PIRUS – Publisher and Institutional Repository Usage Statistics

a PALS 3 project

Developing a global standard to enable the recording, reporting and consolidation of online usage statistics for individual journal articles hosted by institutional repositories, publishers and other entities
(Publisher Metadata and Interoperability Projects 3)

Final Report

January 2009

Project Team:

Tim Brody, University of Southampton
Richard Gedye, Oxford University Press
Ross MacIntyre, University of Manchester
Paul Needham, Cranfield University
Ed Pentz, CrossRef
Sally Rumsey, University of Oxford
Peter Shepherd, COUNTER (Project Manager)

Table of Contents

1. Acknowledgements

This project has been funded by JISC under the auspices of the Publisher Metadata and Interoperability Projects 3 programme and we are most grateful for this support. In particular we would like to express our thanks to Alastair Dunning for his guidance on the project. It is also important to acknowledge the resources devoted to this project by Cranfield University, Southampton University and BioMed Central in setting up and implementing the required tests, as well as the co-operation of the publishers and repositories who provided feedback on the different scenarios.

2. Executive Summary

The aim of PIRUS (**P**ublisher and **I**nstitutional **R**epository **U**sage **S**tatistics) was to develop COUNTER-compliant standards and usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, repository, etc.,) that hosts online journal articles and will enable the usage of research outputs to be recorded, reported and consolidated at a global level in a standard way.

The core objectives did not change as PIRUS progressed, but it became apparent in the course of the surveys and desk research carried out in Phase 1 of the project that some of the proposed approaches would have to be modified to take into account the enormous variety of repository technical systems, organizations, management and content. One size does not fit all and more than one approach will have to be offered to repositories if they are to produce COUNTER compliant usage statistics at the individual article level. Different scenarios were developed that will, the project team believe, allow the majority of repositories, as well as publishers, to provide COUNTER-compliant usage statistics at the individual article level.

The four main outputs of the project are:

a. A proof-of-concept COUNTER-compliant XML prototype for an individual article usage report, Article Report 1: Number of successful full-text article downloads, that can be used by both repositories and publishers. In principle this report could be provided for individual authors and for institutions. In practice, the individual author reports are much easier to generate and are a realistic short-term objective, while the reports for institutions and other entities, such as funding agencies, will be more complex and should be regarded as a longer term objective.
b. A tracker code, to be implemented by repositories, that sends a message **either** to an external party that is responsible for creating and consolidating the usage statistics and for forwarding them to the relevant publisher for consolidation **or** to the local repository server.
c. A range of Scenarios for the creation, recording and consolidation of individual article usage statistics that will cover the majority of current repository installations . Each repository may select the scenario that corresponds to their technology and implementation.
d. Specifying criteria for a central facility that will create the usage statistics where required (for some categories of repository) and collect and consolidate the usage statistics for others

The recommendations of the project team are as follows:

a. To JISC: PIRUS has demonstrated that it is *technically* feasible to create, record and consolidate usage statistics for individual articles using data from repositories and publishers. If this is to be translated into a new, implementable COUNTER standard and protocol, further research and development will be required, specifically in the following areas:
    - Technical: further tests, with a wider range of repositories and a larger volume of data, will be required to ensure that the proposed protocols and tracker codes are scalable/extensible, work in the major repository environments, and can be applied to items other than articles.
    - Organizational: the nature and mission of the proposed central clearing house/houses has to be developed, and candidate organizations identified and tested
    - Economic: assess the costs for repositories and publishers of generating the required usage reports, as well as the costs of any central clearing house/houses; investigate how these costs could be allocated between stakeholders
    - Political: the broad support of all the major stakeholder groups (repositories, publishers, authors) will be required. Subject repositories, such as PubMed Central, which have not been active participants at this stage in the project, will have to be brought on board. Intellectual property, privacy and financial issues will have to be addressed.

    The PIRUS project team recommends that JISC considers funding further research to address the issues described above.

b. To COUNTER: expand the mission of COUNTER to include usage statistics from repositories; consider implementing the new Article Report 1 as an optional additional report; modify the existing independent COUNTER audit to cover new reports and processes.
c. To repositories: subject repositories to participate in the next stage of this project. All repositories should use standard data descriptions for article versions etc.
d. To publishers/vendors: accept, in principle, the desirability of providing credible usage statistics at the individual article level.

In conclusion, PIRUS has shown that it is feasible for repositories and publishers to adhere to common technical standards for measuring online usage, despite the diversity of organizational and technical environments in which they operate **but** also that further work will be required to translate the results of this feasibility study into practical, implementable, scalable solutions.

3. Background

COUNTER was initiated to improve the reliability of usage statistics available for online publications. It has done so by developing Codes of Practice that set standards for the recording, reporting and delivery of vendor usage statistics. Release 1 of the COUNTER Code of Practice for Journals and Databases was published in January 2003. Release 2, published in March 2005, has since been widely adopted by vendors and the resulting usage reports are widely used by librarians. COUNTER's reach was further extended in 2006 with the publication of a Code of Practice covering online books and reference works. COUNTER compliant usage statistics are now available from over 100 vendors for over 15,000 online journals, as well as for a growing number of online books and reference works. Release 3 of the Code of Practice for Journals and Databases is now available on the COUNTER website at http://www.projectcounter.org/code_practice.html, and must be implemented by vendors before 31 August 2009. This Release contains a number of refinements, including a requirement for usage reports in XML format and the implementation of the SUSHI protocol.

