

3.2 ACCESSES

3.2.1 Origins of accesses

There was strong demand from researchers for data on **country** of access. This was seen as the minimum requirement as far as granularity is concerned. More than half of them were interested in having further information on who was accessing their articles. Identification to **institution-level** was generally thought to be desirable and informative because it would probably be enough to enable researchers to track down the individual(s) who was interested in their work. Some people would like to know more: apparently the Stanford Encyclopaedia of Philosophy identifies users down to departmental level and the repository manager (RM) who mentioned this claimed that this is a very useful feature. The question of whether this level of identification might inhibit usage because some users may require anonymity was raised in the RM focus group, and there was a suggestion that the software might incorporate a facility for users to click on a button to indicate they had read an article (i.e. were waiving their anonymity at institution level) but if they didn't click on this that download would remain unidentifiable at institution level. If identification to institution is deemed unwise, then it should be possible to identify and compare at least **domain types** such as ac.uk, gov, edu, edu.au, etc, though it was pointed out by one person that if there are only IP addresses in the logs this won't be easy.

It is very important that human and robot accesses are differentiated and counted separately. There should also be COUNTER-like exclusion of aborted accesses and double-clicks.

3.2.2 Other granularity issues

In general it was felt that the more granular the data the better, so, for example, per-day accesses would be helpful in addition to per-week and per-month counts. Counts for individual days are also useful for helping to understand spikes. Data for 'current month' are not helpful because the month is not complete. Tasmania gets round this by having a 'last four weeks' span as default and can provide any full month on a click.

ISI's *Science Indicators* feature has a quarterly datapoint so if the new service could provide a match for this it would be ideal. A half-life measure, showing usage decay from the time of deposit onwards, would be useful too.

The counts provided must be related to the number of items in the repository so that it will be possible to assess.

The ability to *sort* by institution (if it is indeed possible to identify by institution) would be a good feature, so that particular institutions of interest can be sorted in or out.

Accesses to articles from a particular department or research group could also be useful.

Standard measures of usage *for a repository*, such as average number of downloads per article, total number of downloads and so forth – such as those provided by any ISP for websites – would be useful to repository managers, both for analysing how well their own repository is working (and being able to compare it to other repositories) and so that they can use it for advocacy and education about the local repository.

On-campus (within the home institution) access data would be helpful for teachers tracking usage of their course materials and allied content.

Data on the most common search terms used could be useful (though this is probably more interesting than useful). Edinburgh already provides this statistic for its repository - it returns a count if the same word is used more than 20 times.

Finally, the number of unique users accessing is of interest since it clarifies the issue of multiple downloads by individual searchers.

3.2.3 Weighting

There was little enthusiasm for weighting overall, though almost everyone agreed that the home institution accesses should be filterable because these are seen as 'different' to accesses from elsewhere.

One researcher, who is about to move to another institution and expects to have his articles on both old and new institutional repositories, suggested that being able to see, from one place, the usage of articles that reside on more than one repository would be helpful.

3.3 TRENDS ANALYSIS

The idea of being able to see trends in article usage over time is very popular. Many suggestions were made, as follows:

- Usage of individual articles over time
- Usage of articles from a department or research group over time
- Usage of articles by subject area (for example, UCL assigns subject classifiers to all articles in its repository, relating articles to departments or research groups)
- Usage of articles in comparison to the usage of other articles in the same repository (cumulative)
- Usage by keyword
- An alerting facility that draws attention to unusual usage

The way trends data should be presented generated a considerable amount of discussion, which distilled down to:

- Raw data are essential
- Excel format would be nice, and possibly some other common formats such as SPSS, but these are not *essential* if the next suggestion holds
- A portable data format (e.g. .psv) would be workable as an alternative to providing data in any specific formats

- A graphical output is extremely desirable, for all the suggestions above. There was some discussion around the possibility of providing guidance (and warnings) on the use and interpretation of such graphs

3.4 'LEAGUE TABLES'

The value and desirability of a 'league table' approach to presenting usage statistics is under active discussion in all the repository sites consulted in this exercise. There are two schools of thought. The first is that introducing an element of competition (such as publicly displaying the number of downloads for any article) encourages deposition by authors and increases author interest in the repository. The other school of thought is that some authors are actively discouraged from contributing articles to a repository that displays easily-comparable usage data because their own articles may not be accessed much (for many reasons). There has also been talk in the repository community and within individual institutions - UCL is a prime example here - that 'league table' data might in time become embedded in the RAE exercise. (There is no evidence at all that this may happen but the possibility has arisen as a discussion point and is sounding something of an alarm in certain circles). In short, then, whilst league table approaches are considered to be appropriate in many instances, there is the reservation that they may be open to abuse.

