New Avenues for Electronic Publishing in the Age of Infinite Collections and Citizen Science
B. Schmidt and M. Dobreva (Eds.)

doi:10.3233/978-1-61499-562-3-83

Measuring the Usage of Repositories via a National Standards-based Aggregation Service: IRUS-UK

Ross MACINTYRE[a],[1], Jo ALCOCK[b], Paul NEEDHAM[c], Jo LAMBERT[a]
[a]*Jisc, United Kingdom*
[b]*Evidence Base, Birmingham City University, United Kingdom*
[c]*Cranfield University, United Kingdom*

Abstract. Many educational institutions have repositories for research outputs. The number of items available through institutional repositories is growing, and is expected to continue to do so due to requirements for outputs from public-funded research to be open access. But how much usage are institutional repositories and their individual items getting? The Jisc-funded service IRUS-UK is designed to help institutions understand more about the usage of their institutional repositories. IRUS-UK collects raw usage data from participating repositories and processes these into COUNTER-compliant statistics. This provides repositories with comparable, authoritative, standards-based data and opportunities for profiling and benchmarking. It enables institutions to run reports at both repository level (e.g. total download figures) and at item level. IRUS-UK utilises a robust, multistage ingest process, validating data, stripping out robot and unusual accesses, and filtering out double clicks, to transform raw usage data into COUNTER-compliant statistics. IRUS-UK currently has data from 83 UK institutional repositories (using Eprints, DSpace and Fedora software) and has recorded over 35 million downloads since July 2012. The data from IRUS-UK can be used to provide information for management reporting, for usage monitoring, and for external reporting. Data can be viewed within the online portal, downloaded for further analysis, or harvested using the SUSHI service (NISO Z39.93). IRUS-UK is also working with and contributing to other groups and initiatives involved in a range of activities relating to usage statistics. These include: the Distributed Usage Logging/CrossRef DOI Event Tracker Working Group, OpenAIRE2020 and COAR Working Group.

Keywords. COUNTER, usage statistics, repositories, benchmarking, altmetrics

1. Introduction

Institutional repositories (IRs) have attracted much attention over the last decade and there has been considerable interest in the growing number of repositories and their contents. However, until now there has been a lack of comprehensive information about usage of the resources hosted by IRs.

Although most IRs provide statistics that purport to show usage, you can't count on them – not entirely. Different types of software – out-of-the-box, add-ons, Google Analytics and other third-party solutions – process raw usage data in different ways,

[1] Corresponding author. Jisc, J14A Sackville Building, The University of Manchester, Sackville Street, Manchester M1 3BB, UK; E-mail: ross.macintyre@jisc.ac.uk.

making it impossible to compare like for like across repositories. There is currently no agreed standard to measure usage across repositories.

IRUS-UK, funded by Jisc, is a national aggregation service that responds to this problem by providing standards-based statistics for all content downloaded from participating UK IRs. The service collects usage data from participating repositories, processes the data into COUNTER-compliant [1] statistics and then presents statistics back to originating repositories to be used in a variety of ways. It provides opportunities for benchmarking at a national level by enabling UK IRs to access and share comprehensive and comparable usage data. Some of the underlying technical principles were taken forward from the substantial work done in LANL's Mesur project [2] and in the European Knowledge Exchange Working Group on Usage Statistics [3].

IRUS-UK now provides a nationwide view of the majority of the UK's institutional repositories use, helping demonstrate the importance and value of IRs. There is also potential for the service to act as an intermediary between UK repositories and other agencies.

IRUS-UK is one of a number of Jisc-funded repository and infrastructure services which aims to increase the cost effectiveness of repositories of open access (OA) literature. The service was developed by a consortium involving Mimas (now part of Jisc itself), Cranfield University and Evidence Base at Birmingham City University. The team is also responsible for the development of the Journal Usage Statistics Portal (JUSP) [4], which provides a 'one-stop shop' for libraries to view, download and analyse their journal usage reports from multiple publishers. Consequently, the team members have significant skills and expertise in managing and developing usage statistics products and services.

