Integrating Download Statistics from IRUS-UK into an EPrints Repository

Alan Stiles *ORCID : 0000-0003-3343-1088*

Library Services
The Open University

Introduction

Journals and the authors of their articles are traditionally rated by academic publishers using schemes such as h-index and journal impact factor counts, which tend to relate citation counts and to sales and views of an entire journal issue rather than specific articles. Book chapters tend to be considered only in terms of the whole book, conference items may not be rated at all or may be counted as a part of the conference proceedings, if published.

With repositories each item can be considered as an independent entity, and where the full-text is available to be downloaded, these downloads can be tracked via various mechanisms and provide an interesting metric of the popularity of the specific item. As with the other ratings mechanisms above, this is more of a quantitative rating than a qualitative one, although there is a tendency to substitute the more readily measured score for the difficult to rate one.

Statistic Tracking Mechanisms

Various mechanisms have been used within Open Research Online (ORO - the Open University's Institutional Repository) to track item downloads over time, and many of these are still used to track other, perhaps less significant, repository usage factors.

Google Analytics

Originally used for all of ORO's transactional metrics, Google analytics makes use of various cookies, image files and script-based mechanisms to submit usage data to Google's servers to be analysed, the results of which can then be viewed within a web interface. However, this process presents a number of issues -

- The processing done on the data is hidden and unknown
- Direct accesses of item files, e.g. from search engine results, may not trigger the appropriate scripts to count the download
- Many people configure their web browsers to prevent scripts from running or to block third party cookies, which would prevent the download being counted

IRStats2

Available as a readily installed and configured package from the EPrints Bazaar (a publicly available applications and extensions 'store' for more recent versions of EPrints software), IRStats 2 runs as part of the local EPrints package, utilising the EPrints database to analyse various aspects of repository usage including deposit of new items as well as item downloads. This is perceived as an improvement over Google Analytics as

- it can analyse all download activity from the local database
- it is open source software so all processing is visible
- it can be extended to filter undesired data
- it is easily integrated into the repository, providing graphs and tables of usage data.

Alan Stiles, Library Services, The Open University, MK7 7DA, United Kingdom

IRUS-UK

IRUS-UK stands for Institutional Repository Usage Statistics - for United Kingdom. It came out of the PIRUS project by MIMAS (originally part of the University of Manchester, now part of the Digital Resources division at Jisc, a charitable organisation championing the use of digital technologies in UK education and research.), which aimed to integrate download statistics from both publishers and institutional repositories, but faced significant hurdles with access to usage data from the publishing industry.

It makes use of the PIRUS tracker protocol, which requires a small amount of code to be installed on the local server (easily available from the EPrints Bazaar), which is triggered to run every time an item is downloaded, passing details of the item and the http request header to the IRUS-UK server. IRUS-UK subsequently analyse this data in an attempt to parse out downloads by automated harvesters, and to pull further details of the downloaded item from the originating repository, such as titles, authors and DOIs, via OAIPMH (Open Access Initiative Protocol for Metadata Harvesting). The data from all participating institutions is then cleansed and processed to be better than COUNTER compliant before being presented back to all of those institutions via a web interface on the IRUS-UK website, with a range of reports modelled on SUSHI (Standardised Usage Statistics Harvesting Initiative from NISO) standard reports.

The benefits of using download statistics from IRUS-UK are

- processing fully described
- analyses complete download data
- Low/no local maintenance overhead for COUNTER compliant statistics
- Consistent, independent processing which allows good comparisons and benchmarking

Use of Download Statistics

In February 2013, after a beta version of IRStats2 had been installed and tested over approximately 6 months for stability and accuracy, the ORO team undertook an exercise to disseminate details of these download statistics to a number of the Open University research staff to gauge whether they considered that data to be useful. Approximately 750 currently employed individuals who were named as contributors to work which had been deposited in the 12 months from January 2013 were sent an email with a brief explanation and a link to the page of download statistics for their total contributions to the repository. The number of site visits over the subsequent few days saw a significant spike before returning to levels similar to those prior to the mailing. The ORO team mailbox, from where the email had been sent, saw over 70 replies, none of which were negative in relation to the statistics. 10 of the replies related to items either missing from an individual's publications list or having been attributed to someone in error, whilst the remaining responses varied from basic thanks for the email, through individuals noting how helpful it would be in terms of writing future funding bids and on to a number of replies indicating how the statistics had inspired the authors to ensure their full list of publications were available to download through the repository.

