if you haven't yet implemented it. There are still plenty of issues and initiatives to be resolved, but the ongoing work occurs pretty much in the background. There's a new kid in town now: usage statistics. And there's no escaping it. Open any of the trade publications or speak to anyone in the library, vendor or publisher communities and it's the phrase on everyone's lips.

In an effort to bring people up to date with the many initiatives and efforts in this important area, UKSG organised a seminar, with speakers representing many of the interested parties – publishers, librarians (academic and corporate) and professional organisations (such as ARL and PALS).

The Chair, Richard Gedye of Oxford University Press, outlined some of his own thoughts prior to the presentations, by recalling that ten years ago he attended conferences where librarians would discuss measuring usage. Unfortunately library-user co-operation was not always forthcoming; the task was labour-intensive, and publishers knew even less than the librarians about genuine usage. In theory, electronic journals are changing this, but of course we need to see how we can exploit the sheer volume of data that is available in log files; we need consistency and compatibility. Several of the papers from the meeting are published in full in this issue. This paper summarises the others.

### Developing a code of practice for usage statistics: the work of PALS

*Hazel Woodward, Cranfield University*

Hazel began by updating the 80-strong audience on a recent initiative organised by PALS (the Publishers and Libraries Solutions group of JISC, the PA and ALPSP).

After outlining why both libraries and publishers need usage statistics, Hazel summarised the current state of play by stating that, although things are better than only a year ago, there are still some vendors who do not provide any usage data at all! Also, there are many variations in definitions and inconsistencies in report delivery and formats. Many publishers still overlook the fact that libraries operate in different environments and have different data requirements. Any initiative in online usage statistics must aspire to the three Cs – credibility, consistency and compatibility.

PALS have created a working group to develop a code of practice for vendor based electronic journal and database usage statistics. This code of practice will include guidance on many things including: (a) which data-elements should be measured, (b) data definitions, (c) output report formats/frequency/granularity, and (d) combining usage data reports from a variety of sources. It was intended that this work should also complement existing initiatives from the ARL, ICOLC, NCLIS and others: there's no point in duplicating effort or re-inventing the wheel.

The group's terms of reference are to: (a) review existing work being undertaken in this area, (b) consider the most appropriate mechanism to produce the code of practice, (c) develop the code of practice by the spring of 2002, and (d) promote and gain acceptance of the code of practice.

Much work has been completed since its humble beginnings in September 2000. All pre-existing projects were reviewed and a dialogue established with each one. An international forum was also organised in June 2001 to assess the degree of concern and commitment, to target areas of complexity and to identify areas of agreement and unresolved issues requiring further work. Following this, a number of task forces were created to focus on specific issues such as data processing, reporting and gateways/hosts. Each of these groups comprises key members from academia, primary and secondary publishers, hosts, etc. A key milestone is February 2002 when the Working Group which is steering the project hoped to circulate an initial draft of the code of practice; Spring 2002 should see the code published.

More information on this project can be found at www.usagestats.org

### Usage statistics in the corporate library

*Roger Brown, GlaxoSmithKline (GSK)*

Roger's presentation provided an insight into the issues surrounding usage statistics for corporate libraries. Many corporates make the provision of statistics an obligation within any licences that they sign, and they prefer access to content electronically as this overcomes many of the physical limitations that they have.

GSK have many global licences and the collection and collation of usage data is of paramount importance in the corporate world, where it is used for cost–benefit analyses, budget justifications and business cases. But there are problems and issues that are particularly relevant to corporate libraries.

Access to the content must provide a return on the investment, both at the gross level (all licensed content), as well as at individual title level. Similarly, any licence needs to be flexible as a result of rapidly changing business needs; it's quite possible that journal requirements for November could be completely different to those of January. Therefore, there is a requirement to change content almost 'on-the-fly'.

GSK do not collect and collate their own data but are dependent upon information provided by vendors. They do, however, count usage information from their Intranet, and look to combine the two figures as appropriate. This dependence does cause problems. Not all of the vendors actually provide this data, and of those that do, not all provide a breakdown of usage by month, or even by title. Indeed, only a quarter of the vendors provide all of the information that GSK requires.

Without divulging too much sensitive

corporate information Roger was able to say that online usage had increased some 400% over the last two years – this information related to full-text accesses. This important data is used for budgeting purposes and for managing the move to a 'virtual library'. Some interesting information that was divulged showed the high usage of a core number of titles, representing only 5% of the title coverage. Similarly, GSK were able to establish by examining usage data that a significant number of the 'least used titles' had print subscriptions, whilst a number of the 'most used titles' did not have a print subscription, raising questions with regard to the previous collection strategy.