Until now the most granular level at which COUNTER requires reporting of usage is at the individual journal level. Demand for usage statistics at the individual article level from users has hitherto been low. This, combined with the unwieldiness of usage reports in an Excel environment, has meant that COUNTER has, until now, given a low priority to usage reports at the individual article level. A number of recent developments have, however, meant that it would now be appropriate to give a higher priority to developing a COUNTER standard for the recording and reporting of usage statistics at the individual article level. Most important among these developments are:

- Growth in the number of journal articles hosted by institutional and other repositories, for which no widely accepted standards for usage statistics have been developed
- A Usage Statistics Review, sponsored by JISC under its Digital Repositories programme 2007-8, which, following a workshop in Berlin in July 2008, proposed an approach to providing item-level usage statistics for electronic documents held in digital repositories (1)
- Emergence of online usage as an alternative, accepted measure of article and journal value and usage-based metrics being considered as a tool to be used in the UK Research Excellence Framework (2) and elsewhere.
- Authors and funding agencies are increasingly interested in a reliable, global overview of usage of individual articles
- Implementation by COUNTER of XML-based usage reports makes more granular reporting of usage a practical proposition
- Implementation by COUNTER of the SUSHI (3) protocol facilitates the automated consolidation of usage data from different sources.

COUNTER is in a position to bring together the relevant experts from publishers, repositories, repository systems suppliers, authors and libraries to build on the existing COUNTER standards to develop workable, widely accepted, new standards to govern the recording, reporting and consolidation of online usage statistics for individual articles hosted at a number of different locations.

4. Aims and Objectives

The aim of this project was to develop COUNTER-compliant standards and usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, repository, etc.,) that hosts online full-text journal articles and will enable the usage of research outputs to be recorded, reported and consolidated in a standard way.

The core aim of the project did not change as the project progressed, but it became apparent in the course of the surveys and desk research carried out in Phase 1 that some of the specific objectives would have to be modified to take into account the enormous variety of repository technical systems, organizations, management and content. One size does not fit all and more than one approach will have to be offered to repositories if they are to produce COUNTER compliant usage statistics at the individual article level. Different scenarios were developed that will, the project team believe, allow the majority of repositories, as well as publishers, to provide COUNTER-compliant usage statistics at the individual article level. These scenarios are described in Diagram 1 in Section 6 below.

It was also agreed that the objective should not be to require publishers, repositories, etc. to produce a routine monthly report that lists every article published by them and records the number of times it is downloaded, which would result in vast amounts of data and unmanageably huge reports. Rather, the objective will be to enable the usage data for individual articles, or sets of individual articles, to be collated and made available as and when required. This will provide a much more practical approach.

5. Methodology

The project was divided into three phases, described below.

Phase 1 (August-September 2008): *Survey/Desk research to assess: current practice in the application of individual article identifiers and other metadata; how different versions of individual articles are identified, etc*. This built on work already done on article identifiers under the PALS 2 programme (http://www.jisc.ac.uk/whatwedo/programmes/pals2/counter.aspx). Two surveys were carried out; one by P Needham, which covered repositories; the other by P Shepherd, which covered publishers. This phase of the project was completed on schedule.

Phase 2 (September-November 2008)): *Develop draft usage reports and protocols for the recording and reporting of individual article usage, test this with publisher and repository usage data.* As the results of Phase 1 demonstrated that there is a very wide range of repository configurations, even when they use the same software, such as DSpace, it would be impractical to specify a single approach to the recording, collection and reporting of online usage statistics. For this reason it was decided to develop and test more than one approach, each of which would deliver valid, comparable usage statistics. As a result this phase of the project was not completed until December.

Phase 3 (December 2008): *Taking into account the results of the tests, propose a final format for COUNTER-compliant usage reports, together with supporting protocols, and submit this to COUNTER for approval as a new standard, to be adopted and maintained by COUNTER.* As Phase 2 was not completed until December, the final report and recommendations to COUNTER and other organizations was not completed until January 2009.

The Project Team met a total of 10 times in the course of the project, usually by conference call.

6. **Implementation.** The objectives of this project were:
    a. to understand how repositories and publishers currently identify, record and report online usage at the individual article level.
    b. to develop COUNTER-compliant standards and usage reports at the individual article level that can be implemented by any entity (publisher, aggregator, repository, etc.,) that hosts online journal articles and will enable the usage of research outputs to be recorded, reported and consolidated at a global level in a standard way.

PIRUS was implemented in 3 Phases, as described in Section 5, Methodology, above. The implementation of each Phase is described below.

Phase 1: *Survey/Desk research to assess: current practice in the application of individual article identifiers and other metadata; how different versions of individual articles are identified, etc*

Publisher Survey

The questionnaire was sent to 15 publishers/vendors/hosts. A total of 12 responses were received, either in writing or in telephone interviews. The organizations who responded are indicated by an asterisk in the list below.

