

2006

Download statistics - what do they tell us? The example of research online, the open access institutional repository at the University of Wollongong, Australia

Michael K. Organ

University of Wollongong, morgan@uow.edu.au

Publication Details

This article was originally published as Organ, MK, Download Statistics - What Do They Tell Us? The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia, D-Lib Magazine, 12(11), November 2006. Original journal available [here](#).

Download statistics - what do they tell us? The example of research online, the open access institutional repository at the University of Wollongong, Australia

Abstract

A study was undertaken of download and usage statistics for the institutional repository at the University of Wollongong, Australia, over the six-month period January-June 2006. The degree to which research output was made available, via open access, on Internet search engines was quantified. Google was identified as the primary access and referral point, generating 95.8% of the measurable full text downloads of repository content. Further long-term studies need to be carried out to more precisely identify factors affecting download rates of repository content. This data will assist institutions and faculty in measuring research impact and performance, as an adjunct to traditional bibliometric tools such as citation indexes.

Disciplines

Arts and Humanities | Social and Behavioral Sciences

Publication Details

This article was originally published as Organ, MK, Download Statistics - What Do They Tell Us? The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia, D-Lib Magazine, 12(11), November 2006. Original journal available [here](#).

ARTICLES

D-Lib Magazine November 2006

Volume 12 Number 11

ISSN 1082-9873

Download Statistics - What Do They Tell Us?

The Example of Research Online, the Open Access Institutional Repository at the University of Wollongong, Australia

[Michael Organ](#)

Project Manager, Research Online
University of Wollongong
<morgan@uow.edu.au>

Abstract

A study was undertaken of download and usage statistics for the institutional repository at the University of Wollongong, Australia, over the six-month period January-June 2006. The degree to which research output was made available, via open access, on Internet search engines was quantified. Google was identified as the primary access and referral point, generating 95.8% of the measurable full text downloads of repository content. Further long-term studies need to be carried out to more precisely identify factors affecting download rates of repository content. This data will assist institutions and faculty in measuring research impact and performance, as an adjunct to traditional bibliometric tools such as citation indexes.

Introduction - consideration of new performance measures

As universities and funding bodies worldwide seek to quantify return-on-investment for research output, and more closely monitor individual academic and institutional performance, new forms of bibliometrics are being sought that go beyond journal quality assessment and into the area of research quality and impact [1].

The measurement of research impact is an area with which Australian universities are currently grappling, due to the proposed introduction of a Research Quality Framework (RQF) by the federal government in 2008 [2]. Though based on similar assessment processes in the United Kingdom and New Zealand, the precise details of the RQF are not known, though it is clear a variety of new performance measures are being considered to assess research impact [3].

The onset of the Internet and an ever-changing information technology landscape have provided new statistical sources to supplement the data available from traditional tools such as citation indexes and journal circulation figures. For example, Thomson's Web Citation Index and Google Scholar are both developing statistical packages to monitor Internet usage of research outputs.

A new suite of bibliometric data available to research organisations and funding bodies is the download statistics generated by institutional repositories. As research outputs are increasingly exposed to the web via search engines such as Google, organisations will be looking at, and making use of, institutional repository statistical packages. Studies have already shown that the placement of research papers in open access repositories can increase citation rates by anywhere from 50 to 500% [4]. This is driving vice-chancellors and CEOs to sign-on to these new pieces of research infrastructure, with the promise of improving and enhancing the reputation of their institution and research staff.

It is still early days in regards to the institutional repository movement. Software remains under-developed and sustainable economic models are in their infancy. Whilst return on investment may be in the order of 30:1, the message is yet to be disseminated amongst senior executive and funding bodies as to the real worth of an institutional repository [5]. In Australia, for example, only eighteen of the 38 higher education institutions have an open access repository, and only one of those (Queensland University of Technology) has mandated deposit of research material [6].

As with any new technology it is not clear precisely what effect these repositories will have on the research process, or what conclusions can be drawn from the statistics currently being generated, apart from the broad acceptance of the role they play in improving citation rates, as noted above.

In the current environment of rapid technological change and tightening budgets, Research Online (<http://ro.uow.edu.au>) at the University of Wollongong provides a working example of a research-focused institutional repository. The download statistics generated by Research Online over the six-month period January - June 2006 may reflect general trends and providing pointers for future directions in regards to the measurement of research impact and performance. Few such analyses have been published to date.

