Usage Data – The Academic Library Perspective

Louise Cole

A paper given at the UKSG seminar, 'Usage Data for E-Collections: What Do We Want? What Does It Mean?', London, 7 June 2000

This paper looks at various aspects of usage data collation and provision in the area of e-journals management in academic libraries. While some publishers have begun to make useful data available, there is not yet an acceptable solution for librarians, publishers, and users. Current and future trends are outlined and discussed.

Louise Cole is E-journals Co-ordinator, Brotherton Library, University of Leeds, Leeds LS2 9JT
E-mail: L.Cole@leeds.ac.uk

Overview

This paper is divided into a number of sections:

- E-journal provision at the University of Leeds. A brief look at what we are currently providing.
- What are usage statistics? If we say we want usage data, what do we actually mean?
- Who provides this data? Not only where are we getting statistics from, but also what kind of quality and reliability do we expect?
- What do we need to know? What information will be more useful to us as academic librarians working within collection management policies?
- What are the problems connected with the provision and collation of usage data?
- Are there any solutions to these problems?
- Since this is the 'academic library perspective', what are the advantages for us?
- Are there advantages for publishers too?
- What would be an ideal situation in which all parties could benefit from usage data?
- And a brief conclusion …

E-journal provision at the University of Leeds

We currently provide access to approximately 3,000 electronic journals: Arts 5%, Sciences 46%, Health Sciences 18%, Social Sciences 31%.

We have titles from publishers' sites, aggregrators, portals, as part of databases, and the usual mix.

The subject breakdown is achieved the same way our faculty teams are arranged – that is, Arts, Social Sciences (Business, Law, Economics and Social Science), two sciences teams (Biological Sciences and Physical Sciences), and Health Sciences.

And, of course, we provide remote access to titles where site licences allow it.

All our electronic journals are in the library catalogue and on local web pages. We also have a searchable subject listing. The web address for the University of Leeds library catalogue is http://lib1.leeds.ac.uk and for our subject breakdown page http://www.leeds.ac.uk/library/ejourn/

What are usage statistics?

The simple term 'usage data' can cover a whole range of possibilities:

Which titles are being accessed? Are there certain titles that are more popular than others? What about titles we do not hold in print? Are they the titles you would expect to see well used in their print versions? What about titles which are electronic-only?

Which subjects are most popular for e-journal use? It is popularly thought that in the arts, there is little interest in electronic media – is this true? You would expect to see high usage in STM areas, but what about law, business, theology, and politics? What about titles which are multidisciplinary in subject?

Who is using the e-journals? Again, expectations would suggest the postgraduate researcher, the academic, but what about undergraduate students, university members on placement and students and staff from unrelated academic departments? This can be monitored to some extent at enquiry desks but there is no substitute for hard facts, and here usage data can be very helpful.

Where are they using the e-journals from? At the moment it is unclear whether this level of information can be provided, but it would be useful to know whether people are coming into the library, using computer clusters, accessing from their own desktop whether at work or at home, and so on.

What part of the service is being used? Are tables of contents and abstracts more popular than full-text? What kind of full-text is favoured – html or pdf, for example? What about archives, search and browse facilities, and additional features such as chat rooms and forums?

What is 'popular'? By which I mean which aspects of the service are favoured most highly – the text in advance of the print edition, the search facilities, the additional material available online only?

Who produces usage data?

Moving on to who could (should?) provide the data.

As our numbers of electronic journals grow, I am sure we have all begun to look at ways in which we can monitor their usage. After all, if we spend time trialling our new print titles, why not electronic ones?

In an environment where our collection management divisions are looking at statistics for all aspects of acquisitions, serials, and metadata processes, and when we as workers are providing figures on our daily working activity; we have to look at the 'big picture' in order to make informed decisions on e-journals in relation to such developments as Periodicals Voting Exercises, Resource Allocation Models, external funding, staff time and resources, dedicated course support, and so on.

