

February 2007

Institutional Repositories Workshop Strand Report

Strand title: Usage Statistics

1. Executive Summary

An understanding of the use of repositories and their contents is clearly desirable for authors and repository managers alike, as well as those who are analysing the state of scholarly communications. A number of individual initiatives have produced statistics of various kinds for individual repositories, but the real challenge is to produce statistics that can be collected and compared transparently on a global scale. This report details the steps to be taken to address the issues to attain this capability.

2. Summary of recommendations

1. ACTION: determine practical definition of 'usage'
 - i. Decide meaning of 'use'
 - ii. Produce event-based web-log based format for sharing 'usage events' to deliver many profiles (COUNTER, awstats, JISC Monitoring, DRIVER etc)
2. ACTION: define objects to be counted
 - i. Lobby COUNTER to add article level stats
 - ii. Create vocabulary of academic output types in conjunction with Research Paper Metadata group
3. ACTION: standard reports
 - i. Agree on a small set of standard useful statistical reports that repositories should produce
4. ACTION: agree policies for stats
 - i. Compliance with local laws on e.g. privacy
 - ii. Enhance SHERPA policy tool
5. ACTION: collection and aggregation
 - i. Agree de-spidering process (first draft agreed)
 - ii. Specify issues of aggregation and deduplication for later study
6. ACTION: collation with external sources
 - i. Talk to COUNTER about complex objects
 - ii. Set up COUNTER-IR to shadow the publisher group
 - iii. JISC Project IRS to provide initial COUNTER-style reports
 - iv. Talk to COUNTER about aggregating COUNTER stats at consortium level
 - v. Investigate SUSHI interoperability with repositories and OAI-PMH + OpenURLContextObjects

3. Discussion (including recommendations and items of interest)

The first section gives the context of the discussion, and is written by the rapporteur in order to provide an understanding of the issues facing the group. The second part is the notes of the group discussions, reported according to the issues agreed in the first stages of the workshop.

BACKGROUND: WHY GATHERING REPOSITORY STATISTICS IS NOT SIMPLE

Received wisdom is that Web logs statistics are inaccurate and misleading [cite], and yet they are the basis of many analysis tools, and the foundation upon which many business and marketing decisions are taken.

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 (exposed in the URI naming scheme) 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 we 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!

By contrast, ‘COUNTER’, the publisher standard for statistics reporting [cite], defines a simple download model. The one concession to determine the ‘intentionality’ of the access is to screen out apparent ‘double clicks’: two identical requests for an item that take place within 30 seconds. Any approach to the merging of multiple requests has to be aware of the range of reasons for unintentional downloads, for example

1. double clicking (ie inappropriate use of the hypertext user interface)
2. multiple clicking as an expression of impatience at the speed of delivery
3. requesting the browser to perform a Page Refresh
4. pressing the ‘Back’ button
5. the user forgetting that they had already downloaded the file.
6. an operating system or browser crash that causes the document to be reloaded on restart.
7. the downloaded page was sent to the printer and the download window subsequently closed, but the printout failed.
8. the user has forgotten where the download was saved on the disk.
9. the user accidentally closed the browser window.
10. the user closed the browser window because they thought that they had finished with the paper, but now they need to refer to it again.
11. there is a queue of students at the library workstation, all with the same reading list.

The opposite defect with web logs is the phenomenon of missing downloads, ie genuine reading events that do not result in entries in the web log. Some examples of these are

1. institutional web proxies may cache documents requested in the past and use them to satisfy future requests without re-contacting the original server
2. the paper may be downloaded by a lecturer and then provided for a number of students by printouts, photocopies or local re-hosting of the digital object.
3. many repositories provide RSS feeds which may be aggregated and cached, so that researchers subscribed to the RSS feed take the information from the aggregator and not from the repository
4. many repositories provide email subscription alerts which allow researchers to keep up to date with the contents of the repository, without interacting with the repository

Further, publishers offerings are not held on publicly-available websites, and so the COUNTER standard does not need to address the issue of search engine crawlers, or robot downloads, that will tend to inflate the access statistics (up to half of the apparent downloads from a repository such as eprints.soton.ac.uk may be attributed to crawlers and spiders).

Despite the limitations, an increasing number of repositories have been implementing their own statistics, either from scratch (DSpace ANU or EPrints Sale) or by integrating widely-used generic web stats packages such as awstats (eprints.ecs). The latter have the disadvantage of being web-page-based rather than eprint-aware, and so make it difficult to attributed bibliographic information (title/authors) to the data. The former overcome this problem, but lack a public model on which the statistics are based.

