Open Metrics for Open Repositories

Brian Kelly[*], Nick Sheppard[+], Jenny Delasalle[#], Mark Dewey[*], Owen Stephens[$], Gareth J Johnson[θ] and Stephanie Taylor[*]
* UKOLN, University of Bath, Bath, UK {B.Kelly, M.Dewey, S.Taylor}@ukoln.ac.uk
+ Leeds Metropolitan University, Leeds, UK {N.E.Sheppard@leedsmet.ac.uk}
# University of Warwick, Warwick, UK {J.Delasalle@warwick.ac.uk}
$ Consultant, UK {owen@ostephens.com}
θ UKCoRR/University of Leicester, Leicester, UK {gjj6@le.ac.uk}

ABSTRACT

Increasingly there is a need for quantitative evidence in order to help demonstrate the value of online services. Such evidence can also help to detect emerging patterns of usage and identify associated operational best practice.

This paper seeks to initiate a discussion on approaches to metrics for institutional repositories by providing a high-level overview of the benefits of metrics for a variety of stakeholders. The paper outlines the potential benefits which can be gained from providing richer statistics related to the use of institutional repositories and also reviews related work in this area.

The authors describe a JISC-funded project which harvested a large number of repositories in order to identify patterns of use of metadata attributes and summarise the key findings.

The paper provides a case study which reviews plans to provide a richer set of statistics within one institutional repository as well as requirements from the researcher community. An example of how third-party aggregation services may provide metrics on behalf of the repository community is given.

The authors conclude with a call for repository managers, developers and policy makers to be pro-active in providing open access to metrics for open repositories.

Keywords

repositories, open data, metrics.

1. ABOUT THIS PAPER

The potential benefits of open access to research publications are widely accepted. In addition the difficulties of achieving such benefits, in particular the challenges associated with copyright issues, are also understood and greater emphasis is being placed on publication in open access journals.

However statistics on use of institutional repositories, including information on file formats, download statistics, statistics on metadata usage, etc. should not be constrained by copyright concerns. There are therefore opportunities for repository managers to demonstrate a commitment to openness by providing open access to data related to repository services.

This paper describes activities taking place across the UK repository sector at institutional and national level which are aimed at providing a better understanding of how repositories are being used to influence policy and practice.

The paper concludes by arguing the importance of gathering evidence in order to inform policy decisions and practice. The repository community, with a long-standing culture of promoting openness to support research activities, should be well-positioned to support greater provision and use of open data associated with repository services, whilst acknowledging that the interpretation of data needs to be done carefully.

2. THE NEED FOR METRICS

A performance metric is defined in Wikipedia as "*a measure of an organization's activities and performance. Performance metrics should support a range of stakeholder needs from customers, shareholders to employees. While traditionally many metrics are financed based, inwardly focusing on the performance of the organization, metrics may also focus on the performance against customer requirements and value*" [1].

Metrics for repositories can be used to provide a better understanding of how repositories are being used, which can help to inform policy decisions on future investment, technical policy decisions on enhancements to the technical infrastructure [2] [3]. They are also able to help operational decisions by practitioners as well as being able to demonstrate the value of investment or, if appropriate, inform decisions on deprecating aspects of the services. Metrics are also used to monitor the effectiveness of open access activities[1].

3. THE BIG PICTURE

3.1 Survey of Numbers of Full-text Items

In order to identify lightweight approaches for profiling institutional uses of repository services, the advanced search facility in the ePrints service was used to find numbers for the full-text items for the OPuS institutional repository service at the University of Bath. In June 2011 for a total of 20,210 items there were 1,387 (6.9%) full text items [4]. However, in seeking to use this approach across a wider set of repositories it was found that very few repositories had configured use of the advanced search facility to enable this survey to be carried out.

In the light of the difficulties in consistent implementation of features in ePrints software to carry out such repository profiling activities, it was felt that a more scalable approach would be based on use of a national aggregation service.

3.2 RepUK

RepUK[2] was funded by the JISC and developed at UKOLN in order to monitor patterns of metadata usage within institutional repositories across the UK's higher and further education community. The work involved the development of software for the extraction, analysis and visualization of metadata hosted across all UK repository platforms.