Nonetheless, comparative data (in this case we are talking about easily comparable datasets) are an extremely useful tool for repository managers, particularly in their 'advocacy upwards' work. One example of this is where the repository manager at UCL used the fact that the most downloaded article one month was a humanities thesis and was able to use this fact to persuade the administration at UCL that *all* theses should be deposited as a matter of policy. In general advocacy within specific communities, too, such a league table approach can work very well: in Glasgow, a very highly-downloaded article for a while was a book by an author in the education department, a fact that was communicated to other authors in arts/humanities and that encouraged them to deposit books (note: the book referred to here was on sale all the while, sales did not decline and may have even increased as a result of deposition in the IR).

Overall, it is best to say that opinion on comparative data of this kind is that public usage of them is best left to the discretion of the local repository manager, but that the provision of them to those managers is highly desirable. They have a valuable role to play as an advocacy tool even if not used completely openly in the institution.

There is another variant possible here too, and that is the provision of **deposit** statistics as well as usage statistics. The repository managers suggested this as a means to monitor activity in their repository, to perhaps look at deposition by department or by research group or discipline, and to measure their own repository against others. Thus they were asking for the means to make a comparison of the success of their own local policies with those of other

repositories in the hope of learning more about best practice and policy enhancements.

3.5 POTENTIAL USES OF USAGE STATISTICS WITHIN INSTITUTIONS (BY WHOM, FOR WHAT, HOW, WHEN)

For all the repositories involved in this study usage statistics remain, by and large, the sole domain of the repository manager. As yet, they are not being used by senior administration in universities in any organised way. This seems likely to change: we heard that discussions with senior management are underway at Glasgow over how the repository might return statistics that could be used for measuring research output in various ways and at UCL there has been some reporting of usage data to senior management by the IRM as part of an effort to get an institutional mandate on thesis deposition introduced (successfully).

Changes may well occur over the course of this project and we will monitor the situation regularly and report back to the software development team any movement on this issue. Although it will not materially affect the type of usage statistics required by the community at large, interest and involvement of senior managers may well change relative importance of the different statistics we work on producing.

3.6 USAGE OF METADATA/FULLTEXT OF ARTICLES

There was complete concurrence on this topic: usage data for metadata accesses and for fulltext downloads should be provided, separately and combined. One RM pointed out, however, that the general assumptions underlying explanations of access behaviour (searcher looks at metadata, decides on level of interest, if high enough proceeds to fulltext) may not actually apply in all cases. For example, Google refers searchers directly to the fulltext pdf (at least in EPrints) and Google is becoming more and more represented as a fraction of total referrals. Nonetheless, the usage data for each type of access, and the combined count, are what are demanded by the user community, whatever the caveats about their interpretation.

3.7 USAGE OF ARTICLES FROM NON-IR SITES (i.e. PUBLISHER SITES)

Again, there was consensus on this. Everyone wants to see usage figures for non-IR sites too (and to have them combined with the IR-usage to give a third, total, figure). There was also the suggestion from the repository managers that if it is possible to measure the click-through rates from the IR version to the publisher version those would be useful data.

3.8 OTHER ISSUES

3.8.1 Language-related issues

This topic was not readily applicable to many of the people involved in discussions for this project, but two of the focus groups provided the

opportunity to discuss it in some depth: one had a Dutch representative amongst the researchers and in the Ireland focus group we were able to discuss the issue particularly as it applies to publishing in Irish, which is something that happens fairly commonly in some subject disciplines.

The outcome of these discussions was that it was thought a good idea if articles in a language other than English could be deposited initially in the native language (even just the metadata) and usage data returned for these. The English translations of various fields could be added subsequently. Clearly, there will be many cases where both the native language version and an English version may be deposited and in these cases usage data for each and for the combined total should be provided.

It was suggested that a language tag could be useful (e.g. EN) in IR software.

3.8.2 Data release/privacy

This was a topic that most people felt unable to discuss, simply because it was not something that had arisen or that they felt was applicable to them. At Glasgow, however, we learned that it is something under a great deal of discussion at the moment. – how data from the repository can be used for measuring the University's research output and how the RAE might fit in here. The feeling at this institution was that permission to release the usage data should reside with the institution or individual departments.

3.8.3 Registration to use the service (desirability thereof)

The researchers said they had no particular problems with the idea of registration being required for the service so long as it was simple. The repository managers on the whole felt the same and agreed that it would be possible to authenticate users as members of their own institution. One RM felt it would not work in practice. For example, there was much discussion about instances where legitimate users might wish to access from home or from another IP address.