The publishers/vendors included in the survey were: American Chemical Society***;** American Institute of Physics**;** Atypon***;** BioMed Central***;** EBSCO**;** Elsevier***;** Ingenta***;** Institute of Physics Publishing***;** Nature Publishing Group***;** OUP***;** Ovid***;** Sage***;** Springer***;** Taylor & Francis***;** Wiley Blackwell**.** These publishers/vendors are all currently COUNTER-compliant and were selected to ensure that the sample was representative of the industry in terms of scope, size and geographical location

See Appendix A for a complete list of questions and responses.

While the majority of the publishers who responded to the survey believe that it would be valuable, in principle, to report usage at the individual article level, one or two remain to be convinced that this will really be of benefit. More are concerned about the potentially large volumes of data involved, as well as the practicalities and costs involved in handling it. These concerns were based on the assumption that individual article level reports would mimic the existing COUNTER usage reports, which must be produced monthly, for each customer, for every journal to which they subscribe. The project team has acknowledged that not only would it be unduly burdensome to require publishers to provide such reports at the individual article level, but the volume of data involved would be too great to be managed easily by customers. It would be less burdensome for publishers, and more valuable for institutions and authors, for individual article usage data to be provided as and when required for authors and institutions. The objective should, therefore, be to ensure that publishers and repositories have the capability to generate and process individual article usage statistics for authors and their institutions that are comparable, credible and consistent with the COUNTER standard.

Most respondents apply a unique article identifier, which is then a permanent attribute of the article, and most, but not all, use the DOI. There is a diversity of practice in terms of how versions are tracked and identified.

While the DOI appears to be universally applied by journal publishers, albeit with some variations in practice, for aggregators its use is not so universal, as many of the full-text items they host are not journal articles. As far as journal articles are concerned, however, the DOI appears to be in universal use. The reporting of the usage of individual articles will require that a standard article identifier be set and, at least in the publishing environment, it would appear that the DOI is the strongest contender. Prescribing that it be used will also require that a protocol is specified for its implementation in the publishing process. This protocol will have to cover the following issues;

- article versions to which the DOI is applied
- the stage in the publishing process at which the DOI is applied

Institutional Repository (IR) Landscape

The situation with regard to IRs was investigated through a combination of desk research, discussions within the PIRUS team and beyond, and an email survey circulated via the JISC-REPOSITORIES list.

The first task was to gain a clear picture of software applications in use around the world.

Repository Software Applications

Although a few years old now, the Budapest Open Access Initiative (BOAI) report "A Guide to Institutional Repository Software v 3.0" (4) provides an excellent introduction to repository software applications. It details and compares nine softwares available under an Open Source licence, namely: Archimede, ARNO, CDSware, DSpace, Eprints, Fedora, i-Tor, MyCoRe and OPUS. It includes a useful System Feature & Functionality Table providing summary comparison of the nine applications.

In addition to these Open Source applications, there are also a number of proprietary systems available, including: Digital Commons (BePress) and Digitool (Ex Libris).

A table of the most common Software Applications (Appendix B) was compiled by combining information from the BOAI Guide to Institutional Repository Software, the Registry of Open Access Repositories (ROAR) and software specific documentation.

Collating information from this table shows, globally, four out of five (80%) of listed IRs are based on just five software applications:

DSpace (inc. Open Repository)	37.9%
Eprints	27.3%
Digital Commons	8.3%
OPUS	3.9%
DiVA	2.7%

Two thirds of all IRs appear to be based on just two applications: DSpace and Eprints.

It is worth noting, however, that – for some reason - Fedora repositories appear to be under represented in the ROAR listings.

IR Content Types

IRs typically contain mixed content types including (but not limited to) journal articles, conference papers, theses, working papers, technical reports, project reports, book chapters, presentations, datasets, images.

Therefore, in order to identify which items are articles - and how different versions of articles are identified, it is necessary to take a closer look at metadata usage within IRs.

Metadata

Most of the repository softwares support qualified Dublin Core (qDC) or hold metadata that corresponds to and can be mapped quite easily to qDC.

Metadata elements typically used when cataloguing articles in IRs include:

- Title
- Author(s)
- Abstract
- Journal title
- Volume(Number)
- Pages
- ISSN
- DOI
- Bibliographic citation
- Resource type
- Local identifier

All repositories include Title, Author and Resource type metadata. Desk research (Appendix C), supplemented by the email survey (Appendix D), confirms that many repositories do add citations identifying the published versions of articles in their records.

Looking at the two most popular softwares in more detail:

DSpace

The ‘Title’ field is a free-text entry field.

The ‘Author’ field is composed of two free-text entry sub-fields ‘Last name’ and ‘First name(s)’

The ‘Citation’ field is a free-text entry field. Lack of control means that the contents of this field will be unpredictable across repositories

The ‘Type’ field is selected from a list, which can be configured per installation. There are great variations in values held in this field, including:

- Article
- Journal Article
- Postprint
- Research Paper
- refereed published journal paper

As can be seen, the ‘Type’ field can be overloaded, trying to convey:

- resource type (article),
- academic status (refereed/peer reviewed)
- and publication status (published).

Eprints

The ‘Title’ field is a free-text entry field.