Though there are hundreds of such repositories worldwide, the majority have been in operation for a relatively short period of time (less than two years), and both the software and method of operation are still evolving. Perhaps repository managers have been too busy dealing with software development issues, securing funding and staff, sourcing material, obtaining copyright clearances and implementing an often complicated upload process to engage in open debates on matters of statistics, though the topic is obviously an important one to them. In the case of the University of Wollongong, ongoing statistical analysis of the repository will be vital in assisting with forward planning, especially for the University Library as primary manager of the project.

Research Online - Digital Commons at the University of Wollongong

The University of Wollongong was the first Australian higher education institution to install Proquest's Digital Commons institutional repository software. As a proprietary solution, housed off-site, it differs from the other leader in the field, the open source DSpace, developed by MIT and Hewlett Packard. Corporate IT support from Proquest, a relatively simple upload interface, numerous operating instances in the United States and United Kingdom, a relatively quick installation and implementation process (one month), low cost, and a built-in statistics generation package resulted in the October 2005 decision by senior university executives to support a two-year implementation project, commencing in 2006. Other packages such as DSpace, ePrints and Harvest Road Hive were investigated by a specially formed committee during 2005, however Digital Commons was considered by the University Library and the Research and Innovation Division to offer the best solution, at that time, for Wollongong [7]. Digital Commons was installed locally on 12 December 2005 and the first papers were uploaded on 18 January 2006.

The project had a clear goal from day one: make available, via open access, University of Wollongong research output from the period 2000-2005, with the aim of improving citation rates and enhancing the reputation of the institution and its staff. The implementation team was also asked to monitor the performance of Digital Commons over the two-year project period.

The built-in statistics package would provide timely data on the number of site hits and full text downloads being generated across the site. It was recognised that direct correlation between download statistics and corresponding improved citation rates, as revealed through the aforementioned citation indexes, would be a relatively slow process and subject to a one- to two-year delay, from point of upload to appearance in a relevant index. Precise methods of measuring the impact of Research Online were yet to be developed, and its success, or otherwise, would not simply rest with what the download statistics told us.

Other measures would include acceptance by faculty and senior executive, the number of items uploaded, funding success (i.e., in regards to sustainable funding for the institutional repository service), ease of implementation of the package and goodwill generated on campus between the various parties involved, be they librarians, academics or administrative staff [8].

The impact of the repository would feed into the RQF assessment process and also the metrics used in the compilation of higher educational institution ratings tables, such as the Shanghai 100 [9]. The proposed RQF was a significant driver in the University of Wollongong's adoption of an institutional repository, though it was acknowledged there would be positive spin offs for the institution even if the RQF did not eventuate, with some commentators suggesting that, in the long term, associated bibliometric data could be integrated into faculty workflow and individual performance assessment programs. A local understanding of the statistics generated by Research Online therefore needs to be developed.

Download statistics

Digital Commons provides a relatively simple statistics package, with output presented in Excel files. At the repository level local administrators generate statistics on the number of full text downloads of individual documents and the number of hits on the cover page (i.e., item description or abstract page) that links to the document. These statistics can be further broken down by day, month or year. Reports can also be generated for individual collections or series, and academics can monitor hit and download rates for their own papers.

In regards to where those searches and downloads are coming from, Digital Commons also provides the facility to identify external referrals down to the domain level. Internal, or local, referrals and downloads are not measured. Referral data is only available at the site level, and not at the individual author level. During the study period we were not able to correlate hit and download statistics for individual papers with the location of the referral, though other packages such as ePrints have this feature.

The various statistics capabilities of Digital Commons can be highlighted by looking at the specific case of Research Online. During the six-month period January - June 2006 the number of papers uploaded was 561, comprising predominantly refereed and non-refereed journal articles and conference papers from the disciplines of informatics, engineering and commerce (86%). Following upload, the items were usually discoverable by Google within 24 - 48 hours. As of 30 June 2006 Research Online was also harvested by ROAR, DOAR, ARROW, Oaister, Google Scholar, Yahoo! and Scirus. Arising from this high Internet visibility, over the length of the study period there were 19,447 hits to the site. These comprised 10,661 full text downloads of documents in pdf form, and 8,740 cover page downloads. The remaining hits (46) were related to searches within the site.

During the study period 6.2% of the uploaded documents received greater than 50 full text downloads (FTDs), with the vast majority (79.5%) within the 1 - 50 FTD range (Table 1).