There was also some discussion (inconclusive) around the fact that if there were to be some kind of access-permission setting (such as, say, allowing only the author of an article see certain types of usage figures) this would be confounded in cases of mediated deposition. This is an issue that needs thinking about and the simplest way out of it is to drop the idea of registration altogether. If it is deemed to be very desirable, though, there are these issues mentioned here, at least, to take into account, and similar examples of the complexities that might arise are easy to think of.

3.8.4 Cookies and tracking

There were no specific objections to the use of cookies except from NUIM, where the university disables them. They were not considered desirable, though.

3.8.5 Versioning issues

As with the metadata/fulltext, the most useful way to collect and present statistics where multiple versions of an article exist is to provide usage data

for each version and the combined figure. None of the RMs interviewed do any version-linking for their own repositories and they see the usage figures for each version of an article.

3.9 GATEWAYS AND THEIR REQUIREMENTS

The (ARROW) gateway manager consulted reported much the same set of requirements as the IRMs did, but in addition commented that many universities belong to consortia and are starting to work on a consortial basis with respect to IR usage and monitoring. ARROW, in turn, monitors these, so from its point of view aggregated statistics by consortium would be desirable so that consortial performance can be compared and evaluated.

4. OTHER USAGE STATISTICS PROJECTS AND STUDIES

There is considerable activity building in the area of repository and website usage measures. These seem to be the main studies/activities underway at the moment:

1. Philip Mayr at the University of Bonn has developed measures around website entries (accesses). This project is specifically concerned with websites as opposed to repositories but approaches the issues of search engine queries, backlinks and direct navigation quite interestingly.

Website entries from a web log file perspective – a new log file measure

http://cybermetrics.wlv.ac.uk/AoIRASIST/Mayr_full.htm

2. The current Los Alamos initiative (led by Herbert van de Sompel) Herbert is one of the people listed for consultation during the course of this project, so the latest developments in his work will no doubt be up for discussion then. For now, his latest reporting on the work was made at the recent OAI4 meeting at CERN in Geneva in October and the presentation slides can be found at:

Johann Bollen and Herbert van de Sompel:

A Framework for Assessing the Impact of Units of Scholarly Communication based on OAI-PMH harvesting of usage information

<http://indico.cern.ch/getFile.py/access?contribId=48&sessionId=1&resId=0&materialId=slides&confId=0514>

3. Standardized Usage Statistics Harvesting Initiative (SUSHI)

Developed at Cornell University, SUSHI is a large-scale project managed under the direction of NISO and is developing a standard model for machine-to-machine automation of statistics harvesting. It has multiple collaborating partners, both service providers and data providers. The content providers are two agents (SWETS and EBSCO), ISI and Project Euclid. It focuses on library-level usage of online content.

<http://www.library.cornell.edu/cts/elicestudy/ermi2/sushi/>

4. There is a European initiative for which a workshop is to be held in February in Berlin. An invitation to the team for this present project has been extended and a representative from Southampton should attend.:

Workshop on institutional repositories and enhanced and alternative metrics of publication impact, 20 / 21 February 2006

http://www.esz.hu-berlin.de/index_html-en?set_language=en&cl=en

5. There is also some parallel development going on by Google, focused on website accesses. Details can be found about **Google Analytics** at

<http://www.google.com/analytics/>

5. Appendix: Focus group reports

Below are the notes from the discussions.

IRM = institutional repository manager

R = researcher

Issues addressed in each session/interview:

Issue	Researchers	IR managers	Gateway managers
Origin of accesses	✓	✓	
Other granularity issues	✓	✓	
Weighting according to institution accessing	✓	✓	
Multiple downloads from the same institution	✓	✓	
Cookie-tracking?	✓	✓	
Metadata vs full-text	✓	✓	
Versioning issues	✓	✓	
Trends analysis	✓	✓	
Aggregation of journal and IR usage data	✓	✓	
Value of league tables	✓	✓	
Registration desirable? If so, institutional or central?		✓	
Data release issues: how much does privacy matter to an institution?		✓	
Language issues	✓	✓	
Gateways: what do they require?			✓
Research committee/institution-level usage of data		✓	

Origin of accesses

Glasgow: Country definitely very important (R, IRM). Extension (.com, .ac.uk, etc) helpful (Glasgow already collects this). IP address (or domain name) if possible, so that institution can be identified (R): researchers would follow this up if from an interesting country to see who it is – to check out of interest but also to make sure they were not going to be caught out by a ‘scoop’ (fast developments in unexpected institutions/countries). Suggestion (R): could individual researchers manipulate these data to enable them to keep a particular eye on favourite or specific places/institutions? Other comments: need to have a way of reporting on ‘human’ hits – i.e. filtering out Google bots, robots, etc. Want to know when a hit is due to a human or when a spider, crawler and so on.

