



## Project Document Cover Sheet

Project Information			
<b>Project Acronym</b>	PIRUS2		
<b>Project Title</b>	Publisher and Institutional Repository Statistics 2		
<b>Start Date</b>	1st October 2009	<b>End Date</b>	31 <sup>st</sup> May 2011
<b>Lead Institution</b>	Mimas		
<b>Project Director</b>	Kevin Cole, Mimas		
<b>Project Manager &amp; contact details</b>	Paul Needham, <a href="mailto:paul.needham11@btinternet.com">paul.needham11@btinternet.com</a>		
<b>Partner Institutions</b>	Mimas, COUNTER, Cranfield University, Oxford University Press and CrossRef		
<b>Project Web URL</b>	<a href="http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/">http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/</a>		
<b>Programme Name (and number)</b>	<i>JISC Capital Programme</i>		
<b>Programme Manager</b>	Andrew McGregor		

Document Name			
<b>Document Title</b>	PIRUS2 Final Report		
<b>Reporting Period</b>			
<b>Author(s) &amp; project role</b>	Peter Shepherd and Paul Needham		
<b>Date</b>	06/10/2011	<b>Filename</b>	PIRUS2FinalReport.pdf
<b>URL</b>	<i>if document is posted on project web site</i>		
<b>Access</b>	<input type="checkbox"/> Project and JISC internal		<input checked="" type="checkbox"/> General dissemination

Document History		
Version	Date	Comments
0.1	04/03/2011	Final draft version
0.2	29/07/2011	Extended final draft version
1.0	06/10/2011	Final version

Project Acronym: PIRUS2  
Version: 1.0  
Contact: Paul Needham ([paul.needham11@btinternet.com](mailto:paul.needham11@btinternet.com))  
Date: 06/10/2011



## **JISC Final Report**

# **Publisher and Institutional Repository usage Statistics: The PIRUS2 Project**

## **Final Report**

Authors: Peter Shepherd and Paul Needham

Date: October 2011

## Table of Contents

Acknowledgements .....	4
1 Executive Summary .....	5
2 Background.....	7
3 Aims and Objectives .....	8
4 Methodology .....	8
5 Implementation .....	9
5.1 Development of a prototype service for individual article statistics.....	9
5.1.1 Organizational model .....	10
5.1.2 Economic model.....	15
5.2 Software, standards and protocols development.....	17
5.2.1 Standards and Protocols .....	18
5.2.2 Consolidating usage data from Publishers and Repositories .....	20
5.2.3 Gathering publisher usage events .....	21
5.2.4 Gathering repository usage events .....	21
5.2.5 The PIRUS2 database .....	23
5.2.6 A Prototype Article Level Usage Statistics Portal .....	23
5.2.7 Discussion of technical issues .....	24
5.3 Dissemination and advocacy .....	27
6 Outputs and Results .....	28
7 Outcomes.....	28
8 Conclusions .....	28
9 Implications .....	29
10 Recommendations .....	30
11 References .....	32
12 Appendices.....	33
Appendix A Publisher feedback on proposed Individual Article Usage Reports .....	33
Appendix B Proposed AR1 report, transmitting statistics from publishers to CCH – updated .....	38
Appendix C Proposed consolidated usage report for authors .....	39
Appendix D Proposed publisher-only usage report for authors.....	40
Appendix E Proposed usage report for repositories.....	41
Appendix F Proposed usage report for research institutions .....	42
Appendix G PIRUS2 OpenURL key-pair string initial specification .....	43
Appendix H OpenURL Context Object in an OAI-PMH wrapper .....	44
Appendix I COUNTER robots exclusion list .....	45
Appendix J The PIRUS2 relational database.....	46
Appendix K Screenshots of reports available from the demonstration portal. ....	47
Appendix L PIRUS2 CCH - model for allocation of costs to publishers - scenario A.....	49
Appendix M PIRUS2 CCH - model for allocation of costs to publishers - scenario C .....	50

Appendix N	PIRUS2 CCH - model for allocation of costs to repositories - scenario A .....	51
Appendix O	Report on the results of the Extensions to PIRUS 2 .....	52
	Introduction .....	52
	UK Institutional Repository Usage Statistics Demonstrator .....	52
	Fedora implementation guidelines .....	60
	Publisher survey: economic models for the Central Clearing House .....	70

## Table of Figures

Figure 1.	AR1 example report .....	19
Figure 2.	PIRUS2 demonstration portal home page .....	24
Figure 3.	IRUS-UK demonstrator home page, overall stats .....	54
Figure 4.	IRUS-UK demonstrator home page, monthly stats .....	54
Figure 5.	SUSHI Report response at the individual item level .....	55
Figure 6.	Typical Fedora architectures .....	65

## Table of Tables

Table 1.	PIRUS2 OpenURL key-value pair specification - updated .....	25
Table 2.	Sources of information mapped to PIRUS2 information requirements .....	66

## Acknowledgements

The PIRUS2 Project is part of the Capital Programme funded by JISC.

We would like to thank all those who have helped during the lifetime of this project, and particularly:

- JISC for funding the project;
- Andrew McGregor, JISC Programme Manager, for his advice and support;
- Balviar Notay, JISC Programme Manager, for her advice and support;
- Ben Wynne, JISC Programme Manager, for his advice and support;
- Gary van Overborg and colleagues at ScholarlyIQ for their contributions to the development of the Central Clearing House model
- Paul Smith and colleagues at ABCe for their contributions to the development of the Central Clearing House model
- Members of the Steering Committee:
  - Judith Barnsby (IOP Publishing, PALS3 Metadata and Interoperability Group)
  - Simon Bevan, Senior Project Manager (Cranfield University)
  - Daniel Beucke (OA-Statistik)
  - Amy Brand (Harvard University)
  - Tim Brody (Eprints , Southampton University)
  - Richard Gedye (OUP)
  - Gregg Gordon (SSRN)
  - Bill Hubbard (RSP/SHERPA)
  - Ross MacIntyre (Mimas)
  - Andrew McGregor, Programme Manager (JISC)
  - Paul Needham, Project Manager (Cranfield University)
  - Mark Patterson (PLoS)
  - Ed Pentz (CrossRef)
  - Sally Rumsey (Oxford University)
  - Peter Shepherd (COUNTER)
  - Syun Tutiya (Chiba University)
  - Hazel Woodward, University Librarian, JISC Collections (Cranfield University)
- Participating publishers:
  - ACS Publications
  - Emerald
  - Institute of Physics Publishing
  - Nature Publishing Group
  - New England Journal of Medicine
  - Oxford University Press
  - Springer
  - Wiley
- Institutional Repositories:
  - Bournemouth University Research Online (BURO)
  - Cranfield CERES
  - Harvard DASH
  - University of Edinburgh ERA
  - University of Huddersfield Repository
  - University of Hull Institutional Repository
  - Oxford University Research Archive (ORA)
  - University of Salford Institutional Repository
  - Southampton ECS EPrints Repository
- Members of the Fedora Working Group:
  - Sally Rumsey (Oxford University)
  - Neil Jefferies (Oxford University)
  - Anusha Ranganathan (Oxford University)
  - Chris Awre (Hull University)
  - Richard Greene (Hull University)
  - Steve Bayliss (Acuity Unlimited)

# 1 Executive Summary

The aim of the PIRUS2 (**P**ublisher and **I**nstitutional **R**epository **U**sage **S**tatistics Project 2) was to take forward the outcomes of the original PIRUS project and build on its recommendations, by developing a prototype service (including technical, organizational and economic models for a Central Clearing House) that will enable publishers, repositories and other organizations to generate and share authoritative, trustworthy usage statistics for the individual articles that they host.

While the core objectives of the project did not change as PIRUS2 progressed it became apparent that not all the scenarios for creating the usage statistics envisaged in the original PIRUS project would be practical – at least in the short-term. Moreover, as more detailed feedback was obtained from publishers, repositories and others as the project progressed, it became clear that there remained significant concerns about aspects of the organizational and economic models initially proposed. These have, therefore, been modified in light of this feedback, and areas for further investigation identified. A staged implementation of a global usage statistics service is now envisaged.

PIRUS2 has achieved its overall aims, by delivering a prototype service that meets the following criteria:

- A workable technical model, refined from that proposed in the original PIRUS project with more extensive tests with a larger and more diverse data set
- A practical organizational model based on co-operation between proven, existing suppliers of data processing, data management and auditing services that meets the requirement for an independent, trusted and reliable service. *However, it is clear from a survey carried out at the end of this project that the majority of publishers are not, largely for economic reasons, yet ready to implement or participate in such a service.*
- An economic model that provides a cost-effective service and a logical, transparent basis for allocating costs among the different users of the service. *While this economic model is based on costs that vendors of usage statistics services have validated as reasonable, there is strong resistance from publishers to accepting these costs.*

The main outputs of PIRUS2 are:

- a fully tested prototype aggregated statistics service, comprising usage data and statistics from publishers and institutional repositories, employing agreed first versions of Standards and Protocols; DSpace, Eprints & Fedora Software plug-ins; Software to process and filter OpenURL usage data according to COUNTER rules;
- a proposed business model for the prototype aggregated statistics service, including a list of organizations that meet the required criteria for the central clearing house(s), an assessment of the costs for repositories and publishers and the running the central clearing house(s); proposals for dealing with legal issues, results of market research surveys;
- feedback from authors, publishers, repositories, and research funding agencies on the proposed model for the aggregated statistics service
- an end-of-project seminar to share the results, knowledge and experience acquired in the course of the project with the stakeholder communities

Additional outputs of PIRUS2, as a result of further work described in Appendix O, are:

- further developments to the proposed organisational, economic, political and technical models – based around a more distributed model of feeding usage statistics to the CCH via a number of national or regional agencies, illustrated by:
  - a UK institutional repository usage statistics demonstrator service, which could very cost-effectively consolidate article statistics for all UK IRs and act as a single point of transfer for those statistics to the CCH, and provide extra opportunities to furnish UK IRs with COUNTER-compliant statistics for all their item-types (not just articles), as well as offering opportunities to demonstrate the impact and value of IRs
- a generic set of guidelines for implementation of PIRUS2 functionality across Fedora repositories
- a finding that usage of articles hosted by institutional repositories is rather high. Over the 7-10 month period of the project during which usage data was collected for articles hosted by the 6 participating repositories, there were 527,224 downloads of 6,089 articles; an average of 86 downloads per article

The PIRUS2 project has achieved many of its aims and objectives. However, some outstanding tasks remain to be completed in order to take this work forward and lead to the creation of a global article level consolidated statistics service. There is more work to be done to:

- achieve formal acceptance of a CCH from all stakeholder groups
- roll-out patches or, better still, embed PIRUS2 functionality out-of-the-box into repository softwares

Furthermore, before a fully-fledged, comprehensive usage statistics consolidation service can be launched, a number of issues, beyond the control of this project, still need to be addressed:

- a. SUSHI: the proposed article level reports will need to be endorsed by COUNTER, and extensions to the COUNTER-SUSHI schema – to accommodate required article level metadata elements – will need to be endorsed and adopted by NISO.
- b. ORCID: reliable identification and attribution of individual authors remains problematic, making it – currently – virtually impossible to consolidate usage across multiple articles for any given author. The adoption of the ORCID system, due to launch as a beta service at some point in 2011, “will, from the start, enable 3rd parties to build value added services using ORCID infrastructure”<sup>1</sup>.
- c. Institutional Identifiers: Although identifying institutions is less problematic than identifying authors, nevertheless, the eventual outcomes from the NISO I<sup>2</sup> Working Group<sup>9</sup> will improve the efficiency and potential for interoperability of an article level usage statistics service.

The recommendations of the project team are, therefore as follows:

- a. To JISC: PIRUS2 has developed a costed prototype service that capable of creating, recording and consolidating usage statistics for individual articles using data from repositories and publishers. Further feedback is required, however, to demonstrate with confidence that there is sufficient support for full implementation.
  - Organizational: while it is unlikely that there will be widespread implementation of PIRUS2 by publishers in the immediate future, due to cost concerns, there is a strong case for implementation of ‘IRUS’ the Institutional Repository Usage Statistics service, based on the technical and organizational model proposed in this report. Unlike the publishing world, there are currently no standards for usage statistics from Institutional Repositories. Adoption of the proposed IRUS model would provide, for the first time, such standards. For these reasons, we recommend that JISC should support the implementation of IRUS.
  - Economic: the economic models for supporting the central clearing house are reasonable and should form the basis for going forward, both for publisher and for repositories
  - Political: support for the outcomes of PIRUS2 among publishers and institutional repositories is weak. JISC could play an ongoing role, together with COUNTER, in trying to build this support
  - Statistical: while detailed statistical analysis of usage was not one of the objectives of PIRUS 2, the article download figures for the 6 institutional repositories that participated in the project indicate that usage of articles in repositories is significant and merits more rigorous statistical analysis.

The PIRUS project team recommends that JISC considers funding further research in the short term, while the project has momentum, to address the issues described above.

- b. To COUNTER: expand the mission of COUNTER to include usage statistics from repositories; consider implementing the new PIRUS Article Reports as optional additional reports; modify the independent audit to cover new reports and processes. Use the fact that there is growing demand from authors for individual article usage reports to encourage publishers to provide them, based on the PIRUS2 standards.
- c. To repositories: consider participating in the proposed IRUS service and provide individual item level usage reports

- d. To publishers/vendors: accept, in principle, the desirability of providing credible usage statistics at the individual article level; implement the new PIRUS article reports for their own usage reporting to authors
- e. To repository software vendors/developers: accept, in principle, the desirability of incorporating PIRUS2 tracker functionality into their “out-of-the-box” software

## 2 Background

The most granular level at which COUNTER currently 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, reporting and consolidation 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
- 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 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<sup>2</sup> protocol facilitates the automated consolidation of large volumes of usage data from different sources.
- It should be noted that the outputs of PIRUS (the XML schema for the individual article usage reports, the tracker code and the associated protocols) are already being implemented by publishers and repositories (e.g. PLoS and SURF). It is important that these are fully tested and, if necessary, refined, before they are too widely adopted.

PIRUS2 builds on the work undertaken by the JISC-funded PIRUS project<sup>3</sup>, the JISC Usage Statistics Review and the Knowledge Exchange Institutional Repositories Workshop Strand on Usage Statistics.

The JISC Usage Statistics Review<sup>4</sup> “*aimed at formulating a fundamental scheme for repository log files and at proposing a standard for their aggregation to provide meaningful and comparable item-level usage statistics for electronic documents*”. The Review suggested that “*usage events should be exchanged in the form of OpenURL Context Objects using OAI*” and “*policies on statistics should be formulated for the repository community as well as the publishing community*”, and also noted that “*as an aggregator and an initiator of further development in Great Britain JISC is probably the most suitable actor*”.

The Knowledge Exchange Institutional Repositories Workshop Strand on Usage Statistics<sup>5</sup> detailed steps needed to “*produce statistics that can be collected and compared transparently on a global scale*”. The Workshop made a number of recommendations for action, including: the need for an “*event-based web-log based format for sharing ‘usage events’ to deliver many profiles*” (OpenURL Context Objects). They also made a number of suggestions relating to COUNTER: add article level stats, investigate complex objects, set up COUNTER-IR to shadow the publisher group, and investigate aggregating COUNTER stats at consortium level.

The aim of the JISC PIRUS Project was to develop and extend COUNTER-compliant standards and usage reports beyond the journal level to the individual article level. PIRUS devised a range of



Scenarios for the creation, recording and consolidation of individual article usage statistics that would cover the majority of current repository installations. In keeping with the recommendations of the projects mentioned above, a key component in all the Scenarios was the generation of OpenURL Context Objects for the exchange of usage events (not for link resolving).

Prototype software was created and tested against DSpace and Eprints repositories which sent usage data as OpenURL key/value pair strings:

- **either** to an external party (who would be responsible for creating and consolidating the usage statistics and for forwarding them to the relevant publisher for consolidation)
- **or** to the local repository server where usage data could be exposed via OAI-PMH or alternatively processed to produce reports locally, which could be exposed via SUSHI.
- As an example, PIRUS developed a proof-of-concept COUNTER-compliant XML prototype for an individual article usage report (Article Report 1: Number of successful full-text article downloads), which could be used by repositories, publishers and other stakeholders.

Also, criteria were specified for a central facility (statistics aggregator) that could create the usage statistics where required (for some organizations) and collect and consolidate the usage statistics for others.

Further research and development – technical, organizational, economic, and political - is now required to transform the prototype outputs and standards specified by PIRUS into implementable, widely accepted processes for journal articles.

Looking beyond journal articles, the rules for filtering of ‘raw’ usage data can be applied to logs and log entries, irrespective of the resource type(s) and repository software applications under consideration, to create COUNTER compliant usage data. It is, therefore, pertinent to ask: What can be done with these data? What should be done with these data? To answer these questions, further research is required to develop implementable, widely accepted usage statistics processes and services for all resource types.

### 3 Aims and Objectives

The aim of the PIRUS2 Project is to take forward the outcomes of the original PIRUS project and build on its recommendations, by developing a prototype service (including technical, organizational and economic models) that will enable publishers, repositories and other organizations to generate and share authoritative, trustworthy usage statistics for the individual articles and other items that they host.

In order to achieve this overall aim, the project has sought to meet the following main objectives:

- Develop a suite of free, open source programmes to support the generation and sharing of COUNTER compliant usage data and statistics that can be extended to cover any and all individual items in IRs and SRs.
- Develop a prototype article level Publisher/IR statistics service
- Define a core set of standard usage statistics reports that repositories could/should produce for internal and external consumption

In the course of the project it became clear that publishers and repositories did not want an overly prescriptive set of usage reports. While they want to have access to individual article usage statistics that are prepared and recorded according to a global standard, they want to have flexibility in the reports they deliver to, for example authors, and also the option to combine usage data with other categories of quantitative data, such as citations. The output and recommendations of this project reflect this.

### 4 Methodology

The overall methodology was similar to that which has been used, successfully, in previous JISC sponsored projects, including the original PIRUS project and the EThOSnet project. The work was divided into six work-packages, each led by one partner institution. The primary partners, including work-package leaders formed a Project Management Team that worked in close collaboration with the

Project Manager. Additionally, there was significant horizontal, cross-work-package activity to ensure compatibility and consistency across a number of issues, including technical platforms and business requirements.

To provide wider input into the project, to help in the evaluation of results and to support dissemination and advocacy, a Steering Committee and Publisher Forum were set up and met regularly via conference calls in the course of the project. The Steering Group, chaired by Hazel Woodward, comprised the members of the Project Management Board (PMB), plus representatives of other publishers and repositories. The Publisher Forum was comprised of representatives of the major international scholarly journal publishers, both commercial and not-for-profit, covering all the major scholarly disciplines.

Although the PIRUS2 project is UK-based, its work was international in scope - and built upon international standards and policies. In order to ensure the development of acceptable and viable technical and business models, the project undertook the following activities:

- Developing and implementing an advocacy and dissemination campaign to ensure that the proposed service receives a sufficient level of buy-in from stakeholders so that it can be financially viable in accordance with the business model developed by the PIRUS2 project;
- Developing a robust and scalable technical infrastructure in readiness for a successful move from prototype to 'live' service;
- Ensuring a distinction is made between the PIRUS2 project and the service under development, and how they relate to each other;
- Monitoring and testing, as appropriate, relevant technology trends with a view to improving the technical sustainability of the service being developed and facilitating its interaction with repositories, publishers and statistics aggregators – also taking into account the work being carried out in related areas;
- Making appropriate links with other projects and initiatives in the UK and in a wider, international context (e.g. OA-Statistics) where there is potential for synergy and sharing of experience and good practice;
- Ensuring that an appropriate governance structure is put in place for the proposed service
- Agreeing a set of metrics for the proposed service, and ensuring that the system is capable of providing them.

The critical success factors for a successful implementation of the PIRUS2 project were:

- Close cooperation between COUNTER, CrossRef, publishers, repository software developers, NISO and other interested parties in the UK, Europe and beyond.
- buy-in from a variety of stakeholder groups, in particular repository managers, publishers and potential statistics aggregators
- usability, interoperability and scalability of the prototype aggregated statistics service

The methodology adopted for the project has been designed to meet these success factors.

## 5 Implementation

The project was implemented as six Work-packages (WPs) and focussed around three main areas:

- Development of a prototype service for individual article statistics
- Software, standards and protocols development
- Dissemination and advocacy

### 5.1 *Development of a prototype service for individual article statistics*

The work described in Section 5.2, below, demonstrates that it is *technically* feasible to create, record and consolidate usage statistics for individual articles using data from repositories and publishers, despite the diversity of organizational and technical environments in which they operate. To translate this into a new, implementable COUNTER standard and protocol, further research and development is required, specifically in the following areas:

- **Organizational:** the nature and mission of the central clearing house/houses proposed by PIRUS 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 the Social , 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

These issues were addressed in discussion with the Steering Committee and the Publisher Forum, based on models developed by members of the PMB in consultation with external organizations with much experience in the collection, processing, consolidation and reporting of large volumes of usage data for online publications ( such as ABCe and ScholarlyIQ).

Useful input for the design of the reports that the CCH should generate was obtained from a survey of publishers (See Appendix A), the majority of whom have received requests from their authors for article-level usage statistics. An earlier PLoS survey of their authors already demonstrated that the majority of them find the article-level usage statistics provided by PLoS useful. The purposes for which authors use, or would use these statistics is less clear, although the publishers feel that the main purpose would be to demonstrate the impact of their research to colleagues or to management.

In terms of the format in which author usage reports could be provided, some publishers favour a standardised format, while others prefer more flexibility.

There are two rather strong messages from the publishers. First, that they wish to be the provider of usage data to authors in their publications, and second, flexibility in the format and frequency of delivery of this information is desirable. Furthermore the majority of the publishers think that the usage data should be available for a relatively long period of 10 years plus.