The initial task was to harvest oai:dc metadata in a reliable way over the OAI_PMH protocol and to develop an infrastructure for processing updated entries and newly deposited items. The OpenDOAR directory of open access repositories[3] provided a list of active UK repositories. The records were queried exposing trends within individual repositories and at a national level.

[1] http://repositories.webometrics.info/about_rank.html

[2] http://repuk.ukoln.ac.uk/

[3] http://www.opendoar.org/

3.2.1 Findings

At the time of writing (March 2012) RepUK has harvested 153 repositories and contains 1,654,090 records. From the summary listing[4] we find information for the repositories with the largest numbers of items.

Name	Nos. of Records	Software
UCL Discovery	240,854	EPrints
DSpace @ Cambridge	214,530	DSpace
Visual Arts Data Service	186,210	Unknown*
Leodis - A photographic archive of Leeds	170,667	Unknown*
STFC ePublication Archive	69,153	Cocoon

Table 1: UK's largest repositories

* 'Unknown' means the software does not identify itself via OpenDOAR

Although RepUK provides an overview of repository usage across the UK the main purpose of the service was to provide analysis of metadata usage. A summary of the most popular file formats hosted in the repositories is given in Table 2.

Format	Nos. of Records
PDF	739,900
CML	667,012
HTML	246,172
JPEG	85,972
MS Word	32,320
Plain text	22,836

Table 2: Most popular formats

RepUK also provides a timeline of the numbers of deposits harvested since the first set of items were deposited in 2000 as shown in Figure 1.

In addition to the overview of the repository sector which RepUK provides, the service also provides detailed information related to the metadata harvesting for individual repositories[5]. This information includes use of DC metadata fields for deposits in the University of Bath repository over time, as shown in Figure 2.

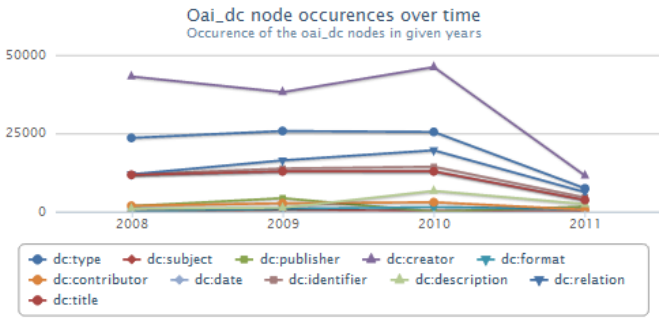

Figure 2: DC usage for University of Bath repository

[4] http://repuk.ukoln.ac.uk/repositories.htm

[5] For example see the harvesting summary for University of Bath repository at http://repuk.ukoln.ac.uk/publicRepositoryRecord.htm?rId=485

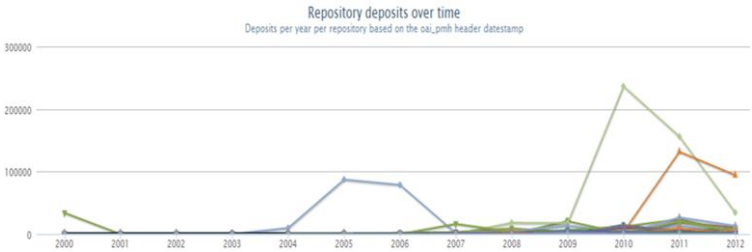

Figure 1: Deposit rates since 2000

3.2.2 Discussion

RepUK has gathered quantitative evidence across the repository sector which can identify patterns of usage. This can help inform policy-making at national, international and institutional levels.

From the data on usage of DC terms we can see that the DC.Rights and DC.Coverage fields are little used ranging from 86,090 records (5%) for DC.Rights down to 9,623 (0.6%) for DC.Coverage. This shows that it would be inappropriate to develop services which require machine-readable information on rights and coverage information. This information may also suggest areas in which the tools and mechanisms for providing metadata are too complex to be used within the sector.