NUIM: Apparently the Stanford Encyclopaedia of Philosophy identifies users down to departmental level and this is very useful (IRM). Upon pressing, it was clear that there is an interesting-versus-useful dichotomy here: the researchers would find this level of granularity interesting and the IRMs would find it useful. Nevertheless, the researchers agreed that the feedback of usage data like these would be an incentive to deposit.

Oxford: Knowing as much as possible about who is accessing work in a repository is useful (some discussion here of plagiarism fears). Identification to departmental level would be enough. This is sufficient to be able to work out who is interested in their work and follow up if need be. One delegate said he’s deposited his work because ‘it’s easier than keep emailing the file’ and is not interested at all in who is accessing it.

UCL: Varying opinions on access data: some would be happy with institution identification only; others wanted it down to department or research group. One of the delegates currently sends her research students to the UCL IR to look at the research activity of the large research group she works in. Would like to know whether an access is by a human or a robot (IRM)

London LEAP: Institution is more useful than country: higher education accesses by institution is the optimum. Access from one’s own IP address could be a bit misleading and it would be good to be able to filter those out of the statistics. Identification down to department level might actually inhibit people from depositing. One suggestion was for identification to be optional, i.e. the reader could click on a button to say they have read/downloaded the article, thus having the option to retain anonymity if required. Some institutions would find it very useful to be able to specifically identify government departments or other similar bodies accessing the content of their repository: this may indicate commercial opportunities that the institution could follow up. Very important to be able to distinguish between robots and humans accessing. Useful to be able to identify and compare ‘ac.uk’, ‘edu’ and ‘.gov’ accesses. The service should define a list of search engines and OAI service providers referring enquirers, and be able to distinguish who types in the URL directly and who comes in via other routes. There was some discussion of the issue of anonymity and its importance, or potential importance in the sense of possibly deterring people from using if not respected: one idea was to provide a guide on ‘How not to use these statistics’ to help protect anonymity.

Tasmania: The problem here is that no-one much has experience with data that (a) goes down to an institution level, and (b) is not very tightly subject

specific. Two important questions need to be asked about such fine-grain tracking: “Why do you want this data?” and “What will you do with it when you’ve got it? How will it affect your behaviour?” There is no value in institutional download data except for very high-volume papers. No question that robots should be excluded, and also COUNTER-like exclusion of double-clicks and aborted accesses as well, etc. Researchers not expected to know about these.

Other access/granularity issues

Glasgow: Should be able to relate the number of papers in a repository to the number of hits; in other words, is increased download activity real in the sense of more hits per article, or is it a function of number of articles?

NUIM: Data on when accesses occur: There should be data on usage in the last month, last year and last quarter – the latter discussed quite extensively and declared not particularly useful for researchers but very useful for IRMs studying usage and compiling reports for management. The ISI’s Science Indicators feature has a quarterly data point, so the new service would match up with this which would be useful (IRM). A ‘half-life’ feature would be useful too – a measure, like the SCI’s ‘cited half life’ capability, of usage decay from the point of deposit over time.

Oxford: Would like to see usage on a weekly/monthly/annual basis, with severe spikes shown so that they can mentally ignore any ‘skewing attempts’. Would like data filterable by domain (e.g. ac.uk or ox.ac.uk) (R). EPrints should have a subject classification, so that it would be possible to see usage by subject (IRM).

UCL: Date of accesses; accesses in the last month and year (R). From an institutional point of view, would need only comparable, overall, statistics (for, say, Russell Group institutions (IRM)).

London LEAP: Should be as granular as possible; usage reported per day is very attractive, because peaks can be identified and explained. Accesses not only to articles of one individual but to those of a research group or department would be very useful. A department might be interested in, say, the accesses to papers from the whole department from Taiwan.

Tasmania: Filtering by domain should be relatively easy to do if you have domain names, but not if you’ve got IP addresses in the logs. Note that data for the current month is not useful because incomplete and not even roughly comparable with a full month. We address this by having a “last 4 weeks” default, plus the summary has a month histogram which can provide any month on a click.