Table 1: Spread of downloads (%), January - June 2006, Research Online

No. of downloads	0	1 - 10	11 - 50	51 - 100	100+
Full text downloads	14.3%	41.2%	38.3%	3.6%	2.6%
Cover page downloads	3.8%	45.5%	47.3%	15%	2%

The difference between the spread of full text downloads and cover page downloads is of note, specifically the fact that during the study period only 3.8% of cover pages had no hits, compared with 14.3% of the equivalent pdf documents.

Whilst length of time on the system is obviously a factor in the number of downloads generated, these figures point to the obvious fact that not all repository content will be subject to the same rate of usage. In regards to Research Online discovery, of the 8,740 cover page downloads, 2,134 (24.5%) were referrals from domains with two letter top-level domains. For example, there were 542 referrals from the 'au' (Australia) domain, 191 from 'uk' (United Kingdom) and 179 from 'in' (India), with a total of 77 countries identified in the statistics provided.

The majority of referrals to the cover page - 6,606 or 75.5% - were from sites without two letter top-level domains, such as those coming from the United States. In regards to full text downloads of repository content, 3,308 (31%) were referrals from sites with two letter top-level domains. Of these, 1,075 referrals from the 'au' (Australia) domain, 245 from 'uk' (United Kingdom) and 190 from 'in' (India), with a total of 79 countries identified. Once again, the majority of referrals to the pdf document - 7,353 or 69% - were from sites without two letter top-level domains.

In regards to country of origin, or referral information, webmasters have long been able to extract such data for individual web pages and sites. However the institutional repository brings this facility into the hands of librarians, repository managers and individual academics, in a relatively simple form. For example, Digital Commons provided statistics relating to specific domain and url. Over the six-month period January to June 2006, the precise url for 5,449 (51.1%) of the full text downloads was known. Of these, 95.8% were from Google and its various domains around the world (e.g., www.google.com - 1770; scholar.google.com - 173; www.google.fr - 139; www.google.ru - 22; www.google.pk - 17). The remaining 4.2% were from sites such as www.scirus.com and www.answers.com.

In regards to discovery of cover pages, or abstracts, the figures vary slightly. According to the available data, 80.9% of cover page downloads were referred from Google domains, and the remaining 19.1% from a variety of sources including Yahoo! (12.1%). These two sets of figures suggested that users accessing Research Online from Google are in the majority of cases going straight to the document pdf, rather than to the cover page. This is perhaps influenced by Google's ranking of the pdf higher than the metadata page [10].

In looking at the monthly download statistics over the full study period, there was a steady increase between January - May, correlating with the continued increase in upload of content and wider diversity of Internet access points coming online. During May there were 3,433 full text downloads from the 404 items then on Research Online. However in June there were only 2,684 full text downloads from 561 items. This discrepancy may be seasonal, and longer-term trend data will assist in clarifying this. One suggestion was that such fluctuations might be connected with the academic year. For example, the northern hemisphere summer holiday break and corresponding mid-year break in the southern hemisphere may impact upon download rates.

In regards to the articles with the most number of downloads - commonly referred to as the most 'popular' items - Research Online clearly identifies these. The top ten downloaded items for the period January - June 2006 are listed in Table 2, both in regards to the number of full text downloads of the pdf and also cover page downloads, with their ranking indicated in brackets.

Table 2: Research Online - Top 10 Full text downloads & cover page downloads, January - June 2006

(NB: Individual ranking given in brackets)

Title	Faculty	Full text downloads	Cover page downloads
Clothing the Soviet Mechanical-Flâneuse	Creative Arts	395 (1)	47 (22)
Modelling the Draganflyer four-rotor helicopter	Informatics	366 (2)	197 (2)
The pros and cons of RFID in supply chain management	Informatics	341 (3)	242 (1)
Strike 1912 – Looking for Australia's earliest workers' film	History	237 (4)	58 (12)
Introducing location-based services into information technology curriculum: reflections on practice	Informatics	179 (5)	48 (19)
A century of the Phillipine Labour movement	History	179 (6)	72 (10)
Simmel, Ninotchka and the Revolving Door	Creative Arts	152 (7)	29 (45)
A Disgrace to our Australian Civilisation: Mothers, Miners and the Commemoration of Mortality in New South Wales	History	149 (8)	25 (89)
A Century of the Labour Movement in Australia	History	146 (9)	46 (25)
The Shooting of William (Billy) McLean	History	130 (10)	30 (64)
Conducting polymer-carbon nanotubes composites	Science	48 (41)	80 (6)
Personal firewall for Pocket PC 2003: Design and implementation	Informatics	38 (69)	76 (7)
What are the benefits in CRM technology investment?	Informatics	103 (15)	76 (8)
Image analysis using line segments extraction by chain code differentiation	Informatics	87 (17)	75 (9)
Visual perceptual process model and object segmentation	Informatics	29 (109)	89 (4)
A fast neural-based eye detection system	Informatics	105 (14)	83 (5)
Japanese technology for aged care	Arts	29 (107)	95 (3)