The ‘Author’ field is composed of two free-text entry sub-fields ‘Family name’ and ‘Given name/Initials’

The ‘Citation’ field is synthesized from a number of other fields which can include: Author(s), Publication date, Title, Journal Title, Volume (Number), Pages, ISSN

Of these, Author, Title and Journal Title are mandatory fields, the others are not.

The ‘Type’ field is selected from a list, which can be configured per installation. However, the field is (almost) always defined as the out-of-the-box value: ‘Article’.

Peer reviewed status and publication status are held as discrete values

While much metadata entered as free text is useful to humans in identifying articles, it is less useful in the context of automatic machine to machine identification. In order to provide usage statistics at the individual article level, it vital that individual articles can be identified accurately to enable aggregation, de-duplication, etc. For that, it is necessary to employ reliably recognisable identifiers.

Identifiers

All repository softwares allocate a local identifier when records are created. Current practices include:

Digital Commons
Digital Commons assigns its own identifier to each record.

DigiTool
Digitool assigns its own identifier to each record.

DiVA
DiVA assigns a URN:NBN identifier to each record. The URNs can be retrieved by using the Swedish Royal Library resolving-service.

DSpace
DSpace, out-of-the-box, employs CNRI's Handle system as a primary identifier, and the vast majority of DSpace-based repositories do use Handles.

Eprints
Eprints assigns its own identifier to each record.

Fedora
"Fedora digital objects are identified within Fedora using a PID (Persistent IDentifier). A PID is case-sensitive and consists of a namespace prefix and a simple string identifier."

[Ref: http://www.fedora.info/definitions/identifiers/]

OPUS
OPUS assigns a URN to each record. The URNs can be retrieved by using the German National Library resolving-service.

arXiv
"Since 1 April 2007 (0704-) All new papers have identifiers with the following form: arXiv:0706.0001
and specific versions are referred to by adding the version number: arXiv:0706.0001v1

In general, the form is arXiv:YYMM.NNNNvV, where YY is the two-digit year (07=2007 through 99=2099, and potentially up to 06=2106)"

[Ref: http://arxiv.org/help/arxiv_identifier]

Identifiers from 1991 through 2007-03 followed the format:

Archive.subject class (where applicable)/YYMMnumber, e.g. math.GT/0309136

Some of these identifiers – URNs, Handles, PIDs – are recognised persistent identifiers, while others are not (though they are persistent as long as the base URL of a repository remains the same). All of these identifiers are important for reliable identification of items, and their retrieval, *within* repositories. However, their use is limited in terms of easy identification of items across and beyond repositories.

Fortunately, in the context of journal articles, there is a global identifier which can be potentially used to match items reliably across different locations: the DOI.

DOIs

The DOI is the most widely used global identifier for articles in the publishing world and all repository softwares are capable of storing DOIs. Desk research and the email survey reveal that many repositories do add DOIs linking locally stored items to articles on publisher websites – where they are available and time permits.

As an example, on 14th January 2009, Cranfield CERES held 707 items corresponding to journal articles. Of those 707 items, 468 (66%) included a DOI in their metadata.

This is encouraging but, clearly, some work needs to be done to increase the percentage of articles held in repositories which include the DOI. In this context it is worth noting that there is a tool provided by CrossRef - the Simple Text Query available at http://www.crossref.org/SimpleTextQuery/ - which can aid in retrieving DOIs given a text citation or list of citations.

It is important to note that the DOI can also be applied to resources other than journal articles so it is vital that the resource type can also be identified unambiguously.

Solutions to identifying articles

As observed previously, DSpace installations tend to overload the 'type' field, trying to convey the resource type (article), the academic status (refereed/peer reviewed) and the publication status (submitted, published) in one field. Eprints, on the other hand, presents these as discrete fields.

The PIRUS team recommends that all IRs should be encouraged to adopt the practice of exposing the resource type, version information and peer-review status of articles as separate metadata elements, as well as adding the DOI where possible.

Resource type
Recommended values for the resource type should be 'article' or 'journal article'.

Article versions

The project team has decided that usage will be counted only for accepted manuscripts and subsequent versions, as only at the point of acceptance for publication in a journal does an article become part of the formal record of scholarship. It was also agreed by the project team that PIRUS should be consistent with the terminology used by the JISC VERSIONS project (http://www.lse.ac.uk/library/versions/VERSIONS_Toolkit_v1_final.pdf), which defines 5 main stages in the life of an article, as well as the recently agreed NISO/ALPSP recommendations on article versions (http://www.niso.org/publications/rp/), which defines seven stages of a journal article.

Table 1: Stages in the publication of an article

NISO/ALPSP Definitions	VERSIONS Definitions	PIRUS Protocol
Authors Original (AO)	Draft	Not counted
Submitted Manuscript Under Review (SMUR)	Submitted Version	Not counted
Accepted Manuscript (AM)	Accepted Version	Counted: Version A
Proof (P)		Counted: Version A
Version of Record (VoR)	Published Version	Counted: Version B
Corrected Version of Record (CVoR)	Updated Version	Counted: Version B
Enhanced Version of Record (EVoR)	Updated Version	Counted: Version B

It was agreed, however, that for the purposes of PIRUS it is not necessary to record and report separately the usage of each of stages in either the NISO/ALPSP definition or the JISC definition. For usage purposes it would be desirable to distinguish between usage of the accepted manuscript/proof and usage of the version of record. While it is desirable that usage of these two broad categories of versions (Table 1, Column 3, Versions A and B) should be separately recorded, consolidated and reported for each article, this is unlikely to be practical for most publishers and repositories in the near future.. Bundled A and B usage reports will, however, be acceptable in the short term.