Metadata versus full-text

Glasgow: Download counts for metadata and full-text essential (R). This is already important for Glasgow in the sense that the repository only has about 20% of its articles showing FT, and current statistics are only applied to the FT downloads, but if these could be compared to those for abstracts there may be a completely different picture (IRM). When articles have only metadata available, enquirers write for a copy of the whole article, which is not

made available because of (a) copyright restrictions (b) because at the preprint stage there is a fear of being ‘scooped’ and (c) this method allows the author to keep an eye on who is interested in an article (N.B. what the stats will show).

NUIM: There was considerable discussion of the issue of researchers opting only to deposit metadata, and of the issue of enquirers then having to email for an article. On the topic of usage feedback, all agreed that there should be data for both metadata and full-text accesses.

Oxford: Usage data for both metadata and full-text downloads.

UCL: It is very important to be able to distinguish usage of metadata and full-text (R).

London LEAP: The statistics should be reported both separately and also combined. Knowing the metadata usage alone is useful and if users have gone on to the full-text that tells more. It also helps to be able to separate out the metadata-only users so that they don’t confound the total figure. Also, if the metadata are accessed a lot, there may be a reason for this (i.e. mistaken accesses).

Tasmania: Both measures have their problems: Metadata accesses are recorded only if the data comes in via the IR search, via a persistent URL link, or a gateway that redirects to the repository metadata page. Not recorded for gateways that link straight to full-text, aggregators that harvest both metadata and full-text, and Googlebot which indexes pdf files. Thus metadata is a slim reed for any judgment as Google searches become a bigger and bigger fraction of all searches. Full-text accesses are presumed to be evidence of intent to read the document, but not so. (Only Googlebot downloads pdf files, so assume these are filtered out.) However, accesses via Google go straight to the pdf at least in EPrints, and so are more like an exploratory metadata access for a searcher with good bandwidth. Perhaps accesses straight to the pdf and not preceded by a metadata access in the last five minutes should be discounted 80%? (tongue-in-cheek). Then a proxy-server gets in to mess this all up, and in Australia’s case it may be the local UTas one, or the Australian AARNet gateway one. Accesses simply may not be recorded.

Weighting according to institution accessing

Glasgow: No requirement for this but interest in the accessing institution and what this means to individual researchers is high.

NUIM: Weighting: could there be some aggregation of usage statistics in a way that reflects the accessing institution’s prestige (R)? During discussion of this point it was agreed that this is fraught with subjectivity issues and that one of the most basic problems is that one institution may be considered extremely prestigious in one discipline yet not so meritorious in another. In the end they agreed that researchers had to make this value judgment themselves using the original, unweighted usage statistics. It was suggested, though, that on-campus (within-institution) access statistics would be very useful for teachers following up the usage of their online teaching materials.

Oxford: Not weighting, but an ability to pull out accesses by the home institution (and perhaps other allied ones) would be useful.

UCL: It would be good to use some sort of filter to pull out the ‘parochial’ aspect, i.e. more detailed statistics for UK usage than from abroad. Being able to see downloads from different university domains would be ideal too.

London LEAP: Important to be able to *sort* the data in various ways, so that individual institutions could be sorted out or in. Or produce this graphically so that one line on the graph represents own-institution accesses.

Tasmania: No interest in institutional weighting, apart from the home institution.

Multiple downloads from the same institution

Glasgow: No specific comments.

NUIM: Could it be shown whether the same people keep returning to look at a document again (R)? We discussed the fact that this would need identification of individual users, something that is unlikely: they agreed that identification of users down to departmental level would be ideal. Then if repeated visits from one department were seen they would follow up and try to find out who it might be.

Oxford: No specific comments.

UCL: No specific comments.

Tasmania: Ask why you are interested in tracing down who is responsible. Let’s take a real case. For example, I have downloaded a paper “Aligning Alignments Exactly” several times in the last few months. Twice on campus, because I lost the first copy and wanted to take away the paper to read carefully on the plane and in the slack moments of a conference I was attending. Once at home (not in uni domain) because it was simpler than carrying it home on a memory stick. My PhD student and the other supervisor no doubt each downloaded it at least once each. Next year’s Honours student will download it as he works on my project parallellizing the AAE algorithm, and so may others do who are ranking the project in their list they’d like to do. Now, what do you conclude seeing all this traffic, but not the reasons? Remember you don’t know who accessed it, especially behind the proxy server (IRM).

Cookie-tracking?

Glasgow: No specific comments.

NUIM: Not considered desirable. In fact the university specifically disables this (IRM).

Oxford: The University permits cookie-tracking. No delegates have any objections.

UCL: No specific comments.