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 graph 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.

The first two are issues of technical convergence, but the latter is a more fundamental issue, requiring agreement on *what items are being counted* and *how these items are counted*, as well as *how they are reported*.

DISCUSSION NOTES

The context for usage statistics and the reasons for promoting and collecting standardized usage statistics are

- licensing issues: consortia and publishers want to investigate value for money issues
- research policy issues: determining impact factors in new oa environment / real value of scholarly output issues, etc.

There are problems of confidence in data and statistics in general

- o the naive use of impact factors and other metrics within scientific community
- o the need to minimize the opportunities for fraud and data manipulation by ensuring transparency of data and processes. It will be necessary to communicate the standards that are used for compiling statistics and to provide auditing procedures

As well as the statistics concerning the usage of public repository contents focused on below, there is also need for private repository management statistics for the following purposes

- o Maintenance – events that disrupt repository flow (outages, upgrades, disasters)
- o Ingest statistics
- o Number of records processed by editorial assistants
- o Revisions/transitions in workflow
- o Number of items deposited per researcher

One of the key issues is for an understanding of the kinds of usage that we are seeking to measure. The background section above lays out our wish to be able to measure ‘academic engagement’ with an object, but it is clear that this is not practical as there is no way of tracking the ‘attention paid’ to a downloaded object.

- *Is the COUNTER definition / standard adequate?* Counter tracks all downloads, with a single exception in the form of a timing rule for ignoring ‘double clicks’. These does not correspond to our wish to track ‘meaningful’ downloads only, but it should be supported by IR’s in order to aggregate data with publisher statistics.
- What other activities / standards etc. are relevant here? AWStats, is the most used public web log analysis software. It also defines a timing rule of ignoring accidental multiple downloads. (A 1-hr period by default as opposed to the COUNTER 30-second rule.)

The second key issue is an understanding about what we are counting. Do we wish to count downloads of individual files (the full texts only) or the abstract pages as well? What about items with multiple bitstreams?

- o Granularity of data (journals, articles, etc.)
How can IR’s aggregate statistics to the journal (and/or book) level?
Should IR’s amalgamate all stats for individual bit streams or should we take the max. of aggregates?

Action: lobby COUNTER to add article level to it stats (COUNTER++) – talk with advisory board members!

Action: Research Paper Metadata group needs to define a breakdown of documents into

Kommentar [So1]: I would list this and other actions in the section ‘outcomes’

meaningful sets (peer reviewed vs. non, doc type, oa and non-oa, etc.)

Also as work package in Project 1: Data Model -- set up data model for what IR's should collect

- List of working definitions
- What is the standard/agreed set of statistics?

Actions: consult with relevant players (IR-community plus below list) together to articulate their policies / procedures

This list should include: COUNTER, JISC Monitoring Unit, AWStats, GASCO (DE) et al.

Question: what is NL and DK doing here?

(roundtable or workshop initiated by KE or partner orgs.) as a first step for defining common policy / standard for IR's

Actions: create filters that can deliver e.g., COUNTER++ compliant stats (downloads per journal per month, turnaways, e.g.)

Actions: create filters to deliver defined breakdown of meaningful sets

- Policy issues: which policies for collection and use of usage statistics are in place?
 - Is access data sharing OK?
 - Sharing with aggregating services under what circumstances?
 - What stats could be presented?

IR should have a policy on statistics! This should be a human-readable document that sets out the general principles

- commercial reuse of data (OA as goal)
- compliance with local laws
- which standards /auditing procedures used in creating statistics
- obvious way to differentiate between usage of OA and non-OA objects
- who controls the use of the data

Action: enhance sherpa policy tool with section on IR statistics

<<http://www.opendoar.org/tools/en/policies.php>>

As work package in Project 1.

- how is collection and aggregation of data done (repository environment/ecology)

<<http://www.ebase.ucc.ac.uk/docs/jiscnesli2summaryeb.pdf>>

Robots: need to eliminate robots – should this be done at a data or service level? Both are possible, there are indicators that it might be more efficient on a service level. On a service level, you can spot trends not visible locally. Also, creates less work for local IR's.

(e.g., user agent, ip e.g. in awstats, >100 items per day, etc.)

How to deal with spam referrers – referrer validation?

