Repository Statistics: What Do We Want to Know?

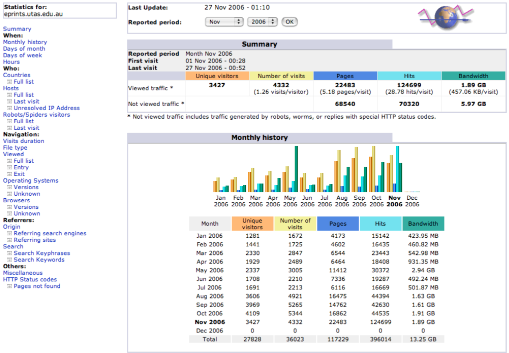
Leslie Carr, University of Southampton. lac@ecs.soton.ac.uk
Tim Brody, University of Southampton. tdb2@ecs.soton.ac.uk
Alma Swan, Key Perspectives. a.swan@talk21.com

Keywords: repository statistics, analytics.

Background

One of the factors motivating institutional repositories is the potential efficiency gains in various parts of the scientific and scholarly publishing cycle: visibility, dissemination, use and impact. Researchers want their work disseminated and used, and need it to be cited. Institutions want to increase their visibility, and funders want to maximise the effect of their investments. Both citations and downloads are relevant evidence for the use of research, and many recent studies have shown that in a variety of communities, download figures for open access papers are strongly correlated with subsequent citations.

Download data is being logged by every repository as a by-product of the Web requests they receive. This raw data is being and turned into useful download statistics for individual papers and users by a growing number of institutional repositories, thematic repositories (e.g. RePEc) and OAI services (e.g. Citebase). However, there is no consensus over what data needs to be collected, what filtering mechanisms are appropriate, and what analyses are useful for academics in various disciplines.



Statistics for: eprints.utas.edu.au

Last Update: 27 Nov 2006 - 01:10
Reported period: Nov 2006 OK

Summary

Reported period: Month Nov 2006
First visit: 01 Nov 2006 - 00:28
Last visit: 27 Nov 2006 - 00:52

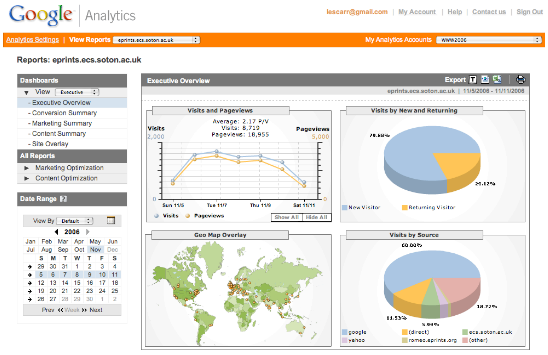
	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Viewed traffic *	3427	4332 (1.26 visits/visitor)	22483 (5.18 pages/visit)	124699 (28.78 hits/visit)	1.89 GB (457.06 KB/visit)
Not viewed traffic *			68540	70320	5.97 GB

* Not viewed traffic includes traffic generated by robots, worms, or replies with special HTTP status codes.

Monthly history

Month	Unique visitors	Number of visits	Pages	Hits	Bandwidth
Jan 2006	1281	1672	4173	15142	423.95 MB
Feb 2006	1441	1725	4602	16435	460.82 MB
Mar 2006	2330	2847	6544	23443	542.98 MB
Apr 2006	1929	2489	6464	18408	931.35 MB
May 2006	2337	3005	11412	30372	2.94 GB
Jun 2006	1708	2210	7336	19287	492.24 MB
Jul 2006	1691	2213	6116	16669	501.87 MB
Aug 2006	3606	4921	16475	44394	1.63 GB
Sep 2006	3969	5265	14762	42630	1.61 GB
Oct 2006	4109	5344	16862	44535	1.91 GB
Nov 2006	3427	4332	22483	124699	1.89 GB
Dec 2006	0	0	0	0	0
Total	27828	36023	117229	396014	13.25 GB

Figure 1a: AWStats



View Reports: eprints.ecs.soton.ac.uk
My Analytics Accounts: WWW2006

Reports: eprints.ecs.soton.ac.uk

View: Executive

eprints.ecs.soton.ac.uk | 11/5/2006 - 11/11/2006

Average: 2.17 P/V
Visits: 8,719
Pageviews: 18,955

View By: Default

2006

Jan	Feb	Mar	Apr	May	Jun
Jul	Aug	Sep	Oct	Nov	Dec

	S	M	T	W	T	F	S
→	29	30	31	1	2	3	4
→	5	6	7	8	9	10	11
→	12	13	14	15	16	17	18
→	19	20	21	22	23	24	25
→	26	27	28	29	30	1	2

Figure 1b: Google Analytics

The most common approach to gathering statistics on repository usage has been the application of common web sites analysis tools, such as AWStats (Fig 1a) which is hosted on the repository's local web server and Google Analytics (Fig 1b) which is hosted externally by Google. However, both of these techniques suffer from the limitation that they treat the website as a whole, with a broad selection of information available about general usage, but the information about specific pages (in the case of a repository, specific eprints or bitstreams) is very much reduced.