2. Background to Development of the Service and PIRUS2

IRUS-UK builds on the work of the successful Jisc-funded PIRUS2 project [5], which demonstrated how COUNTER-compliant article-level usage statistics could be collected and consolidated from publishers and institutional repositories. The primary aims and objectives of PIRUS2 were to assess the feasibility of and develop the technical, organizational and economic models for the recording, reporting and consolidation of usage of journal articles hosted by publishers, institutional repositories and subject repositories.

PIRUS2 achieved its aims by delivering a prototype statistics aggregation service, comprising:

- usage data and statistics from publishers and institutional repositories
- a practical organizational model based on co-operation between data processing suppliers
- data management and auditing services that meet the requirement for an independent, trusted and reliable service
- an economic model that provides a cost-effective service and a logical, transparent basis for allocating costs among the different users of the service.

PIRUS proposed the establishment of a global central clearing house (CCH) to deliver such a service. Unfortunately, it became clear from a survey conducted at the end of the

project that the majority of publishers were not, largely for economic reasons, yet ready to implement or participate in such a service. Nevertheless, this work has been used to inform the development of a COUNTER Code of Practice for Articles. Furthermore, the project found that usage of articles hosted by institutional repositories was substantial. As a result of this, a second set of aims and objectives emerged: to develop the technical, organizational and economic models for the standardized recording and reporting of usage at the individual item level – regardless of content type – for items hosted by institutional repositories and subject repositories (IRUS).

To support these extra objectives, a secondary demonstrator service was developed, which focused solely on repositories. It revealed that significant numbers of other item types (theses, conference papers, reports, etc.) were also being regularly downloaded. This additional work ultimately led to the establishment of IRUS-UK, which adheres to both the COUNTER Codes of Practice (Articles & e-Resources).

3. IRUS-UK Usage Statistics Portal

The IRUS-UK service provides a single gateway for libraries to access statistics relating to usage events recorded within their IR. In particular, it contains COUNTER-compliant usage statistics for each participating UK higher education institution's IR (institutional repository). The service, underpinned by a MySQL database, comprises:

- A web user interface (written in PHP)
- Downloadable reports
- An initial API
- A SUSHI (Z39.93) server

All institutional members of the UK Access Management Federation [6], whether or not their institutional repository is an IRUS-UK participant, can log in to the IRUS-UK portal and view the statistics and reports listed below.

3.1 Summary Reports

IRUS-UK provides a number of summary tables and reports which give an overview of downloads from our participating repositories. You can see:

1. An overall summary of downloads for all participating repositories.
2. Total number of downloads for each individual participating repository.
3. A breakdown of repository participation and number of downloads by selected countries in the UK (England, Scotland, Wales).
4. A breakdown of repository participation and number of downloads by platform used (DSpace, Eprints or Fedora).
5. Numbers of each type of item downloaded and number of downloads of each type of item for all participating repositories.
6. Numbers of each type of item downloaded and the number and percentage for each item type which have DOIs available in the metadata that we harvest.
7. An analysis of the data ingest process for each repository showing raw data, exclusions for robots and double clicks, and the resulting number of downloads showing in IRUS-UK.

3.2 Usage Reports

The usage reports available include:

1. *'Item Report 1'* provides the number of successful item download requests by month and repository identifier for a selected repository.
2. *'Item Report 2'* provides the number of successful item download requests by month and item type for a selected repository.
3. *'Article Report 4'* provides the number of successful article downloads by month for participating repositories. The report can be filtered to limit the results to a selected journal or repository. It can be run for an individual month or over a number of months
4. *'Book Report 1'* provides the number of successful book downloads by month for a selected repository. It can be run for an individual month or over a number of months.
5. *'Book Report 2'* provides the number of successful book section downloads by month for a selected repository. It can be run for an individual month or over a number of months.
6. *'Electronic Thesis or Dissertation Report 1'* provides the number of successful thesis or dissertation download requests by month and repository identifier for a selected repository. For each thesis or dissertation, it shows the item URL, EThOS ID (British Library's Electronic Theses Online Service) [7] if available, title, author and total downloads by month and in total for the period selected. It can be run for an individual month or over a number of months.
7. *'Journal Report 1'* provides the number of successful Full-Text Article Requests by Month and Journal for participating repositories. The report can be filtered to limit the results to a selected journal or repository. It can be run for an individual month or over a number of months.
8. *'Repository Report 1'* enables you to view the number of successful item downloads by month for all participating repositories. The report can be filtered to limit the results to a selected item type, Jisc Band and/or Country.