An outstanding issue to be resolved here is which metadata element should be used to expose this information – there is no standard as yet.

Peer review status
Again, an outstanding issue to be resolved here is which metadata element should be used to expose this information – there is no standard as yet.

Subject Repositories

While this survey covered only institutional repositories, the project team recognised that a growing proportion of online usage of articles is taking place in subject repositories, of which PubMed Central is a noteworthy example. It is, therefore, important to ensure that such subject repositories are aware of PIRUS, that any emerging new standard is relevant to them and would be supported by them. In the course of this project three major subject repositories (PubMedCentral, ArXiv and the Social Science Research Network) were consulted and provided the following feedback:

- none had a problem with the technical approach proposed by PIRUS
- two had concerns on privacy issues, but only if individual users of articles could be identified, which is not proposed in this project
- all three want to be kept informed of the next steps to be taken by PIRUS and are interested, in principle, in being involved

Phase 2 (September-November 2008)): *Develop draft usage reports and protocols for the recording and reporting of individual article usage, test this with publisher and repository usage data.*

Publisher situation

As the majority of online journal publishers are already compliant with the COUNTER Code of Practice and are providing online usage statistics at the journal level, there are a number of obvious steps that can be taken to deliver online usage statistics at the individual article level. These are:

- require the DOI to be used as the unique article identifier by all publishers
- require the DOI to be implemented in a standard way by all publishers and made a permanent attribute of an article at the same stage in the publishing process
- specify the versions of articles whose usage may be counted. This is likely to be the Stage 2 (author manuscript accepted for publication in the journal) and Stage 3 (final published manuscript) as defined by the PEER project (ref)
- prescribe a format and associated protocols for the recording and reporting of usage at the individual article level. (See Article Report 1 in Section 7 below)

Repository situation

The repository situation with regard to usage statistics is an area which has been the focus of considerable interest in the last year. Of particular relevance to PIRUS, are the findings discussed in two reports produced by:

- The JISC Usage Statistics Review Project (1)
- The Knowledge Exchange Institutional Repositories Workshop Strand on Usage Statistics (5)

Both of these projects were in agreement on the need for a fundamental format or scheme for repository log files to overcome the many variations between repository softwares. For this reason, both projects recommended the use of OpenURL Context Objects exposed in XML as an ideal candidate for a normalized format. It should be noted, while the most common application of OpenURL is to provide appropriate copy resolution, that is not the intention here. Instead, the idea is to make use of an existing, accepted standard format which is already capable of holding and exposing metadata required for statistical purposes, rather than trying to invent a completely new, arbitrary standard. Link resolving using OpenURLs may be a complex issue, but the construction of an OpenURL is actually a simple task.

The PIRUS team is in full agreement with this recommendation. However, in view of the wide range of repository softwares currently implemented - and the different ways they operate, it is impractical to propose a single approach that will work in all situations. For example, currently, when a full-text article is downloaded:

- In DSpace, a java servlet (BitstreamServlet.java) is invoked, which returns the requested file and generates a DSpace log entry
- In Eprints, a Perl module is invoked, which rewrites a cosmetic URL to an internally useful one which returns the requested file and a database access log entry is generated

As a practical solution to overcoming these variations, the project team proposes that the following scenarios (Diagram 1, below), in which there are three possible routes to generating standardized usage statistics, will cover most repository situations that are envisaged.