The figures reveal that there is no definitive correlation between the two download types. For example, the article 'Clothing the Soviet Mechanical Flâneuse' ranks number 1 in regards to full text downloads (395), but only number 22 in regards to cover page hits (47). This indicates that primary access to the document, once it is discovered, is via the pdf.

Against this, 'Modelling the Draganflyer four-rotor helicopter' ranks number 2 in both fields, with 366 full text downloads and 197 cover page downloads. In this latter case, we know from anecdotal evidence that this paper was set as a class reading by an academic in Holland during April 2006, thereby accounting in part for the relatively large number of hits. However such results could also be related to how the item appears on the Google search screen, and whether the cover page or pdf link are therein most prominent. In any instance, it is difficult to be precise due to the number of variables in the search and discovery process. In addition, these statistics do not by themselves offer an

explanation as to why particular articles rank so highly, raising the question: What is the cause of high (or low) download rates? Is it the inherent quality of the article, or perhaps the standing of the author? Is it related to the length of time in which the item is available on open access, or has a lecturer or professor set the item as a required course reading and directed students to download it, as was the case with the Draganflyer article? Does the paper possess a generic title, which is easily picked up by search engines and gives rise to a higher ranking? In such a case the high hit rate may have nothing to do with quality or impact.

Similar questions were asked at the University of Otago, New Zealand, when between November 2005 and March 2006 a suite of 220 School of Business papers generated 18,744 full text downloads from 80 distinct countries via the ePrints repository software [11]. A detailed explanation for this high download rate was not forthcoming at the time and points to the need for further studies in this area. The aforementioned questions highlight the pitfalls in drawing conclusions from a limited data set. Nevertheless, the Research Online download statistics are of use.

What do the download statistics tell us?

The need to compare and analyse statistical data across institutional repositories has been recognised, as has the desirability of collating download statistics for individual articles from institutional and discipline-based repositories and publisher databases in order to present a true picture of their popularity or otherwise as determined by hits [12]. Publisher concern over the negative impact on their hit rates by the existence of archive copies is a hindrance. With Google a significant research tool for students and an increasing number of academics, the ability to have one's output discovered by it in a quick and efficient manner is of primary importance.

In the case of Research Online, the download statistics primarily indicate that institutional repository software packages such as Digital Commons achieve their goal of exposing research output to Internet search engines such as Google and Yahoo! The important role Google plays in the research and discovery process has become apparent. Whether the University of Wollongong figures are similarly reflected in the download statistics of other institutional repository packages such as ePrints and DSpace is not known to the author, though the dominance of Google is most likely universal.

Another obvious finding from this limited study is the fact that, due to Research Online, full text downloads of research output occurred that would not otherwise take place. They also occurred in addition to that which is being generated by publisher online databases, personal web sites and discipline-based repositories.

Beyond this, the data provided reveals a number of developing user behaviours. It is clear that researchers are accessing Research Online in a variety of ways. Primarily they are coming via Google. In the majority of instances they go from Google to the document pdf, rather than to the abstract or cover page. This is reflected in the number of full text downloads as compared to the number of cover page downloads (10,661 : 8,740).

At present the number and range of material on Research Online is limited, and more detailed assessments of the download statistics need to be made covering a wider variety of material and over a longer time period. For the present, the numbers available tell an interesting story. They point to the success of the repository software and open access protocols in making research output available on the Internet. Beyond this, the repositories themselves may have impacts that were never foreseen. For example, it has been observed that academics and researchers may be influenced by their download statistics to alter the direction of their research [13]. On the basis of relatively low download statistics for a particular strand of research, they could decide to pursue a more popular strand, as identified to them by higher download rates. This would be a reflection of the increasing trend for research initiatives to be driven by business imperative and government policy, rather than as an expression of pure research. The usefulness of institutional repositories in this regard may assist with ensuring their long-term financial sustainability.