The feedback from the PLoS survey and from this publisher survey indicates that authors, on the whole, would value having available usage statistics for their own articles and that demand for this is likely to increase. There is less agreement on the format in which publishers and authors would like to have this information, but this is not unusual. The situation was similar when the COUNTER usage reports for librarians were launched in 2003. At that time the decision was taken to launch the usage reports and refine them as a result of further feedback based on usage. This approach has proven successful and cost-effective over the years. A key objective at this stage must, therefore, be to ensure that the usage data and the associated metadata are captured at a sufficiently granular level to allow flexibility on the creation of the reports.

### **5.1.1 Organizational model**

The CCH must be reliable, flexible, scalable and cost-effective. Reliable: because authors, institutions and other organizations worldwide will be basing decisions on its outputs which must be trustworthy. Flexible: because the CCH will have to accept usage data in a variety of formats from a range of organizations and must also be able to deliver a range of routine and customisable usage reports. Scalable: because the volumes of data to be handled will be huge (see below). Cost-effective: because repositories will not use the facility if the tariffs are too high.

This organizational model is based on the collection, COUNTER-style processing and reporting of article level usage data and makes the following assumptions with regard to the total volumes of data involved:

- 40 million overall DOIs
- 1.5 million new articles published annually
- 1.5 billion requests for article usage per year
- 48 requests per second for article level usage (burstable)

After analyzing the roles and functions of the CCH, as well as the demand and challenges associated with delivery of such services, it is recommended that the central role be filled by an organization with extensive experience with accommodating large volumes of data, as this will help ensure that costs can be minimized. Keeping costs to a minimum will, in turn, encourage a higher degree of adoption which is critical for the success of the PIRUS2 initiative.

Methods for reducing costs include leveraging an already proven infrastructure that is reliable, scalable and maximizes performance for high volumes of data. The data should be horizontally partitioned as well as aggregated properly to ensure the highest degree of performance for both the collection and reporting of article level usage data. Proper data management and layering of services will also be crucial in order to reduce hardware costs associated with growth and volume.

While it is envisaged that a single CCH will be responsible for gathering, processing and consolidating the usage statistics globally, the option to have a series of national data processing centres that would consolidate the usage statistics from institutional repositories prior to forwarding the resulting reports to the CCH, should be kept open. Given the small size and local mission of many institutional repositories, as well as legal and cultural differences between countries, this may lower the barrier to support of the CCH by such organizations.

#### **5.1.1.1 Role of the Central Clearing House**

The CCH will have two broad roles. First, to collect, consolidate and process usage data from repositories and publishers. Second, to create and distribute usage reports to authorised parties (mainly publishers and repositories). The CCH Demonstrator, described below, has shown, using test usage data from publishers and from repositories, that it is technically feasible. How the CCH will function in these two roles is described below.

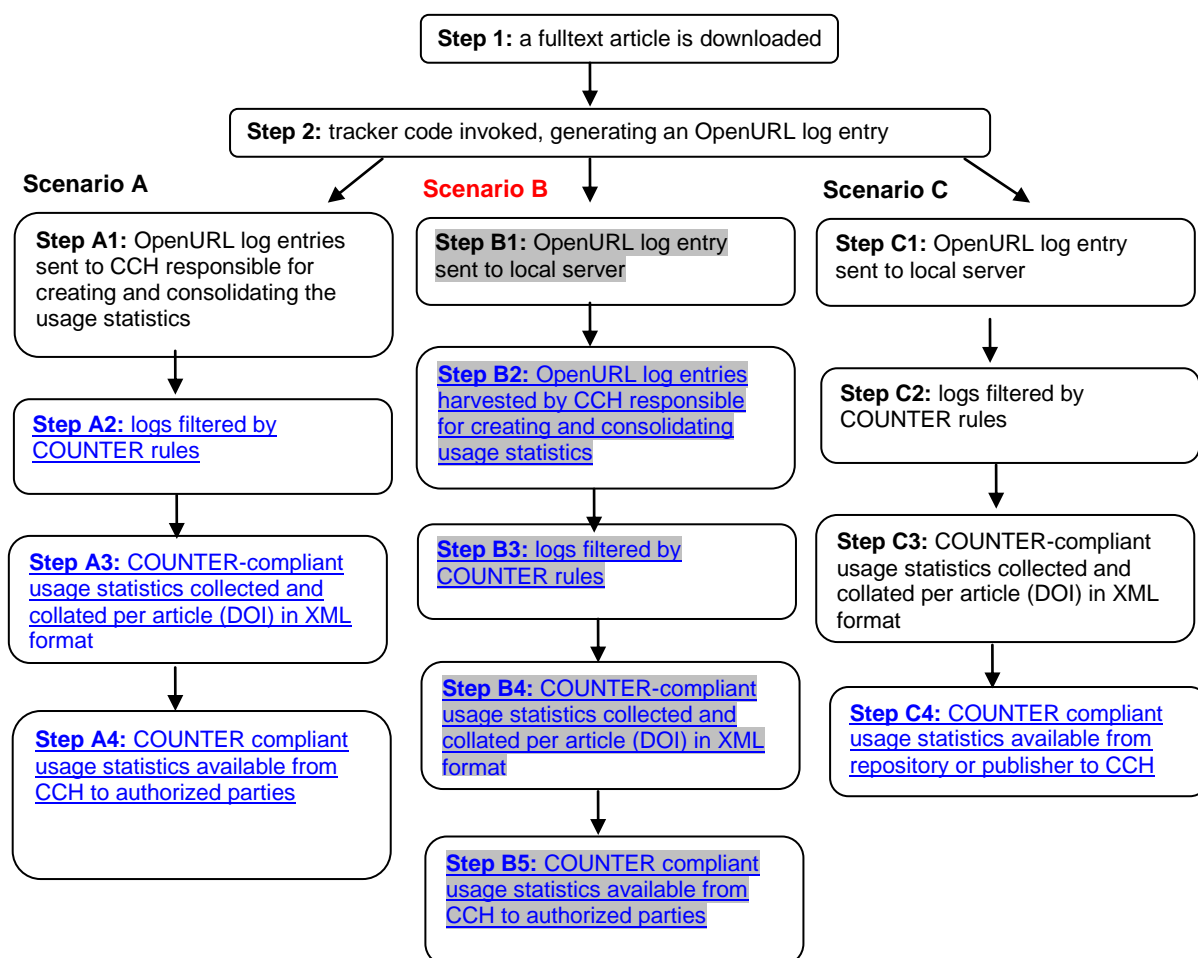
##### Collection, consolidation and processing of usage data

One recommendation of the original PIRUS project was that the CCH should be able to support three scenarios (A, B and C in Scheme 1, below) for the collection of usage data:

Scenarios A and B are likely to be prevalent among institutional repositories, while Scenario C will be prevalent among participating publishers and larger repositories. In the course of testing the repository usage data, however, the project team came to the view that we should drop Scenario B above and adopt Scenario A (i.e. the tracker (push) approach rather than use the OAI (pull) approach) for repositories that cannot implement Scenario C (the great majority). It would cut down processing effort at the CCH, and remove any real need for additional auditing at the repository level - if we allow exposure via OAI, we would need to introduce some auditing routines to ensure the integrity and validity of the usage data. There's nothing to stop repositories also exposing usage data/stats via OAI, if they want to for purposes other than ours.

Steps highlighted in blue text will take place in the CCH.

##### Scheme 1



### 5.1.1.2 Scenarios for the Central Clearing House

As described above, the CCH will adopt two models for the processing of usage statistics:

Scenario A: consolidated processing (applies to most repositories and to some publishers)

- Relies on all journal article downloads invoking a tracker code that sends data to a single big bucket
  - Consolidated usage reports can be generated by the CCH
- Single data **standard**, not necessarily data tool
  - Requirements can be met by various counting or analytics packages
  - Compliance with the standard can be checked by the “data gathering” audit
- All data in one place allows mining - deeper insights into data and easy integration of other projects, e.g. JUF
- Publishers who use this option could lose control of own data and report compilation
  - Terms and Conditions could handle some aspects of this
- All steps are auditable:
  - Data gathering
    - Process of sending data packet to bucket
    - Profile of data packet – does it meet standards?
  - Counting
    - Correct interpretation of data packets received
  - Compilation of usage reports
    - Correctness, completeness
- Audit overhead lower due to standard system

Scenario C: distributed processing (applies to most publishers and some repositories)

- Relies on repositories and publishers gathering data in own buckets
  - Publishers
    - count and produce own usage reports according to the specifications of Article Report 1.
  - Repositories
    - count and produce own usage reports and send reports to CCH OR
    - send data to CCH who count and produce usage reports (and return to repositories)
  - CCH sends repository reports to publishers
- All steps are auditable:
  - Data gathering
    - Process of sending data packet to bucket
    - Profile of data packet – does it meet standards?
  - Counting
    - Correct interpretation of data packets received
  - Compilation of usage reports
    - Correctness, completeness
- Many possible risk areas due to multiple supply points

It has been agreed that, in view of the technical challenges that the CCH faces, its strong dependency on other initiatives, such as ORCID<sup>6</sup> (the Open Researcher & Contributor ID) and institutional identifier and the requirements for publishers to re-engineer some of their processes, it may be prudent to implement the CCH in two Stages:

Stage 1: gather and consolidate usage data only from repositories and provide the usage statistics generated by the CCH to publishers and other authorised bodies ( i.e. largely Scenario A, but also with a capability to support Scenario C, as some larger repositories may want to take this approach)

Stage 2: and collect usage data from publishers that wish to use the CCH service for this purpose (Scenario C)

#### Creation and distribution of Usage Reports

The CCH will require a capability to create a range of standard usage reports that will meet the needs of authors, publishers, research institutions and repositories. Examples of the reports that the Demonstrator has shown can be produced are listed in Appendix K. It should be noted, however, that these examples demonstrate the range of usage reports that the CCH will have the capability to deliver and the specific reports that will be listed on the menu will be refined in discussion with user groups.

Reports from the CCH will include:

- usage reports for publishers
- usage reports for repositories
- usage reports for research institutions
- usage reports for research funding agencies

It is not envisaged that individual authors would have direct access to the CCH; rather, they would receive usage reports for their articles via the relevant publisher or institution.

#### Usage reports for publishers

Authors increasingly want to know how much usage the articles they publish are receiving. These reports will allow the publisher to report usage on individual articles by author and provide such reports to the authors themselves. By incorporating usage statistics from participating repositories and aggregators, publishers will not only be able to provide a more comprehensive overview of usage, but will also be able to determine which authors publish the most frequently used articles, as well as the repositories and aggregators where usage is highest for a particular article or group of articles. A proposed format for the consolidated author reports may be found in Appendix C.

In the short-term, it should be possible for publishers to provide a standardized report of usage just from their own platform to authors – without having to wait for the formation of a CCH. A proposed format for the publisher-only author reports may be found in Appendix D.

#### Usage reports for repositories

Institutional repositories are increasingly under required to demonstrate the value they provide to their parent institution. The PIRUS2 usage statistics will allow them to provide reliable, consistent usage statistics for the articles that they host, which can be compared with other organizations that host the same articles. An example of a proposed PIRUS2 usage report for repositories is provided in Appendix E.

#### Usage reports for research institutions

Institutions use a range of metrics to demonstrate the productivity and impact of their research. Citation data are already widely used and the capability to generate this is part of the offering of Elsevier's and Thomson Reuters products. The CCH will have the capability to generate usage-based reports for institutions, provided author identifiers (ORCID) and institutional identifiers are fully implemented by participating organizations. Examples of the types of usage report that the CCH will be able to generate for research institutions is provided in Appendix F.

#### Usage reports for research funding agencies

The response from research funding agencies on the value of individual article usage statistics has been mixed. Most see them as a potentially useful metric, but not an essential part of their toolkit for measuring the value or impact of the research that they fund. Authors are, however, required to demonstrate the value and impact of the research they carry out and submitting article usage reports in the standard formats specified in Appendix C or Appendix D in their project reports to funding agencies will provide a new indicator of the value and impact of their research.

#### **5.1.1.3 Register of Repositories:**

The CCH will need to maintain a list of the sources of usage data. To this end, an official Register of Repositories will need to be created. One condition for inclusion on the Register is that repositories must implement an officially approved PIRUS2 capability.

As a number of services already maintain data about repositories (e.g. ROAR<sup>7</sup>, OpenDOAR<sup>8</sup>), it would be worth investigating whether an existing register could fulfil this function in some way, and thus reduce duplication of effort.

#### **5.1.1.4 Business requirements:**

The usage data should be controlled by the participants (i.e. contributors of the data)

Some participants may require access to the data in its granular form

The participants can decide whether to compile the usage reports themselves or to delegate that role to the CCH

Usage reports must state the sources from which they have been compiled to ensure transparency

Delivery mechanisms for usage reports should also include:

- Web services with usage data provided in XML and JSON formats
- Web 2.0 based technologies (ex. JQuery and cross-site AJAX) to provide interactive reports and graphs that can then be published by the requestor within their domain – graphing tools must be extensible, support call-outs and be mobile friendly
- Data extracts in custom formats or direct access to a mirrored instance of the database

High availability service where hardware is redundant and geographically dispersed

Layered infrastructure with a clear separation of services for proper load balancing

- Collection Service
- Aggregation Service – usage data warehouse
- Reporting Service

Each layer must be scalable and operate independently of one another

Automated monitoring of the data transmitted (inbound and outbound)

#### **5.1.1.5 Other requirements/issues**

An approach for updating the repository software. Following further discussion, rather than a plug-in, the project team thinks that it would be better to build PIRUS2 functionality into the core functionality

of the repository softwares. Then, at each new software release, the developers would have to make sure the new version was PIRUS2-compliant *only once*, as opposed to hundreds of institutions around the world having to apply the same change *hundreds of times*. (Testing has already demonstrated that applying plug-ins to heavily customised repositories is problematic.) A situation is envisaged where the functionality is always in the softwares, by default, and just needs to be switched on or off in a config file.

1. it would be worth including the Institutional Identifier in the specification. NISO is working on setting a standard – I<sup>2</sup> - for the institutional identifier<sup>9</sup>
2. usage associated with an author's articles based on work done at a particular institution should not move with the author when they change institutions. Work done at a particular institution should continue to be credited to that institution. The institutional affiliation metadata in a particular article does not change when an author moves; it is a permanent attribute of that article
3. the ORCID researcher identifier has been included in the specification for Article Report 1

#### 5.1.1.6 Candidate organizations for role in the Central Clearing House

Candidate organizations should meet the following criteria:

- Experience with handling large volumes of online usage data and associated metadata
- Willingness to work within the proposed PIRUS organizational structure
- Credibility with publishers, repositories and other providers of usage data

#### 5.1.1.7 Confidentiality/Intellectual property

The CCH should abide by the following principles:

- The “bucket” of usage data should be controlled by the participants - they can decide whether to compile the usage reports themselves or to delegate that role to the CCH
- Usage reports must state the sources from which they have been compiled to ensure transparency
- Usage reports must state the sources from which they have been compiled to ensure transparency

### 5.1.2 Economic model

#### 5.1.2.1 Revenues

The economic model that would be required to support Stage 1 of the CCH has been developed in discussion with existing, large-scale providers of online usage statistics services. We have identified the following possible sources of revenues to support the CCH:

- membership fees that give members the right to use the services of the CCH
- transaction-based fees:
  - from repositories and publishers who use the CCH to create usage statistics from their raw log files
  - from publishers, who obtain usage statistics from the CCH for consolidation into their own usage reports
  - from publishers who submit their usage statistics for consolidation with usage data from other sources
  - from organizations (e.g. Thomson ISI or Elsevier (SciVal)), who would use the data from the CCH to enhance the citation and usage based performance reports that they provide to institutions.
  - from institutions, who want reliable usage reports for content produced by their researchers and departments

#### 5.1.2.2 Costs

##### Collection Services Costs

- The following table outlines the costs associated with data collection and processing and is only relevant to the fulfilment of Scenario A above:

Description	Frequency	Cost
Setup and development	One-time	\$57,300
Infrastructure	Monthly	\$8,400



Operations	Monthly	\$6,700
------------	---------	---------

- Setup and development costs for the collection of usage statistics should be mainly absorbed by the CCH. Ideally, the CCH should already have a good portion of the development and setup already completed by leveraging its existing infrastructure and services related to collecting and processing usage data. The only requirements for contributing data would then be membership.
- Web 2.0 services, such the CCH's interactive reports which can be embedded seamlessly into the requesting organization's web site and accessed by their constituents may also be considered as an additional form of revenue at a fixed rate – possibly incorporated into the membership fees (example provided below in Membership Fee Schedule) or within the transaction-based fees.

Membership Fee Schedule

Total Revenue*	Annual Fee	Annual Fee (Including Web 2.0 Services)	Annual Fee (including harvesting of usage statistics)
<\$250k	\$470	\$670	\$890
\$250-500k	\$970	\$1400	\$1900
\$500k- 1 Million	\$1,700	\$2,400	\$2,900
\$1 – 5 Million	\$4,700	\$7,400	\$7,900
\$5 – 10 Million	\$9,700	\$12,400	\$16,900
\$10-50 Million	\$15,700	\$19,400	\$26,900
\$50-100 Million	\$25,700	\$35,400	\$44,900
>\$100 Million	\$38,700	\$54,400	\$66,900

\*Total revenue shall include all revenue from all divisions and will be self-categorized by members.

Reporting Services Costs

- The following table outlines the estimated costs associated with reporting services:  
:

Description	Frequency	Cost
Setup and development	One-time	\$58,500
Infrastructure	Monthly	\$4,200
Operations	Monthly	\$4,400

- The transaction-based Fee Schedule below is based on the requested volume of article downloads; however, other transaction schedules to be considered include factoring the number of articles as well as downloads or the number and size of the reports. It should be noted that the schedule below to publishers is specific to the requirements of the PIRUS2 project for article-level metrics.

Transaction-based Fee Schedule

Total Article Downloads Per Month*	CPM Article Downloads Per Month	Monthly Fee
< 500 Thousand	\$0.80	\$400
500 Thousand - 1 Million	\$0.75	\$750
1 - 5 Million	\$0.70	\$3,500
5 - 10 Million	\$0.65	\$6,500
10 – 20 Million	\$0.60	\$12,000
20 – 40 Million	\$0.55	\$22,000
40 – 50 Million	\$0.50	\$25,000
50 – 70 Million	\$0.45	\$31,500
70 – 100 Million	\$0.40	\$40,000

100+ Million	\$0.35	\$0.35 / 1,000 downloads
--------------	--------	-----------------------------

\*Total article downloads per month represents the total number of article downloads requested for all articles in the scope of the reports requested. For example, an organization requesting usage statistics for 2 million downloads to consolidate into their own usage reports would equate to a \$3,500 monthly fee.

#### Example CCH charges to repositories and publishers.

##### Repositories

Appendix N provides examples of costs for typical institutional repositories, based on Scenario A in Scheme 1 above. Based on realistic assumptions about the number of full-text article downloads, the annual cost to a typical repository of using this service will be in the range £800-£2,000.

##### Publishers

The majority of publishers are likely to implement Scenario C in Scheme 1 above, i.e. they will be responsible for producing the final, article-level usage reports and will make these available to the CCH for harvesting. In this scenario, the publishers will pay an annual membership fee to the CCH, based on annual revenues as outlined in the Membership Fee schedule above and summarized in Appendix M. This annual fee will range from \$890 for the smallest publishers to \$66,900 for the very largest publishers.

While it is likely that most currently COUNTER compliant publishers would generate their own usage reports, many may find it more cost effective to use the CCH for this purpose. For this reason we provide in Appendix L. An estimate of the CHH costs for the implementation of Scenario A for a small, medium and large publisher. It should be noted that where the CCH is already generating the existing COUNTER usage reports for a publisher, the incremental costs of providing the PIRUS article reports will be significantly reduced

## **5.2 Software, standards and protocols development**

Work on developing 'software, standards and protocols' to support the prototype article level service, described above, was undertaken under the auspices of WP4.

The aim was to demonstrate that is technically possible to capture full-text article download usage events – by employing relevant protocols and standards - from various sources (publishers, repositories, etc.), and to consolidate those downloads to show the overall usage of articles.

Key objectives were:

- to achieve a means of providing normalised COUNTER-compliant statistical data at the individual article level for the main institutional repository softwares;
- to build on work done between publishers and the JISC community to provide a reliable basis of exchange of usage data by adopting the emergent standardised methods and protocols throughout, particularly SUSHI, OAI-PMH and OpenURL context objects;
- to develop a prototype Article Level Usage Statistics Portal to demonstrate the feasibility of consolidating/aggregating publisher and institutional repository usage statistics

The original PIRUS project identified three scenarios – with supporting protocols and standards - for the transmission of usage data and statistics:

Scenario (A): when a full-text article is downloaded, a message – raw usage data - is pushed out to a remote server

- protocol: 'tracker' – analogous to a server-side 'Google Analytics' for full-text article downloads
- standard: OpenURL key-value pair strings (URLs)
- candidate organisations: most repositories and some small publishers

**Scenario (B):** as full-text articles are downloaded, records of raw usage data events are stored locally and made available for harvesting by a remote server, on demand

- protocol: OAI-PMH – a protocol already familiar to repositories
- standard: OpenURL context objects (XML)
- candidate organisations : repositories

**Scenario (C):** as full-text articles are downloaded, records of raw usage data events are stored locally. Usage data is processed according to COUNTER rules, and made available for harvesting by a remote server, on demand

- protocol: SUSHI (Standardized Usage Statistics Harvesting Initiative Protocol) –familiar to publishers
- standard: proposed COUNTER-compliant AR1 report
- candidate organisations: publishers

In the context of PIRUS2, a ‘protocol’ defines a set of rules to send and receive messages between computers on the internet; while a ‘standard’ defines the rules for the content of those messages.

The standards and protocols, central to the project, are discussed in the following section.

## 5.2.1 Standards and Protocols

### 5.2.1.1 OpenURL standard

The use of the OpenURL for the purpose of exchanging usage data was first suggested by the MESUR project<sup>10</sup>.

#### OpenURL Context Objects

Work on OpenURL Context Objects (XML) has been taken forward in Europe under the ‘Knowledge Exchange’ –an initiative involving members of DEFF, DFG, JISC and SURFfoundation, as well a number of projects funded by those bodies (including PIRUS2 and OA-Statistik),

Leveraging that existing work, the OpenURL Context Object Specification<sup>11</sup> defined by OA-Statistik was employed as a basis for work undertaken supporting scenario (B).