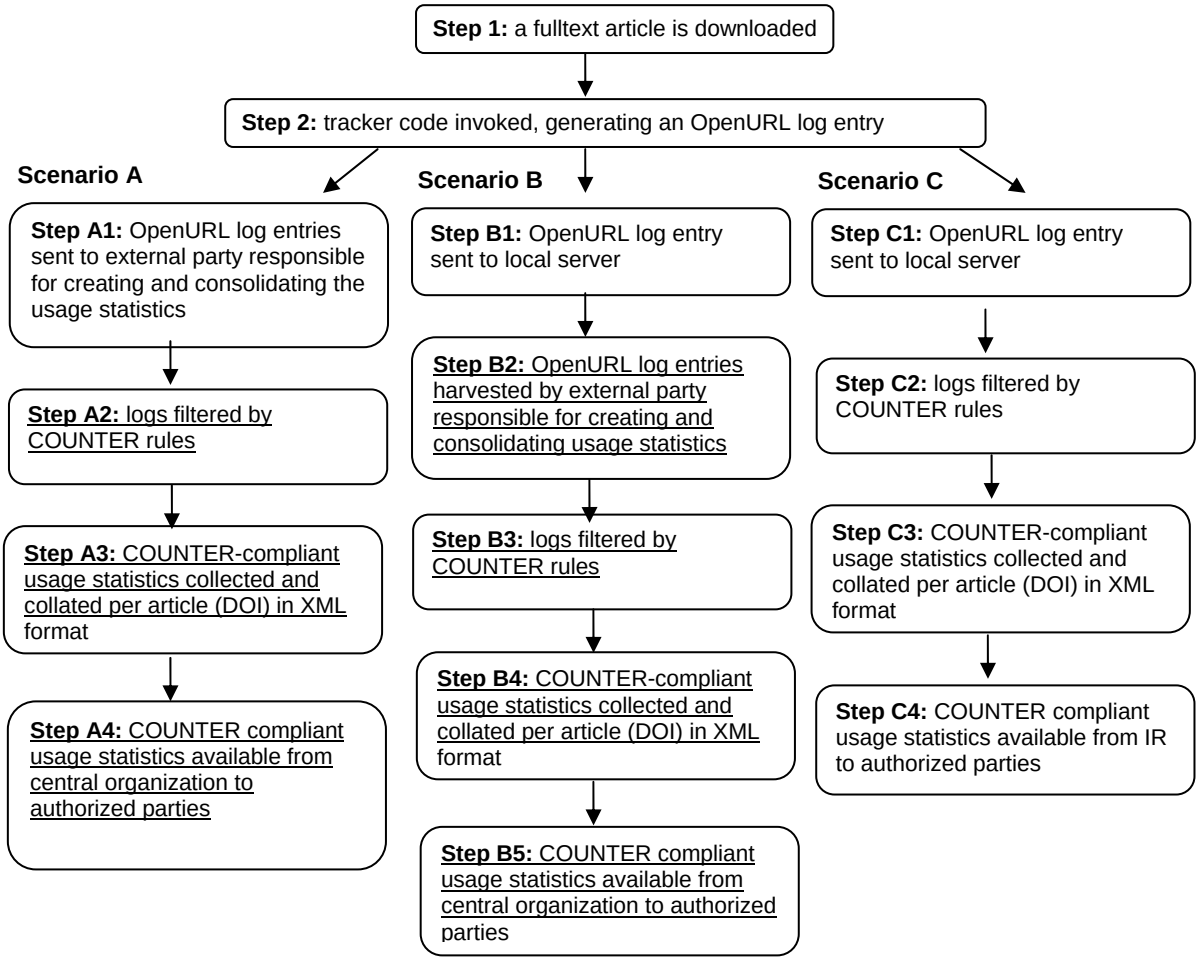

Diagram 1: Proposed approaches to recording and reporting usage statistics for repositories

The steps in Diagram 1 where the text is not underlined take place within the local institution hosting a repository. Those where the text is underlined are handled by an external party.

Step 1
The scenarios illustrated in Diagram 1 above all have a common genesis: an agent, e.g. a human user or a robot, accesses a link which leads to the downloading of an article.

Step 2
The download event triggers different actions, depending on the software in use. For example:

- In DSpace, when the java servlet (BitstreamServlet.java) is invoked, for a fulltext article download, in addition to the DSpace log entry, it would generate an OpenURL Context Object log entry
- In Eprints, again a database access web log entry is generated. However, an additional script running in a polling style loop would check recent log entries and, for a fulltext article download, generate an OpenURL Context Object log entry

While the detail would vary for other software applications, the outcome would remain the same: *in every case* - for fulltext article downloads - an OpenURL Context Object log entry would be generated.

To allow for the different requirements and capabilities of the organisations and institutions involved, as well as the variety of software applications in use, three alternative scenarios for processing the OpenURLs are now suggested and briefly described:

Scenario A
In scenario A, the generated OpenURLs are transmitted directly to a server hosted by a third party (central clearing house). This method is similar the Google Analytics model, however, unlike Google Analytics, server-side code is employed rather than JavaScript, as it doesn't really address the issue of logging PDFs and various other file types.

Once received, the external party filters the entries according to COUNTER rules and produces COUNTER-compliant reports which can be made available to authorized parties.

The project team have tested this scenario against both DSpace (at Cranfield University) and Eprints (at Southampton University). In both cases, the team successfully transmitted log entries to a server at BioMed Central.

Scenario B
In scenario B, the generated OpenURL entries are sent to a server hosted locally at the institution, which then exposes those entries via the OAI-PMH for harvesting by an external third party.

Once received, the external party filters the entries according to COUNTER rules and produces COUNTER-compliant reports which can be made available to authorized parties.

Though the project team have not tested this scenario, the OAI-PMH is well understood within the Institutional Repositories environment and would be relatively simple to implement.

Scenario C
In scenario C, again, the generated OpenURL entries are sent to a server hosted locally at the institution. In this scenario, however, all the processing necessary to filter the entries and create the COUNTER-compliant reports takes place locally.

Once created, the COUNTER-compliant reports can be made available to authorized parties via SUSHI.

The project team prototyped and tested this scenario against the DSpace installation at Cranfield University. The tests demonstrated that it possible to create a COUNTER-compliant XML report, and the output can be viewed at http://cclibweb-1.dmz.cranfield.ac.uk/pirus/AR1.xml. (All that would remain to complete the scenario is to add a SUSHI-wrapper around the report – something that, unfortunately, was not possible within the timescale of this project but is, never the less, eminently achievable.)

The scenarios described in Diagram 1 will cover the great majority of repositories and implementing one of them will enable repositories to generate online usage statistics for individual articles that are not only consistent with those generated by publishers, but that can also be consolidated to provide a total online usage figure per article.