London LEAP: Don’t use cookies unless you have to.

Tasmania: Cookies may be useful for determining “sessions” and handling “double-clicks, “aborts”, etc. Otherwise, forget it. An IR does not want to track an individual.

Versioning issues

Glasgow: No specific comments.

NUIM: There should be data for the usage of each and every version of an article, if present in more than one version. It is useful to be able to see if early versions are being used once a later or final version is available. Cumulative data?

Oxford: No specific comments.

UCL: No specific comments.

London LEAP: Overall, it would be useful to have usage data for each version. Some articles/theses come as several separate chapters and these would each have their own usage instances. These repository managers do not do any version-linking themselves.

Tasmania: I currently see stats for each version. Aggregating would be preferable, but even that has its pitfalls. Suppose someone sees my preprint, goes to the metadata, and sees that there is a later version available. They go straight to the newer full-text. This looks like Googling to the analyser which doesn't know about the relationship. There is no simple answer.

Trends analysis

Glasgow: Definitely desirable to be able to track usage over time. Example given of peaks and troughs in a research field which need monitoring to see where the subject is going (R), i.e. would use IR to monitor the field as well as specific authors' contributions.

NUIM: Would like to have trends data by: articles from departments (IRM), articles from individuals (IRM), institution (IRM) and nation or larger (IRM). The researchers are only really interested in usage data over time for their own articles. The format of these was discussed: ideally, the basic data should be exported in a portable format (e.g. comma-delimited) so that they can be used in a variety of software types (R), and there should also be a graphical representation as well (IRM, R).

Oxford: Would like to be able to see trends for individual articles and for articles from a research group (i.e. some sort of user-defined grouping of authors) and department.

UCL: Would like to focus on a single article and give its usage context in the form of a comparison to the usage/growth of others in the same repository. Again, would like to be able to look at usage trends by subject area (UCL's subject classifiers relate to department, research group, etc). One idea suggested was that the service should be able to answer the question 'How does my paper relate to others?' with a facility like Amazon offers, where it gives 'people who have bought this book also bought...'. A graphical interface is ideal and perhaps there could be a three-dimensional view with usage shown as peaks of different heights. Also, just raw statistics would be useful – plain download stats over the longterm. An alerting facility when something out of the ordinary happens would be good. Ability to plot the usage of an article against the usage of similar articles (cumulative). Downloads by keywords.

London LEAP: Definitely want the raw dataset plus Excel format and perhaps something like SPSS. Graphical representation would be very nice, and a health warning about interpretation of the graphs on a separate 'crib' sheet would be advisable. As well as graphs, a portable data format such as

PSV would be useful. Repository managers may well collect the data using this and devise demo presentations for their users. It would also be in the spirit of the thing if the statistics show which repositories are working well or not, and this would also be good to use with end users. The managers would use these data for advocacy s they want data at the top end for sensitivity. Standard measures would be useful too so that managers can plug in a 'Statistics on this repository' service. Other suggestions: average downloads per article, number of downloads, etc, such as an ISP provides as standard for any website. Dynamic report on weblogs – N'ham's School of Modern Languages has foregone any royalties it would get on selling conference proceedings papers because it prefers to get the usage and dissemination. **Tasmania:** Trend analysis over disciplines, or all an author's works, or the repository may be useful. Otherwise leave the trend analysis by document to the author. We provide a monthly access breakdown at UTas, and it would be minor work to package this up into a .csv file. We've done it before for other purposes (IRM).

Aggregation of journal and IR usage data

Glasgow: Definitely required, but would want to see the contributions of the individual components as well as the whole (R, IRM).

NUIM: Yes please! The library is setting up an SFX interface (apologies if the terminology is wrong there) so that the users at NUIM search all the library's databases (publisher sites, aggregator sites, etc) *and* the IR simultaneously (IRM).

Oxford: Yes, combined data would be very useful (R, IRM).

UCL: Combined data desirable but want the separate elements to be available too, i.e. make sure the extra (journal) data are not lost in the whole (R, IRM). These data are unlikely to be useful for cancellation policies (IRM – who is also the science librarian!)

London LEAP: Aggregation of these is ideal but how easy is it to do? It would be nice to measure the click-through from the repository version to the journal version.

Tasmania: Yes please, but can we get data down to the article?

Value of 'league tables'

Glasgow: Under discussion. IRM collects usage statistics (hits) and these are fed back to individual researchers privately to inform them of the results of having articles in the repository, and anecdotally it is reported that these act as an incentive for researchers (IRM). Researchers concur with this (R). No plans to produce publicly-available 'league tables' as there is debate on the desirability of this.

