Interoperable Repository Statistics Project

Project evaluation

Prepared by:

Key Perspectives Ltd
48 Old Coach Road
Playing Place
TRURO
TR3 6ET

CONTENTS

Page

1. INTRODUCTION

The aim of the project was to produce an interoperable statistics package that measures repository usage. The tool would have the capability to filter out web crawler visits that distort the true usage picture.

The package is now released in beta. It has three components:
- A web log database
- An OAI interface to enable data sharing with analytical packages that measure usage
- A package that measures usage and creates reports of the download activity from a repository

At the beginning of the project Key Perspectives carried out a user requirements study to find out what the main stakeholder groups – repository managers and researchers – would find most useful when measuring and reporting usage from their repository. Both groups demonstrated strong interest in having a tool that could do this and there was a great deal of commonalty between them with regard to what they would like it to do. The major differences were that repository managers saw it not only for usage measurement alone but also as a tool for advocacy and for senior management to use for research measurement, whereas the researchers were very focused on using it to find out who is using their work, wanting as much detail as possible in this respect. Interestingly, whereas the repository managers hedged on using the proposed tool too overtly to create a competitive environment around their repository (e.g. highlighting the most-downloaded articles) for fear this would actually discourage researcher involvement, the researchers themselves were rather keen on the notion of league tables and see value in them. The implementation of such features is a matter for each individual institution, of course, so we decided they should be made available as part of the software functionality, with the implementation decision left at local level.

Perhaps not surprisingly, overall the repository managers were the more sophisticated in comprehending what the software might be able to offer them and what would be possible. At this point we made the decision to use this stakeholder group for the final evaluation, then to make the software available to them to implement on their own repository and encourage them to gather end user responses to provide us with further feedback.

Now, as the software is released, we have been carrying out an evaluation exercise with repository managers to ascertain how well it measures up against their requirements and expectations.

Key Perspectives

2. METHODOLOGY

When the software was ready for use, a guide was produced showing the functionality of the tool and the types of report that it could generate. It was prepared in the form of a walk-through of the analysis and interpretation that may be made using the tool.

The guide was sent to nine repository managers (the evaluation group), most of whom had been part of the initial user requirements study and had at that time given their views on the proposed software. We asked them to study the guide and then we asked them a set of questions to find out whether they liked the functionality, the outputs and the general look-and-feel of the product.

3. RESULTS

3.1 Stakeholder requirements

The initial user requirements study indicated positive demand for a number of indicators to be provided by the software. These are listed in Table 1 below. There are notes in the right hand column to indicate the degree of technical difficulty:

Feature requested	Notes
Repository usage:	
Total downloads	Possible
Average downloads over a period	Possible
Origin of accesses:	
- by country	Possible
- by institution/domain type	Possible
- by department/research group	Extremely difficult
Download activity:	
- quarterly	Possible
- per month	Possible
- per week	Possible
- per day	Possible
Trends analysis:	
Usage of individual articles over time	Possible
Usage of articles by specific user (e.g. domain name) over time	Possible
Usage of articles by subject area	Difficult
Article usage:	
Views of metadata (hits)	Possible
Views of full-text (downloads)	Possible
Combined totals	Possible
Overall usage of article:	
Usage from repository	Possible
Usage from publisher site	COUNTER-compliancy
Usage from other sources	Extremely difficult
Aggregation of usage from different sources	Extremely difficult

Table 1: stakeholder requirements list

The final software provided most of these features plus some additional ones. The evaluation group were asked a series of questions about the final functionality as demonstrated in the walk-through guide provided and their responses are reported below.

Key Perspectives

3.2 Current repository usage analysis activity

The DSpace repository software includes a usage statistics package: four of the respondents have DSpace repositories and two of them are using this. Both are also looking at implementing Google Analytics (possibly this says something about the DSpace package). Technicians at one of the other DSpace repositories have built their own analysis software and are currently upgrading it. Two repositories use AWstats. One (an EPrints repository) is currently examining options for a usage analysis package and two do not use any analysis tool at present. One of these is a DSpace repository with staff looking at the DSpace tool with a view to implementing it. Finally, it is instructive to know here that one repository is being used solely as an internal reporting tool for senior management at the moment, though there are plans to give it a more outward-facing role, positioning it as a dissemination tool and engaging researchers by demonstrating how it can further their own self-interests. The respondents from this institution therefore have an interesting and slightly different perspective on usage analysis at the moment, which was welcome in this study.

3.3 Comparison of current usage analysis packages with the IRS software

The DSpace statistics package users all said that IRS does much more than their current package. One of them, currently evaluating Google Analytics, commented that Google Analytics does more than IRS in total but is not specifically repository-oriented, so it doesn't report the same useful things (such as the ten most downloaded papers). One respondent said that AWstats focuses on web usage which compares unfavourable with IRS's focus on the usage of eprints.