The team proposes that the OpenURL Context Objects created by the tracker code used in the Diagram 1 Scenarios should ideally contain the following information:

- Date/time of access
- DOI
- Article format (e.g. html or pdf)
- Article version (though not all repositories are in a position to offer this data at the moment)

- Abuse data (data collected to detect and eliminate misuse)
    - IP address or hash thereof
    - user agent
    - Other activity that occurred in the same session

It will, however, be important to impress on organizations involved that collecting IPs will be for statistical purposes only, that the data will be securely stored and will not be shared. (This will be a particularly sensitive issue for pharmaceutical companies and for many US government departments). These concerns will be addressed by hashing IPs so that usage patterns will be detectable, but actual IPs will be anonymised.

One of the reasons that COUNTER has been so widely implemented is that it does not place and undue financial or organizational burden on participating publishers. It is important that this philosophy is also extended to repositories to encourage participation. For this reason repositories will not be required to fill in fields such as ISSN, as consolidation of usage by journal is mainly of interest to publishers, who could allocate this information further 'downstream' as needed.

The requirement for a central clearing house/houses

It is clear from the scenarios described in Diagram 1 that a central clearing house/houses will be required to receive the usage data from repositories, generate the resulting usage statistics and enable the publisher to consolidate these with its own usage statistics to provide a total usage figure. It would be premature to identify specific organizations at this stage, but the project team think that there are existing, established organizations that already have most of the capabilities that will be required. A central clearing house would have to meet the following criteria, as a minimum:

- be independent, and trusted by the major stakeholder groups (authors, libraries, publishers, funding agencies)
- have a proven capability to receive, store and process the relevant metadata and to generate usage statistics
- be able to handle large volumes of metadata and usage statistics

7. Outputs and Results

The four main outputs of the project are:

a. A proof-of-concept COUNTER-compliant XML prototype for an individual article usage report, Article Report 1: Number of successful full-text article downloads, that can be used by both repositories and publishers (Appendix E). This prototype is consistent with Release 3 of the COUNTER Code of Practice (6) and may be found at: http://cclibweb-1.dmz.cranfield.ac.uk/pirus/AR1.xml In principle this report could be provided for individual authors and for institutions. In practice, the individual author reports are much easier to generate and are a realistic short-term objective, while the reports for institutions and other entities, such as funding agencies, will be more complex and should be regarded as a longer term objective.

b. A tracker code, to be implemented by repositories, that sends a message **either** to an external party that is responsible for creating and consolidating the usage statistics and for forwarding them to the relevant publisher for consolidation **or** to the local repository server.

c. A range of Scenarios for the creation, recording and consolidation of individual article usage statistics that will cover the majority of current repository installations (See Diagram 1 above). Each repository may select the scenario that corresponds to their technology and implementation.

d. Specifying criteria for a central facility that will create the usage statistics where required (for some categories of repository) and collect and consolidate the usage statistics for others

These outputs are consistent with the original objectives of the project and take into account the very diverse range of technical and organizational configurations for repositories that currently exist. If implemented by COUNTER they will enable online usage statistics for individual articles to be created, reported and consolidated at global level, irrespective of source and according to the same standards.

The result would be a substantial further enrichment of the COUNTER data at a much more granular level that will provide, for the first time, authors, publishers, institutions and funding agencies with comparable usage statistics at the individual article level.

8. Outcomes

This project is the first to attempt to set a standard for measuring online usage that would apply to both publishers and repositories, as well as the first to set such a standard for recording and reporting usage of individual articles. This would not have been possible without the advances already made by COUNTER and we have explored new territory. We have also learned some important lessons about the practical challenges involved in setting standards for repositories, which currently exhibit a diversity of technological and organizational configurations.

PIRUS had very specific objectives, which have been met and, indeed, exceeded. All it has done, however, is demonstrate that it is *technically* feasible, even in the current, extremely diverse repository environment, for a standardised set of online usage statistics to be generated for individual articles. Organizational, intellectual property and political issues have yet to be fully addressed and it should be noted that the surveys carried out in Phase 1 showed that not all publishers and not all repositories enthusiastically endorse the principle of reporting usage at the individual article level.

9. Conclusions

Individual article usage statistics are a potentially valuable tool for several important stakeholders involved in research and the dissemination of its outputs, notably:

- researchers/authors, who are interested in monitoring online usage of their publications and understanding what this means
- repositories, who are interested in the usage of the items they hold, to help assess the value of making these items available , and to demonstrate the cost-effectiveness of the investment in the repository
- research institutions, who are competing for research funds and are under pressure from, for example, research assessment exercises to demonstrate the value of the research and researchers that they support
- funding agencies: who are seeking more quantitative, transparent ways of assessing the performance and impact of the research projects that they fund

While providing individual article usage statistics for authors/researchers for their own publications is relatively straightforward, aggregating these usage statistics for repositories, research institutions and funding agencies is significantly more challenging and a number of technical (eg, volume of data), organizational (eg, scalability and consolidation), economic (eg, allocation of costs) and political (eg, confidentiality) issues, which will take time to resolve.