Alan Stiles, Library Services, The Open University, MK7 7DA, United Kingdom

Subsequent messages from most of those authors contained full text for existing items, in some cases requesting to replace an embargoed proof or version of record (VOR) with an openly accessible author's accepted manuscript (AAM).

At around the same time as this exercise took place (January 2013), we also implemented the IRUS tracker into the repository, becoming the thirteenth repository to participate in IRUS-UK (82 participant repositories as of April 2015). Throughout the remainder of 2013, and continuing through to the present day, we have run comparisons between the reports we see from or local IRStats and those we see through IRUS-UK.

The numbers tend to be broadly similar, but on occasion there is a definite divergence between the numbers. This divergence has been discussed between the participating repositories, IRUS-UK and the IRUS Community Advisory Group, a group of representatives from the participating institutions whose aim is to direct the future development of the IRUS service and have been attributed to various causes. Chief amongst these are the different processing of the two systems to identify automated malicious downloads, instances where downloads of auxiliary files may have been counted by one system but excluded from the other and different time windows for the systems to exclude duplicate downloads (e.g. 2 or more downloads of a specific item from the same client IP address).

Satisfied that the download statistics were broadly in line with each other, accounting for the above exceptions, we have moved to providing our regular reporting of downloads based on the reports from IRUS-UK as they are independently processed, can be easily compared to all the other IRUS-UK participants and reports are available to provide many of our requirements such as top item downloads and most downloaded author data.

Integrating IRUS-UK with the Repository

Including the tracker code in the repository was relatively straightforward, requiring just a small amount of investigation and configuration to ensure the tracker pings were reaching the IRUS-UK servers via the Open University's web proxy servers.

Whilst the results of the data processing are available to authorised human reading via the reports on the IRUS-UK website, these are not so accessible for automated machine reading and reuse. To deal with this use case, IRUS-UK have implemented an API (V1) which allows various data to be requested and returned in the form of JSON data structures.

These data are returned relatively quickly, depending on the exact nature and size of the response and can be readily used to produce graphical representations of the data, for example by using Google-Charts functions.

Whilst comparing the formats of the various responses from the IRUS-UK API it was realised that the current item being used for testing returned different data depending on whether the request was submitted using an OAI identifier or a DOI. Further examination of these and other items illustrated that the API was returning the download statistics for a duplicate of the item, where it was available from another institutional repository. A subsequent conversation

with IRUS-UK developers confirmed that, with the V1 API, the DOI search would look through the IRUS-UK index to find the first recorded item with that DOI, and then return those statistics, so whilst OAI identifiers would have a one to one relationship with items in the IRUS-UK dataset, there may be a one to many relationship between a DOI and IRUS-UK items, but there was no way to specify which IRUS-UK item was returned for a given DOI within the existing API.

With the discovery of DOIs returning data from other repositories, it was realised that this data could be used to further enhance the display of download statistics by including details of these 'external' downloads in the displayed data. A simple example would be to include them as a stacked bar chart.