Homegrown solutions have been used by various repositories and are processed and presented to the local community in various ways. For example as download statistics for each eprint broken down by countries of origin, or as a league table of most-accessed documents and most-read authors. Such statistics may be made public, or may be used internally by repository management as evidence of the effectiveness of the repository. While satisfying a local need to provide usage figures, these statistics open up further problems for the repository community

(a) they are not sharable, often provided in graphical form, or available to authorised users only
(b) there is no agreement as to how to share such information (standard statistics, standard locations, standard formats)
(c) there is no agreed baseline for comparison of the statistics.

Furthermore, received wisdom is that Web logs statistics are inaccurate and misleading, despite being the basis of many analysis tools, and the foundation upon which many business and marketing decisions are taken. It is true that there are well-rehearsed problems with the interpretation of web logs; naïve summaries of web sites based on unscreened totals of user clicks lacking careful pruning and any form of contextual information are indeed unhelpful. However, Institutional Repositories are not arbitrary web sites; they are well-designed information resources, with a regular and interpretable structure and a well-defined objective. They contain a large array of 'records' or 'items', each of which hosts links to component bitstreams; this regularity and the interpretation of each item playing an objectified and understood role in the scholarly and scientific communication environment means that a repository avoids many of the problems of general web site analysis.

There are still issues of interpretation that make analysing the usage of repositories difficult; however rejecting web log statistics out of hand because they are less than 100% accurate is an over-reaction. Web logs contain evidence of user interaction with the repository and its contents; we should maintain a cautious and balanced approach to the interpretation of such evidence, but we should not eschew it. Web logs do not contain ALL the evidence that we would like, nor is all the evidence that they do contain about human users. They also record evidence of search engine crawler interaction with the repository, and as the number of such engines and the depth of their coverage increases, the burden of services like Google and Yahoo increases dramatically.

In an ideal world authors would like to know how much "genuine academic usage" each item in a repository achieved. Authors are less interested in how many "clicks" a paper attracted than in how many researchers "engaged with the contribution" of the paper. A paper that is intentionally 'downloaded' and even 'printed' as part of a trawl of the latest papers on a topic may in fact be discarded after a brief glance at the title. Not only is this impossible to determine, but even if the paper is subsequently cited it still may not have been read!

Overall, there has been no consensus building yet among stakeholders over what data needs to be collected, what filtering mechanisms are appropriate, and what analyses are useful for academics in various disciplines and for the publishing industry who have a natural interest in the usage of Open Access versions of their subscription-articles. However, appropriate analyses of the usage of a repository should be able to answer the questions of multiple stakeholders for multiple purposes.

- For a researcher: how much attention is my research receiving and from whom? What items are currently receiving most/least attention? How have people found out about my work? Am I generating the same amount of attention as my peers/colleagues/competitors?
- For a research manager: what marketing activities and web site restructurings boost the audience for the institution's research? Which items are getting disproportionate attention, and why?
- For a repository manager: is the repository being used in line with its business objectives? How many items are being deposited/downloaded, how often, of what type and by what part of the repository's constituency?
- For repository designers: How are repositories used? What information seeking behaviour is evidenced in the logs? What internal navigation or searching features are evidenced? How should measured repository usage inform the design of next generation repositories?
- For open access campaigners and repository funders: Which repositories are live and effective, and how is this related to their policy and administration? In the long term, how is the pattern of downloads correlated to the pattern of citations?
- For library funders and publishers: how do the download figures for an author's eprint contribute to the total usage of the published article as measured by the publishers' holdings.

Implementation

The JISC *Interoperable Repository Statistics* project (irs.eprints.org) addressed these issues. Its first task was to design an open API to collect usage data from repositories was relatively straightforward. The concept of an *access* can be neatly captured by the *OpenURL ContextObject* schema, and this can be delivered by the OAI-PMH protocol. Making the sharing (technical interoperability) of the usage information build on an existing protocol that all repositories already support, reduces any barrier to adoption and the expense of supporting this new repository feature.