Publishers and intermediaries

There are many of these already providing usage data, for eg. American Chemical Society, Annual Reviews, Synergy, IDEAL, Science Direct and SwetsNet

Quality

Level of statistics, and how they are accessed, varies widely. However much we would like a standard form of presentation, the current situation ranges from detailed information available on the publisher's website, through to a very basic presentation of login numbers sent out monthly. It is far easier to look at detailed figures provided in tabular format and broken down into a number of headings, than to navigate through an enormous list of statistics arranged by numbers of logins only.

In-house usage data:

Of course we can collate some figures in-house, but these would be very basic – counting website hits or use of links from the catalogue. Once someone has left our pages/services, we would still need to rely on others to allow us to see where they are going on someone else's site. All we would be able to say for certain is that a person followed a link to the home page of a service or a title. From there they could have spent a long time searching and downloading full-text, they might have viewed abstracts, but they might simply have decided the site was of no interest to them and moved on.

If we had to think of a number of things we needed to know about e-journal usage, what would they be?

Which sites are being accessed?

Just making a link is not enough. However much effort we put into selecting a title, cataloguing it, publicising it on our web pages, what we really need to know is – is anyone actually using the resource?

This is particularly important if we are paying for access. We need to know if we are getting value for money, or whether we are wasting resources. In a time when we are all aware of a squeeze on funds, we need to allocate the money we have in the most appropriate way.

What are all these electronic titles, databases, packages of resources, being used for and how is the information they contain being exploited?

This can be quite important – often access to contents and abstracts is free, while access to full-text requires registration, a parallel print subscription, or payment of a fee. If we find that a large number of users are simply looking at contents, do we have an argument for withdrawing a subscription and simply making the service available as part of a 'contents only' set?

If one aspect of a site is not being used, we need to look at the reasons why. Without usage data, we do not know what is happening. Therefore we cannot act on our knowledge and our professional awareness is diminished. Perhaps, after all, it is a lack of publicity on the part of the library. Perhaps it is the user interface that turns people off – in which case, do we need to rethink our training programmes?

Which titles are being looked at – particularly important when we have purchased everything from one publisher online. Are they all being used?

Numerous deals, particularly those instigated by NESLI, in recent years have had the additional 'advantage' of having a number of 'electronic-only' titles added on.

In many cases these are titles we have previously cancelled due to unpopularity or rising costs. It would be interesting to know whether titles we do not hold in print are proving popular electronically, and to what extent. This could not only influence collection management policies for the future, but could also indicate whether such deals are the way forward, as many commentators have speculated.

If the hybrid library is going to become a reality, if electronic-only collections are to expand in the way some observers have speculated, we need to have solid knowledge and reasoning behind our decisions. This is, I would argue, particularly important if we are getting involved in three-year long deals with 'no-cancellation' clauses.

Which subject areas are the most popular amongst our users?

It is generally perceived that scientists are the academic group more likely to use e-journals (e-collections), but what about health workers, academics in business and management, law, or politics, those working in the arts fields?

The availability of full-text publications in resources such as ABI-Global have increased the profile of the social science e-journal, but has this trend in provision meant an increase in access? Indeed, without proactive marketing of these services by the library, is there an awareness of them? Is it a fact or myth that academics in the arts still shy away from electronic information? What about services such as Literature Online?

Issues surrounding remote access

It is clear that one distinction that needs to be addressed is the access to resources by users on-campus, either from library PCs or their desktops, and users accessing resources from home or placement by means of user-id and password. Some of this can be monitored locally, such as remote logins to our webpages, but when it

comes to a journal title mounted on a publisher's website, for example, or a database accessed through a password, how can we determine who is accessing the resource and from where?

If concrete usage data was made available which indicated a high level of usage off-campus, perhaps that would persuade publishers that cumbersome registration procedures and dissemination of unique ids was not the solution to 'who should be allowed access'.