The institution with usage analysis software created in-house reported that this software generally does the same as IRS but IRS can generate bespoke reports which the local software cannot.

Staff at the repository where usage analysis packages are being evaluated reported that their specification had listed almost all the IRS capabilities as options they wished to have.

3.4 Presentation

The presentational aspects of IRS were very well thought-of, variously described as 'good', ' great' and 'excellent', although one person though the graphs were not as good as they could be and don't measure up to some of the presentational aspects of the modern graphical web environment. Another said that presentation of some of the statistics is somewhat basic, citing the Download Count HTML as an example. Despite these gripes about detail, though, there is

general agreement that the presentation and features of IRS are really good overall.

The 'Download Dashboard' is a particular success. More than half the respondents specifically mentioned this and praised its clarity and the thought that had gone into presenting the statistics in this way.

People appreciated the level of detail the package provides. One commented that graphs, charts and summarised data are essential for report-writing and that it was good to have these generated at the touch of a button, even if options are limited to predefined formats. The ability to generate of bespoke reports is much appreciated and the user interface for this is considered to be very user-friendly, though one person queried what the Top Ten Search Terms referred to (whole repository, or just for the article concerned?).

One person commented that the package is easy to embed in the repository.

3.5 Does IRS measure the right things?

All respondents gave an unequivocal 'yes' to this question, but some also added further comments, suggestions and questions, some of which are already provided by the software:

- Measuring hits (metadata-only reads) as well as full-text downloads would be useful. There is merit in knowing how many times the metadata/abstract are viewed because users may be routing from these to the publisher's version of an article and it would be useful to know what volume of traffic hits the metadata only.
- Add more detail in the Highest Climbers area; for example, offer monthly comparisons so that it is possible to look at usage over short periods.
- Make it possible to analyse activity by community (e.g. research group) [provided]
- Make it possible to modify the Top Ten tables to show the Top Twenty or Top Hundred
- Show the download count with the item itself [already provided]
- Make it possible to add another author to the analysis, i.e. analyse two authors at once
- Provide a COUNTER-compliant presentation [provided]
- Make it possible to count downloads with one button click
- Make it possible to count supporting items (e.g. datasets, video files, Powerpoint files) as well as the full-text document

3.6 Usefulness of the IRS tool

Key Perspectives

Every respondent agreed that the IRS tool would be useful for their repository. As well as this general agreement, there were some specific points made and these can be summarised as follows:

- IRS will be especially useful when talking to senior administrators about using the repository for administrative purposes [two people]
- IRS will be very useful for research management because it can be used across schools, research centres and individuals (this comment is from the institution that is currently using the repository only for administrative purposes)
- IRS will be extremely useful in advocacy work for the repository [two people]
- IRS will be very useful when the repository manager tries to appeal to academic egos(!)
- The respondent currently evaluating usage statistics packages said he would recommend that IRS is considered as one of the packages being evaluated by the procurement staff

3.7 Additional comments from the test community

Most of the respondents offered additional comments and these are reported verbatim here:

"I think stats are going to play an increasingly important role for IRs. However, they need to be meaningful and transparent. The goal of how useful a user actually found an item would of course be useful (as opposed to just downloading it and then forgetting about it). Bit of a pipe dream though. The importance of stats will be for RAE (when they tell us what's going to happen) and institutional administrators at all levels - not just top brass. Massaging the academic ego is a good one - helps encourage use of the repository by showing who is popular.

We all have to be careful when dealing with stats though - who we tell what to. If academics think the whole purpose of them is for senior management to have a 'big brother' eye on them, they won't be very popular. Handled well though, they will be extremely useful.

Stats need to be very flexible in what they will show. Downloads by individual, dept etc. depending on who wants them."

"Ideally it would not be written in Perl, relying on proprietary graphing tools. From a DSpace perspective, we are trying to move to a 100% Java solution. Going back to Perl goes against this stated aim of the DSpace platform. Java with a free library like jfreechart would be a better solution from our perspective."

"Things that we struggle with are:

1) National or International comparisons. ISI Thompson charge 10K USD + for their citation averages UK and International and even if you bought them it would be a very depressing experience. Any insights that can be gained into how well we are doing in

Key Perspectives

comparison to others would useful.Making available top line data (with their permission) for e-prints repositories using this tool would provide something.

2) Finding a metric that would measure quality in multidisciplinary areas or areas not otherwise defined by ISI would be helpful too if other E-Print users would share the data. Downloads could be such a measure.

3) Articulating the value of measures of usage has proved a challenge here at University X. We are finally getting to grips with Impact Factors and so forth. Case studies and explanatory material that put forward benefits limitations and applications of data would help."