#### OpenURL key-value pair strings

PIRUS2 has lead work on developing the standard for OpenURL key-value pair strings, used in scenario (A).

The OpenURL log entries are based on a subset of the NISO OpenURL 1.0 standard *KEV ContextObject Format*. An important requirement is that the OpenURL strings must be URL encoded, with key-value pairs separated by &. The elements to be transmitted, initially suggested by PIRUS2, are given in Appendix GAppendix B.

### 5.2.1.2 COUNTER-compliant AR1 report standard

Currently, COUNTER-compliant usage statistics report at the Journal level, e.g. Journal Report 1 (JR1) Report: Number of Successful Full-Text Article Requests by Month and Journal

Typically, publishers:

- Log download events as they occur
- Periodically process log entries according to COUNTER rules:
  - Stripping out robot accesses
  - Eliminating double click entries
  - Converting raw usage data to COUNTER-compliant monthly statistics
- Resulting statistics are shared with authorized parties via
  - MS-Excel/CSV files – manually downloaded
  - The SUSHI protocol – machine to machine

In order to extend the COUNTER-compliant reports to article level, PIRUS2 devised a tentative first version of an Article Report: AR1 Report: Number of Successful Full-Text Article Requests by Month and DOI, in MS-Excel (see Figure 1. AR1 example report, below) and XML formats.

Journal	Print ISSN	Online ISSN	Publisher	Platform	Article title	First Author Surname	Article Version	DOI	Jan-09	Feb-09	Mar-09	Total
Totals for all articles												
African Affairs	0001-9909	1468-2621	Oxford Journals	HighWire	Article title 1	Surname	Version of Record	10.1093/afraf/adn001	11	21	23	55
African Affairs	0001-9909	1468-2621	Oxford Journals	HighWire	Article title 2	Surname	Accepted Manuscript	10.1093/afraf/adn002	40	65	3	108
African Affairs	0001-9909	1468-2621	Oxford Journals	HighWire	Article title 3	Surname	Version of Record	10.1093/afraf/adn003	25	42	31	98
Toxicological Sciences	1096-6080	1096-0929	Oxford Journals	HighWire	Article title 4	Surname	Version of Record	10.1093/toxsci/kfp001	59	61	37	157
Toxicological Sciences	1096-6080	1096-0929	Oxford Journals	HighWire	Article title 5	Surname	Version of Record	10.1093/toxsci/kfp002	90	82	101	273

**NOTES**

- Columns A, D, E, I and J++ are mandatory
- Data is required for either Print ISSN or Online ISSN, but both may be provided if desired
- 'Article title' and First Author Surname' data are required only if the DOI is not available
- Article Version - highly recommended but optional - uses the terms proposed and defined by the "NISO/ALPSP Working Group on Versions of Journal Articles" ([http://www.niso.org/workrooms/jav/Recommendations\\_TechnicalWG.pdf](http://www.niso.org/workrooms/jav/Recommendations_TechnicalWG.pdf)) but alternative version names (or urls) are acceptable. Each version of an article should have its own separate record (row), showing its usage.
- Usage data should:
  - include successful full text requests (HTML plus PDF)
  - include Accepted Manuscript, Proof, Version of Record versions
  - exclude Authors Original and Submitted Manuscript Under Review versions
  - exclude internal use by publisher and host, downloads from LOCKSS caches and by robots listed at

Figure 1. AR1 example report

The development of a proposed COUNTER-compliant AR1 report is intended to permit publishers to transmit usage statistics to a central clearinghouse, supporting scenario (C),

### 5.2.1.3 The tracker protocol

When a user downloads a full-text Journal Article from any given system, that system pings a remote server (Central Clearing House) transmitting an OpenURL key-value pair string.

Two examples of OpenURL key-value pair strings transmitted:

```
138.250.13.22 - - [17/Oct/2010:04:04:44 +0100] "GET /tracker/?url_ver=Z39.88-2004&req_id=e02db545fbefd7d19bf24302a57f93ac&req_dat=Mozilla%2F5.0+%28compatible%3B+Googlebot%2F2.1%3B+%2Bhttp%3A%2F%2Fwww.google.com%2Fbot.html%29&rft.artnum=http%3A%2F%2Fdspace.lib.cranfield.ac.uk%2Fhandle%2F1826%2F3228&svc_val_fmt=info%3Aofi%2Ffmt%3Akev%3Amtx%3Adc&svc.format=application%2Fpdf&svc_dat=Unknown&rfr_id=dspace.lib.cranfield.ac.uk&url_tim=2010-10-17T03%3A04%3A42Z&rft_id=info%3Adoi%3Ahttp%3A%2F%2Fdx.doi.org%2F10.1016%2Fj.istr.2008.10.006 HTTP/1.1" 200 635 "-" "Java/1.6.0_21"
```

```
152.78.189.11 - - [20/Feb/2011:04:16:47 +0000] "HEAD /tracker/?url_ver=Z39.88-2004&url_tim=2011-02-20T04%3A16%3A45Z&req_id=urn%3Aaip%3A219.219.127.3&req_dat=Mozilla%2F4.0+(compatible%3B+MSIE+6.0%3B+Windows+NT+5.1%3B+)&rft.artnum=http%3A%2F%2Fprints.ecs.soton.ac.uk%2Fid%2Fprint%2F8671&svc.format=application%2Fpdf&rfr_id=prints.ecs.soton.ac.uk&rft.date=2003-12&rft.aulast=Schwanecke&rft.volume=91&rft.atitle=Broken+Time+Reversal+of+Light+Interaction+with+Planar+Chiral+Nanostructures HTTP/1.1" 200 - "-" "EPrints 3.2.5 (Stollen) [Born on 2011-01-17]"
```

These URL-encoded strings are not easily readable to the human eye but, containing standard elements, they are easy enough to process programmatically.

#### 5.2.1.4 The OAI-PMH protocol

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) “provides an application-independent interoperability framework based on *metadata harvesting*. There are two classes of participants in the OAI-PMH framework: *Data Providers* administer systems that support the OAI-PMH as a means of exposing metadata; and *Service Providers* use metadata harvested via the OAI-PMH as a basis for building value-added services.”<sup>12</sup>

Although intended for metadata harvesting, it is possible to re-purpose and use this protocol as the mechanism for harvesting usage events. In this context, the OAI-PMH acts as the transmission protocol wrapped around OpenURL Context Objects (described above). An example is given in Appendix H.

#### 5.2.1.5 The SUSHI protocol

The SUSHI protocol “defines an automated request and response model for the harvesting of electronic resource usage data utilizing a Web services and COUNTER-compliant usage statistics reports are framework”<sup>2</sup>

The protocol acts as a wrapper for transmitting COUNTER-compliant reports. In the case of PIRUS2, the protocol would transmit XML versions of AR1 reports.

### 5.2.2 Consolidating usage data from Publishers and Repositories

With a major technical aim of PIRUS2 being to demonstrate the ability to consolidate usage statistics from disparate sources for an article, a fundamental question is - How do we match usage events from publishers and repositories?

The answer is the DOI. It is the key metadata element, essential for consolidation purposes, and is mandatory in the proposed business model.

While it is impossible to give exact figures, we have ascertained from CrossRef that in excess of 95% of currently published articles have DOIs. Additionally, all the major publishers have deposited all of their backfiles back to volume 1, issue 1 for all their journals. Consequently, there is a total of 39.6 million journal article DOIs from 23,000 journals going back as far as 1665 (in the case of Philosophical Transactions from the Royal Society). As it is estimated that around 50 million articles have been published since 1665, we can say that around 80% of all articles ever published have a DOI. As the total number of articles published is growing by 1.5-2 million (3-4%) per annum, and virtually all of these have DOIs, the percentage of the archive that has DOIs is also growing and should be over 90% within the next 5 years.

So, the vast majority of journal articles can be individually recognized with certainty. Also, most repositories add DOI metadata to (some, but not all of) their records pertaining to articles. Where both publisher and repository statistics have supplied a DOI, identifying a match is easy and certain.

But how can publisher and repository statistics be consolidated where a repository hasn't catalogued the DOI?

Fortunately, in all cases, the Article title and first author surname are available in repository records. It is possible, in many cases, to use those to query and retrieve the DOI from:

- The CrossRef database using a variety of synchronous and asynchronous methods<sup>13</sup>
- The PIRUS2 database – where an article has been previously identified and is already known to the system

The query matching rate is dependent on a number of factors, including comprehensiveness of metadata in the CrossRef and PIRUS2 databases, matching technology, and data quality. It is difficult to put an exact figure on the success rate of author/title matching (the more sophisticated the matching algorithms, the higher the success rate) but we estimate that, realistically, matching can work in 85%+ of cases.

The best solution to the author/title matching problem is, of course, to avoid it altogether - by making sure that, where available, the DOI is routinely catalogued in the repository in the first place!

### 5.2.3 Gathering publisher usage events

The AR1 (see above) is the report format used for most of the data supplied to us by participating publishers. For input to the project, we accepted MS-Excel files from publishers. In the real world, data would be gathered using SUSHI

SUSHI was not considered appropriate, at this stage. To implement an extended SUSHI service incorporating article level reports would be technically challenging and time-consuming – for both publishers and the project. Furthermore, the AR1 standard is not yet an agreed **COUNTER** standard and is still under development.

#### 5.2.3.1 Publisher usage data

Data was processed and loaded into the PIRUS2 database, using a mix of manual processing and Perl scripting. At the end of the process we had usage records from publishers pertaining to 581,556 Articles across 537 Journals, representing 93,729,498 downloads across 2,743,839 records.

Data from the publishers was also used to populate the PIRUS2 database with Journal and Article Authority tables.

#### 5.2.3.2 Participating publishers

AR1 reports were received from ACS, Emerald, IOP, Nature, Oxford Journals, Springer and Wiley. In the case of Nature and Oxford Journals we were able to re-purpose data originally supplied for the UKSG-sponsored Journal Usage Factor project.

### 5.2.4 Gathering repository usage events

The original PIRUS project described, in some detail, how there are many different repository softwares, both Open Source, e.g. CDSware, **DSpace**, **Eprints**, **Fedora**, i-Tor, MyCoRe, OPUS; and Proprietary, e.g. Digital Commons (BePress), Digitool (Ex Libris).

Across the various systems, there are great variations in the way they work, different programming languages and platforms, and different methods of logging download events.

However, despite the underlying differences, there are a number of common requirements in terms of cataloguing items put into repositories. IRs commonly catalogue metadata including: **Title**, **Author(s)**, Abstract, Journal title, Volume(Number), Pages, ISSN, **DOI**, Bibliographic citation, **Resource type**, Local identifier. And all repositories investigated included **Title, Author and Resource type** metadata in their records.

The key to overcoming the underlying differences is to get usage data out in a standard manner as described in the section on standards and protocols above.

When considering repositories, we quickly put Scenario C (SUSHI) to one side:

- AR reports still under development, not yet a COUNTER standard
- The technology is complex and unfamiliar to repositories
- There are considerable auditing cost and data preservation implications in producing ready-made COUNTER-compliant reports

So, we turned our attention to Scenarios A & B, where repositories share raw usage data and the audit and preservation burdens sit with the Central Clearing House.

With so many repository softwares extant, as a project, we couldn't carry out development on all of them. Following on from PIRUS outcomes, we decided to focus on DSpace, GNU Eprints and Fedora - all open source, and comprising the underlying software for around two-thirds of repositories.

#### 5.2.4.1 Repository software plug-ins/extensions to expose article level usage data.

Repository extensions were developed for:

- DSpace – developed by @mire
- Eprints – developed by Tim Brody, Southampton University
- Fedora – developed by Ben O’Steen, Oxford University

Links and downloads to these extensions are available on the PIRUS2 project web site

### DSpace

Patches were developed for DSpace v1.6.2 for both the tracker and OAI-PMH, scenarios A & B.

We tested both approaches and found that both worked well. However, for a number of reasons, we decided to adopt just the tracker approach for wider testing:

- For simplicity – it is easier and less expensive for a service to adopt a single workflow instead of multiple workflows
- Data privacy – the tracker pushes usage data straight to a central clearing house, without having to worry about authorization/authentication issues which would have to be considered for OAI-PMH
- Looking to future auditing and preservation implications, the tracker is a simpler and cheaper option to implement

### Eprints and Fedora

Plug-ins were developed for both Eprints and Fedora employing the tracker approach.

#### **5.2.4.2 Repository usage data**

The PIRUS2 server logs have been receiving, on average, 16-18MB of raw usage data a week from participating repositories.

Using a series of Perl scripts, we processed the usage data from those logs:

- Filtering according to COUNTER rules to eliminate Robots and Double clicks
- Processing into monthly statistics
- loading into the PIRUS2 demonstrator database

As the main aim of the demonstrator was to show that it is possible to consolidate publisher and repository statistics, we focussed on loading data from repositories that matched existing records – based on DOIs - for articles from our participating publishers. This represented 31,272 downloads across 5,574 records where *publisher and repository statistics have been successfully consolidated!*

The remaining repository usage data – where a DOI was not supplied - was queued up for further processing.

We have a run a series of tests where we have been able to retrieve DOIs for some of the repository data from the CrossRef database, using their Article Title/Author surname query facility. This confirms that more of the repository data can be matched with publisher data and loaded into the database.

However, the Article Title/Author surname query did not successfully return DOIs in every case. This still leaves a hardcore of repository data with no DOI available to enable simple matching with publisher records. This needs further consideration.

#### **5.2.4.3 Participating repositories**

The following repositories installed the relevant PIRUS2 extensions to their repository software and participated in providing usage data to PIRUS2:

DSpace:

- Cranfield CERES
- Harvard DASH
- University of Edinburgh ERA

Eprints:

- Bournemouth University Research Online (BURO)
- University of Huddersfield Repository

- University of Salford Institutional Repository
- Southampton ECS EPrints Repository

Fedora:

- Oxford University Research Archive (ORA)

An attempt was also made to install and trial the Fedora plug-in at the University of Hull – but the attempt was unsuccessful. See Section 5.2.7.6, below.

### 5.2.5 The PIRUS2 database

For the PIRUS2 we decided to use a MySQL relational database to store usage data received from publishers and repositories. MySQL is open source and a technology familiar to the team.

The database was designed to support the business model initially proposed - where article level statistics are consolidated from both publishers and repositories.

The Statistics table stores the monthly usage events for articles, supported by and related to Article, Supplier (publishers, repositories), Platform (e.g. Eprints, Insight) and Journal authority tables.

The main tables and some of the relationships between those tables are shown diagrammatically in Appendix J.

### 5.2.6 A Prototype Article Level Usage Statistics Portal

The demonstrator is a proof of concept - an arena intended to illustrate the technical feasibility of gathering, consolidating and re-exposing usage events from disparate sources.

The demonstrator comprises:

- A web user interface, written in PHP
- Downloadable example reports

It sits behind an authentication/authorisation barrier to allay privacy concerns. But, for authorized users, it is the window into the PIRUS2 MySQL database, described in the section above, which holds the data gathered from publishers and repositories:

The main features of the portal are:

- The home page (see Figure 2 below) provides summary information:
  - the number of articles and journals indexed
  - overall totals of download events recorded from publishers and repositories
- The Search facility makes it possible to find individual articles or groups of articles by:
  - DOI
  - Title/Author
- Journals related to the articles can be browsed
- A number of reports can be generated:
  - AR1j – this is a variant of the AR1, with usage events restricted to one journal at a time to reduce the report sizes. Its main purpose was to allow easy cross-checking that the data exposed from the PIRUS2 database matched the original data supplied by publishers
  - AR2 – this is a report intended for article authors – showing usage consolidated from publishers and repositories for their articles
  - AR1ir – this is a report intended for Institutional Repositories – it contains COUNTER-compliant usage statistics synthesized from their own raw usage data
  - IR DOI Update – another report intended for IRs – supplying the DOIs for articles where the repository record does not currently hold them



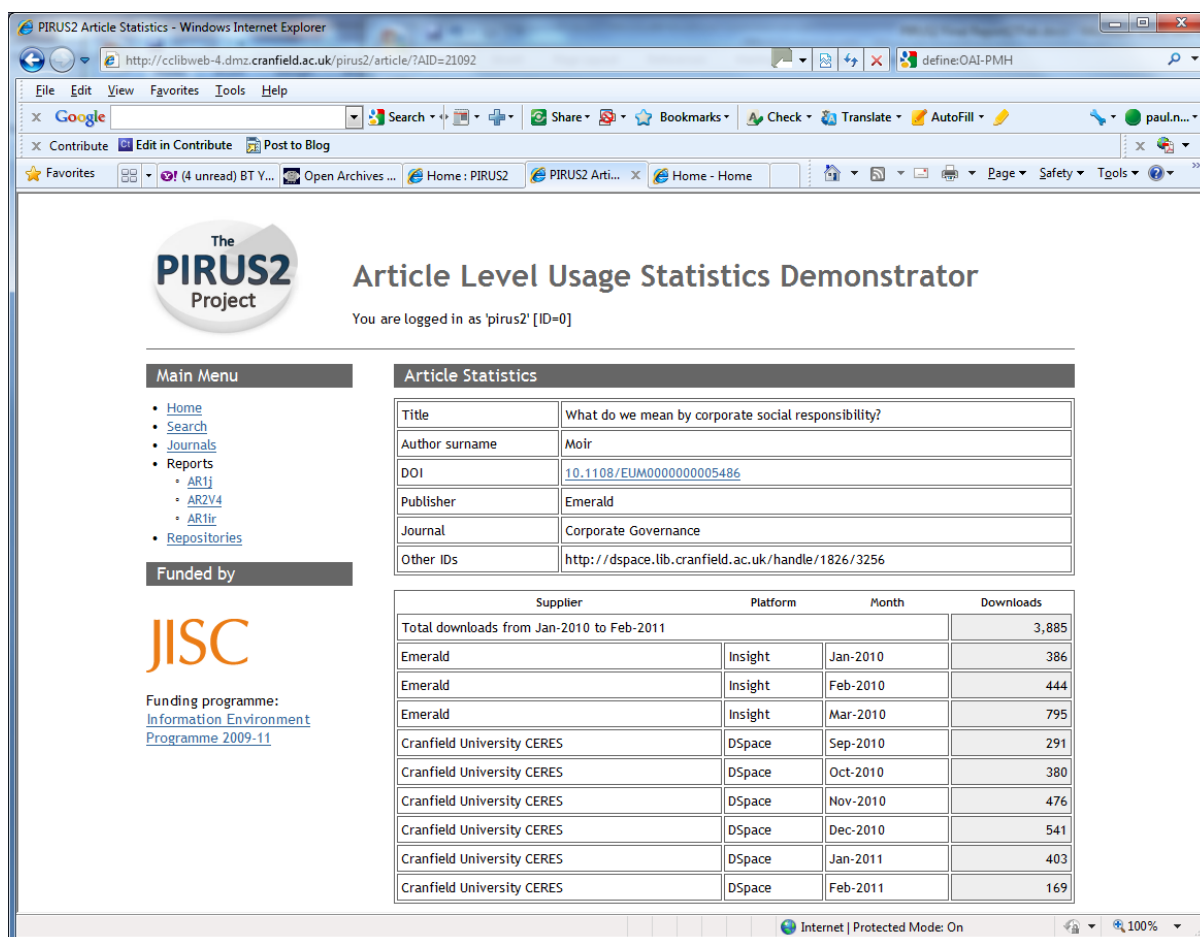


Figure 2. PIRUS2 demonstration portal home page

Each of the reports can viewed in a web page in the portal or downloaded for use locally as MS-Excel/CSV files (See Appendix K for some screenshots of the various reports).

Of course, in the real-world, these reports would also be available via SUSHI and other Web Service protocols.

## 5.2.7 Discussion of technical issues

### 5.2.7.1 Consolidation of usage statistics for journal articles

Consolidation of usage statistics for articles from publishers, repositories and other entities is possible! Using the DOI the matching process is certain and easy.

### 5.2.7.2 SUSHI protocol

SUSHI is frequently mentioned - throughout this report - as the key protocol for transmission of usage statistics in the proposed service. However, before it can be used, work will need to be carried out to extend SUSHI to support article level statistics reports. This has not been possible, so far, as the proposed article level reports have been in a state of flux and subject to re-definition and refinement during the entire course of the project.

Once there is agreement between COUNTER and relevant stakeholders regarding the formats desired article level reports, developing and extending SUSHI to the article level will require a project in its own right.

### 5.2.7.3 OpenURL key-value pair strings

During testing, it became apparent that that we could simplify the OpenURL key-value pair string specification – there is no need to endlessly transmit either the DOI or Article title and Author

Surname and other supplementary metadata (the latter adding considerably to data volumes and to the difficulty of parsing the OpenURL string).

Instead, it would be better and simpler just to supply the OAI-PMH identifier, instead, which (in conjunction with the OAI baseURL of the repository) would allow us to look up and load the available metadata for an article from the source repository.

This lookup would only have to be performed once for each article, thereafter the article metadata would already be in our system.

The updated specification (see Table 1, below) would be much simpler to implement in repository softwares than the initial specification given in Appendix G.

Element	OpenURL Key	OpenURL Value (example)	Notes
	url_ver	Z39.88-2004	Identifies data as OpenURL 1.0. String constant: Z39.88-2004 (Mandatory)
Timestamp	url_tim	2010-10-17T03%3A04%3A42Z	Date/time of usage event (Mandatory)
Client IP address	req_id	urn:ip:138.250.13.161	IP Address of the client requesting the article (Mandatory)
UserAgent	req_dat	Mozilla%2F4.0+%28compatible%3B+MSIE+7.0%3B+Windows+NT+5.1%3B+Trident%2F4.0%3B+GoogleT5%3B+.NET+CLR+1.0.3705%3B+.NET+CLR+1.1.4322%3B+Media+Center+PC+4.0%3B+IEMB3%3B+InfoPath.1%3B+.NET+CLR+2.0.50727%3B+IEMB3%29	The UserAgent is used to identify and eliminate, by applying COUNTER rules, accesses by robots/spiders (Mandatory)
Article OAI identifier	rft.artnum	oai:dspace.lib.cranfield.ac.uk:1826/936	(Mandatory)
MIMEtype	svc_format	application%2Fpdf	(Highly Recommended)
Source repository	rfr_id	dspace.lib.cranfield.ac.uk	(Mandatory)

**Table 1. PIRUS2 OpenURL key-value pair specification - updated**

#### 5.2.7.4 Article Report 1 (AR1)

Here again, it became apparent that we could simplify and compact the AR1 report for transmission of usage statistics from publishers to the central clearinghouse, suggested in section 5.2.1.2, above.