The extra processing involved in sending two requests to IRUS-UK, one by OAI identifier for the item held in ORO, and one by DOI for the external repository, then the processing to compare these to determine they were not the same set of statistics, and then render them in some form was increasing the processing delay to unacceptable levels to achieve this via AJAX calls on demand, at the time the item abstract pages were loaded. It would also present a load issue to the IRUS-UK server if it was receiving tens or even hundreds of thousands of API requests per day for each of the participating repositories. As the data is effectively static, only being added to in batches and not changing historically unless significant reprocessing was necessary, it made sense to store a copy of the relevant data locally. In doing so it would significantly reduce the load on the IRUS-UK server as well as reducing the delay to render the data for the end user.
Testing and discussions with IRUS-UK are currently underway to determine the most appropriate mechanism for updating the local data-cache, with consideration being given to retrieving all relevant repository data in one batch, or capturing requests for specific items abstract pages and triggering a server process to request that specific and much smaller dataset so it will be updated for the next viewer whilst the current requester will see the existing data from the cache which may be missing the most recent figures. Capturing the request to trigger the update of the dataset is relatively straightforward due to the need to use the local EPrints server as an intermediary for the data request in order to avoid cross-site scripting restrictions. It would also permit the updates to be scheduled to happen during a time agreed with IRUS-UK when the servers had lower load and could cope with the requests more readily.

Subsequent to the discovery of the potential difference between DOI and OAI identifier statistics, it was realised that ORO holds a significant number (of metadata only records which include a DOI and that these could also be used to retrieve download statistics for these items if they were available through an alternative repository. The idea with these is that seeing interest in the article may encourage the author to seek out a copy of the full text which can be deposited locally, or even that the external repository could ultimately be identified through IRUS-UK and a copy of the full-text pulled into ORO automatically, although this would require further enhancements from IRUS-UK.

Alan Stiles, Library Services, The Open University, MK7 7DA, United Kingdom

Future developments

IRUS-UK are currently undertaking a piece of work to incorporate a SUSHI-Lite compliant API as a demonstrator for the new standards described in the recent Technical Report from the NISO SUSHI-Lite Working Group (the membership of the working group contains both the author of this paper and the lead developer of IRUS-UK). Once this work has been implemented it may provide an improved mechanism for obtaining the download statistics for all IRUS-UK items related to one DOI and with the potential to identify the URIs of the item in external repositories, so the local functions to process the JSON responses will be amended to use this path to the data rather than the current V1 API.

This change may also provide the opportunity to retrieve and incorporate item usage statistics for the version of record, directly from publishers, although use and display of these data will likely be subject to contracts.

Having the data in a local store may also allow the display of IRUS-UK COUNTER compliant statistics on a per-author and per-faculty base.

Further investigation will also be undertaken into how the data is displayed, with consideration to including the graphs directly into the static abstract pages rather than using AJAX methods to dynamically incorporate them, as the data tends to update on a similar interval so it may not require the client scripting solution.

There is also the potential for improved filtering, both of the data received by IRUS-UK and within EPrints itself to remove downloads by bots and spiders from the statistics. Standard crawlers such as search engines and recognised harvesters like CoRe obey instructions from the EPrints server regarding the frequency of their activity and what items they may access, and will identify themselves clearly during their crawling activity and thus are simple to remove from the statistics, but there is an ever-present element of malicious spidering which represent themselves as normal web browsers but ignore instructions regarding access limits and so add excessive load to the servers, either deliberately attempting to cause the repository services to fail, or for any one of a number of alternative reasons. This is obviously not something that affects just institutional repositories, so there is a large global effort ongoing to reduce and filter this undesired traffic, however there is no clear solution evident. For this reason, the IRUS-UK statistics, although better than COUNTER compliant, should be perceived as a ‘best-effort’ to report the repository usage, and considered more in terms of relative trends than absolute numbers.