Are there any titles which consistently do not allow access, viewing, printing etc.?

Particularly if we are paying large amounts of money, I see this as essential. It would be interesting to see some data on denials and failure rates, either collected locally or obtainable from publishers' and/or aggregators.

If we can pin down the sites, which are consistently underperforming, again we are better informed about where to spend our resources in the future. A good site with a good reputation will gain more users.

If we have access from two places – e.g. IDEAL and ingenta, which is the most popular. This assumes we publicise both means of access to the same level.

Indicating which site is the most used could tell us a lot about both the service provided, and our own library users. It would also allow us to target our publicity and training programmes more specifically. There are also differences in service between intermediaries – does this affect their use?

How often should statistics be released?

This has often come up in the literature when discussing the provision of usage data – the average seems to be monthly, which is long enough to have some meaningful figures, and not too long to be overwhelming. There should also be an option to customise the statistics, for example by combining months together. This would enable us to track the usage of a service over a semester or an academic year.

Level of information

What should be provided? Some services send out just a list of numbers indicating hits which is really a little primitive when we know what technology can do in the 21st century. A little better is a breakdown of how many hits were recorded on each part of the site – how many PDFs were viewed, downloaded, printed, and so on.

We need to liaise with other library sections such as reader services and with academic departments to find out what specifically is needed. It is likely everyone will require something different, but in the end what this all comes down to is are we spending our money as effectively as we think we are? And if we are not, what should we be spending it on instead?

And lastly, a nod to some of those providing excellent usage data already

Here I would have to mention Jstor at MIMAS, American Chemical Society, Proceedings of the National Academy of Sciences and an increasing number of Highwire titles. All available monthly and all fully detailed.

Some problems with usage data

The publisher holds statistics, but 'not in a form which can be made public'.

There is a need for all major publishers providing access on the Internet to their journals to allow usage statistics to be made available to subscribing institutions. After all, in many cases we are paying for the service. Even if not, the information is not held locally on our websites, the publishers' hold it. I have heard many times on telephoning a publisher that 'we can see x number of people logged in today', but when asked to give out that information, they won't. I think we, as academic librarians, are justified in saying 'why not?'

Statistics are made available in a very basic form – either as lists of logins, bald statements of articles viewed and/or printed without any breakdown, etc.

As I have said previously, technological advances mean we can all do better than screens of computer generated unreadable figures. That may be fine for computer programmers, but we busy librarians need something more useful to interrogate and work with.

Assume popularity

When a title is seen to be so widely regarded and respected, there seems little point in monitoring

its usage. For a title like Nature that cuts across so many disciplines, this could be true, but then again there are many other aspects to consider (and of course, trends do change).

Statistics are sometimes withheld, it seems, so the publisher can claim a site/title is more popular than it actually is. This seems pointless, even if a publisher claims it justifies removing the electronic version of a title at a later stage.

Proxy caches

I am sure we are all familiar with this problem, particularly when we try to make links to sites based in the USA. It is a peculiarly UK academic institutional problem that the JANET cache exists and, as such, complicates our access procedures. Not only does the use of a proxy complicate access by IP (often an institutional IP address is unrecognised and access is denied), it makes some statistics of on-campus usage unreliable.

Collecting the data ourselves

It may be that it is seen to be more practical for a library or institution's overstretched systems departments to collate statistics on usage of e-journals. Should we be doing it ourselves?

Moving on to the problems associated with different types of e-journal:

Usage statistics for journals available freely:

Understandably, this is not possible for the vast majority of these titles. There is no registration procedure and no IP address recognition. A solution to this from our point of view could be by logging the number of accesses to a local webpage for that title, or the number of accesses through the direct link to resource provided in the online catalogue. But it also remains less likely we would want to make use of data made available in this way.