NUIM: Not explored.

Oxford: Not a good idea on public pages but OK to be able to produce them privately. There would be an issue of divisiveness if they appeared in public (R). Thought to be a very useful feature (IRM). Being able to have a measure of citations would be much more useful (R).

UCL: These could be open to abuse, and the reasons given were: There is the issue of some articles being in the IR longer and so likely to get more downloads – this needs to be controlled for. Some articles are in very popular journals and will be downloaded a lot: this does not necessarily reflect a greater merit than other articles that are not downloaded so much. It was generally considered that such tables might somehow be used for the RAE, which would be a bad thing; conversely, they could be used to attract users to the IR which would be very good. Would it be possible to produce league tables by discipline? (R).

The IRM had something to say on this: he thought it might be divisive to use this tool as part of the public face of an IR, but acknowledged that from an IRM's point of view it would be extremely useful. He suggested such data might be made available to the IRM for each repository and the 'public on/off switch' could be left to those individuals to decide what is appropriate for their own local needs. He himself has used the UCL weblog figures politically, in that he was able to show that the most downloaded article one month was an arts/humanities thesis, and used this to persuade the 'powers-that-be' that a mandate should be in place on depositing all UCL theses. Suggested that there might be provision of author-confidential league-table data but that there would be 'shocking administrative overheads' associated with this to do with access control.

London LEAP: Maybe it would be best, from the point of view of authors, if only the *top* of the table were shown (so as not to discourage those at the bottom!) but repository managers would like to have the whole thing and perhaps publish openly the 'top ten'. No automatic 'naming and shaming'. The usefulness of these statistics is borderline for this project and they are best left at repository manager level rather than made open further down to users. They are open to abuse. Managers may report back to departments individually, rather than using whole-repository statistics, but this is a local decision. Deposit statistics would be useful as well as hit statistics. It is currently possible to search by department in EPrints but not to get statistics by all departments all at once. This would be a very useful enhancement.

Tasmania: There are several issues conflated here. A public author ranking feature would definitely be counterproductive in Australia, except to the top 10+ authors. Such ranked lists are often stable with time, and affected by no of papers and discipline (eg Chem cv lists tend to be 3x longer than CS lists). I can see that IRMs would find such data useful politically, but maintain that they are generally a bad idea. One of the principal reasons why we rejected using University of Queensland stats package and wrote our own. A short-term ranking of articles, or even a long-term one, seems to be politically inoffensive. It is the identification of a person and the aggregation that creates the problem. Ranking within a classification (eg LC #) might be even more acceptable. We also need to be aware that the issues are confused by the types of readers on the Internet. High downloads may be associated with popular (ie public) digestible content, rather than scholarly interest.

Registration desirable? If so, institutional or central?

Glasgow: No specific comments.

NUIM: The issue of registration was not warmly welcomed by the researchers though they would tolerate it if it were simple, such as only requiring their email address as the username. Ideally, if registration to the new service is going to be required, institutions should be able to register all valid users en masse in the way they do for other services.

Oxford: No objections to registration if it is simple (like being able to use your email address)

UCL: No major objections to registration but it *is* another barrier and may put some users off (R). It could technically be linked to the UCL authentication database but there would be problems with this.

London LEAP: Authenticating users as members of an institution is easy but they can't be identified after that. Shibboleth would be better than EPrints' own system here. The problem is letting users access their *own* papers' data because of mediated deposit – it would be difficult to identify the actual author(-as-user) if someone else deposited the article. If some usage statistics should be restricted to author only, it is the responsibility of the repository manager to work out a solution. It could be a *push* rather than a pull solution (to a secure location). Repository managers require a protected level of report which they decide locally how to enable.

Tasmania: Unacceptable. It wouldn't work.

Data release issues: how much does privacy matter to an institution?

Glasgow: Under intense discussion at senior management level. Librarian is to make a presentation to VC/PVCs at the end of September on this – how data can be used for measuring research output, for RAE, etc. (L).

Permissions for release of data should be preferably set by each institution or department. Develop software to allow complete feedback, but leave the permissions to local level (R, IRM).

Language issues

NUIM: (Some researchers at this institution publish in Irish, though not any of those attending the discussion session). The option to be able to deposit initially in the native language – even just metadata - and then later add English translations of fields (perhaps incrementally) was thought to be a very good idea.

UCL: a language tag would be useful (e.g. EN). There are difficulties with special characters when depositing (Google ignores all of these). An English abstract, at least, is required but if an article is in another language then both language version should be deposited and usage data for both would be needed. If special characters are possible in IRs then some sort of fuzzy searching should be used to deal with this.