Journal title, print and online ISSNs, article title and author metadata can be retrieved – as a one off operation - from CrossRef by using the DOI. There is no need to endlessly re-transmit these fields. A simpler updated version of the proposed AR1 is shown in Appendix B.

#### 5.2.7.5 Robots

Robots were filtered according to the COUNTER standard list of exclusions (see Appendix I), current at the time.

It is noteworthy that, as a result of work undertaken by PIRUS2 and European colleagues under the Knowledge Exchange initiative, the COUNTER robots exclusion list has recently been updated and expanded to incorporate a much more comprehensive list of exclusions, available as part of the Release 3 Code of Practice<sup>14</sup>.

#### 5.2.7.6 PIRUS2 Repository software extensions

For project purposes we oversaw development of patches and plug-ins for DSpace, Eprints and Fedora. These required individual institutional repositories to download and install them to participate in PIRUS2.

The patches for DSpace and plug-in for Eprints worked well and proved easy to install across all the participating DSpace and Eprint repositories.

The plug-in for Fedora was successfully installed and worked well at the University of Oxford. However an attempt to use the plug-in at the University of Hull was not successful.

Fedora is highly flexible and customizable, and local implementations can vary enormously. The plug-in was developed on the assumption that the local Fedora system was already configured and customized to output data to a local log of usage events triggered by a download request – however, at Hull this was not the case.

Further work has been carried out, by a PIRUS2 Fedora Working Group, to investigate the possibilities of enabling PIRUS2 functionality across a wider community of Fedora repositories. (See Appendix O.)

There is no guarantee that these extensions will continue to function, when new versions of repository software are released. And, even if they do, the individual repositories will still have to reinstall the extensions again after a software upgrade.

If/when article level reporting becomes adopted as a global COUNTER standard, it will be vital that this new functionality is embedded into the core software of all major systems, open source and proprietary, and available 'out-of-the-box', requiring only configuration changes to switch on or off. However, at this stage, this must be regarded as a long term aspiration.

#### **5.2.7.7 PIRUS2 repository usage data**

We have collected - and are still collecting - a considerable amount of usage data from repositories. Only a small portion of this data was used in proving that consolidation of publisher and repository usage statistics is possible.

However, the availability of this (growing) dataset presented an opportunity explore the possibilities of developing a UK institutional repository usage statistics service (IRUS-UK). Such a service could:

- Collect raw usage data from UK IRs for all item types within repositories
- Process those raw data into COUNTER-compliant statistics and return those statistics back to the originating repositories for their own use
- Give JISC (and others) a nation-wide picture of the overall use of UK repositories – demonstrating their value and place in the dissemination of scholarly outputs (“Making scholarly statistics count in UK repositories”)
- Offer opportunities for benchmarking
- Potentially act as an intermediary between UK repositories and other agencies – e.g. global central clearinghouse, national shared services

So, we have re-purposed and re-used usage data collected by PIRUS2 to create a new demonstrator based just on UK Institutional Repository usage data – IRUS-UK. This has:

- Required a modest re-design of the existing PIRUS2 database (to remove the journal authority table and potentially accommodate multiple resource types, though at this stage most data still only pertains to articles)
- Re-used much of the code used in the Perl scripts to ingest data into the existing PIRUS2 database
- Required a re-write of the web interface to illustrate possibilities (though again much of the underlying code has been re-used)
- (See Appendix O for more detail)

The outcomes from this demonstrator look very promising, so the next logical steps would be to undertake a project to:

- Further develop the technical model
- Widen the scope to all resource types held within UK IRs and extend participation to a larger number of UK repositories

- Carry out research into the organizational, economic and issues that would need to be addressed in order to establish whether a national service would be feasible
- Explore the place of a national IR usage statistics service in the context of
  - a global central clearing house service for article level statistics
  - existing registers of repositories
  - EThOS, the BL one-stop shop for theses
- Consider co-ordinating efforts with SURF and others (perhaps through the Knowledge Exchange initiative) who are carrying out similar work with a view to establishing a network of interoperable national repository usage statistics services - national centres are a natural unit of administration, and working within national boundaries would mitigate many of the problems of differing privacy laws across countries

### **5.3 Dissemination and advocacy**

The objectives of the dissemination and advocacy strategy were: to inform all the major stakeholder groups (repositories, publishers, authors, funding agencies) about the project and the progress being made; to actively involve these stakeholder groups in the PIRUS2 as it developed; to make the wider scholarly information community aware of the project; and to secure broad support for the project outcomes and recommendations.

The channels used to achieve these objectives were:

PIRUS2 website ( maintained by Cranfield University), which provided the basic information on PIRUS2, including the Project Plan and also provided regular updates on progress and was a major channel for promoting, and collecting registrations for, the End of Project Seminar

PIRUS2 Steering Committee: this 17-member group included representatives of publishers, institutional repositories and subject repositories, as well as members of the project team itself. Chaired by Dr Hazel Woodward, it met 6 times during the course of the project and provided feedback on all aspects of the project..

PIRUS2 Publisher Forum: Chaired by Peter Shepherd, this group meet regularly throughout the project and provided a very useful forum for discussing the proposed organizational and economic models for the CCH, helped develop these models and also provided useful feedback on formats of the proposed article usage reports.

Presentations at appropriate conferences and workshops in the course of the project. A special effort was made to arrange presentations at conferences attended by the main stakeholder groups. Particularly noteworthy in this respect were the Open Repositories 2010 in July 2010 in Madrid and the 30<sup>th</sup> Annual Charleston Conference in Charleston, SC, USA.

Surveys of key stakeholder groups. Surveys were carried out of publishers, repositories and funding agencies at appropriate stages in the project.

Articles in appropriate periodicals and other publications, including Against the Grain, Serials and Learned Publishing

End of Project Seminar (23 February 2011, London). This attracted 71 delegates from the publisher, repository, research funding and library worlds. Feedback on the seminar was very positive and it provided an excellent forum for discussing and refining the project outcomes.

It was clear by the end of the project that, while the dissemination strategy was very successful in getting the project widely known, further advocacy work is required among publishers and repositories to convince them of the value of providing article level usage statistics. Acceptance of the CCH, a central outcome of PIRUS2, will be very dependent on its costs and charges, as both constituencies are very sensitive to these. The economic model for the CCH only became clear towards the end of the project, once the organizational structure and data flow issues were settled. More time is needed to test and refine this model.

## 6 Outputs and Results

The main outputs of PIRUS2 are:

- a fully tested prototype aggregated statistics service employing agreed first versions of Standards and Protocols; DSpace, Eprints & Fedora Software plug-ins; Software to process and filter OpenURL usage data according to COUNTER rules;
- a set of reports on the business model for the prototype aggregated statistics service, including a list of organizations that meet the required criteria for the central clearing house(s), an assessment of the costs for repositories and publishers and the running the central clearing house(s); proposals for dealing with legal issues, results of market research surveys;
- feedback from authors, publishers, repositories, and research funding agencies on the proposed model for the aggregated statistics service
- an end-of-project seminar to share the results, knowledge and experience acquired in the course of the project with the stakeholder communities

Additional outputs of PIRUS2, as a result of further work described in Appendix O, are:

- further developments to the proposed organisational, economic, political and technical models – based around a more distributed model of feeding usage statistics to the CCH via a number of national or regional agencies, illustrated by:
  - a UK institutional repository usage statistics demonstrator service, which could very cost-effectively consolidate article statistics for all UK IRs and act as a single point of transfer for those statistics to the CCH, and provide extra opportunities to furnish UK IRs with COUNTER-compliant statistics for all their item-types (not just articles), as well as offering opportunities to demonstrate the impact and value of IRs
- a generic set of guidelines for implementation of PIRUS2 functionality across Fedora repositories
- a finding that usage of articles hosted by institutional repositories is rather high. Over the 7-10 month period of the project during which usage data was collected for articles hosted by the 6 participating repositories, there were 527,224 downloads of 6,089 articles; an average of 86 downloads per article

## 7 Outcomes

The PIRUS2 project is the first to propose not only a standard for measuring online usage that would apply to both publishers and repositories, but also a supporting organizational structure for recording, consolidation and reporting usage of individual articles. While the project has provided a workable technical model for doing so, questions remain, especially at a time of more limited funding, whether there is sufficient support for taking forward the organizational and economic models that have been proposed. While there is strong evidence that authors appreciate and use individual article usage statistics where they are available, there is no evidence that they would be prepared to pay for this service. Publishers and repositories both have concerns about the potential costs and there continues to be tension between these two stakeholder groups.

Most of the major objectives of PIRUS2 have been met. It has developed a workable technical prototype for recording, consolidating and reporting global usage in a standardised way at the individual article level, based on data from a variety of sources operating in a very diverse technological environment. Organizational, economic, intellectual property and political issues have also been addressed and a business model proposed for the CCH that will be required. .

## 8 Conclusions

PIRUS2 has demonstrated that:

- It is technically feasible to create, consolidate and report usage at the individual article level based on usage data from a range of sources based on different platforms
- A practical technical/organizational model for a Central Clearing House that can handle the large volumes of usage data and associated metadata that are involved
- An economic model that provides a rational and cost-effective basis for allocating the costs of the CCH among the repository, publisher and other clients that would use it

While the project has also shown that Individual article usage statistics are a *potentially* valuable tool for several important stakeholders involved in research and the dissemination of its outputs, most have yet to be convinced that this is essential information, rather than simply nice-to-have information:

- researchers/authors: are interested in monitoring online usage of their publications. Evidence from PLoS and other surveys has shown that authors find this individual article usage reports useful.
- Publishers appreciate that providing reliable usage statistics at the individual article level will enhance the service they offer to their authors, many have reservations about implementing a standard that also applies to repositories
- while repositories, 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, they appear to be reluctant to incur the costs associated with adhering to a global standard
- research institutions, are increasingly required to demonstrate the value of the research and researchers that they support, and appreciate the potential value of article usage statistics to achieve this
- funding agencies: who are seeking more quantitative, transparent ways of assessing the performance and impact of the research projects that they fund

The following broad conclusions can, therefore, 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
- there is a workable, cost-effective technical/organizational/business model for the CCH facility
- flexibility will be required in the usage reports output by the CCH, so that its users can integrate usage data with other categories of metrics to provide insights into the reach, impact and value of research articles
- further advocacy will be required to persuade the major stakeholder groups that the prototype developed in this project should be fully implemented.

## 9 Implications

This work has the following policy implications:

- a. For COUNTER: further improvements and extensions to the COUNTER Code of Practice will be required. 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 participation of COUNTER in the strategic management of the CCH.
- 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 benefit from such a standard and accept that there will be additional costs for doing so.
- c. For Authors: credible and transparent global usage statistics on an individual article level provide authors with a new metric that allows them to see how their research outputs are being used. Authors should take these into account, along with citation data and other measures.
- d. For Publishers/Vendors: PLoS has found that its authors welcome credible usage statistics for their articles, There is evidence that they want to have access to such data for their other articles and are likely to put pressure on publishers to participate in the process to provide it. Providing individual article usage statistics would give 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, implement new standards such as ORCID, 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 e.g. 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, but also to decide whether to support related standards, such as ORCID, which are important for this project

Furthermore, before a fully-fledged, comprehensive usage statistics consolidation service can be launched, a number of issues, beyond the control of the project, still need to be addressed:

- d. SUSHI: the proposed article level reports will need to be endorsed by COUNTER, and extensions to the COUNTER-SUSHI schema – to accommodate required article level metadata elements – will need to be endorsed and adopted by NISO.
- e. ORCID: reliable identification and attribution of individual authors remains problematic, making it – currently – virtually impossible to consolidate usage across multiple articles for any given author. The adoption of the ORCID system, due to launch as a beta service at some point in 2011, “will, from the start, enable 3rd parties to build value added services using ORCID infrastructure”<sup>15</sup>.
- f. Institutional Identifiers: Although identifying institutions is less problematic than identifying authors, nevertheless, the eventual outcomes from the NISO I<sup>2</sup> Working Group<sup>9</sup> will improve the efficiency and potential for interoperability of an article level usage statistics service.

## 10 Recommendations

The PIRUS2 project has achieved many of its aims and objectives. However, a number of outstanding tasks remain to be completed in order to take this work forward and lead to the creation of a global article level consolidated statistics service. There is more work to be done to:

- achieve formal acceptance of a CCH from all stakeholder groups
- further developments to the proposed organisational, economic, political and technical models – based around a more distributed model of feeding usage statistics to the CCH via a number of national or regional agencies, illustrated by:
  - a UK institutional repository usage statistics demonstrator service, which could consolidate article statistics for all UK IRs and act as a single point of transfer for those statistics to the CCH, and provide extra opportunities to furnish UK IRs with COUNTER-compliant statistics for all their item-types (not just articles), as well as offering opportunities to demonstrate the impact and value of IRs
- roll-out patches or, better still, embed PIRUS2 functionality out-of-the-box into repository softwares

The recommendations of the project team are, therefore as follows:

- a. To JISC: PIRUS2 has developed a costed prototype service that capable of creating, recording and consolidating usage statistics for individual articles using data from repositories and publishers. Further feedback is required, however, to demonstrate with confidence that there is sufficient support for full implementation.
  - Organizational: while it is unlikely that there will be widespread implementation of PIRUS2 by publishers in the immediate future, due to cost concerns, there is a strong case for implementation of ‘IRUS’ the Institutional Repository Usage Statistics service, based on the technical and organizational model proposed in this report. Unlike the publishing world, there are currently no standards for usage statistics from Institutional Repositories. Adoption of the propose IRUS model would provide, for the first time,

such standards. For these reasons, we recommend that JISC should support the implementation of IRUS.

- Economic: the economic models for supporting the central clearing house are reasonable and should form the basis for going forward, both for publisher and for repositories
- Political: support for the outcomes of PIRUS2 among publishers and institutional repositories is weak. JISC could play an ongoing role in trying to build this support.
- Statistical: while detailed statistical analysis of usage was not one of the objectives of PIRUS 2, the article download figures for the 6 institutional repositories that participated in the project indicate that usage of articles in repositories is significant and merits more rigorous statistical analysis.

The PIRUS project team recommends that JISC considers funding further research in the short term, while the project has momentum, to address the issues described above.

- b. To COUNTER: expand the mission of COUNTER to include usage statistics from repositories; consider implementing the new PIRUS Article Reports as optional additional reports; modify the independent audit to cover new reports and processes. Use the fact that there is growing demand from authors for individual article usage reports to encourage publishers to provide them, based on the PIRUS2 standards.
- c. To repositories: consider participating in the proposed IRUS service and provide individual item level usage reports
- d. To publishers/vendors: accept, in principle, the desirability of providing credible usage statistics at the individual article level; implement the new PIRUS article reports for their own usage reporting to authors
- e. To repository software vendors/developers: accept, in principle, the desirability of incorporating PIRUS2 tracker functionality into their "out-of-the-box" software



## 11 References

- <sup>1</sup> <http://www.orcid.org/content/orcid-beta-scope>
- <sup>2</sup> <http://www.niso.org/workrooms/sushi>
- <sup>3</sup> <http://www.jisc.ac.uk/whatwedo/programmes/pals3/pirus.aspx>
- <sup>4</sup> <http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/usagestatisticsreview.aspx>
- <sup>5</sup> <http://www.driver-repository.be/media/docs/KEIRstrandreportUsageStatisticsFINALFeb07.pdf>
- <sup>6</sup> <http://www.orcid.org/>
- <sup>7</sup> <http://roar.eprints.org/>
- <sup>8</sup> <http://www.opendoar.org/>
- <sup>9</sup> <http://www.niso.org/workrooms/i2>
- <sup>10</sup> <http://www.mesur.org/MESUR.html>
- <sup>11</sup> [http://www.dini.de/fileadmin/oa-statistik/projektergebnisse/Specification\\_V5.pdf](http://www.dini.de/fileadmin/oa-statistik/projektergebnisse/Specification_V5.pdf)
- <sup>12</sup> <http://www.openarchives.org/OAI/openarchivesprotocol.html>
- <sup>13</sup> [http://www.crossref.org/help/Content/04\\_Queries\\_and\\_retrieving/CrossRef%20Data%20Formats.htm](http://www.crossref.org/help/Content/04_Queries_and_retrieving/CrossRef%20Data%20Formats.htm)
- <sup>14</sup> [http://www.projectcounter.org/code\\_practice.html](http://www.projectcounter.org/code_practice.html)
- <sup>15</sup> <http://www.orcid.org/content/orcid-beta-scope>

## 12 Appendices

### *Appendix A Publisher feedback on proposed Individual Article Usage Reports*

This summarises the feedback from publishers on the kind of individual article usage reports that they would find useful as an outcome of PIRUS2. Feedback was sought from major scholarly publishers that between them cover all the main disciplines, and who also represent an international geographical spread, as well as commercial and not-for-profit organizations. The following publishers responded to the survey:

- ACS Publications
- Emerald
- IOP Publishing
- Nature Publishing Group
- OUP
- Springer
- Wiley

#### **Introduction**

PLoS, the Public Library of Science, have been providing their authors with individual article usage statistics, covering usage on their own platform, since 2009. In a recent author survey they included questions about these article level metrics. This is the first large scale survey of author attitudes to having access to this information, and I list below the relevant questions and the responses from authors in PLoS Biology:

1. How aware were you that Article Level Metrics existed (prior to reading it here)?  
- 75% responded that they were either 'very' or 'moderately' aware (33% very, 42%moderately)
2. Is it clear what information is available under each article tab?  
- 82% responded that they were either 'very' or 'moderately' aware (39% very, 43% moderately)
3. How useful do you find article-level metrics?  
- 71% responded that they were either 'very' or 'moderately' useful (40% very, 31% moderately)

In addition there were 131 free text comments received in the survey, of which representative examples are listed below.

- 1) I will include the data on my CV. I have also already used it to see how popular our article is compared to other work.
- 2) I would like to be able to SEARCH for the highest rated/downloaded/cited articles in my area (Ecology).
- 3) I love watching my metrics go up!
- 4) For my annual merit review, I listed the number of article views and downloads to show the impact of my paper.
- 5) One of PLoS' strengths may also be a limitation: Because the articles are open access, I suspect that more people may have received my paper e.g., by email than directly from PLoS One's website, therefore skewing the metrics statistics.
- 6) I have not yet thought much about Article-Level Metrics, so do not have any useful suggestions.
- 7) I support the idea, but have not used the data.
- 8) Really cool, all journals should do

More information on the PLoS individual article metrics may be found on the PLoS website at: <http://article-level-metrics.plos.org/>

While the evidence from PLoS indicates that authors appreciate and use the individual article usage statistics to which they have access, we were interested in obtaining some insights from the established journal publishers as to how their authors would respond to receiving individual article usage statistics from the publisher.

The PIRUS2 technical model that has been developed has demonstrated much flexibility in terms of the types of individual article usage reports, provided the required usage data and associated metadata is available. An important challenge, therefore, is to prioritise the types of report it should deliver, based on input from authors, publishers and repositories.

In the PIRUS2 Demonstrator, to which you already have had access, you will find an example of the basic information on individual article usage that can be provided from the system we have set up. This information has been generated from the test usage data that you provided to us. In addition to this information on the most used articles in each journal, we envisage providing additional author-oriented usage reports and provide an example of one of these in Question 3 below.

## Summary of survey results

### Question 1: Have you received requests from your authors for article-level usage statistics for any of your journals?

a) Yes 7

b) No 0

#### Comments:

- Interest is increasing strongly
- Although we have not received many requests, the number of requests is increasing particularly for our OA journals. This will be increasingly important for younger scientists looking to differentiate themselves and their work in some way. We have in the past provided authors with statistics relating to usage compared to other papers in the same journal.
- Compared to the amount of articles we publish the requests are very low. Our internal surveys showed that authors would find these statistics interesting and nice-to-have, but not necessarily a must-have.

### Question 2: For what purposes do you think authors would use article-level usage statistics? (Responses are ranked in order of preference)

- to demonstrate impact of research to colleagues and management 1
- to compare usage with other authors in the field 2
- for general interest 3

#### Comments

- We do have some concerns about gaming, for example, we often receive requests via help desk from authors requesting article statistics to justify a visa to travel to the US. A PIRUS CoP would need to provide adequate safeguards regarding this issue. We also feel that there should be no requirement to provide statistics to any individual or organization other than authors as we see the provision of these figures, and potential value-adds associated with them, as an author benefit and an area for competition between publishers.
- Every user in the research cycle will be interested, from the author, his/her tutor right up to the heads of the university and funding agencies.

### Question 3: Do you think that your authors would find the format proposed below for the article usage statistics useful?

Source of usage	Jan-09	Feb-09	Mar-09	Total
Publisher	152	226	143	521
Host 1	23	31	29	83
Host 2	15	20	18	53
Host 2	10	15	12	37
<b>Total</b>	<b>200</b>	<b>292</b>	<b>202</b>	<b>694</b>

NOTES

- Article title data is highly recommended but optional
- Usage data should:
  - include: successful full text requests (HTML plus PDF)
  - include: Accepted Manuscript. Proof. Version of Record versions
  - exclude: Author's Original Manuscript and Submitted Manuscript Under Review versions
  - exclude: internal use by publisher and host. downloads from LOCKSS caches. and by robots

a) Yes 4

b) No 3

**Comments:**

- Some way of summing Author downloads would be useful.
- Subject to how authoritative it might, how all data is stored and where it is accessed. It makes most sense to make it accessible from the author article information on the publisher site then linking out to some external data repository
- Publishers should decide how and where to make available (e.g. on website)
- Data not imperatively be in excel format
- Important that an Industry Standard should be applied (like COUNTER) to make figures comparable
- Authors would prefer one report for all their articles
- The concept of providing these statistics to authors is useful as long as the data is standardised and transparent. The particulars would need to be worked out.