Action: define set of general guidelines IR's should use to eliminate robots

First Draft:

1. does user agent match wget et al.
2. does ip/user ever access /robots.txt
3. does ip match AWStats list of robots
4. does IP download > 1000 items per day (# and timeframe to be defined)

De-duplication: how is it done, where is it done? – this should be done as an external service (also see LANL bX, Google Scholar, CiteSeer on de-duplication, SURF WG on Persistent Identifiers)
Question for plenary

- how is data collated with external sources (publisher data, etc.)

Action: talk with counter to be able to support more complex object structures (codeword counter++) – but also beyond article: proceedings, chapters in edited volumes, e-learning objects

Project (IRS): provide an example for applying COUNTER style stats to IR objects

Action: talk with publishers through KE WS on NL strand on usage and statistics: 1. counter should be able to deliver stats on article level; 2. usage stats should be aggregated on a consortium level

Action: investigate interoperability between OpenURL context objects in OAI-PMH and sushi
OVERLAP: OAI-PMH Group?

- o how is this data exchange btw. publishers and repositories (in both directions)
Technical aspects might be solved with above issues.
Policy challenge remains: how to make sure also publisher information can be aggregated by external services?
- o what does the division of responsibility look like btw. a repository and an external service?
 - what should be exposed?
 - what is practical / realistic to expect?

Relevant Initiatives:

- DINI (DE): proposed infrastructure for Germany based on link resolvers and webserver-logs
- Blackbox (Los Alamos): implemented in LANL and U. of Cal. System, very interesting for Metrics layer; next phase is MESUR project;
- IRS (UK) Stage 1 (Unique ID, Date stamp, requested URL, requester IP, Referring URL) → Stage 2 (Unique ID, Date Stamp, Hostname, Organization (if available), country, doc type (boolean abs. / full text), referrer zone, query/search engine, query term, referrer url)
- Google analytics

4. Outcomes

This group acknowledges the fundamental importance of shared, community-interpretable statistics; this represents a key achievement because all previous initiatives in this area have addressed the scope of a single repository.

Having achieved various successes with individual statistics projects, this group agrees the need to carefully examine the underlying model of repository use on which any statistics are built, namely (a) what counts as a 'use' and (b) what items or features of a repository are to be counted.

The group identified a list of meaningful actions in this area

Suggested Project 1:

IR's should be able to deliver a "fundamental format" for weblog file

In addition there should be a process format that should be "normalized" to be able to deliver multiple profiles, e.g., COUNTER, AWStats, JISC Monitoring Unit (UK), GASCO (DE).

This format should also be

- adequate for long-term storage of data
- include ways to anonymize users / be shibboleth compliant
- distinguish between internal url's and external uri's
- Identification of Users/Sessions - Global identification? (Shibboleth)

Suggested Partners: DINI Stats Project, Groningen, Southampton (others?), University College London

... tbc

5. Annexes (including supporting papers, referenced materials, list of participants)

Questionnaire

The questionnaire on repository practices with respect to statistics (conducted amongst the group participants) is available on the SURF Groepen website, and not reproduced here for reasons of space.

List of Participants

Role	Name	Affiliation	Email
lead moderator	Dr. Peter Schirmbacher	Humboldt University, Berlin	schirmbacher@cms.hu-berlin.de
moderator 1	Frank Scholze	University of Stuttgart	frank.scholze@ub.uni-stuttgart.de
moderator 2	Dr. Leslie Carr	University of Southampton	lac@ecs.soton.ac.uk
Knowledge Exchange/DFG	Dr. Max Vögler	DFG	max.voegler@dfg.de
Participants DFG	Ulrich Herb	Saarland University and State Library	u.herb@sulb.uni-saarland.de
Participants DFG	Dr. Andreas Hübner	GeoForschungsZentrum Potsdam	huebner@gfz-potsdam.de
Participants JISC	Adam Field	University of Southampton	af05v@ecs.soton.ac.uk
Participants SURF	Henk Ellerman,	Groningen University	h.h.ellermann@ub.rug.nl
Participants SURF	Marlon Domingus	Leiden University	domingus@library.leidenuniv.nl
Participants DEFF	Peter Søndergaard	Roskilde University Library	psø@ruc.dk

References

<http://www.projectcounter.org/>
http://www.projectcounter.org/r2/COUNTER_COP_Release_2.pdf
http://www.niso.org/committees/SUSHI/SUSHI_comm.html
<http://irs.eprints.org/>
<http://www.mesur.org/>
<http://www.dini.de/veranstaltung/workshop/oaimpact/>

DFG

JISC

DEff

SURF