4. THE INSTITUTIONAL PICTURE

If we take the view that "*two key purposes of a repository are (1) maximising access to research publications and (2) ensuring long-term preservation of research publications*" [2] metrics repository managers should be interested in may include:

- total number of records, both full-text and metadata only;
- number of records that include (openly accessible) full-text output (raw figure and a proportion of total records);
- number of times metadata records are accessed;
- number of times full-text items are accessed;
- how records are accessed (browsing the repository, search engine referral, referral from an OAI-PMH aggregation).

In light of such interests there are technical, pragmatic and ideological issues to consider. One limitation of repositories across UK HE from an original arXiv[6] conception, of holding, disseminating and preserving full-text research outputs, is that they have become "diluted" by metadata records for which it has not been possible to procure full-text or copyright does not permit deposit. Of the 155 institutional repositories in the UK currently listed on OpenDOAR, only 18 of the 75 responses from an RSP survey have full-text-only services. Given the communication issues in promoting self-archiving and/or mediating full-text deposit, the vast majority of repository managers take a policy approach to content. They record metadata as a matter of course and advocate the value of green open access to their research communities whenever the opportunity presents itself.

From a technical perspective, the availability of metrics differs not only across software platforms, but also across different implementations of the same software. Moreover to some extent the availability will depend on the technical ability of repository staff or the availability of in-house expertise. EPrints is the most common repository software in the UK and many implementations include the option to search for full-text only

[6] http://arxiv.org/

from the advanced search form[7]. ePrints is also well supported with the IRStats plug-in[8] and many ePrints installations do now incorporate download data alongside metadata records[9]. However, there are repositories running on a wide range of other systems, both open source (ePrints, DSpace, Fedora) commercial (intraLibrary, Equella, DigitalCommons) as well as commercial/open source hybrid (OpenRepository).

At Leeds Metropolitan University the repository is based on commercial software (intraLibrary[10]) for which there is no equivalent of the IRStats plug-in. Instead, Google Analytics[11] has been implemented to track use of the repository including full-text downloads. This is achieved by applying Google tracking code to the download link. Google Analytics is a powerful tool capable of generating date-limited usage data including visits, unique visitors, page views, traffic source (search engines, referring sites), countries and territories (which can be visualised on a map). In addition more experienced users can drill down to generate more nuanced data such as tracking individual users' routes to a given PDF for example - whether from a search engine, third-party aggregator, or by searching / browsing the repository itself.

The current process is to review, manually transcribe and disseminate relevant data on a monthly basis by simply adding it to a HTML page[12] which lacks the dynamic nature of the IRStats plug-in for ePrints. The data disseminated in this manner currently includes total number of visits and national origin of those visitors, total number of records added and how many of those records include the full-text output, total number of full-text downloads (again lacking the dynamic and granular functionality of IRStats to incorporate this data on individual records) and the top-ten viewed items (which may not be full-text but can be used as an advocacy tool).

Arguably, "hybrid" repositories of full text and metadata are becoming *de facto* research management systems, particularly at institutions that are not research-intensive. As they often contain more metadata records than full text[13]; usage data will be important to the institution and the wider community, both to illustrate gross research activity and relative Open Access to full-text research outputs. Increasingly, however, the trend is towards dedicated research management systems (commercial solutions include Atira Pure, Symplectic Elements and Converis) that, properly implemented, can complement an Open Access repository as part of an institutional research management infrastructure. Leeds Metropolitan University, for example, is in the process of integrating Symplectic Elements with intraLibrary, aiming to make it easier to maintain a constant, up-to-date picture of research activity across the institution as well as upload full-text outputs directly to the repository from the Symplectic interface (see the JISC funded RePOSIT Project[14]). Via its API, the system will also facilitate dynamic bibliographies from researchers' and departmental web pages including, where available, links to the full-text in the repository. In turn this should improve Search Engine Optimisation (SEO) and bring more traffic to the repository; it also raises the possibility that repository policy be reviewed and become full text only.

It remains to be seen whether the integration of a cohesive research management system comprising an Open Access repository as a component of a research management system rather than as a discrete, disconnected system will increase the rate of self-archiving and concomitant internet traffic to that full-text research output. However the success or otherwise of this and similar approaches at other institutions can only be ascertained if repository mangers are pro-active in openly disseminating the most sophisticated metrics possible from their own repositories.