Tasmania: Difficult to answer from a continent in which English is the sole official language. However, most of our neighbours don't write in English, so it is a problem we're interested in.

Gateways: what do they require?

A means of monitoring how different consortia of universities compare and contrast with one another.

Research committee/institution-level usage of data

Glasgow: Under discussion at senior management level.

Other suggestions/comments and additional information

NUIM: Could there be the facility for commentary to be added to articles in IRs? (*APS: This suggestion mirrors the 'Trackback' feature just added to arXiv, I think*).

Glasgow University:

IRM collecting download statistics – overall numbers, where they are from in the sense of coming in via Google, Scirus, OAI, etc. But they are cumulative and not giving snapshots at time intervals.

Most of the content in the IR is from STM departments. This is because they are 'more engaged in getting exposure to their work' and it is easier to work with them at this stage. Content is rarely deposited by individual authors; the library staff are collecting it in batches from departments (metadata) because each department has a research record database in ReferenceManager, Endnote or some bespoke format (e.g. computer science). The Eprints team then enters the metadata into Eprints and adds the full-text where permitted by the author. A feasibility study has just been completed on a central publications database for the whole university. EPrints is considered robust enough to cope with this, and departmental databases are expected to be fed into it, filtered and processed and regular extractions to be made to feed into the EPrints-based IR.

On deposition: 'There is a non-trivial amount of work involved in producing an author postprint version' (R): this researcher, a computer scientist, composes in LaTeX and puts a lot of effort into making his postprint look *unlike* the publisher's version. To do this he concentrates on the author/affiliation list, where the publisher's version is very characteristic of and standardised for that publisher. This can take up to an hour per paper.

There was a little discussion of how the collected statistics might be interpreted and extrapolated for users.

Books (and, soon, book chapters) are also being added to the IR. Example: Julia Preece's book on education in Botswana; has had hundreds of downloads, to the author's delight (even though the book is for sale). The inclusion of books is encouraging the arts faculty to view the IR favourably.

NUIM: Quite soon it is hoped that every institution in Ireland will have its own IR. There is work going on now to develop an all-Ireland IR search service.

Tasmania: The source of enquiries via search engines would be extremely valuable to the IRM. I get some of this from *awstats*, but it is not aggregated nor longitudinal. People who use LaTeX can be a pain. Sometimes the pdf they produce is unreadable by anyone (a font and macro problem I believe). Fortunately they are a minority, mostly in Computer science and a few science disciplines, and if pressed can solve the problems themselves (assuming that a mandatory policy is in place). Other than LaTeX people the postprint is simple. The best solution is for the author to post whatever they have (Word, OpenOffice, pdf, ps, etc, and let the Library/IRM deal with conversion to pdf. More could be done to promote mounting a sample book chapter and index as in Amazon (but the metadata for the whole book). Fruitful area for IRM work.

Participants in focus groups or individual interviews

Glasgow:

William Nixon (Glasgow EPrints manager)
 Joan Keenan (administrative assistant)
 Lesley Drysdale (technical issues)
 David Manlove (computer scientist)

NUIM:

Suzanne Redmond-Maloco (Chief Librarian and IRM)
 Agnes Neligan (librarian)
 Alan Rogers (engineer)
 Bill Lannigan (experimental physicist)

Oxford:

Johanneke Systema (librarian and IRM)
 Dave Waters (Earth Sciences)
 John Baines (Oriental Studies: Egyptology)
 Keith Gillow (Mathematics)

UCL:

Martin Moyle (Science Librarian and IRM)
 Sanjay Rana (Centre for Transport Studies)
 Ulrich Tiedau (Dutch)
 David Boniface (Epidemiology & Public Health)
 Laura Vaughan (Architecture/Urban Transformations[!])

London LEAP consortium:

Martin Moyle (UCL)
 Sally Rumsey (LSE)
 Russell Burke (KCL)
 Barbara Cumbers (Birkbeck)
 Colin Rennie (SOAS)

University of Tasmania:

Arthur Sale (IRM)
 Peter Vamplew (Computer science)
 Vishv Malhotra (Computer science)

Others:

Elisavet Chantavaridou (University of Macedonia, Greece)
 Paula Callan (QUT)
 Marie Suhre (Idaho National Laboratory, USA)
 Theo Andrew (Edinburgh University)
 Scott Yeadon (Australian National University)
 Philipp Mayr (University of Bonn)
 Debbie Campbell (ARROW Gateway)