**Question 4: How would you prefer to provide article usage statistics to your authors?**

a) Directly from our own website 5

b) Via a central service 2

c) Via the author's institution 1

**Comments:**

- There are lots of caveats here – the need for confidentiality, ease of access, authority of the data, how and where the data is hosted; unbiased, independent source of data endorsed by the publisher.

- Industry standard statistics would be ideal though.

**Question 5: How frequently do you think authors would like to view article usage statistics?**

- a) Daily              1    
b) Weekly             1    
c) Monthly            1    
d) Annually           0    
e) On demand        4

Comments

- We think our authors would like to be able to retrieve or receive usage statistics on a monthly basis – triggered by their request. We also think that they would like the statistics to show monthly usage – see comments below.
- Thinking about a workable system from our perspective, we might envision a system where an author could request usage data for an article on demand, and would see usage on a monthly basis for the previous 2 years. We wouldn't be able to commit to such a system until we have a better understanding of whether it can be developed. In the beginning we think authors would want to view the data frequently, then might lose interest and only check periodically (unless perhaps they were trying to game the system, in which case they might want to see the immediate effects).
- Self service will be essential, figures to be updated monthly
- Frequency will depend on where the author is in their research cycle. To make it most useful it would also require some kind of comparative analysis with similar authors in the same research area.

**Question 6: For how many years after an article has been published do you think individual article usage statistic should be made available?**

- a) 2 years   1    
b) 5 years       
c) 10 years          2    
d) Other

- For the humanities, 15-20 years would be ideal
- Ongoing
- Our Full Text Access figures show that 80% of usage is to current and previous 5 years, and 95% to current and previous 10 years. 10 years would catch classic papers in SSH.
- This may be discipline-dependent but in some fields the impact of some studies peaks more than 5 years after publication.

Comments

- This data should be available in some form to all researchers at any point in future. Current usage (say the recent 2-5 years) should be updated in real time and then updated annually/semi-annually thereafter.
- It should be handled like citations
- Capturing and storing usage data on an article-by-article basis is quite costly – especially given that some authors may want retrospective information, others

prospective, and still others some combination of the two. The 2-year time frame seems a reasonable one to preserve the detailed information which could be provided to authors

**Question 7: Do you think that the availability of individual article usage statistics from PLoS will increase the demand from authors for such statistics for articles published in your journals?**

YES                    4

NO                      3

Comments:

- Ambitious academics will seize on whatever they can.
- It already is!
- But this is not true for all subject areas (e.g. SSH).

**Discussion and Conclusions**

The feedback obtained in this survey indicates that all of the participating publishers have received requests from their authors for article-level usage statistics. An earlier PLoS survey of their authors already demonstrated that the majority of their authors find the article-level usage statistics provided by PLoS useful. The purposes for which authors use, or would use these statistics is less clear, although the responses to Question 2 above show that publishers feel that the main purpose would be to demonstrate the impact of their research to colleagues or to management.

In terms of the format in which author usage statistics could be provided, the publisher response to the proposed report format (Question 3 above) was mixed, with some publishers favouring a standardised format, while others prefer more flexibility. One publisher said that authors would prefer a report that listed all of their articles and their usage (which would, in fact, be possible in the proposed report).

There are two rather strong messages from the publishers. First, that they would like to be the provider of usage data to authors in their publications, and second, flexibility in the format and frequency of delivery of this information is desirable. Furthermore the majority of the publishers think that the usage data should be available for a relatively long period of 10 years plus.

Opinion is divided as to whether the availability of article level usage statistics from PLoS will increase demand for such statistics from other publishers. Those publishers with a strong presence in biomedicine think it will, while those with a stronger presence in the social sciences and humanities think it will not.

**Recommendations**

The feedback from the PLoS survey and from this publisher survey indicates that authors, on the whole, would value having available usage statistics for their own articles and that demand for this is likely to increase. There is less agreement on the format in which publishers and authors would like to have this information, but this is not unusual. The situation was similar when the COUNTER usage reports for librarians were launched in 2003. At that time the decision was taken to launch the usage reports and refine them as a result of further feedback based on usage. This approach has proven successful and cost-effective over the years. A key objective at this stage must, therefore, be to ensure that the usage data and the associated metadata are captured at a sufficiently granular level to allow flexibility on the creation of the reports.

## **Appendix B Proposed AR1 report, transmitting statistics from publishers to CCH – updated**

<b>PIRUS Article Report 1: Number of Successful Full-Text Article Requests by Month and DOI</b>				
<Customer>				
<Vendor>				
<Platform>				
Date run:				
yyyy/mm/dd				
DOI	Jan-09	Feb-09	Mar-09	Total
Totals	225	271	195	691
<DOI 1>	11	21	23	55
<DOI 2>	40	65	3	108
<DOI 3>	25	42	31	98
<DOI 4>	59	61	37	157
<DOI 5>	90	82	101	273
etc....				
NOTES				
1. Report type: service to service, e.g. Vendor: Oxford Journals, Customer: Clearinghouse				
2. DOI is mandatory				
3. Usage data should:				
a) <b>include</b> successful full text requests (HTML plus PDF)				
b) <b>include</b> Accepted Manuscript, Proof, Version of Record versions				
c) <b>exclude</b> Authors Original and Submitted Manuscript Under Review versions				
d) <b>exclude</b> : internal use by publisher and host, downloads from LOCKSS caches, and by robots				

## Appendix C Proposed consolidated usage report for authors

PIRUS Article Report 2 :Number of Successful Full-Text Article Requests by Author, Month and DOI, consolidated from different sources									
<Publisher>									
<Publisher Platform>									
<Author name>									
<Author Identifier>									
<Institutional Identifier>									
Date run:									
yyyy/mm/dd									
Source of usage	Article Title	DOI	Publication Date	Journal	Pre-2011	Jan-11	Feb-11	Mar-11	Total
Publisher	<Article title 1>	<DOI>	<yyyy/mm/dd>	<Journal>	5109	152	226	143	5630
Host 1					0	23	31	29	83
Host 2					0	15	20	18	53
Host 3					0	10	15	12	37
<b>Total</b>					<b>5109</b>	<b>200</b>	<b>292</b>	<b>202</b>	<b>5803</b>
Publisher	<Article title 2>	<DOI>	<yyyy/mm/dd>	<Journal>	3289	352	456	245	4342
Host 1					0	23	31	29	83
Host 2					0	15	20	18	53
<b>Total</b>					<b>3289</b>	<b>390</b>	<b>507</b>	<b>292</b>	<b>4478</b>
etc....									
NOTES									
1. Report type: service to end-user									
2. Author Identifier may be the publisher's own author identifier; the ORCID Identifier will be the preferred option once it is implemented)									
3. All columns are mandatory									
4. Usage data should:									
a) <b>include:</b> successful full text requests (HTML plus PDF)									
b) <b>include:</b> Accepted Manuscript, Proof, Version of Record versions									
c) <b>exclude:</b> Author's Original Manuscript and Submitted Manuscript Under Review versions									
d) <b>exclude:</b> internal use by publisher and host, downloads from LOCKSS caches, and by robots									



## Appendix D Proposed publisher-only usage report for authors

PIRUS Article Report 3: Summary of Successful Individual Full-text Article Requests for an author, by month and DOI								
<Publisher>								
<Publisher Platform>								
<Author name>								
<Author Identifier>								
<Institutional Identifier>								
Date run:								
<yyyy/mm/dd>								
Article	Publication date	Journal	Pre-2011	Jan-11	Feb-11	Mar-11	Total	
<Article title 1>	<DOI> <yyyy/mm/dd>	<Journal>	5109	200	292	202	<b>5803</b>	
<Article title 2>	<DOI> <yyyy/mm/dd>	<Journal>	3241	183	197	152	<b>3773</b>	
<Article title 3>	<DOI> <yyyy/mm/dd>	<Journal>	1109	54	66	32	<b>1261</b>	
<Article title 4>	<DOI> <yyyy/mm/dd>	<Journal>	24976	665	782	322	<b>26745</b>	
etc....								
NOTES								
1. Report type: service to end-user								
2. Author Identifier may be the publisher's own author identifier; the ORCID Identifier will be the preferred option once it is implemented)								
3. All columns are mandatory								
4. Article requests should:								
a) <b>include</b> : successful full text requests (HTML plus PDF)								
b) <b>include</b> : Accepted Manuscript, Proof, Version of Record versions								
c) <b>exclude</b> : Author's Original Manuscript and Submitted Manuscript Under Review versions								
d) <b>exclude</b> : internal use by publisher and host, downloads from LOCKSS caches, and by robots								

## Appendix E Proposed usage report for repositories

PIRUS Article Report 4: Number of Successful Repository Full-Text Article Requests by Month and DOI						
<b>&lt;Institutional Identifier&gt;</b>						
<b>&lt;Repository name&gt;</b>						
<b>Date run:</b>						
<b>&lt;yyyy/mm/dd&gt;</b>						
Repository Identifier	Title	DOI	Oct-10	Nov-10	Dec-10	Total
	Totals for all articles		72	73	44	189
<b>&lt;Repository identifier 1&gt;</b>	<b>&lt;Article title 1&gt;</b>	<b>&lt;DOI 1&gt;</b>	21	15	5	41
<b>&lt;Repository identifier 2&gt;</b>	<b>&lt;Article title 2&gt;</b>	<b>&lt;DOI 2&gt;</b>	17	17	8	42
<b>&lt;Repository identifier 3&gt;</b>	<b>&lt;Article title 3&gt;</b>	<b>&lt;DOI 3&gt;</b>	19	22	12	53
<b>&lt;Repository identifier 4&gt;</b>	<b>&lt;Article title 4&gt;</b>	<b>&lt;DOI 4&gt;</b>	15	19	19	53
<b>etc....</b>						
NOTES						
1. Report type: service to service						
2. All columns are mandatory						
3. Article requests should:						
a) <b>include:</b> successful full text requests (HTML plus PDF)						
b) <b>include:</b> Accepted and Publisher versions						
c) <b>exclude:</b> Pre-prints						
d) <b>exclude:</b> robots						

## Appendix F Proposed usage report for research institutions

PIRUS Article Report 5: Number of Successful Full-Text Article Requests for a Research Institution by Author, Month and DOI, consolidated from different sources											
<Publisher>											
<Publisher Platform>											
<Institutional Identifier>											
Date run:											
yyyy/mm/dd											
Source of usage	Article Title	DOI	Author	Author identifier	Publication Date	Journal	Pre-2011	Jan-11	Feb-11	Mar-11	Total
Publisher	<Article title 1>	<DOI>	<Author name>	<Author Identifier>	<yyyy/mm/dd>	<Journal>	5109	152	226	143	5630
Host 1							0	23	31	29	83
Host 2							0	15	20	18	53
Host 3							0	10	15	12	37
<b>Total</b>							<b>5109</b>	<b>200</b>	<b>292</b>	<b>202</b>	<b>5803</b>
Publisher	<Article title 2>	<DOI>	<Author name>	<Author Identifier>	<yyyy/mm/dd>	<Journal>	3289	352	456	245	4342
Host 1							0	23	31	29	83
Host 2							0	15	20	18	53
<b>Total</b>							<b>3289</b>	<b>390</b>	<b>507</b>	<b>292</b>	<b>4478</b>
etc....											
NOTES											
1. Report type: service to end-user and/or service to service, e.g. Web service between publisher server and institutional CRIS											
2. Author Identifier may be the publisher's own author identifier; the ORCID Identifier will be the preferred option once it is implemented)											
3. All columns are mandatory											
4. Usage data should:											
a) <b>include:</b> successful full text requests (HTML plus PDF)											
b) <b>include:</b> Accepted Manuscript, Proof, Version of Record versions											
c) <b>exclude:</b> Author's Original Manuscript and Submitted Manuscript Under Review versions											
d) <b>exclude:</b> internal use by publisher and host, downloads from LOCKSS caches, and by robots											

## Appendix G PIRUS2 OpenURL key-pair string initial specification

Element	OpenURL Key	OpenURL Value (example)	Notes
	url_ver	Z39.88-2004	Identifies data as OpenURL 1.0. String constant: Z39.88-2004 (Mandatory)
DOI	rft_id	info:doi:http%3A%2F%2Fdx.doi.org%2F10.1016%2FS0022-460X%2803%2900773-9	DOI of the article. The value should be normalised to the second format given above and prepended with 'info:'. (Mandatory if available)
Client IP address	req_id	urn:ip:138.250.13.161	IP Address of the client requesting the article (Mandatory)
UserAgent	req_dat	Mozilla%2F4.0+%28compatible%3B+MSIE+7.0%3B+Windows+NT+5.1%3B+Trident%2F4.0%3B+GoogleT5%3B+.NET+CLR+1.0.3705%3B+.NET+CLR+1.1.4322%3B+Media+Center+PC+4.0%3B+IEMB3%3B+InfoPath.1%3B+.NET+CLR+2.0.50727%3B+IEMB3%29	The UserAgent is used to identify and eliminate, by applying COUNTER rules, accesses by robots/spiders (Mandatory)
Article identifier	rft.artnum	http://hdl.handle.net/1826/58	(Mandatory)
MIMEtype of downloaded file	svc_format	application%2Fpdf	(Recommended)
Article version	svc_dat	Accepted+version	In practice, most repositories won't have this data. But, in case they have, see Note 1. (Optional)
Source repository	rfr_id	dspace.myu.edu	See Note 2. (Mandatory)
<b>Elements given below are <i>only</i> required if DOI is not available</b>			
Article title	rft.atitle	3D+bulk+measurements+of+the+force+distribution	(Mandatory)
First author family name	rft.aulast	Needham	(Mandatory)
ISSN	rft.issn	1359-6640	(Highly Recommended)
Journal title	rft.jtitle	Geoderma	See Note 3. (Optional)
Volume	rft.volume	4	
Issue	rft.issue	11	

## Appendix H OpenURL Context Object in an OAI-PMH wrapper

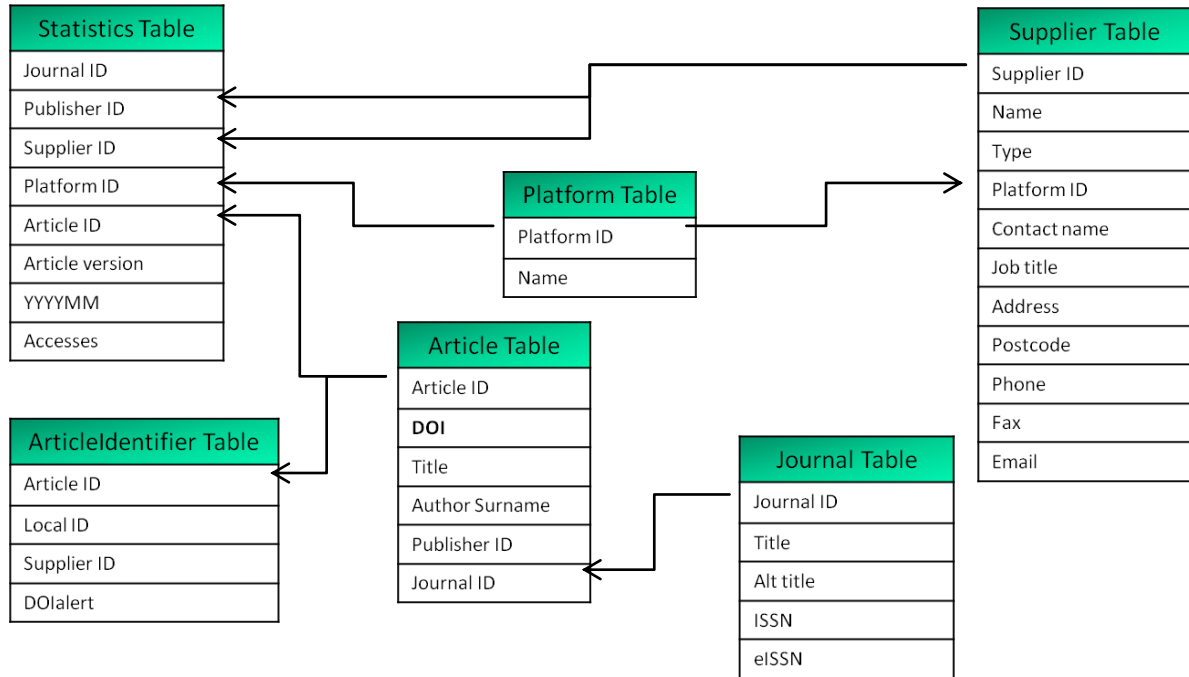
```
<?xml version="1.0" encoding="UTF-8" ?>
- <OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2010-04-26T11:11:49Z</responseDate>
  <request metadataPrefix="statistics" verb="ListRecords">http://dspace-test.central.cranfield.ac.uk:8080/counter-oai/request</request>
- <ListRecords>
- <record>
- <header>
  <identifier>oai:dspace-test.central.cranfield.ac.uk:11/2010-04-23T14:09:34.748Z</identifier>
  <timestamp>2010-04-23T14:09:34Z</timestamp>
  <setSpec>hdl_1826_20</setSpec>
</header>
- <metadata>
- <context-object timestamp="2010-04-23T14:09:34Z">
- <administration>
  - <oa-statistics xmlns="http://dini.de/namespace/oas-info">
    <document_size>276939</document_size>
    <format>application/pdf</format>
    <service>http://dspace-test.central.cranfield.ac.uk:8080</service>
  </oa-statistics>
  </administration>
- <referent>
  <identifier>http://dspace-test.central.cranfield.ac.uk:8080/bitstream/1826/21/1/Municipal%2520waste%2520compost%
    2520as%2520a%2520daily%2520cover%2520material-2004.pdf</identifier>
  <identifier>oai:dspace-test.central.cranfield.ac.uk:1826/21</identifier>
</referent>
- <requester>
- <metadata-by-val>
  <format>http://dini.de/namespace/oas-requesterinfo</format>
</metadata-by-val>
- <metadata>
  - <requesterinfo xmlns="http://dini.de/namespace/oas-requesterinfo">
    <hashed-ip>5f251bb0770d91cab8ecdb3bc1ecb246</hashed-ip>
    <hostname>nc-013-146.dmz.cranfield.ac.uk.</hostname>
    <user-agent>Mozilla/4.0 (compatible; MSIE 8.0; Windows NT 5.1; Trident/4.0; .NET CLR 1.0.3705; Media Center PC 4.0;
      IEM83; .NET CLR 2.0.50727; .NET CLR 3.0.4506.2152; .NET CLR 3.5.30729; InfoPath.2; OfficeLiveConnector.1.4;
      OfficeLivePatch.1.3; .NET CLR 1.1.4322)</user-agent>
  </requesterinfo>
</metadata>
</context-object>
</metadata>
</ListRecords>
</OAI-PMH>
```

## ***Appendix I COUNTER robots exclusion list***

Alexandria(\s\+)\prototype(\s\+)\project  
Arachmo  
Brutus\AET  
Code(\s\+)\Sample(\s\+)\Web(\s\+)\Client  
dtSearchSpider  
FDM(\s\+)\1  
Fetch(\s\+)\API(\s\+)\Request  
GetRight  
Goldfire(\s\+)\Server  
Googlebot  
httpget\â~5\2\2  
HTTTrack  
iSiloX  
libwww\perl  
LWP\:\Simple  
lwp\trivial  
Microsoft(\s\+)\URL(\s\+)\Control  
Milbot  
MSNBot  
NaverBot  
Offline(\s\+)\Navigator  
playstarmusic.com  
Python\urllib  
Readpaper  
Strider  
T\H\U\N\D\E\R\S\T\O\N\E  
Teleport(\s\+)\Pro  
Teoma  
Web(\s\+)\Downloader  
WebCloner  
WebCopier  
WebReaper  
WebStripper  
WebZIP  
Wget  
Xenu(\s\+)\Link(\s\+)\Sleuth

## Appendix J The PIRUS2 relational database

The diagram below illustrates the tables in the PIRUS2 database and indicates some of the key relationships between them.