Usage statistics for journals available as print add-ons, IP address recognition:

This kind of information should be easy for publishers to supply. They might argue that since we are getting the online access 'free' (not actually free, but bundled in with the print cost), it is not something we should expect. However, if we can be identified via our IP address it is not an impossibility. It would also be a compensation for those times we have to re-register for access despite still having a current subscription. Some titles make statistics available as part of their own PR exercise, or because they are proud of providing a good customer service (PNAS is a good example).

Usage statistics for journals available as print add-ons, user-id and password recognition:

This is a little more complicated if we are talking about a range of user-ids and passwords. If there is just one, the situation is much the same as with those titles accessible by IP recognition. Multiple user-ids cause problems with registration and access, and can be misleading if someone forgets the id and has to register twice, appearing as two users.

Usage statistics for journals available with additional fee for online access, IP recognition:

For titles which require us to part with additional sums of money, usage statistics should be part of the service supplied. This is now the case with a number of titles available through Highwire. Even basic statistics such as numbers of TOCs, abstracts, and full-text (HTML or PDF) accessed is preferable to nothing at all. The type of statistics provided by Jstor are a model example to the type of information which is useful to an institution.

Usage statistics for journals available with additional fee for online access, user-id and password recognition:

These can have the same problems as with print-add-ons accessed by the same method, but often one id is allocated for all users. This should identify the institution and, again, it should be easy for publishers to gather information on usage.

Usage statistics for journals available electronically only, at a fee:

Again, it depends whether access is by simple IP or by username and password, but for some of these titles, the financial outlay can be quite high. It should be included in the package and/or site license for the publisher or intermediary to provide quality data on the use of the service.

Possible solutions

Just a few thoughts here: Usage data should be included as an integral part of all agreements. If it becomes the norm for one publisher or provider, so it should be for all.

There has been talk for many years of interfaces becoming more alike, in order to offer a more user-friendly service to library users. Services like SwetsNet and ScienceDirect are already making this a reality – and just as we can expect to see more like interfaces, it is probable we will see statistics from different outlets in similar formats.

And of course, good solid usage data would prove that a service was being used. We could then use that to argue a deal should continue, or that a particular publisher's titles should be favoured.

Why does the academic library need e-journal usage data?

- To help in the selection and cancellation of print titles. If we know what is being used electronically, perhaps we can move over from our print collections. At the moment this is less likely due to the amount of print add-ons, or the high cost of electronic only titles, but there is a possibility in the future.
- To access which titles prove popular to library users.
- To pinpoint potential problems – if there are many accesses to PDF but there are few printing facilities, what will be the effect on users? What should we be doing to counteract these problems?
- To see which sites are being used the most – one would expect sites also available remotely to have high usage but is that really the case?
- To help decide whether to continue with packages which include a large amount of electronic-only titles. These may look good in theory but are they in practice?
- To be able to publicise key titles more effectively within the library, especially in areas such as management and law which are not traditionally thought of in the same light as STM subjects.
- To encourage computer-literacy amongst library users. Persuading them to look for an issue online when it is out on loan or unavailable on the shelves may be a bonus for staff at counter and enquiry points.

What are the advantages for the publisher?

- Good usage statistics raise the profile and reputation of the site.
- Can use statistics to encourage use and tailor services to meet user needs.
- They can see what is being used, when and to what level. This enables them to provide a better service in the future. For eg., if HTML articles are being accessed as much or more than PDF versions, there is an argument for keeping both.
- They can see which subject areas are beginning to prove popular and that may influence which titles are later made available.
- And of course, if something is popular, a publisher can justify making a charge for it.

A quick look at the ideal world:

Publishers should be able to provide data on demand and tailored to individual institutions.

A combination of IP and username access should be provided and justified by being monitored.

Statistics should be reliable and useful.

There should be a strong input both from the library and the user for what kind of usage data is wanted.

Finally … what is so good about usage data? Collected and used properly, it is an advantage for us all:

- Publishers find their services are used.
- CMS sections can make informed decisions.
- Users get better and more relevant services.