3.3 Item Type Usage Reports

We map the hundreds of different item types used by our participating repositories to a core set of 25 item types*: Art/Design Item; Article; Audio; Book; Book Section; Conference Papers /Posters; Conference Proceedings; Conference or Workshop Item – Other; Dataset; Exam Paper; Image; Learning Object; Moving Image; Music/Musical Composition; Other; Patent; Performance; Preprint; Report; Show/Exhibition; Text; Thesis or Dissertation; Unknown; Website; Working Paper*.

All the original item types are stored so that items can be subsequently remapped if necessary. The choice of the 25 item types was informed by a major piece of work [8] examining both metadata guidelines for repositories and actual use of item types. In the IRUS-UK portal you can see for each item type the number of items downloaded and the number of downloads.

3.4 Search for Usage of an Individual Item

One can search for words or phrases in the title or author for all repositories or for an individual repository and for all item types or a specified item type. Search results include basic metadata, a link to the item in the host repository, numbers of downloads and additional statistics relating to the item.

3.5 Check Items with DOIs

IRUS-UK extracts DOIs from downloaded item metadata and provides two tables:

1. A summary for each item type of the number and percentage that have DOIs across all repositories.
2. A breakdown of article DOI availability by repository.

3.6 Robot Usage and Double Clicks

In order to produce COUNTER-compliant usage statistics, IRUS-UK excludes downloads by robots and double clicks on individual items. A table provides an analysis of the ingest process for each participating repository. We have a position statement on the treatment of robots and unusual usage [9] and are undertaking further work to refine this process.

3.7 Report Formats

The reports are made available both for human use and direct machine to machine use:

- Each report can be viewed in a web page in the portal or downloaded for use locally as MS-Excel/CSV files.
- The reports are available via the SUSHI protocol for incorporation into local institutional ERMs, or for automatic gathering for use in other national/global services.

3.8 Ingest Scripts

The ingest scripts, based on the original scripts devised by PIRUS2, have been significantly enhanced and refined through several iterations, adding:

- daily granularity instead of the original monthly granularity
- ‘separation of concerns’ to make the ingest more robust, and to simplify development and maintenance of the scripts.
- improved validation of incoming data
- additional filtering of robots and abnormal usage over and above the minimum specified by COUNTER

Data received for participating repositories gets stored in daily log files. The log for any given day is usually processed the following day.

There is currently a three step daily ingest process:

1. A Perl script parses the logs; processes entries from recognised IRs; sorts and filters entries following COUNTER rules to remove robot entries and double-clicks; filters entries using additional IRUS-UK filters; consolidates raw usage data for each item into daily statistics; and outputs to an intermediate file.
2. A Perl script processes the intermediate file output from Step 1; using the OAI identifier associated with item, it looks up each item against the Item Authority table in the IRUS DB to see if it is already known to the system; if a known item, it retrieves the existing IRUS Item Identifier; if the item is new – as yet unknown to IRUS – the script adds a stub-entry to the Item Authority table – minting a new IRUS Item Identifier and adding the repository identifier, platform and OAI identifier to the table with the rest of the metadata set to 'unknown' at this stage; finally, the script adds the download statistics associated with each IRUS Item Identifier to the Daily Statistics table.
3. A Perl script obtains the "unknown" metadata for new items: it queries the DB to find the 'known unknowns' using the OAI identifiers; issues OAI-PMH GetRecord calls to retrieve OAI_DC metadata; parses the OAI records; updates the metadata – Title, Author, Item Type, etc., in the Item Authority Table in the DB; and additionally maps the Item Type, as given by the source repository, to a smaller (more manageable list) of IRUS Item Types.