Yet the statistics available on Research Online also reveal that some of the best performing (i.e., most popular) items are from areas that do not figure in the traditional citation indexes, such as the creative arts and history. Download statistics provide a powerful tool for repository managers and librarians to sell the importance of this innovative technology to faculty and funding authorities. This will perhaps be their most important use.

Acknowledgements

In the compilation of this article I would like to thank my colleagues Helen Mandl, Natalie Keene and Lucia Tome for their advice and commenting on an early draft of this article. Mention must also be made of Arthur Sale, Steven Harnard and Susan Gibbons for their inspiration.

References

1. L. Tome and S. Lipu. Indicators of Journal Quality, Research & Development Discussion Paper no.6, University of Wollongong Library, 2004, 14p.
2. *Research Quality Framework*, Department of Education, Science and Training (DEST), Australian Government, Canberra. [web site, accessed 6 November 2006]. url: <<http://www.dest.gov.au/>>.
3. Hon. J. Bishop, *Knowledge Transfer and Engagement Forum - Keynote Address by the Minister for Education, Science and Training*, Sydney, 16 June 2006. [web site, accessed 6 November 2006]. url: <<http://www.dest.gov.au/Ministers/Media/Bishop/2006/06/b001160606.asp>>.
4. C. Hajjem, S. Harnard and Y. Gingras, 'Ten-year cross-disciplinary comparison of the growth of open access and how it increases research citation impact', *IEEE Data Engineering Bulletin*, 25(4), 2005, 39-46; 'The effect of open access and downloads ('hits') on citation impact: a bibliography of studies', The Open Citation Project, 6 June 2006. [web site, accessed 6 November 2006]. url: <<http://opcit.eprints.org/oacitation-biblio.html>>.
5. J. Houghton, 'Scholarly Communication Costs and Benefits: the role of repositories'. Presentation to The Successful Repository conference, Brisbane, 29 June 2006. [Powerpoint presentation, available on web site, accessed 6 November 2006]. url: <<http://www.apsr.edu.au/presentations/successful.html>>.
6. T. Cochrane, 'How can a Repository Contribute to University Success?' Presentation to The Successful Repository conference, Brisbane, 29 June 2006. [Powerpoint presentation, available on web site, accessed 6 November 2006]. url: <<http://www.apsr.edu.au/presentations/successful.html>>.
7. *Content Management System Evaluation Team*, University of Wollongong, 2005. [web site, accessed 6 November 2006]. url: <<http://ir.uow.edu.au/>>.
8. B. Weaver, 'Success is in the eye of the beholder'. Presentation to The Successful Repository conference, Brisbane, 29 June 2006. [Powerpoint presentation, available on web site, accessed 6 November 2006]. url: <<http://www.apsr.edu.au/successful/weaver.htm>>.
9. *Academic Ranking of World Universities 2005*, Institute of Higher Education, Shanghai Jiao Tong University, China. [web site, accessed 6 November 2006]. url: <<http://ed.sjtu.edu.cn/ranking.htm>>.
10. L. Carr and A. Sale, *IRS: Interoperable Repository Statistics, A proposal to Activity Area (iv) Pilot Services of the*

call for projects in the JISC Digital Repositories Programme, March 2005, 12p. Submission by the University of Southampton, Key Perspectives Ltd., the University of Tasmania, Long Island University and the COUNTER Project. [web page, accessed 6 November 2006]. url: <<http://irs.eprints.org/about.html>>.

11. N. Stanger and G. McGregor, 'Hitting the Ground Running: Building New Zealand's first publically available institutional repository', *The Information Science Discussion Paper Series*, Number 2006/07, March 2006, 10p.

12. S. Harnad, 'Re: Self-archiving, journal usage and cancellations', American Scientist Open Access Forum, 8 October 2005. [list posting, accessed 6 November 2006]. url: <<http://www.ecs.soton.ac.uk/~harnad/Hypermail/Amsci/4846.html>>.

13. S. Gibbons, 'Making a Repository a Success with your Academic Staff'. Presentation to The Successful Repository conference, Brisbane, 29 June 2006. [Powerpoint presentation, available on web site, accessed 6 November 2006]. url: <<http://www.apsr.edu.au/presentations/successful.html>>.

Copyright © 2006 Michael Organ

[Top](#) | [Contents](#)
[Search](#) | [Author Index](#) | [Title Index](#) | [Back Issues](#)
[Previous Article](#) | [Next Article](#)
[Home](#) | [E-mail the Editor](#)

[D-Lib Magazine Access Terms and Conditions](#)

doi:10.1045/november2006-organ