The following broad conclusions can be drawn as a result of this project:

- common technical standards for measuring usage can be set for repositories and publishers, despite the diversity of organizational and technical environments
- further work will have to be done to translate the results of this feasibility study into practical, implementable solutions that work for all stakeholder groups

10. Implications

This work has the following implications for COUNTER and for the wider community:

a. For COUNTER: further improvements and extensions to the COUNTER Code of Practice will be offered. The existing COUNTER Code of Practice is designed only for publishers/vendors. If developed further and taken up by COUNTER the outputs of this project will be the first standards set by COUNTER for repositories. This significant expansion of COUNTER's strategic role would require modifications to the current Codes of Practice, with new reports and an expansion of the audit.

b. For Repositories: there are few common standards among repositories covering usage statistics; yet repositories are being required to produce and even publish usage statistics. For these to have any credibility they must be produced to a common, accepted standard. Repositories would be wise to adopt some such standard, whether it is set by COUNTER or not.

c. For Authors: credible and transparent global usage statistics on an individual article level will provide authors with a new metric that allows them to see how their research outputs are being used.

d. For Publishers/Vendors: it is likely that, if authors become aware that it is possible for them to obtain credible, global usage statistics for their articles, they will want have access to such data and will put pressure on publishers to participate in the process. Providing individual article usage statistics would provide publishers with an opportunity to further cement relationships with authors. Any requirement for reporting usage at the individual article level will also increase the need for vendors to standardise their implementation of DOIs, clearly define and identify different versions of articles, etc.,

e. For Funding Agencies: metrics used for the evaluation of research are currently heavily citation-based. The early results from the UKSG-sponsored Journal Usage Factor (7) indicate widespread support among authors and publishers for usage-based metrics as a supplement to citation-based metrics in, for example, the UK Research Excellence Framework (REF) (2). The availability of credible usage statistics for individual articles at the global level will further increase pressure on funding agencies to take usage into account as a measure of the impact of research outputs.

f. For Research Institutions: the inclusion of individual article usage statistics as a measure within a modified REF would require research-based institutions to collect and report such data for their own authors.

g. For the Data providers: a standard way to define and store metadata, in eg the Dublin core, will be required

h. For the Industry as a whole: if usage statistics for individual articles are to be consolidated and reported globally, data will have to be collected centrally and a capability to do this will have to be supported. The industry as a whole has to decide whether, in principle, it wishes to support such a capability – which could be an extension of an existing organization. This decision could only be made once the technical and organizational specifications, together with the associated costs, have been worked out in detail.

11. Recommendations

The recommendations of the project team are as follows:

a. To JISC: PIRUS has demonstrated that it is *technically* feasible to create, record and consolidate usage statistics for individual articles using data from repositories and publishers. If this is to be translated into a new, implementable COUNTER standard and protocol, further research and development will be required, specifically in the following areas:
   - Technical: further tests, with a wider range of repositories and a larger volume of data, will be required to ensure that the proposed protocols and tracker codes are scalable/extensible, work in the major repository environments, and can be applied to items other than articles
   - Organizational: the nature and mission of the central clearing house has to be developed, and candidate organizations identified and tested
   - Economic: assess the costs for repositories and publishers of generating the required usage reports, as well as the costs of any central clearing house/houses; investigate how these costs could be allocated between stakeholders
   - Political: the broad support of all the major stakeholder groups (repositories, publishers, authors) will be required. Subject repositories, such as PubMed Central, which have not been active participants at this stage in the project, will have to be brought on board. Intellectual property, privacy and financial issues will have to be addressed.

   The PIRUS project team recommends that JISC considers funding further research to address the issues described above.

b. To COUNTER: expand the mission of COUNTER to include usage statistics from repositories; consider implementing the new Article Report 1 as an optional additional report; modify the independent audit to cover new reports and processes.
c. To repositories: subject repositories to participate in the next stage of this project. All repositories should use standard data descriptions for article versions etc.
d. To publishers/vendors: accept, in principle, the desirability of providing credible usage statistics at the individual article level.

12. References

1. **JISC Usage Statistics Review:** http://www.jisc.ac.uk/publications/publications/usagestatisticsreviewreport.aspx
2. HEFCE: Research Excellence Framework http://www.hefce.ac.uk/Research/ref/
3. Standardized Usage Statistics Harvesting Initiative (SUSHI) http://www.niso.org/workrooms/sushi
4. Budapest Open Access Initiative, Institutional Repository Software http://www.soros.org/openaccess/software/
5. Knowledge Exchange Institutional Repositories Workshop Strand on Usage Statistics**:** http://www.knowledge-exchange.info/Admin/Public/DWSDownload.aspx?File=%2FFiles%2FFiler%2Fdownloads%2FIR+workshop+1617+Jan+2007%2FNew+reports%2FKE_IR_strand_report_Usage_Statistics_Sept_07.pdf.

6. COUNTER Code of Practice for Journals and databases, Release 3 http://www.projectcounter.org/r3/Release3D9.pdf
7. Usage Factor project http://uksg.org/projects

13. Appendices

Appendix A: Survey of Publishers
Appendix B: Repository Software Applications Table
Appendix C: UK Research Institutional or Departmental Repositories
Appendix D: JISC Repositories Email Survey Responses- Summary
Appendix E: Article Report 1 XML Report Example for a Single Article

PIRUS Final Report

Shepherd, Peter T.

2009-01