In addition to the daily ingest scripts, additional scripts are run every few days to add journal information to article records and a further script is run at the end of each month to consolidate the Daily Statistics for that month into a Monthly Statistics table.

3.9 Robots and Unusual Usage

The starting point for eliminating robots and machine accesses from the raw usage data being collected was the COUNTER robots exclusion list. The list contains a set of regular expressions (regexes) of User agents to exclude and is described by COUNTER as a 'minimum' requirement. However, as the service has taken on-board more repositories, it has become obvious that the list is not comprehensive enough to exclude all robots and unusual usage. The problem became most obvious – and acute – when the London School of Economics (LSE) joined IRUS-UK and apparent download figures rose dramatically. At that point, analysis of the data identified a number of further exclusions not in the COUNTER list, including half a dozen user agents and two IP ranges used by Baidu Spider (which User Agent exclusion would not identify). Consequently, the service supplemented the COUNTER exclusions with:

- a set of additional IRUS-UK filters employing the newly identified User Agents and IP ranges.
- an additional check to exclude data where a single IP has exceeded a daily threshold for downloads –unless identified as a legitimate source of high download levels, e.g. an organisational proxy server.

These filters do work reasonably well, but the team was still convinced that more could be done to eliminate even more suspect usage. So, work was commissioned jointly by IRUS-UK and COUNTER to devise an 'adaptive filtering system' – a set of algorithms that will allow the service to dynamically identify and filter out unusual usage/robot activity. The work was undertaken by Information Power Limited, who have supplied a

report and an initial set of scripts of use by IRUS-UK. The results of that work will be assessed, tested, refined and applied to the service in the next phase of development. The information has been shared with the community, via COAR Interest Group 'Usage Data and Beyond' [10] and has led to the formation of a COUNTER Working Group on Robots.

3.10 Updated Tracker Protocol Specification

The specification for this is quite brief and straightforward:

- When a user clicks on a link to (i.e. downloads) a file from a Repository with the tracker protocol in operation, an OpenURL log entry is sent to a remote server for further processing.
- The OpenURL log entry should be based on a subset of the NISO OpenURL 1.0 standard KEV ContextObject Format. The OpenURL string must be URL encoded, with key-value pairs separated by '&'.

The initial specification used by IRUS-UK – based on PIRUS2 work - was designed to work at 'item' level. This is quite adequate for most items which contain a single file; however, there are a proportion of items that may have multiple files associated with them, i.e. the work is divided into chapters or contains appendices or other supplementary materials. In order to accommodate such items and allow reporting at a finer granularity in a future iteration of the service, the team has devised an updated specification containing an extra metadata element – the fileURL – to be transmitted from repositories to the IRUS-UK server (Table 1).

Table 1. Tracker Protocol

Element	OpenURL Key	OpenURL Value (example)	Notes
OpenURL version	url_ver	Z39.88-2004	Identifies data as OpenURL 1.0. String constant: Z39.88-2004 (Mandatory)
Usage event datestamp	url_tim	2010-10-17T03%3A04%3A42Z	Date/time of usage event (Mandatory)
Client IP address	req_id	urn:ip:138.250.13.161	IP Address of the client requesting the article (Mandatory)
UserAgent	req_dat	Mozilla%2F4.0+%28compatible%3B+MSIE+7.0%3B+Windows+NT+5.1%3B+Trident%2F4.0%3B+GoogleT5%3B+.NET+CLR+1.0.3705%3B+.NET+CLR+1.1.4322%3B+Media+Center+PC+4.0%3B+IEMB3%3B+InfoPath.1%3B+.NET+CLR+2.0.50727%3B+IEMB3%29	The UserAgent is used to identify and eliminate, by applying COUNTER rules, accesses by robots/spiders (Mandatory)
Item OAI identifier	rft.artnum	oai:dspace.lib.cranfield.ac.uk:1826/936	(Mandatory)
FileURL	svc_dat	https://dspace.lib.cranfield.ac.uk/bitstream/1826/936/4/Artificial_compressibility_Pt2-2005.pdf	(Mandatory)
HTTP Referer	rfr_dat	http://www.google.co.uk/url?sa=t&rct=j&q=http%20referer&source=web&cd=4&sqi=2&ved=0CeoQFjAD&url=http%3A%2F%2Fwww.whatismyreferer.com%	(Mandatory) The HTTP header field that identifies the address of the webpage (i.e. the URI) that