Roger summarised his presentation by highlighting that not all vendors provide usage data, not all of it is detailed enough, it's not always current enough, it's not always comparable, and there are credibility issues (as a result of browser problems). His suggested solutions were to work closely with publishers and, more importantly, to support any initiatives that were addressing these issues.

**Auditing publisher usage statistics**
*Jerry Cowhig, Institute of Physics Publishing*

Jerry was attending another conference in London but was happy to spare the time to make a presentation regarding his own efforts in the area. He has been instrumental in preparing a plan to publish audited usage statistics – an initiative he calls *EAJUS* (Electronic Article and Journal Usage Statistics).

He began by outlining IOPP's own experience with usage statistics, where they have been made routinely available since the launch of their Electronic Journals service in 1996. Two recent experiments – making access to all online content free to all during the last two months of 2000, and (titles since January 2001) free access to the current issue of all – have also offered rich usage data for analysis. It is no exaggeration to say that more and more people are reading the titles as usage doubles each year.

Jerry pointed out that current online usage statistics are good but they exist in isolation. They also look at how content is used and accessed for single publishers only, with each publisher usually gathering, storing and providing the data using their own proprietory methodologies. In other industries, such as magazine publishing, TV, radio, etc., this information is audited and compared, and often published in directories such as BRAD. So why don't we use and publish this type of 'readership' data? Jerry suggests that it may simply be that there's no heritage of reader research and making this type of information available within the STM journal publishing community. And he posits that STM journal publishing 'may be the only major publishing industry that has no decent statistics'.

Existing data is provided for a single publisher only, and it's provided in confidence for an institution or consortium only. What's more there is an ever-increasing number of methodologies, with absolutely no independent validation. As an example, compare this to the magazine industry, where measurement systems are clearly defined, and results are independently measured (or submitted) and subsequently audited. Many publishers are included and compared and the results are published.

There are suggestions that impact factors serve this purpose. However, impact factors don't address readership, as they only provide an indicator of citation data. Jerry advocates the introduction of a single number that gives a journal score – for readership. This would, of course, need to be audited. It might be calculated in a number of ways; the important thing is that it's consistent and used by everyone. For example, one might use the average number of full-text downloads per paper, per title, in the first year after publication, or the average number of full-text downloads per month, per title, in the first month after publication.

Jerry concedes that there are imperfections in this process, but maintains that they will exist for everyone – a level playing field – and as long as everyone is aware of them, then they can be addressed as technologies and methodologies improve over time. But why would publishers do this? Jerry outlined a number of important reasons, among them: this data is valuable to customers, it will support digital libraries, it will help publishers recruit and motivate authors, it's valuable to authors/faculties, it will demonstrate the high use of e-journals – and customers will demand it!

Jerry's presentation also addressed the costs and logistics of creating an infrastructure to address this initiative, and concluded that it was very affordable, especially if publishers were to contribute a nominal annual fee to support it and make the data available to all. In concluding, he pointed out that the electronic age now means that much data is now available, though not shared, and that librarians quite obviously wish to pool statistics. It would be useful to create a new industry measure; a top-line usage 'score' for a journal title that's audited and published.

**A publisher perspective – working together to understand usage**
*David Sommer, Blackwell Publishing*

David provided an excellent overview of usage statistics from a publisher's perspective, looking at many of the issues raised in earlier presentations and detailing how publishers use the usage information that is available to them. He stated that Blackwell Publishing, like many publishers, previously had very little data with regard to print journal usage and readership.

New online data offers new insights, with publishers able to identify the most popular articles, journals or subject areas, and also look at a geographical or market breakdown of usage. This in turn can influence a publisher's editorial policy. Content usage can also be linked to functionality usage, providing insights into website design and navigation: how are users navigating to the content, what is the ratio of browsing:searching:linking, and so on. Likewise, an analysis of geographical use and response times can help with decisions on mirror sites and global caches.

However, David stressed the dangers of over-reliance on or overinterpretation of online usage statistics. Whilst the information can provide concrete data on what precisely is being accessed, isolated data is of little use. This information is only one element from a variety of sources, and it's critical to examine usage over a period of time and to identify trends. More importantly, the data is not a substitute for talking to people and provide a context for any numbers – what does high usage mean?

David also reminded the audience of the various initiatives that are under way, or have already taken place – the PALS Working Group, ICOLC guidelines, the ARL E-metrics project, etc. He summarised by stressing that publishers are committed to providing data, but that there is a clear need for standards – hence a lot of the ongoing work – and all stakeholders must be involved for anything to succeed. And he concluded by reiterating that there needs to be continuing dialogue between all interested parties.