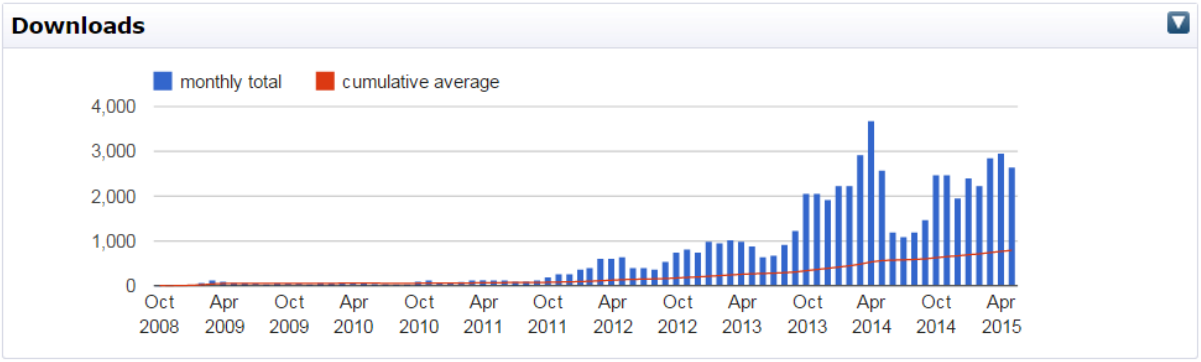

Image 1 – example chart of download statistics from IRStats2 package in eprints

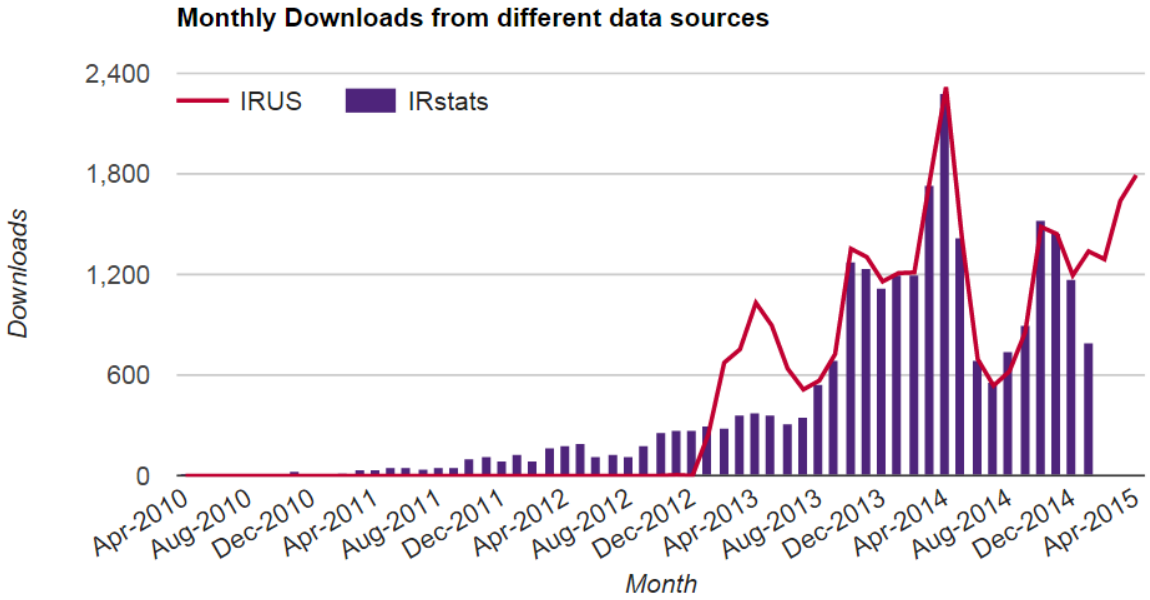

Image 2 – example graph generated with data from IRUS-UK API compared against IRStats2 data showing differences in counts after processing. Created using Googlecharts functionality