5. THE RESEARCHERS' REQUIREMENTS

The public availability of article metrics may have an effect on repository deposits by authors. Publishers like PLoS [5] and the subject specialist Arxiv repository [6] display article-level metrics along with the record describing the article. Institutional repositories (e.g. see WRAP example[15]) may do the same, but authors may be anxious to see visitor numbers aggregated and displayed in total each time, from all locations and versions of the article.

Such an aggregation of metrics will be difficult to achieve: there are many different ways of counting the number of visitors to an article or its record and it would require the repository managers and publishers to share data. However PLoS are already showing visitor numbers from PubMed Central on article metrics records, as well as their own. The JISC-funded PIRUS Project[16] has investigated metrics issues for sharing journal articles, building on the work of the COUNTER Project[17].

Without aggregated measures, researchers may refrain from depositing in open access repositories, in order to maximise visitor numbers at their preferred location. Authors are also likely to find it convenient to handle metrics from one source rather than from many sources. However, relatively few publishers display such article-level metrics publicly or even provide them to authors. This presents an opportunity for repositories to engage with authors.

UK academics are to have their performance measured through the REF (Research Excellence Framework) 2014 exercise, which will include an element of assessment of "impact". Research Councils UK also have a Pathways to Impact expectation[18]. What these two impact expectations have in common is the reach of research beyond the academic sphere. Demonstrating such reach might involve new kinds of metrics for online activity relating to all kinds of research outputs. Blogs and tweets and slideshows can be bookmarked and added to favourite collections, 'liked' on Facebook and re-tweeted and commented on. Such activity can be an indicator of value beyond simple viewing and is likely to involve services from beyond the academic sector.

[7] See http://eprints.lincoln.ac.uk/cgi/search/advanced and http://nectar.northampton.ac.uk/cgi/search/advanced

[8] IRStats plug-in for ePrints - http://files.eprints.org/722/

[9] Example of ePrints record incorporating download data - http://eprints.uwe.ac.uk/11667/

[10] IntraLibrary is a learning object repository repurposed to also manage research outputs - http://www.intrallect.com/index.php/intrallect/products

[11] http://www.google.com/analytics/

[12] http://repository.leedsmet.ac.uk/main/monthly_stats.php

[13] See RSP survey at http://www.rsp.ac.uk/pmwiki/index.php?n=Institutions.HomePage

[14] http://www.jisc.ac.uk/whatwedo/programmes/inf11/jiscdepo/reposit.aspx

[15] http://wrap.warwick.ac.uk/933

[16] http://www.jisc.ac.uk/whatwedo/programmes/inf11/pirus2.aspx

[17] http://www.projectcounter.org/

[18] http://www.rcuk.ac.uk/kei/impacts/Pages/home.aspx

The altmetrics manifesto[19] describes the importance of web metrics and how such metrics may have many roles in the assessment and assurance of quality of information. The altmetrics web site links to tools which are under development and which aggregate metrics from multiple sources to web artefacts, such as Total-Impact[20].

From the researcher's perspective, publishers and repository managers should support social activity in relation to journal articles, measure the activity and report back on it to authors. They should aggregate and allow the aggregation of such activity measures, in relation not only to the article in all its versions, but also in relation to other artefacts which are linked to the journal article.

6. THIRD-PARTY SERVICES

Open repositories offer some level of content reuse. Most institutional repositories offer mechanisms to facilitate harvesting of metadata, and some full-text content. Specifically, the Open Archives Initiative Protocol for Metadata Harvesting guidance states "*It is expected that aggregators, caches, proxies and other third party repositories will emerge*" [7].

The creation of such third-party services has perhaps not been as substantial as originally envisaged. In 2008 Hubbard wrote "*The search party didn't turn up*"[21]. Despite this slow growth, there are now a number of services, including generic web search engines such as Google, which harvest both metadata and content from institutional repositories.

Some third-party services, such as Google Scholar[22] offer only 'search' services, and do not serve content to users directly, preferring to redirect users to the source repository to view content. Others, such as CiteSeerX[23] and CORE, cache copies of content and offer users the choice of using the cached copies, which affects the source repository metrics.