In order to give repository managers a choice of service modalities, the project looked at providing both an OAI service for providing analyses of statistics and also a repository-hosted software module for providing locally-produced and locally-incorporated analyses, graphs and visualizations.

Both these approaches are shown in figure 2. Repository users might be able to see a full report of the usage of their articles *if they were to go to the Citebase web site.* In addition, Citebase offered thumbnail usage graphs for the repository to include directly in its web pages. Consequently, simply by making its log files available for harvest (to select services), a repository (based on whatever software) may gain a valuable usage analysis service without the overhead of having to run any processing internally.

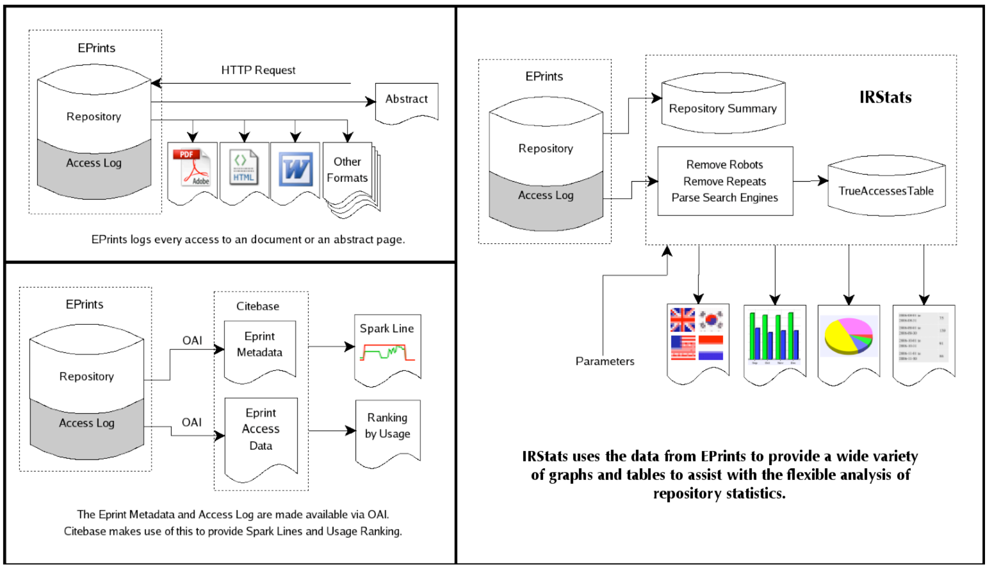

EPrints
Repository
Access Log
HTTP Request
Abstract
PDF
Adobe
HTML
Other Formats
EPrints logs every access to an document or an abstract page.
EPrints
Repository
Access Log
OAI
OAI
Citebase
Eprint Metadata
Eprint Access Data
Spark Line
Ranking by Usage
The Eprint Metadata and Access Log are made available via OAI.
Citebase makes use of this to provide Spark Lines and Usage Ranking.
EPrints
Repository
Access Log
Repository Summary
IRStats
Remove Robots
Remove Repeats
Parse Search Engines
TrueAccessesTable
Parameters
IRStats uses the data from EPrints to provide a wide variety of graphs and tables to assist with the flexible analysis of repository statistics.

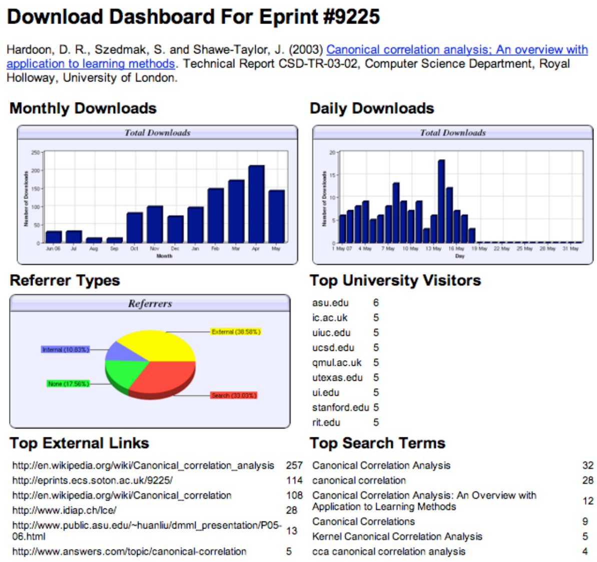
Figure 2: IRS software and the Repository

In addition, the repository may choose to process the log table internally by performing its own crawler filtering and data transformation (using tools such as the GeoIP library) to create another database table that is oriented towards user queries. This construction of this final table is designed to efficiently create reports that allow individual users to see aggregated overviews of the downloads of all their papers, broken down by year/week/month/quarter (displayed in tabular or chart form or exported as a spreadsheet) or to drill down to individual accesses from specific sites in responses to specific query terms from an external search engine. It is this flexibility which the stakeholders (above) need, so that they can explore the as yet unknown whys and wherefores that contribute to the use of their intellectual outputs.