## Appendix K Screenshots of reports available from the demonstration portal.

**AR1j Report – cross check to confirm that exposed data matches original data provided:**

Platform	Title	DOI	Jan-09	Feb-09	Mar-09	Apr-09	May-09	Jun-09	Jul-09	Aug-09	Sep-09
Totals for all articles			361	923	973	1488	1526	1988	1975	1735	2657
nature.com	&alpha;1B-Adrenoceptors mediate ad	10.1111/j.1745-7254.2008.00727.x	0	1	1	0	1	3	0	0	2
nature.com	&alpha;1D-Adrenergic receptor insens	10.1038/aps.2009.160	0	0	0	0	0	0	0	0	0
nature.com	&alpha;2,6-hyposialylation of c-Met at	10.1038/aps.2009.84	0	0	0	0	0	0	12	8	8
nature.com	&beta;-Naphthoflavone protects mice	10.1038/aps.2009.156	0	0	0	0	0	0	0	0	0
nature.com	&beta;-Sitosterol sensitizes MDA-MB-	10.1111/j.1745-7254.2008.00761.x	0	1	1	1	1	2	1	1	5
nature.com	2-Amino-nonyl-6-methoxyl-tetralin m	10.1038/aps.2009.157	0	0	0	0	0	0	0	0	0
nature.com	5-HT1A/7 receptor agonist excites card	10.1111/j.1745-7254.2008.00745.x	1	1	1	1	0	2	0	0	1
nature.com	7-Chloroarctonine-b as a new selective	10.1038/aps.2009.113	0	0	0	0	0	0	0	0	25
nature.com	A comparative study of outcomes of id	10.1038/aps.2009.132	0	0	0	0	0	0	0	0	0
nature.com	A novel acute anemia model for pharm	10.1038/aps.2009.161	0	0	0	0	0	0	0	0	0
nature.com	A novel anti-cancer effect of genistein	10.1111/j.1745-7254.2008.00831.x	0	2	0	6	5	0	2	1	6
nature.com	A novel high-throughput format assay	10.1111/j.1745-7254.2008.00748.x	1	2	6	3	2	2	1	4	6
nature.com	A rat model for studying neural stem c	10.1038/aps.2009.151	0	0	0	0	0	0	0	0	0
nature.com	A sesquiterpene quinone, dysidine, fr	10.1038/aps.2009.5	0	0	25	13	7	7	3	4	6
nature.com	Ability of alpha-lipoic acid to reverse	10.1111/j.1745-7254.2008.00790.x	0	1	1	2	1	0	4	3	4
nature.com	Acetamide-45 inhibited hyperresponsi	10.1111/j.1745-7254.2008.00846.x	0	1	3	2	3	0	1	0	3
nature.com	Acute pulmonary inflammation is inhib	10.1111/j.1745-7254.2008.00899.x	1	6	1	1	1	0	0	1	7
nature.com	Additive evaluation of chelic acid vor	10.1038/aps.2009.42	0	0	0	0	0	0	0	0	0

**AR2 Report to authors showing consolidated usage:**

Source of usage	Platform	Jan-10	Feb-10	Mar-10	Apr-10	May-10	Jun-10	Jul-10	Aug-10	Sep-10	Oct-10	Nov-10	Dec-10	Jan-11	Feb-11	Total
Emerald	Insight	17	14	20	0	0	0	0	0	0	0	0	0	0	0	51
Bournemouth University BURO	Eprints	0	0	0	0	0	0	0	3	39	42	44	37	37	15	217
Totals for all articles		17	14	20	0	0	0	0	3	39	42	44	37	37	15	268



Project Acronym: PIRUS2  
 Version: 1.0  
 Contact: Paul Needham ([paul.needham11@btinternet.com](mailto:paul.needham11@btinternet.com))  
 Date: 06 /10/2011

### AR1ir Report showing COUNTER-compliant statistics for a repository:

Local ID	Title	DOI	Sep-10	Oct-10	Nov-10	Dec-10	Jan-11	Feb-11	Total
1	Article Report 1ir (R0.1), Number of Successful Full-Text Article Requests by Month and DOI								
2	Cranfield University CERES								
3	Date Run 28/02/2011								
4	Local ID								
5	Totals for all articles								
6	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/4199">http://dspace.lib.cranfield.ac.uk/handle/1826/4199</a>	Managing the transition <sup>™</sup> - supplier management in internation	12	23	15	5	5	1	61
7	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3310">http://dspace.lib.cranfield.ac.uk/handle/1826/3310</a>	A configurational approach to the dynamics of firm level knowle	13	17	17	8	12	6	73
8	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1879">http://dspace.lib.cranfield.ac.uk/handle/1826/1879</a>	A dielectric sensor for measuring flow in resin transfer moulding	7	19	22	12	14	2	76
9	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/2228">http://dspace.lib.cranfield.ac.uk/handle/1826/2228</a>	A review of the erosion of thermal barrier coatings	22	15	19	19	47	6	128
10	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1910">http://dspace.lib.cranfield.ac.uk/handle/1826/1910</a>	A role-based perspective on leadership as a network of relation	23	17	16	11	5	8	80
11	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/2655">http://dspace.lib.cranfield.ac.uk/handle/1826/2655</a>	A taxonomy for selecting global supply chain strategies	68	89	100	86	60	38	441
12	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/2734">http://dspace.lib.cranfield.ac.uk/handle/1826/2734</a>	A taxonomy of highly interdependent, supply chain relationship	25	24	28	29	21	8	135
13	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/4312">http://dspace.lib.cranfield.ac.uk/handle/1826/4312</a>	Aligning business leadership development with business needs: 10.1108/02621710810901273	18	24	23	14	17	8	104
14	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3020">http://dspace.lib.cranfield.ac.uk/handle/1826/3020</a>	Aligning Distribution Center Operations to Supply Chain Strateg	38	51	45	42	57	24	257
15	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3407">http://dspace.lib.cranfield.ac.uk/handle/1826/3407</a>	Alliances and Networks: Creating Success in the UK Fair Trade M	32	33	50	29	25	14	183
16	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1043">http://dspace.lib.cranfield.ac.uk/handle/1826/1043</a>	An evaluation of styles of IT support for marketing planning	28	51	50	36	27	9	201
17	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3013">http://dspace.lib.cranfield.ac.uk/handle/1826/3013</a>	An exploration of warehouse automation implementations: cost	31	36	78	44	43	10	242
18	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3021">http://dspace.lib.cranfield.ac.uk/handle/1826/3021</a>	An exploratory framework of the role of inventory and warehou	17	32	29	31	23	11	143
19	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/2670">http://dspace.lib.cranfield.ac.uk/handle/1826/2670</a>	An integrated model for the design of agile supply chains	6	9	9	2	83	20	129
20	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3308">http://dspace.lib.cranfield.ac.uk/handle/1826/3308</a>	An interview with Bernard Rethore, Emeritus Chairman, Flowser	7	12	4	4	5	3	35
21	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1927">http://dspace.lib.cranfield.ac.uk/handle/1826/1927</a>	An interview with Sir John Parker, Chairman, National Grid	14	17	7	2	4	1	45
22	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1198">http://dspace.lib.cranfield.ac.uk/handle/1826/1198</a>	An open-path, hand-held laser system for the detection of meth	24	36	22	11	26	10	129
23	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/8817">http://dspace.lib.cranfield.ac.uk/handle/1826/8817</a>	Analysis of skin tissues spatial fluorescence distribution by the	8	5	7	7	3	7	37
24	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/942">http://dspace.lib.cranfield.ac.uk/handle/1826/942</a>	Back to the workplace: How organisations can improve their sup	15	29	30	14	12	5	105
25	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/948">http://dspace.lib.cranfield.ac.uk/handle/1826/948</a>	Benchmarking and library quality maturity	18	16	14	21	19	7	95
26	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/2666">http://dspace.lib.cranfield.ac.uk/handle/1826/2666</a>	Building the Resilient Supply Chain	104	114	120	77	136	47	598
27	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/4304">http://dspace.lib.cranfield.ac.uk/handle/1826/4304</a>	Case study: meeting the demand for skilled precision engineers	11	18	12	4	3	3	51
28	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1028">http://dspace.lib.cranfield.ac.uk/handle/1826/1028</a>	Chairman and chief executive officer (CEO): that sacred and secr	31	37	54	33	28	7	190
29	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1917">http://dspace.lib.cranfield.ac.uk/handle/1826/1917</a>	Chairman of the board: demographics effects on role pursuit	16	9	16	6	6	3	56
30	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/2721">http://dspace.lib.cranfield.ac.uk/handle/1826/2721</a>	Chaos Theory: Implications for Supply Chain Management	58	40	48	44	33	14	237
31	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1377">http://dspace.lib.cranfield.ac.uk/handle/1826/1377</a>	Characterization of the response of fibre Bragg gratings fabricate	14	21	13	10	8	6	72
32	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1024">http://dspace.lib.cranfield.ac.uk/handle/1826/1024</a>	Constructing a professional identity: how young female manage	51	41	55	29	28	12	216
33	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/1911">http://dspace.lib.cranfield.ac.uk/handle/1826/1911</a>	Consultant's role: a qualitative inquiry from the consultant's per	32	34	42	23	39	27	197
34	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3312">http://dspace.lib.cranfield.ac.uk/handle/1826/3312</a>	Corporate Social Responsibility in Supply Chains of Global Brand	128	110	196	145	156	41	776

### IR DOI Update report supplying DOIs to allow repositories to update and enhance their own records:

Repository DOI update	DOI
1	Cranfield University CERES
2	DOI
3	DOI
4	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3256">http://dspace.lib.cranfield.ac.uk/handle/1826/3256</a>
5	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/3407">http://dspace.lib.cranfield.ac.uk/handle/1826/3407</a>
6	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/885">http://dspace.lib.cranfield.ac.uk/handle/1826/885</a>
7	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/886">http://dspace.lib.cranfield.ac.uk/handle/1826/886</a>
8	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/887">http://dspace.lib.cranfield.ac.uk/handle/1826/887</a>
9	<a href="http://dspace.lib.cranfield.ac.uk/handle/1826/954">http://dspace.lib.cranfield.ac.uk/handle/1826/954</a>
10	
11	
12	
13	
14	
15	

## **Appendix L PIRUS2 CCH - model for allocation of costs to publishers - scenario A**

### **CCH receives and processes the publisher log files and generates the usage reports**

*Note: these costs will be significantly lower where the CCH is already producing the existing COUNTER usage reports for a publisher*

#### 1. Assumptions

- a. the basis for the tariffs for publishers will be a combination of total revenues and article download activity
- b. A **Large Publisher** is defined as one with i) annual revenues in excess of \$100 million and ii) more than 100 million full-text article downloads per month (Assume 150 million per month for the calculation)
- c. A **Medium Publisher** is one with i) annual revenues of between \$5 and \$10 million and ii) between 20 and 40 million full-text article downloads per month (Assume 30 million per month for the calculation)
- d. A **Small Publisher** is one with i) annual revenues of less than \$1 million per annum and ii) less than 500k downloads per month (Assume <500k per month for the calculation)

#### 2. Costs provided by the contractor are broken down as follows:

(detailed figures available in the contractor proposal)

- a. Annual Membership Fee ( \$38,700 for a Large Publisher; \$9,700 for a Medium Publisher; \$1,700 for a Small Publisher)
- b. Reporting Services Costs ( Setup cost : \$58,500; annual infrastructure and operational costs: \$103,200)

	<b>Large Publisher</b>	<b>Medium Publisher</b>	<b>Small Publisher</b>
Annual membership fee	\$38,700	\$9,700.00	\$1,700

#### 3. Allocation of annual costs

Membership fee	\$38,700	\$9,700	£1,700
Reporting services			
Y1	\$808	\$808	£808
Y2 +.....	\$515	\$515	£515
Transaction-based fee costs	\$630,000	\$264,000	£4,800
<b>Total</b>			
<b>Y1</b>	<b>\$669,508</b>	<b>\$274,508</b>	<b>£9,008</b>
<b>Y2+.....</b>	<b>\$669,215</b>	<b>\$283,915</b>	<b>£8,715</b>

## ***Appendix M PIRUS2 CCH - model for allocation of costs to publishers - scenario C***

### **Publisher creates the usage reports, which are harvested by the CCH**

*Note: publishers will be charged only an annual membership fee, based on annual revenues, for this service*

#### 1. Assumptions

- a. A **Large Publisher** is defined as one with annual revenues in excess of \$100 million
- b. A **Medium Publisher** is one with annual revenues of between \$5 and \$10 million
- c. A **Small Publisher** is one with annual revenues of less than \$1 million per annum

#### 2. Costs are broken down as follows:

(detailed figures available in Section 5.1.2 of the Final Report)

	<b>Large Publisher</b>	<b>Medium Publisher</b>	<b>Small Publisher</b>
3. Allocation of annual costs			
Membership fee	\$66,900.00	\$16,900.00	\$2,900.00
<b>Total</b>	<b>\$66,900.00</b>	<b>\$16,900.00</b>	<b>\$2,900.00</b>

## **Appendix N PIRUS2 CCH - model for allocation of costs to repositories - scenario A**

### 1. Assumptions

- a. the basis for the tariffs will be the same for all categories of organization (repository, publisher, etc.)
- b. smaller organizations will have lower membership fees
- c. 25% of total global usage (1.5 billion full-text article requests per annum) will take place in institutional repositories ( 375 million article requests per annum, or 31.5 million per month)
- d. the total universe of academic repositories is 1300 institutions
- e. the more active repositories will participate first in the CCH; assume they represent 12.5% of total global usage ( 187.5 million article requests per annum, or 15.6 million per month)

### 2. Costs provided by the contractor are broken down as follows:

( detailed figures available in the contractor proposal)

- a. Membership fee schedule (\$470 per institutional repository)
- b. Reporting Services Costs ( Setup cost : \$58,500; annual infrastructure and operational costs: \$103,200)
- c. Transaction-based fee costs ( based on the level of activity outlined in 1c above, \$12k per month, or \$144k per annum)

### 3. Allocation of annual costs per repository (US\$)

	Scenario 1 (100 repositories)	Scenario 2 (500 repositories)
Membership fee	\$470.00	\$470.00
Reporting services		
Y1	\$1,615.00	\$323.00
Y2 +.....	\$1,030.00	\$206.00
Transaction-based fee costs	\$1,440.00	\$288.00
<b>Total</b>		
<b>Y1</b>	<b>\$3,525.00</b>	<b>\$1,081.00</b>
<b>Y2+.....</b>	<b>\$2,940.00</b>	<b>\$964.00</b>

## **Appendix O Report on the results of the Extensions to PIRUS 2**

### **Introduction**

Following the successful PIRUS2 End of Project Seminar on 23 February the project team identified several issues that could usefully be further explored and developed before completion of the project. These were:

- Development of a prototype UK institutional repository usage statistics demonstrator (IRUS)
- More detailed analysis of article usage taking place in repositories
- Guidelines for the implementation of Fedora
- The organizational/economic model proposed for the CCH: at the end of project seminar, both publishers and repositories expressed concern about the level of the proposed tariffs for the CCH

JISC agreed that there was value in such work and permitted an extension of the work to be completed by the end of May 2011. Tasks carried out to address these issues further extended work already undertaken on Workpackages 3, 4, 5 and 6 of the project.

### **UK Institutional Repository Usage Statistics Demonstrator**

Institutional Repositories have attracted much attention over the last decade, and there has been great interest in the growing number repositories and the number of items held in those IRs. However, overall, very little of value has been said about the *usage* of those items. And, we not aware of any major studies or initiatives that have proved the impact and value of IRs, as yet.

Increasingly, IRs do provide statistics purporting to show usage within repositories, but different softwares (and indeed individual repositories) process raw usage data in different ways making it impossible to compare like for like across repositories – there is, currently, no agreed standard in place to measure usage within and across repositories.

PIRUS is still collecting a considerable amount of usage data from repositories. Only a small portion of this data was used in proving that consolidation of publisher and repository usage statistics is possible. Now, the availability of this (growing) dataset presents an opportunity to explore the possibilities of developing a UK institutional repository usage statistics service (IRUS-UK). Such a service could:

- Collect raw usage data from UK IRs for all item types within repositories
- Process those raw data into COUNTER-compliant statistics and return those statistics back to the originating repositories for their own use
- Give JISC (and others) a nation-wide picture of the overall use of UK repositories – demonstrating their value and place in the dissemination of scholarly outputs (“Making scholarly statistics count in UK repositories”)
- Offer opportunities for benchmarking
- Potentially act as an intermediary between UK repositories and other agencies – e.g. global central clearinghouse, national shared services

So, we have re-purposed and re-used usage data collected by PIRUS2 to create a new demonstrator based just on UK Institutional Repository usage data: IRUS-UK This new demonstrator:

- Required a slight re-design of the existing PIRUS2 database (to remove the journal authority table and accommodate multiple resource types, though at this stage most data still only pertains to articles)
- Re-used much of the code used in the Perl scripts to ingest data into the existing PIRUS2 database, including COUNTER filtering of robots and double clicks. However, it is important to note that:
  - we have changed the ingest process to emulate a new process in line with the updated OpenURL specification (see Table 1, page 25), which only requires provision of the OAI identifier of the item – used to enable metadata lookup via the OAI-PMH GetRecord call) - instead of direct inclusion of various bibliographic metadata

- elements. This reduces the size of the transmitted messages and simplifies parsing of the entries
- changed the configuration of the tracker at Cranfield University to transmit messages **for all item types** – not just articles
- Item types catalogued at the IR level (which vary from system to system) have been mapped to a set of IRUS item types to provide consistent terminology within the demonstrator
- for this demonstrator, we have made no attempt to lookup the DOI from CrossRef for items of the 'article' type, though this could/should be re-incorporated at a later date
- Required a re-write of the web interface to illustrate possibilities (though again much of the underlying code has been re-used)

### Participating repositories

Using the existing extensions to their repository software, the following IRs participated in providing usage data to IRUS:

DSpace:

- Cranfield CERES
- University of Edinburgh ERA

Eprints:

- Bournemouth University Research Online (BURO)
- University of Huddersfield Repository
- University of Salford Institutional Repository
- Southampton ECS EPrints Repository

### The IRUS-UK Demonstrator

The demonstrator is a proof of concept - an arena intended to illustrate the technical feasibility of gathering, consolidating and re-exposing usage events from UK IRs.

The demonstrator comprises:

- A web user interface, written in PHP
- Downloadable example reports

It sits behind an authentication/authorisation barrier to allay privacy concerns. But, for authorized users, it is the window into the IRUS MySQL database, which holds the data gathered from repositories.

The main features of the portal are:

- The home page provides summary information:
  - The overall stats tab, shows the number of items indexed and overall totals of download events recorded from repositories (see Figure 2/Figure 3 below)
  - the monthly stats tab, shows downloads per repository per month for a given time period (see Figure 4 below)
- The Ingest stats page shows how the COUNTER filtering of double clicks and robots has affected the number of raw usage events
- The Itemtype stats page breaks down the number of items and downloads by item type, e.g. articles vs. conference items vs. Theses, etc.
- The Search facility makes it possible to find individual items or groups of items by Title/Author

The pages, currently, presented are merely illustrations (scratch the surface) of some of the possibilities for 'slicing and dicing' the data held in the IRUS database. Much more could be done to offer additional metrics, particularly to benchmark performance across repositories, and it would be profitable to investigate synergies between IRUS and the existing registers of repositories. With further development effort, it would be possible to create a wide range of additional views into the data, depending on user requirements - yet to be determined!

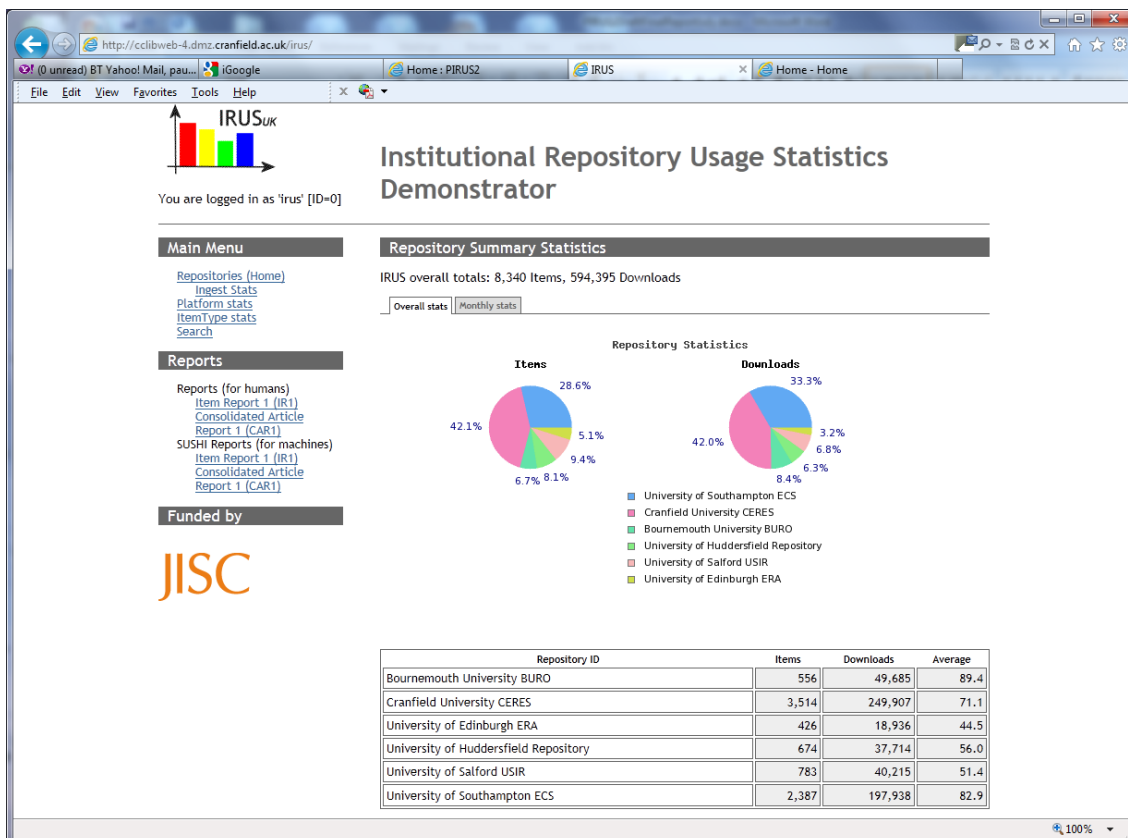


Figure 3. IRUS-UK demonstrator home page, overall stats

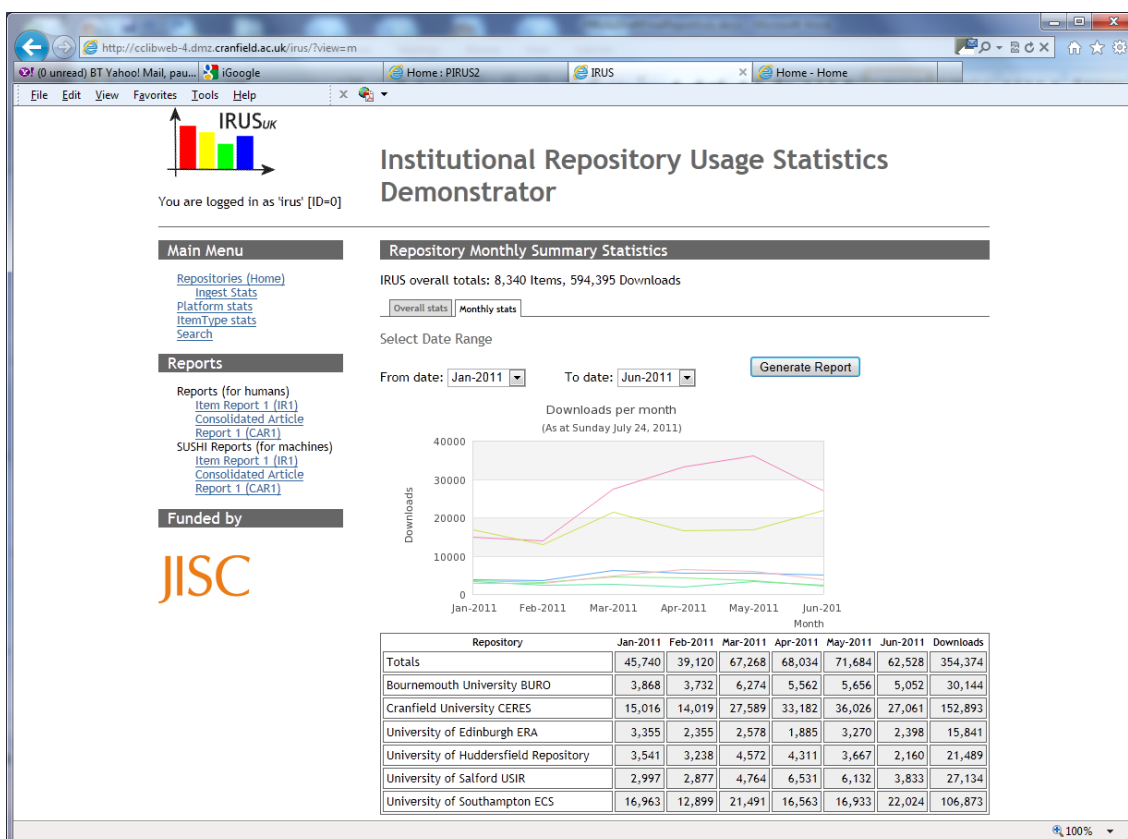


Figure 4. IRUS-UK demonstrator home page, monthly stats

Also, examples of reports that can be generated by the system and retrieved for use in other contexts are given:

- Item Report 1 (IR1): Number of Successful Item Download Requests by Month and Repository Identifier
  - this is a report intended for institutions, which provides them with COUNTER-compliant download statistics for items held in their IR
- Consolidated Article Report 1 (CAR1): Number of Successful Monthly Article Download Requests by DOI and Repository Identifier
  - this is a report intended for use by the proposed Central Clearing House. Only items identified as 'articles' and with a known DOI are included in this report

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

- Each of the reports can be viewed in a web page in the portal or downloaded for use locally as MS-Excel/CSV files (just as librarians have gathered COUNTER statistics from publishers for the last decade)
- More importantly, the reports have been made also available via the SUSHI protocol for incorporation into local institutional ERMs, or for automatic gathering for use in any number of other national/global services (See Figure 5 below). *This is, as far as we know, the first and only working example of a SUSHI service providing statistics at the level of the individual item!*
- with further effort, of course, it would also be possible to develop an IRUS API (or other web services) to embed statistics into other services

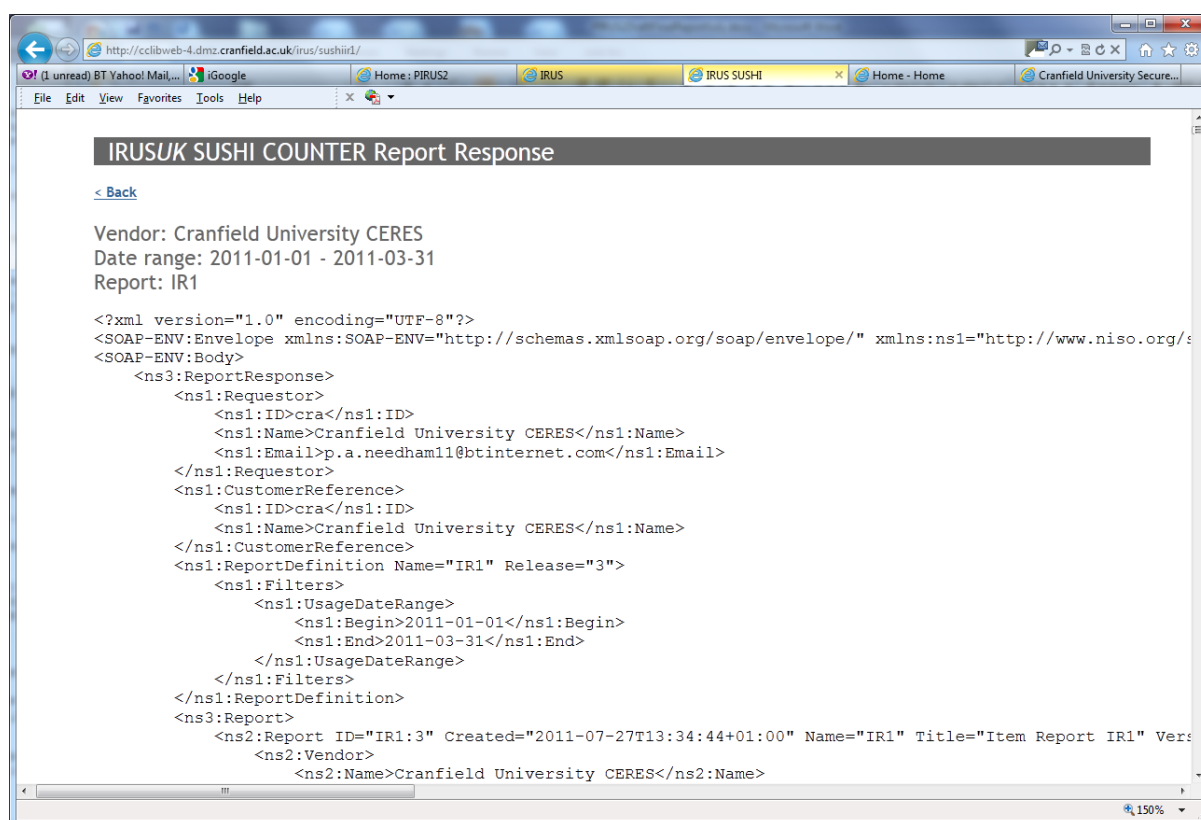


Figure 5. SUSHI Report response at the individual item level

## Repository usage

As observed above, much attention has been paid to the growing number of IRs and numbers of items within those IRs. Repository advocates have been debating whether IRs are competition for traditional publishing, or a supplement to traditional publishing, whilst publishers have rather tended to regard IRs as an irrelevance, or as a source of annoyance, a diversion from the serious world of publishing.



All of this debate has taken place in the absence of comprehensive hard evidence regarding the actual usage of items in IRs.

PIRUS has been collecting usage data from the participating IRs for the last 7 to 10 months (depending when each individual IR signed up to participate in the project). Overall, as at 24th June 2011, the IRUS demonstrator holds usage statistics for **8,340 items** representing an impressive **594,395 downloads** across a number of different item types. Of those items, 6,089 are 'articles' which between them have been downloaded 527,224 times.

Taking a closer look at the six month period, Jan-Jun 2011:

Repository	Jan-2011	Feb-2011	Mar-2011	Apr-2011	May-2011	Jun-2011	Downloads
Totals	45,740	39,120	67,268	68,034	71,684	62,528	354,374
Bournemouth University BURO	3,868	3,732	6,274	5,562	5,656	5,052	30,144
Cranfield University CERES	15,016	14,019	27,589	33,182	36,026	27,061	152,893
University of Edinburgh ERA	3,355	2,355	2,578	1,885	3,270	2,398	15,841
University of Huddersfield Repository	3,541	3,238	4,572	4,311	3,667	2,160	21,489
University of Salford USIR	2,997	2,877	4,764	6,531	6,132	3,833	27,134
University of Southampton ECS	16,963	12,899	21,491	16,563	16,933	22,024	106,873

Total downloads between the 6 IRs are 354,374. With the exception of the Cranfield figures where data has been collected for all item types in the last couple of months, these figures pertain to article downloads. Subtracting the non-article figures leaves a download figure of approximately 290,000 for articles.

This equates to an average of around 8,000 article downloads per month per repository. Using some very rough, rule-of-thumb calculations, scaled up to 100 UK IRs, that would represent close to 800,000 article downloads from UK IRs per month.

Based on these figures, we would argue that **repositories are significant players in the dissemination of scholarly outputs and should be taken seriously!**

## Repository survey

10 'key', influential repositories managers were identified and targeted with a brief survey, which described: the PIRUS2 aims and objectives; the role of the Central Clearing House; the PIRUS2 CCH model for allocation of costs to repositories, based on scenario A in which the CCH receives and processes repository log entries and generates the usage reports:

### 1. Assumptions

- 25% of total global usage (1.5 billion full-text article requests per annum) will take place in institutional repositories ( 375 million article requests per annum, or 31.5 million per month)
- . the total universe of academic repositories is 1300 institutions
- the more active repositories will participate first in the CCH; assume they represent 12.5% of total global usage ( 187.5 million article requests per annum, or 15.6 million per month)

2. Costs provided by the contractor are broken down as follows:  
(detailed figures available in the contractor proposal)

a. Membership fee schedule (\$470 per institutional repository)

b. Reporting Services Costs ( Setup cost : \$58,500; annual infrastructure and operational costs: \$103,200)

c. Transaction-based fee costs (based on the level of activity outlined in 1c above, \$12k per month, or \$144k per annum)

3. Allocation of annual costs per repository (US\$)

		Scenario 1 (100 repositories)	Scenario 2 (500 repositories)
Membership fee		\$470.00	\$470.00
Reporting services			
	Y1	\$1,615.00	\$323.00
	Y2 +.....	\$1,030.00	\$206.00
Transaction-based fee costs		\$1,440.00	\$288.00
<b>Total</b>	<b>Y1</b>	<b>\$3,525.00</b>	<b>\$1,081.00</b>
	<b>Y2+.....</b>	<b>\$2,940.00</b>	<b>\$964.00</b>

**Responses to Questions**

Of the 10 repository managers included in the Survey, 5 provided responses and their responses to each question are summarised below:

**Question 1: Do you find the overall level of the proposed tariffs reasonable?**

Yes 1 + 1 for scenario 2  
No 2

Comments

- Unsure. This would have to be scrutinised by senior management. Budget is tight at the moment. I anticipate we would have to compile a comprehensive cost benefit analysis
- We do, but we aren't sure they are affordable due to the cuts we are having to make in the budgets. However, my hope would be that with the help of the PIRUS project etc. we could sell the idea to the Research Office to see if they would cover the costs in the run up to the REF.
- The amounts seem high for any individual repository to take on, albeit that the cost comes down if more people sign up, although I accept that for a statistical service they are not that high (cf. journal statistics services). Nevertheless, it needs to be clear what added advantages signing up for this service would bring over the use of Google analytics etc that many repositories currently use so that it is clear what value is being purchased. For the UK, I could envisage JISC having a central arrangement that institutions signed up to. Given the description of the model, what criteria will be used to determine how much each repository pays?
- While we do see the benefits of a system similar to the one you are proposing for the CCH and for the project at a whole, in the current climate this is a level of cost that we would find very hard to justify to the University at large.
- I think it's important to stay very close to the \$1K barrier for libraries. I suspect that Libraries would need to think carefully about agreeing to pay up to \$3.5K and many would reject the proposal. This makes advocacy all the more important for us to reach 500 repositories from the start

**Question 2: Do you find the proposed model for allocation of costs reasonable?**

Yes 3  
No 1

Comments:

- This is a question I can't answer for certain yes or no. The division of costs for a membership fee and then additional costs for services is a sensible one. But it seems that if the plan remains to implement for Repositories first and then the Publishers after then the full cost of the setup would fall on the Repositories rather than being shared by all stakeholders. This would seem to be of disproportionate benefit to the Publishers as the true value of the system will only be realised once all stakeholders are contributing their data. Also, please correct me if I'm reading this incorrectly, but I am unsure of the rationale for charging for the production of reports as well as for the processing of our data as you receive it.
- It is unclear why there is a separate membership and reporting fee, and what the distinction is. Why not just have a reporting fee?
- From my personal perspective this seems a reasonable model. Again, it would have to go to senior management for scrutiny if we were to consider participating.

**Question 3: Have you any other comments on the proposed role for the CCH, or on the tariffs associated with it?**

- In terms of the costs for the CCH, these are not the only costs that would need to be accounted for by any repository joining the scheme. There would be associated costs to do with the set up of the plug-ins for the repository, not an easy task in a repository configured for custom metadata that does not represent the EPrints standard. Also as I understand it for Peter Shepherd's talk at the end of PRIUS2 workshop a way of identifying authors would need to be used, ORCID in this case, a project that is not yet complete. I can see definite benefits for the system and for the project but I simply don't see a way we could fund it at this level of subscription costs.
- The dilemma for many repositories is a chicken and egg one. To justify the costs of a statistical service they need content to measure. But one of the benefits of having standard statistics is to attract more content because many don't have enough to justify expense on assessing it. There almost needs to be a simple way in to highlight value that can then be used to demonstrate worth and the case for signing up to a fuller service.
- Nothing further to add – apart from our support for this the success of the model! However, I'm wondering if, as a hosted Repository with Eprints there would be a way of incorporating the costs into our maintenance contract?
- I would imagine that most institutions would want to consider carefully the benefits to them of committing to this annual outlay. Therefore, if it were to be taken forward, it might be worth setting out a paper on the benefits, risks, pros and cons of participating in PIRUS. Times are really tight financially, so an extremely good case has to be made before money is released for any extras.

**Follow-up Question**

Each of the 5 respondents were invited to view the IRUS demonstrator. The estimated costs (£200-£400 per annum) and the rationale for the proposed service were explained. Respondents were asked **would you find this a more attractive proposition?**

Yes 5

- I have discussed the below with the financial powers that be in the Library and the kinds of costing you mention for a UK wide service would be a much more manageable proposition for us at the moment. As I mentioned in the survey I am very interested in the kinds of data that might come out of a project of this kind and scaling down the project will give me a chance to be able to demonstrate the value of the statistics generated by the project.
- A follow-up, I like this a lot better, certainly. The current presentation is very user-friendly and clear (though a I guess a download raw data option would be helpful alongside the Counter reports).

- I think this is an excellent idea - and sounds like a nice add on to JUSP? We'd certainly back this.
- Looks good. What I would want is the ability to embed statistics in my own repository. To me that's where the value lies, hence where the sustainability would come from.
- The proposed cost structure for the IRUS system sounds like excellent value for money and I suspect most institutions would be happy to afford the cost. Such low levels of payment would hopefully mean comprehensive buy-in from the HE community for a system that gives both accurate and comparable usage data, a measure of impact across the research community, and a system, the population of which, would be in the control of institutions.

## Conclusions

Although the survey was carried out on a small scale, the findings are line with feedback from the end of project seminar and other anecdotal evidence we have encountered.

- The original proposed model received a mixed reception from repository managers. While the costs were generally considered *reasonable*, there were concerns that *the costs would not be affordable*, and would be hard to justify at a time when budgets are being squeezed.
- The model built around IRUS-UK, with its greatly reduced associated costs and clearer UK focus, was greeted with universal enthusiasm and approval.

## Tentative costings for an IRUS-UK service

The following gives a provisional estimate of annual costs to run an IRUS service, from a technical perspective:

### PIRUS2 - model for allocation of costs to IRUS -scenario A (IRUS receives and processes the repository log files and generates the usage reports)

#### 1. Assumptions

- a. the basis for the tariffs for repositories will be item download activity
- b. IRUS will be treated as a **Small Publisher**, i.e. one with annual revenues of less than \$1 million per annum

Two cases will be considered:

A: 60 participating repositories; a total of 10 million downloads per annum

B: 120 participating repositories; a total of 20 million downloads per annum

#### 2. Costs provided by the contractor are broken down as follows:

	Case A	Case B
IRUS Annual membership fee to CCH	\$1,700.00	\$1,700.00

#### 3. Allocation of annual costs for IRUS

Membership fee		\$1,700.00	\$1,700.00
Reporting services			
Y1		\$808.00	\$808.00
Y2 +.....		\$515.00	\$515.00
Transaction-based fee costs		\$9,000.00	\$18,000.00
<b>Total</b>	<b>Y1</b>	<b>\$11,508.00</b>	<b>\$20,508.00</b>
	<b>Y2+.....</b>	<b>\$11,215.00</b>	<b>\$20,215.00</b>
<b>Cost per repository</b>			
Y1		\$191.80	\$170.90

Y2+...	\$186.92	\$168.46
<b>4. Cost to transmit AR1 from IRUS to CCH</b>		
Annual fee	\$2,900.00	\$2,900.00
Allocated cost per repository	\$48.33	\$24.16
<b>5. Total cost per repository</b>		
Y1	<b>\$240.13</b>	<b>\$195.06</b>
y2+	<b>\$235.25</b>	<b>\$192.62</b>

There are likely to be other costs associated with an IRUS service, e.g. administration, liaison with repositories. We have not considered these in detail, yet, but are confident that the overall cost per repository is unlikely to exceed £400 per annum.

### ***Fedora implementation guidelines***

The Fedora daemon developed for PIRUS2 by the University of Oxford worked well - at Oxford - and, over a period of time, successfully transmitted usage data to the PIRUS2 test server. However, an attempt to re-use the daemon at the University of Hull was unsuccessful and revealed that the implementation was too specific to the Fedora system at Oxford. So, it became clear that it would be impossible to deploy it more widely across other Fedora repositories without further development/customisation.

Consequently, PIRUS2 assembled a working group of Fedora experts and users to consider how the work might be taken forward to devise a more generic solution to logging usage in Fedora repositories.

The working group, coordinated by the project manager, comprised:

- Sally Rumsey (Oxford University)
- Neil Jefferies (Oxford University)
- Anusha Ranganathan (Oxford University)
- Chris Awre (Hull University)
- Richard Greene (Hull University)
- Steve Bayliss (Acuity Unlimited)

The work of the group was carried out via a series of email exchanges and teleconferencing.

### **Aim**

The aim of the group was to devise a set of Fedora Implementation Guidelines and suggested tools to enable Fedora repositories to transmit item usage data to a third party statistics consolidation service whenever an individual item (article, thesis, conference paper, report, video, etc.) download/access occurs.

### **Objectives**

The objectives of the group were to:

- To understand how the Oxford daemon works
- To understand why it couldn't easily be used at Hull
- Brainstorm/discuss ways of taking the work forward and coming up with potential solutions for a generic implementation of PIRUS2 functionality in Fedora IRs
- create a set of Fedora Implementation Guidelines and suggested tools – or, at least, come up with a road map which would lead to practical guidelines and tools within the foreseeable future

The following sections describe the findings of the group.

## PIRUS2 Daemon implementation at the University of Oxford

### ***Platform and dependencies***

The PIRUS2 Daemon is written in Python, a platform-independent programming language that runs on Linux/Unix systems and Windows.

The Daemon has the following dependencies

- Supervisor (Linux/unix tool for process monitoring and control)
- Redis – written in C, runs under Linux although there are reportedly unofficial ports that run under Cygwin and MinGW on windows
- Redis python client
- Simplejson – JSON encoder/decoder; Python

Although Python runs on multiple platforms, the above dependencies mean that the Daemon will only currently run on Linux/Unix systems

The source code available at <https://github.com/benosteen/PIRUS2Daemon>

The PIRUS2 Daemon is a distributed system for pushing usage data via Open-URL Context-Objects to a PIRUS2 compliant endpoint. This was implemented to work with the Oxford University Research Archives (ORA).

ORA is a web service served by a python based web server running on top of a fedora repository. For the purposes of being able to push usage statistics to a PIRUS2 compliant endpoint, the web server logs messages to a queue in another server, every time a resource is accessed. The format of this message is the standard Apache combined log format (see example below).

```
163.1.203.75 - - [26/May/2011:17:43:53 +0100] "GET
/objects/uuid%3Adcbaea88-2430-4520-9d09-68a52d3400f0 HTTP/1.1" 200 -
"http://ora.ouls.ox.ac.uk/" "Mozilla/5.0 (Windows NT 5.1; rv:2.0.1)
Gecko/20100101 Firefox/4.0.1"
```

The PIRUS2 plugin provides two key functionalities – parsing the log-line and gathering the open URL parameters. The functions *parseline* and *get\_openurl\_params* found in *plugins.ora\_utils* perform these base functions.

The PIRUS2 daemon consists of three different python files - *pirus2.py*, *logfromqueue.py* and *broker.py*. These 3 files are run under Supervisor, a process control system, that allows us to monitor and control processes.

Of these 3 files, *pirus2.py* is the main function which gets the log-line, calls the functions *parseline* and *get\_openurl\_params*, constructs the open URL and sends it to the Pirus endpoint using http.

The supervisor configuration file *worker\_pirus2.conf* spawns 3 PIRUS2 processes. The functionality of *pirus2.py* is detailed below.

- Listen to the queue *pirus2*.
- Read the different parameters from the configuration file *loglines.cfg*
  - Read the plugin name and import it (import plugin *plugins.ora\_utils* which has the functions *parseline* and *get\_openurl\_params*)
  - Read the time for delay on fail (*pauseonfail* = 3600) or set it to a default value of 300 seconds.