These issues suggest the need for metrics relating to third-party services. Firstly, the presence of repository metadata or content in third-party repositories may offer some measure of 'reach' [8]. Secondly third-party repositories could offer statistics on usage of content that could be accessed by source repositories; an area in which CORE is currently working. The PIRUS2 Project [9] which examined the collection of article level usage metrics may have guidance to offer in this area. Finally, as suggested at [10] metrics collected by third-party services could help provide a deeper understanding of trends across the sector.

These types of measurement and the requirements echo the difficulties of tracking any product, organisation or concept across the web. In recent years, a range of services have been established for this purpose, often offering detailed analytics as the 'premium' aspect of a 'freemium' offering. For tracking URLs, bit.ly offers statistics on specific URLs that have been shortened using the bit.ly service[24]. Solutions such as Topsy Analytics[25] offer tracking of keywords on Twitter, and others such as SproutSocial[26] offer tracking across multiple social media and web channels. As open repository content becomes more integrated with the web, these approaches to gathering metrics become more relevant.

[19] http://altmetrics.org/manifesto/

[20] http://total-impact.org/

[21] See notes on "Developing Research Repositories" at http://www.meanboyfriend.com/overdue_ideas/2008/07/developing-research-repositories/

[22] http://scholar.google.com/

[23] http://citeseerx.ist.psu.edu/

[24] http://bit.ly/

[25] http://analytics.topsy.com/

[26] http://sproutsocial.com/features/social-media-analytics

7. CONCLUSIONS

This paper has described reasons why metrics for repositories are needed for a variety of purposes and stakeholders. It has outlined approaches which are being taken across the sector for providing metrics for the various stakeholders.

There are concerns for UK repository managers that metrics which may be of greatest value for operational and strategic purposes may be sidelined by demands from senior management for those that merely offer a volumetric assessment.

In addition to the technical approaches, the paper has argued that repository managers should be pro-active in showing a willingness to provide open access to repository metrics. This is felt to be consistent with the culture of openness which underpins those involved in the provision and support of open access repositories.

ACKNOWLEDGEMENTS

JISC is gratefully acknowledged for their support to the UKOLN Innovation Support Centre and the JISC-funded projects mentioned in this paper.

REFERENCES

[1] *Performance Metrics*, Wikipedia, http://en.wikipedia.org/wiki/Performance_metric

[2] *How Do We Measure the Effectiveness of Institutional Repositories?*, Kelly, B., UK Web Focus blog, 14 June 2011, http://ukwebfocus.wordpress.com/2011/02/24/how-do-we-measure-the-effectiveness-of-institutional-repositories/

[3] *Evaluating Repository Annual Metrics for SCONUL*, Johnson, G. J., LRA, https://lra.le.ac.uk/handle/2381/9421

[4] *A Pilot Survey of File Formats in Institutional Repositories*, Kelly, B., UK Web Focus blog, 14 June 2011, http://ukwebfocus.wordpress.com/2011/06/14/a-pilot-survey-of-file-formats-in-institutional-repositories/

[5] *Who Shares? Who Doesn't? Factors Associated with Openly Archiving Raw Research Data*, Piwowar, H., PL:OS One, http://dx.plos.org/10.1371/journal.pone.0018657

[6] *Comparing webometric with web-independent rankings: a case study with German universities*, Thamm, M. and Mayr, P., arXiv, http://arxiv.org/abs/1105.2443

[7] *Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Aggregators, Caches and Proxies*, http://www.openarchives.org/OAI/2.0/guidelines-aggregator.htm

[8] *Reach (advertising)*, Wikipedia, http://en.wikipedia.org/wiki/Reach_(advertising

[9] *PIRUS2 Final Report*, http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-download_wiki_attachment.php?attId=170&download=y

[10] *The Value of Statistics: Guest Blog Post by Brian Kelly*, Kelly, B., Jorum blog, 13 February 2012, http://www.jorum.ac.uk/blog/post/27/the-value-of-statistics-guest-blog-post-by-brian-kelly