It was this version of the software (rather than the OAI service as prototyped by Citebase) which became the main focus of the IRS project, and resulted in the IRStats package that allows repository managers to incorporate it into their repositories under their own control.

IRStats Package

IRStats is an open source software package for analysing the use of items in institutional repositories. This package includes components for sharing repository usage data and for analysing it locally to produce graphs, charts and tables to include in the host repository. The package reads the repository web logs and creates a database of access events. That database is used to respond to OAI queries from download analysis services (on a pre-arranged basis to maintain necessary privacy) and also used to drive the local analysis software. The local analysis components provide a palette of charts, tables and graphs for the repository manager to include in the local site or to provide as private management reports as required.



Download Dashboard For Eprint #9225

Hardoon, D. R., Szedmak, S. and Shawe-Taylor, J. (2003) Canonical correlation analysis; An overview with application to learning methods. Technical Report CSD-TR-03-02, Computer Science Department, Royal Holloway, University of London.

Monthly Downloads

Daily Downloads

Referrer Types

Top University Visitors

asu.edu	6
ic.ac.uk	5
uiuc.edu	5
ucsd.edu	5
qmul.ac.uk	5
utexas.edu	5
ui.edu	5
stanford.edu	5
rit.edu	5

Top External Links

http://en.wikipedia.org/wiki/Canonical_correlation_analysis	257
http://eprints.ecs.soton.ac.uk/9225/	114
http://en.wikipedia.org/wiki/Canonical_correlation	108
http://www.idiap.ch/lce/	28
http://www.public.asu.edu/~huanliu/dmml_presentation/P05-06.html	13
http://www.answers.com/topic/canonical-correlation	5

Top Search Terms

Canonical Correlation Analysis	32
canonical correlation	28
Canonical Correlation Analysis: An Overview with Application to Learning Methods	12
Canonical Correlations	9
Kernel Canonical Correlation Analysis	5
cca canonical correlation analysis	4

Figure 3: The Download Dashboard Visualisation

The range of analyses encompasses simple access counts (how many times was this item downloaded), league table functionality (what are the top 10 most downloaded items / authors) and more sophisticated calculations (which items are the ‘highest climbers’ in the league tables). A simple deployment of the statistics might be to include a monthly download graph on each item’s abstract page; a more extensive use of the package could make an aggregate collection of statistics available from different perspectives. One such pre-prepared example is provided by IRStats – the so-called “Download Dashboard” (see figure 3) intended to give the author of an item a comprehensive picture of its accesses and the reason for those accesses.

The selection of visualisations is provided by a form interface that allows the user (or repository manger/designer) to select the individual record or collection aggregate whose bitstream downloads are to be analysed, the time period over which the analysis occurs and the visualisation required (the graph, chart or table). The result is an HTML fragment which can be incorporated into a portal, user’s web page or (more likely) a repository.

Conclusions

This project aimed to achieve a mechanism for capturing, using and sharing download information about repository resources in a way that was comparable between repositories and repository platforms. Statistics services are a natural area for repositories to move into. At one level they provide the repository management with a comprehensive understanding of the use of their environment, the impact of policy or functional updates and provide their institution’s senior management evidence of the return on investment. At another level they provide researchers and authors with valuable information on the interest in their intellectual work and (crucially) insight into how their work is being promoted or referred to, and how their own digital profile is manifest.

In an environment where the Web is a platform for research, it is essential that researchers and managers are given the tools to understand the channels through which their own work is made available, the visibility and effectiveness of their distribution mechanisms. An increasing number of studies (Hitchcock 2004-7) demonstrate the link between increased downloads and increased citations; in a scholarly culture where citation increasingly defines research quality which in turn controls research funding, repository statistics are not merely interesting commentaries on surfing habits in the scholarly community, but vital business intelligence.

References

EPrints Software: www.eprints.org

Hitchcock, S. (2004-7)The effect of open access and downloads ('hits') on citation impact: a bibliography of studies. http://opcit.eprints.org/oacitation-biblio.html

IRS Project Site: irs.eprints.org

IRStats Software: http://trac.eprints.org/projects/irstats