		2F&ei=zIBCU6fbEoOqhQf67YcwBg&usg=AFQjCNFt-KmqneTZfEb6OxjPZlD4ogiJcQ&sig2=wZJYkoWgNScNjgxRbRs29w&bvm=bv.64125504,d.ZWU	linked to the resource being requested. The 'HTTP Referer' is used to help identify and eliminate accesses by robots/spiders.
Source repository	rfr_id	dspace.lib.cranfield.ac.uk	(Mandatory)

3.11 Tracker Code

An Eprints Tracker plug-in, developed by Eprints Services, for Eprints 3.2.x and 3.3.x. is available from the 'Eprints Bazaar' [11].

Patches are available for DSpace, developed by @mire, for versions 1.8.1, 1.8.2, 1.8.3, 3.1, 3.2, 4.1 and 4.2. The patches are available on request.

Fedora implementations require bespoke code to be developed by the repository implementers. Example implementations exist for Java and RubyGem for Hydra.

The team conducted a small scale trial involving the OAPEN Library [12], which runs on ARNO repository software. The software has been modified to include IRUS Tracker functionality and is successfully transmitting OpenURL messages about e-book downloads – demonstrating the ease of applying the technical solution and its transferability beyond institutional repositories.

Discussions have been initiated with Atira (now part of Elsevier) to see if it is feasible to add Tracker functionality to the PURE Portal software.

4. Benefits of a Shared Service and Community-driven Developments

In theory, every institution could produce its own COUNTER-compliant statistics for its repository. The rules for eliminating robot accesses and double-clicks and for counting downloads are not that difficult to understand or implement. However, there is more to COUNTER compliance than simply following the COUNTER Code of Practice. In order to become truly COUNTER compliant, it is necessary to go through a regular auditing process. By the time of registration, annual membership and report auditing fees are taken into account, this can potentially cost several thousands of pounds per year per IR. By collecting and processing download data into COUNTER statistics on behalf of IRs, IRUS-UK can substantially reduce these costs; in this scenario, only IRUS-UK itself needs to be audited, the individual IRs do not.

Additionally, IRUS-UK is in a position to act as an intermediary between UK IRs and other agencies, such as OpenAIRE [13], which has an interest in obtaining usage statistics for research outputs funded under the European Seventh Framework Programme (FP7). Having a single point of access to FP7 article statistics for the UK will be a lot easier to manage than collecting those statistics from all the relevant individual repositories.

IRUS-UK data can be used for a number of different purposes, some of which have been highlighted in a series of use cases. The use cases are summarised below.

4.1 Reporting to Institutional Managers

It is useful for institutional managers to understand the usage of items in the institutional repository. This might include, for example, obtaining high-level statistics of total downloads from the repository, or gaining an understanding of the items that have higher downloads. The usage statistics within IRUS-UK can be used to report on downloads for institutional managers. A user can find out the total downloads from each repository, or can use the focused reports for more granular information such as downloads by item (to help identify items receiving high numbers of downloads), downloads by item type, downloads of Electronic Theses and Dissertations, and number of downloads from all participating repositories, which can be filtered by item type, Jisc band, or country (or a combination).