"For the referrers table, when the referrer is Google, Yahoo,... is it possible to present immediately the search terms used ? If so, the users will not have to analyse the url to detect the search terms used (for example : Referrer : www.google.co.uk ; count 10 ; search term : suitcases sculptures)

The total downloads is very useful, but is it possible to add the last month (or quarter) mean?

How can we make the difference between internal downloads and external? And in a larger view, how can we have a view on "auto-downloads" (authors who downloads himeself his own papers in order to increase his download number !). Maybe using the downloads from the same IP, the downloads from internal IP range.... ?

It would also be very helpful to integrate a citation analysis of the publications in the repository."

"At this stage, installation instructions are poor… eg see http://trac.eprints.org/projects/irstats/wiki/Download. We worked out that the page provides an 'instruction' to get the software, but in fact there is no explicit instruction. It is necessary to have the 'subversion' client in order to access the install package.

We're slightly uncomfortable that there are commercial products as part of IRStats (even though it will work with OS equivalents). A techie responsible for downloading the software may not have access to any budget (even though it's very cheap $15 + $99) and so testing may be held up or limited. It would be helpful if they added a tarball (*.tar.gz) of the current install package to simplify installation. Perhaps JISC could purchase the two costed pieces of software for the community."

4. SUMMARY EVALUATION

4.1 Delivery against stakeholder requirements

Below we reproduce the table that appeared in section 3.1. It shows which features originally appearing on stakeholders' 'wishlists'. It now has a new column on the right showing which features were programmed into the beta version of the software.

Feature requested	Notes	Included in beta version
Repository usage:		
Total downloads	Possible	Yes
Average downloads over a period	Possible	Yes
Origin of accesses:		
- by country	Possible	Yes
- by institution/domain type	Possible	Yes
- by department/research group	Extremely difficult	No
Download activity:		
- quarterly	Possible	No
- per month	Possible	Yes
- per week	Possible	Yes
- per day	Possible	Yes
Trends analysis:		
Usage of individual articles over time	Possible	Yes
Usage of articles by specific user (e.g. domain name) over time	Possible	Yes
Usage of articles by subject area	Difficult	Yes
Article usage:		
Views of metadata (hits)	Possible	No
Views of full-text (downloads)	Possible	Yes
Combined totals	Possible	No
Overall usage of article:		
Usage from repository	Possible	Yes
Usage from publisher site	COUNTER-compliancy	No
Usage from other sources	Extremely difficult	No
Aggregation of usage from different sources	Extremely difficult	No

Table 2: Stakeholder requirements and whether they were included in the beta version

In addition to those features deemed by stakeholders to be desirable, the software includes the following bespoke reporting features:

- Search by specific dates
- Search terms used by searcher
- Referrer (e.g. web search engine, repository home page)

Key Perspectives

4.2 IRS scorecard

Table 3 below gives an overall indication of the user response to the IRS product, based on responses from the repository manager test panel. This is just a generalised-score exercise based on our understanding of user response rather than a true empirical assessment, but it usefully serves to summarise the overall level of reception in the user community for IRS. IRS is scored out of ten for each factor.

Factor	Response	Comments
Comparison with usage analysis package currently used	8	Compares very favourably with the DSpace tool and has all the features required as optimal by one team searching for a statistics package
Presentation and design	9	Generally agreed to be very good. Download Dashboard particularly valued. Graphical display could be improved
Number and type of usage measures covered	8	Excellent range of measures possible and bespoke generation particularly valued. Some user requirements not included. Some suggestions for more measures offered by the test community
Overall usefulness	9	The best there is currently

Table 3: overall scorecard for IRS

4.3 Conclusions and recommendations

Overall, the software has been well-received by the test community. It is clear that there is a need for good statistical analysis software for measuring repository usage and IRS has appeared at an opportune time. The DSpace package does not measure up and repository managers are writing their own packages or looking hard for off-the-peg ones with good functionality. IRS fits the bill to a very great extent. Its presentation is generally agreed to be very good, though there are some suggestions from the test community about small improvements. In terms of functionality, there is agreement that IRS provides a whole range of useful functions, well thought-through. We conclude that IRS has fulfilled its intention to provide a good repository usage statistics package for the community and that the community has found it most acceptable.

There are suggestions that the graphical presentation might be improved a little, and there are some suggestions from the test community about additional

Key Perspectives

functionality that might be added. We recommend the IRS team studies these suggestions with a view to incorporating the most appropriate of them in the next version of the software.

We also recommend that the issue of using commercial software as part of the package is examined. If there are open source alternatives, as the respondent suggests in section 3.7, then the availability of these should be made clear on the IRS website.