Alan Stiles, Library Services, The Open University, MK7 7DA, United Kingdom

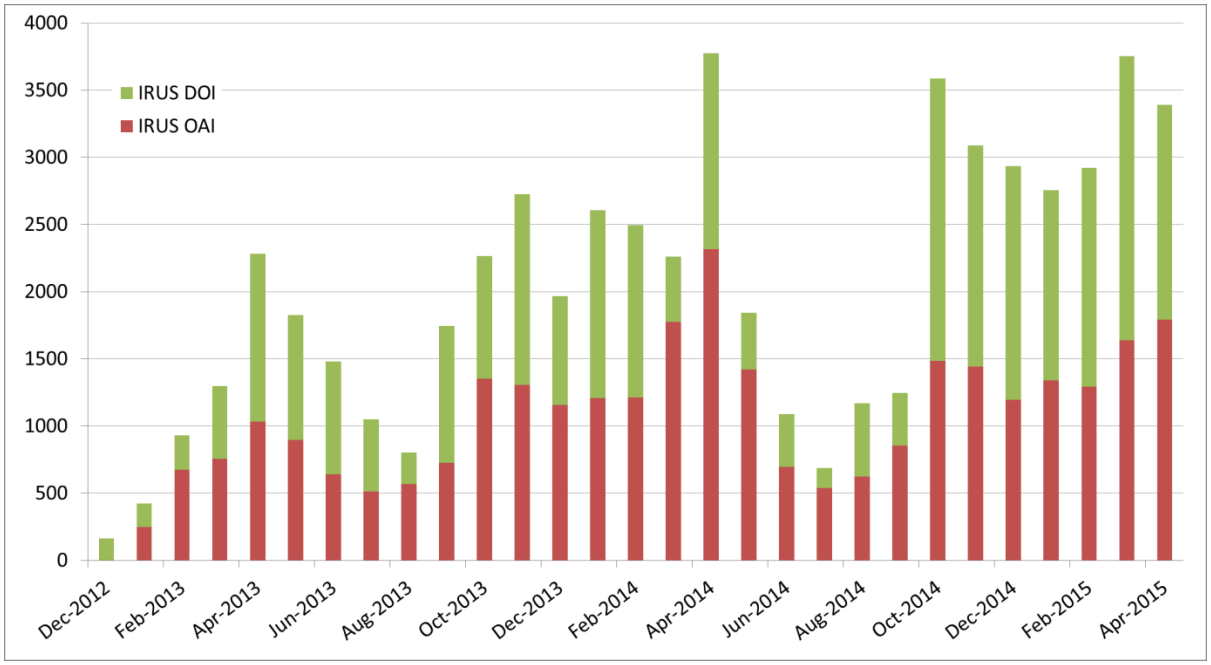

Image 3 – example chart combining data from IRUS UK for one publication downloaded from multiple repositories

Alan Stiles, Library Services, The Open University, MK7 7DA, United Kingdom

Appendix A - Acronyms

AJAX	Asynchronous Javascript and XML (a dynamic web technology)
API	Application Programming Interface
CoRe	Connecting Repositories - a Jisc service for Institutional Repositories
COUNTER	Counting online Usage of Networked Electronic Resources
DOI	Digital Object Identifier (A unique and persistent identifier for digital objects)
EPrints	Open Source repository software developed by Southampton University
IP	Internet Protocol (IP address: identifier used for electronic communications)
(P)IRUS	(Publisher and) Institutional Repository Usage Statistics
JSON	Javascript Object Notation - a hierarchical way to represent data structures
MIMAS	Manchester Information & Associated Services
NISO	National Information Standards Organization (U.S. standards body)
OAI	Open Access Initiative
OAIPMH	Open Access Initiative Protocol for Metadata Harvesting
ORO	Open Research Online
SUSHI	Standardized Usage Statistics Harvesting Initiative
URI	Uniform Resource Identifier

Appendix B – IRUS-UK filtering process

IRUS-UK processing involves filtering out robots based on known-malicious IP addresses, and comparing the user agent string presented to a list of previously identified automated robot strings - one list provided by COUNTER plus an extra list of user agents identified as suspicious by IRUS-UK, and then removing multiple downloads of an item from a single IP address within a specified time window, as well as multiple downloads (over a threshold) from several different repositories by a single IP address within a specified time window.

Alan Stiles, Library Services, The Open University, MK7 7DA, United Kingdom