4.2 Reporting to Researchers

Researchers are often interested in knowing the usage statistics of their items in the institutional repository; this could be for review purposes, for reporting to Research Councils, or just for curiosity. Additionally, a researcher may be involved in dissemination or publicity (e.g. a conference presentation) that refers to their research, and they may wish to see if this has resulted in an increase in the number of downloads of the item. In the IRUS-UK search a user can search for a specific item to report on. The result shows monthly downloads of the item (since a download was recorded in IRUS-UK) and daily downloads for the last month.

4.3 Benchmarking

Being able to benchmark institutional repository statistics is valuable, both with an institution's own data (to look at trends), and with other institutions (to allow comparisons to be made and to provide a wider context within which to interpret the performance of an institutional repository). The standardised COUNTER-compliant statistics available through IRUS-UK enable reliable benchmarking, both with an institution's own data for longitudinal analysis, and with other institutions within IRUS-UK. Users can view total downloads recorded by IRUS-UK for all participating repositories, or can look at monthly data for all participating repositories to look at trends. Repositories can also be filtered by different groupings (Jisc band or Country) or filter by item type (for example downloads of Articles) and filters can be combined.

4.4 Supporting Advocacy

The data within IRUS-UK can be used to support advocacy by sharing headline download figures from all participating repositories, reporting on an overall total number of downloads from a repository since joining IRUS-UK, showing monthly download figures (and trends), identifying items with high levels of downloads, gathering statistics on downloads of different item types within a repository, or sharing downloads for particular researchers or research areas. These statistics can then be used in a number of different ways including presentations, newsletters, blog posts, reports, social media, meeting updates, etc. These may be focused specifically on the performance of a repository, or, more broadly, on Open Access advocacy.

5. Conclusion

IRUS-UK provides a usage statistics service for UK repositories, based on the COUNTER standard, which enables them to expose credible, authoritative and trustworthy usage figures for item downloads, on the same basis as – and therefore comparable with – the majority of publishers, in an extremely cost-effective manner.

By providing a nationwide view of UK repository usage, it also benefits national organizations such as Jisc and SCONUL, and offers opportunities for benchmarking as well as the ability to act as an intermediary between UK repositories and other agencies. We hope that IRUS-UK will act as a model which can be adopted in other countries and regions around the world.

Finally, it may help to inform the current debate, taking place in the absence of reliable or comprehensive usage data, about the value of repositories and their place and significance in the dissemination of OA research literature.

References

[1] COUNTER Code of Practice: http://www.projectcounter.org/code_practice.html (accessed 18 May 2015).
[2] J. Bollen, H. Van de Sompel, An Architecture for the Aggregation and Analysis of Scholarly Usage Data, http://arxiv.org/abs/cs/0605113 (accessed 18 May 2015).
[3] Knowledge Exchange: http://www.knowledge-exchange.info/ (accessed 18 May 2015)
[4] JUSP: https://www.jusp.mimas.ac.uk/ (accessed 18 May 2015).
[5] PIRUS2 Final Report: http://www.projectcounter.org/News/Pirus2_oct2011.pdf (accessed 18 May 2015).
[6] UK Access Management Federation: http://www.ukfederation.org.uk/ (accessed 18 May 2015)
[7] British Library eTheses Online Service: http://ethos.bl.uk/ (accessed 18 May 2015)
[8] IRUS-UK Item Type Mappings: http://www.irus.mimas.ac.uk/help/toolbox/IRUS_item_type_report_v3.3.pdf (accessed 18 May 2015)
[9] Position statement on the treatment of robots and unusual usage: http://www.irus.mimas.ac.uk/news/IRUS-UK_position_statement_robots_and_unusual_usage_v1_0_Nov_2013.pdf (accessed 18 May 2015)
[10] Confederation of Open Access Repositories Interest Group: Usage Data and Beyond: https://www.coar-repositories.org/activities/repository-interoperability/usage-data-and-beyond/ (accessed 18 May 2015)
[11] Eprints Bazaar http://bazaar.eprints.org/ (accessed 18 May 2015)
[12] OAPEN-UK: http://oapen-uk.jiscebooks.org/ (accessed 18 May 2015).
[13] OpenAIRE: http://www.openaire.eu/ (accessed 18 May 2015).