- Read the time for ratelimit (ratelimit = 0.3) or set it to a default of 1 second. This is used to rate-limit the function calls to *pirus2.py*
- Pop the access-log-line from the *pirus2* queue
- Parse the log-line using the function *parseline* in *plugins.ora\_utils*
  - The function *parseline* expects a json-encoded dictionary containing the access log line

```
def parseline(jmsg) where:
```

```
jmsg is a JSON encoded string, the default being:  
{'service':'service_unique_id', 'logline':'line from the logger  
that is to be parsed'}
```

- Split the log line based on the defined pattern
- Check if the return code in the log line is 200 or 304 and the id of the resource accessed is of interest
- Lookup the values for "*title, host, version, family, issn, eissn, doi, collection, content\_type*" in the Fedora repository for the object
- *parseline* returns a dictionary of terms that it was able to extract from the log line and the lookup fields
- Construct the open URL using the function *get\_openurl\_params* in *plugins.ora\_utils*. This function also checks to see if the line relates to an item Pirus is interested in (download of a full text journal article) or if it contains the minimum metadata pirus requires.
  - This function constructs the open URL in line with its requirements, using the values in the dictionary of terms returned by *parseline*.
  - The function *get\_openurl\_params* expects an instance of the config parser (containing a parsed version of the configuration file *loglines.cfg*, the name of the relevant section in the configuration file (in our case *pirus2*) and the dictionary of terms obtained from the *parseline* function. The configuration file is used to define custom variables needed by open URL like *url\_version, service-date, referer id...*

```
def get_openurl_params(c, worker_section, pl) where:
```

```
c - ConfigParser instance, containing a parsed version of  
'loglines.cfg' so you can include whatever variables in this that  
you require
```

```
worker_section - this will be the 'pirus2' section of the  
configuration for the worker that uses this plugin. For custom  
configuration data, please use a section prefixed with your  
plugin's name to curb collisions)
```

```
pl - the parsed dictionary that came from 'parseline' above
```

- The function should return a dictionary containing all the parameters that are to be sent to a PIRUS2 endpoint or an empty dictionary if the log line does not meet Pirus2's criteria
- Analyze the response from the above function and act accordingly
  - \*\* If the access-log-line is not an article, push the line to the *otherlog* queue
  - If it is an article and the parameters have been obtained, send the open url parameters to the Pirus endpoint URL.
    - Encode the URL parameters

- Make a URL request
- \*\* If the response is successful, push the line to the success queue *articlelog*
- If the response is not successful, push the line back to the same queue, wait for *delay\_on\_fail* time and retry

\*\* After processing the access-log-line, the log line is pushed back into one of two queues – *articlelog* or *otherlog*. The function *logfromqueue* reads these other two queues and writes the messages into a log file, for record keeping.

- The relevant section from the loglines configuration file is copied below.

```
[worker_pirus2]
listento = pirus2
repository_plugin = plugins.ora_utils
# OpenURL default details
endpoint_url = http://cclibweb-4.dmz.cranfield.ac.uk/tracker/
url_ver = Z39.88-2004
rfr_id = ora.bodleian.ox.ac.uk
# Expected HTTP status for success
success = 200
# Timeout for request
timeout = 60
# pause for 3600 seconds (1 hr) if fail to push request
pauseonfail = 3600
# Rate-limit (seconds in between requests per process)
ratelimit = 0.3
# Where to pass on loglines on success (comment out to ignore)
success_queue = articlelog
# Where to pass on loglines that aren't relevant to pirus2
other_queue = otherlog
stdout_logfile = workerlogs/pirus2.log
```

**logfromqueue.py** reads the queue *articlelog* or *otherlog* and writes the line to a logfile (logs/articles.log / logs/other.log).

There are two supervisor workers that call this function – *logger\_articlelogger* and *logger\_otherlogger*. The relevant lines from supervisor configuration files and sections of the loglines configuration file *loglines.cfg* are copied below. Each of these workers start just 1 process.

```
logger_articlelogger.conf
./logfromqueue.py %(process_num)s logger_articlelogger

[logger_articlelogger]
listento = articlelog
logfile = logs/articles.log

logger_otherlogger.conf
./logfromqueue.py %(process_num)s logger_otherlogger

[logger_otherlogger]
listento = otherlog
logfile = logs/other.log
```

**broker.py:** The other process that is started by supervisor is *broker.py*, which for now does nothing significant, but can be used if the access rates increase significantly, thereby increasing the number of access log lines, thus requiring us to fan out more processes to parse these log lines and construct open URL parameters.

*broker.py* pushes messages from the *listento* queue to the *fanout* queues, both of which are in the



Project Acronym: PIRUS2  
Version: 1.0  
Contact: Paul Needham ([paul.needham11@btinternet.com](mailto:paul.needham11@btinternet.com))  
Date: 06 /10/2011

*loglines* configuration file. At present, this is just reading from the *loglines* queue and pushing to the *pirus2* queue.

The *fanout\_status\_queue* (*broker\_temp*) is used to get the name of the next fanout queue to push to, if one is configured. In order to achieve this, all of the fanout queue names are pushed into the *fanout\_status\_queue*.

The supervisor configuration file *worker\_broker.conf* calls this functions. It starts three processes. The relevant section from the *loglines* configuration file is copied below.

```
[worker_broker]
listento = loglines
command = ./broker.py
fanout = pirus2
fanout_status_queue = broker_temp
# Time in seconds to sleep if there is nothing on the queue
idletime = 1
stdout_logfile = workerlogs/broker.log
```

### **PIRUS2 Daemon implementation at the University of Hull**

On a practical front, the team at Hull were able to install the PIRUS2 daemon; however they weren't sure what to point it at to enable it to do its job properly. The logs that they had in Fedora didn't seem to relate to what the daemon was looking for so they could be analysed for COUNTER purposes.

As a consequence, it proved impossible for Hull to transmit usage data to PIRUS, and it became obvious that it couldn't be rolled out more widely, in its present form.

### **Fedora background**

Fedora Commons is a digital object repository. Unlike out-of-the box repository systems such as DSpace and EPrints which provide both the data storage and user interface layers, Fedora provides only the data storage layer and does not provide a user interface. Implementers of Fedora provide their own user interface, either developing this from scratch or by re-using/adapting existing Fedora web interfaces and frameworks.

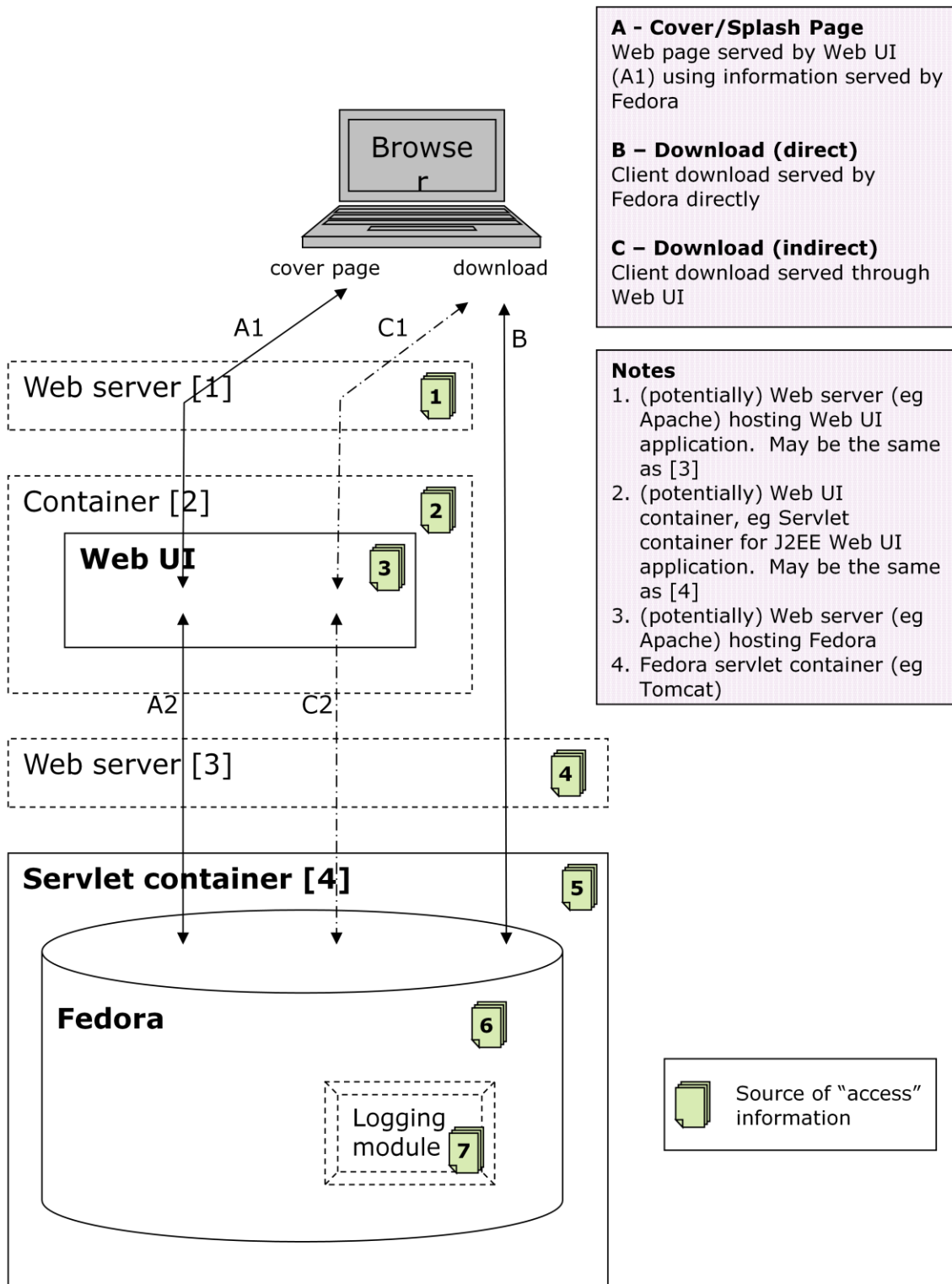
This fundamental difference between a repository based on Fedora and those based on DSpace or EPrints provides some challenges in logging access/downloads, significantly that there is no single consistent point of collecting access/download information.

Fedora repository implementations vary in their architectures (see Figure 6 below).

A web user interface is responsible for providing user navigation and access to repository content. This will include generation of "cover/splash pages" for repository content [A], and for providing access to download content (which is the area of interest to PIRUS).

Downloads may be served indirectly through the web UI [C] with Fedora serving the content behind the scenes, or may be served directly through the Fedora REST API [B]. Potentially both routes for content downloads may coexist within the same repository implementation, and there may be more than one web interface over a single Fedora repository installation. Furthermore, direct access [B] to repository content may be available outside of the web UI, for instance via download links provided by harvesters and aggregators.

# PIRUS2 - Fedora Access Logging



**A - Cover/Splash Page**  
 Web page served by Web UI (A1) using information served by Fedora

**B - Download (direct)**  
 Client download served by Fedora directly

**C - Download (indirect)**  
 Client download served through Web UI

**Notes**

1. (potentially) Web server (eg Apache) hosting Web UI application. May be the same as [3]
2. (potentially) Web UI container, eg Servlet container for J2EE Web UI application. May be the same as [4]
3. (potentially) Web server (eg Apache) hosting Fedora
4. Fedora servlet container (eg Tomcat)


 Source of "access" information

Figure 6. Typical Fedora architectures

Access/download information can be provided by various components in the overall architecture, and an analysis of these is presented in Table 2:

Source		Scenario	Usage event information						
No.	Description		IP Address	Session ID (eg cookie)	Username	Datestamp	MIMEType	User Agent	Resource identifier
1	Web UI web server logs	B	No	No	No	No	No	No	No
		C	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
2	Web UI container logs	B	No	No	No	No	No	No	No
		C	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
<b>3</b>	<b>Web UI application logs</b>	B	No	No	No	No	No	No	No
		C	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
4	Fedora web server logs	B	<b>Yes</b>	<b>Yes</b>	Maybe [1]	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
		C	No [2]	No [2]	Maybe [3]	<b>Yes</b>	<b>Yes</b>	No [2]	<b>Yes</b>
<b>5</b>	<b>Fedora servlet container log</b>	B	<b>Yes</b>	<b>Yes</b>	Maybe [1]	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
		C	No [2]	No [2]	Maybe [3]	<b>Yes</b>	<b>Yes</b>	No [2]	<b>Yes</b>
6	<b>Fedora log file</b>	B	No	No	Maybe [1]	<b>Yes</b>	No	No	<b>Yes</b>
		C	No	No	Maybe [3]	<b>Yes</b>	No	No	<b>Yes</b>
7	<b>Fedora access logging module</b>	B	<b>Yes</b>	<b>Yes</b>	Maybe [1]	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>
		C	No [2]	No [2]	Maybe [3]	<b>Yes</b>	<b>Yes</b>	No [2]	<b>Yes</b>

**Table 2. Sources of information mapped to PIRUS2 information requirements**

**Notes:**

- Scenarios B and C refer to the diagram – B: Direct download from Fedora; C: Indirect download via Web UI
- Sources in **Bold** are ones that would be expected to be present in all Fedora implementation architectures.
- Source repository information is not included in the table, it's assumed that the Daemon configuration would provide this
- "Yes" and "No" are not absolute. "Yes" indicates that information is potentially available at this level, but may or may not be captured in log files. "No" indicates the most likely scenario, but potentially the information could be passed through by layers above this level.
- Provision of resource metadata (title, author, etc; eg through OAI-PMH) is not explored

[1] Depends if Fedora requires authentication

[2] As access is from Web UI, not from client

[3] Fedora may be accessed using a "system" login; or Web UI and Fedora authentication/authorisation may be integrated

Significantly it should be noted that where content is served indirectly through the web user interface, certain information required by PIRUS2 may not be present in either Fedora's log files or those of the J2EE application server hosting Fedora – for instance the IP address recorded will be the IP address of the machine providing the web UI rather than that of the user downloading the content, and the user name may be a generic system login used by the UI to connect to Fedora. Similarly logs collected from the web UI tier will not include direct downloads via the Fedora API.

Fedora has no direct capabilities for access logging. However it has an extensible and modular architecture which facilitates the development of a "pluggable" access logging module, perhaps

building on the existing JMS capabilities of Fedora. This scenario is included in Table 2 (row 7) above, however from analysis of the various options it was noted that this provides no additional information over the existing servlet container logs.

Analysis of the various options in comparison with the architectural scenarios has shown that in order to support PIRUS2 there is a need to collect logging information in more than one place; and a combination of web UI logs and the Fedora servlet container logs would provide the necessary detail.

### **Requirements for a generic Pirus2 Daemon**

The end requirement for a Fedora daemon is that it is able to transmit the information given in the updated PIRUS2 OpenURL key-value pair specification (see Table 1, p25) to a remote third-party server, when an item (the content – not the record view) is accessed/downloaded from the Fedora repository.

### **Choice of software**

#### **A: Software platform – implementation**

The Oxford Daemon will only run on Linux/Unix systems due to its dependencies, and will not run on Windows. It requires the installation of Python and several dependent components which systems administrators/datacentre managers may be resistant to.

#### **B: Sustainability – ongoing maintenance**

Sustainability, particularly ongoing maintenance, of the PIRUS2 Fedora component is an important factor in determining the way forward, and this has an impact on decisions on the software platform/programming language of the component. Although Python skills are not uncommon it is suggested that as Fedora is written in Java this would be a more sustainable platform for the future, as Fedora implementation sites are likely to have Java skills and therefore there is more likelihood of building a community of maintainers.

In attempt to take this forward, the group sent an email to the UK&I Fedora and JISC-REPOSITORIES lists asking for feedback to indicate which of the following software platforms would be acceptable for such a module, and where more than one platform is acceptable to indicate their preference(s):

- a) Python
- b) Perl
- c) Java
- d) Other (please specify)

And, to ask: would you be willing to contribute usage statistics using such a module?

Sadly, there was not a single response!

### **Architectural requirements**

Three fundamental architectural components have been identified for providing transmission of information from Fedora repositories to PIRUS2:

#### **A: Log file entry collection**

- collection of individual entries from log files as they are written, and sending of these log entries to a log queue for subsequent processing by C.
- Processing of historic log file entries (i.e. before the software was installed, before the log watcher was started, dealing with failure of the log watcher component) and transmission of these log entries to the log entry queue

#### **B: Log message queue**

- Persisting log entries transmitted by A, so they can be subsequently picked up and processed by C

#### **C: Log entry processing and transmission to PIRUS2**

- Picking up entries from the queue in B
- Parsing entries

- Combining entries from different sources and working out which one to use (e.g. direct vs. indirect download)
- Working out the type of resource (needed for PIRUS, but not needed for IRUS!)
- Determine the OAI identifier (and/or optionally a URL for a DC record if the repository doesn't do OAI-PMH)
- Constructing and sending the message to PIRUS2

#### **Notes:**

- the Oxford Daemon has these components
- an asynchronous processing requirement is assumed, hence the need for a queue. Particularly if the web application sits on a different box, and for the case of multiple web applications over the same repository; and to cope with failure of software components
- B and C logically sit on the same box – but potentially A sits on a different box; particularly if there are different boxes for the Fedora repository and the web UI
- Architectural implications from the above on the choice of platform:
  - (A) may need to be installed on a platform that does not have java installed, e.g. a separate web UI on a different box running e.g. Ruby on Rails (e.g. Hydra) – potentially there could be a barrier to installing Java. If there were sites not willing to install Java we should ensure that the log message queue protocol/standards are not only Java to enable people to write their own log watchers in a non-Java stack and still be able to send the messages to (B)
  - (B, C) could logically sit on the Fedora server (along with an A for that server) – so Java is not an issue. However, it is worth noting that Oxford's preference was for a separate box for this; with a lightweight footprint and therefore Java was not a preferred option. But of course they can just continue using the system they have already
- We believe that repositories would have OAI-PMH and we can therefore rely on it; but an alternative is included, i.e. providing a URL for access to the resource's DC record in any case

#### **Outline functional description**

As already noted, there is no existing access logging component within Fedora, but this does not necessarily mean something needs to be built from scratch. It is possible to use existing facilities.

Given that most of the users are using Fedora's rest API to access fedora, then there are the logs from Tomcat / Jetty / Apache - depending on how one has installed Fedora. As a java webapp, Fedora needs to be installed within a web server like Apache-Tomcat or Jetty - and access logs are provided by all of these web servers.

An alternative approach - to add access logging as a plug-in module without needing to modify any core Fedora code – was considered. However, we concluded that a logging component built for Fedora would offer no additional value over the log files already provided by the application server, (Tomcat, Jetty etc.).

If using tomcat, for example, it would just be necessary to enable access logging in Tomcat. This involves editing one config file (server.xml). This would provide standard access logs, including the remote IP address / hostname and user-agent.

Since the format of the logs is a standard, the same log file parser could be used, irrespective of how Fedora is installed and accessed – through Tomcat / Jetty / using a proxy in Apache.

Following this route, steps to meet this requirement are as follows:

#### **1. Gather access log entries**

- Collect individual log entries from both the Fedora application server container and from the web user interface(s)
- Store these entries in a queue/store for later processing

**Notes:**

- If using multiple access logs (access log from web UI and from the Fedora servlet), there is a need to remove duplicate records from the access log of the Fedora repository.
- Options for gathering log entries:
  - (A) a simple file copy of logs – e.g. copy logs when the logging system rotates them
  - (B) a “log watch” type of implementation that watches for individual entries when they arrive and transmits them.
  - The Daemon does (B). (A) is probably not that simple in practice, i.e. logic to determine when a new log is started and the old one has been rotated

**2. Parse and Merge log file entries and identify the entries to use**

- Take the entries from the queue/store and merge these
- Parse the entries
- Identify if there is more than one log entry for the same event, and select the appropriate one

**3. Identify if entry is of interest to PIRUS (text-based article)**

- *Could do the first level of filtering just based on request URL in the log line*
- *In a repository containing different kinds of items – text, video..., Oxford would need to determine the type of record. For this we would either identify the type from the DC datastream for the record or check if the record belongs to the collection ora:articles in the rels-ext. Other repository implementations may have different ways of identifying content type, including the collection the resource belongs to, the Fedora content model of the item, or other metadata stored with the object. This component needs therefore to be pluggable. It also may need to cope with atomistic vs. compound object models (ie (1) multiple Fedora objects representing the same resource, ie separate Fedora objects for different representations – PDF, text, etc with a parent “resource” object and (2) a single Fedora object for the resource.*

**Notes:**

- Required step for PIRUS
- Step not required for IRUS which will accept all item types

**4. Determine the OAI identifier used to construct the URL for the OAI-PMH DC record**

- (optionally provide an alternative – a URL for a DC record where OAI-PMH isn't implemented).
- Information from the log file
  - IP address of client that made the request
  - User id of the person requesting the resource
  - Time of request
  - request line and HTTP status code (to identify the resource requested)
  - Referer
  - User-agent
- The OAI identifier, if different from the identifier of the object, in most cases could be mapped using a simple text conversion.

**5. Send to Pirus2**

**Conclusions**

The work of this group has gone a long way towards contributing to understanding the general requirements for enabling PIRUS2/IRUS functionality in a Fedora repository.

It is clear that, because of the nature of Fedora, there is not a 'one size fits all' solution.

The Oxford Daemon is architecturally sound and would provide a good starting point for further developments. However, there are potential barriers in its general application:

- sustainability - who's going to have the (python) skills to keep it going
- and implementation platform issues, whether sysadmins would be happy to install the software and dependences (and, it is strictly non-Windows, which potentially could be a show-stopper in itself, though Windows is not a common platform for Fedora)

The Java route addresses these - but there is likely to be resistance to this in some cases. For example, at Oxford one of their goals was to run this on a separate small-footprint box, which is one reason for the Python approach; Java could be considered as a bit 'heavyweight' for this. There are of course other potential platforms, but in the absence of any feedback (e.g. the email lists), it's difficult to recommend one - and Java does generally seem to be the lowest common denominator.

It is unlikely that a global Central Clearing House for article statistics will be established in the short term, so there is actually no immediate, urgent need for a global solution.

However, if there were to be further research into and development of a UK Institutional Repository Usage Statistics service, this would present an opportunity to:

- develop solutions to implementing IRUS functionality – based on the outline functional description - in the small number of active Fedora IRs in the UK
- build a series of case studies of implementations at these repositories
- and build a knowledge base of commonalities and applicable techniques for the wider Fedora community

## ***Publisher survey: economic models for the Central Clearing House***

### **a. Introduction**

One of the principal outcomes of the PIRUS2 project has been a proposed organizational and economic model for the Central Clearing House (CCH). At the End of Project Seminar on 23 February this was one of the main subjects for discussion, with both the publisher and repository representatives expressing concern about the level of tariffs proposed for publishers and repositories to use the services of the CCH. Having received this feedback it was agreed that we would review the economic model for the CCH, as well as the proposed tariffs.

As a first step we reviewed the overall costs and tariffs proposed with 2 other suppliers of similar usage statistics services, who have confirmed that the overall level of costs are very reasonable and could not, realistically, be significantly lowered.

Having validated the overall level of the costs, we reviewed the proposed tariffs and have agreed lower 'entry level' tariffs, both for Small Publishers and for Repositories.

### **b. PIRUS2: Role of the Central Clearing House**

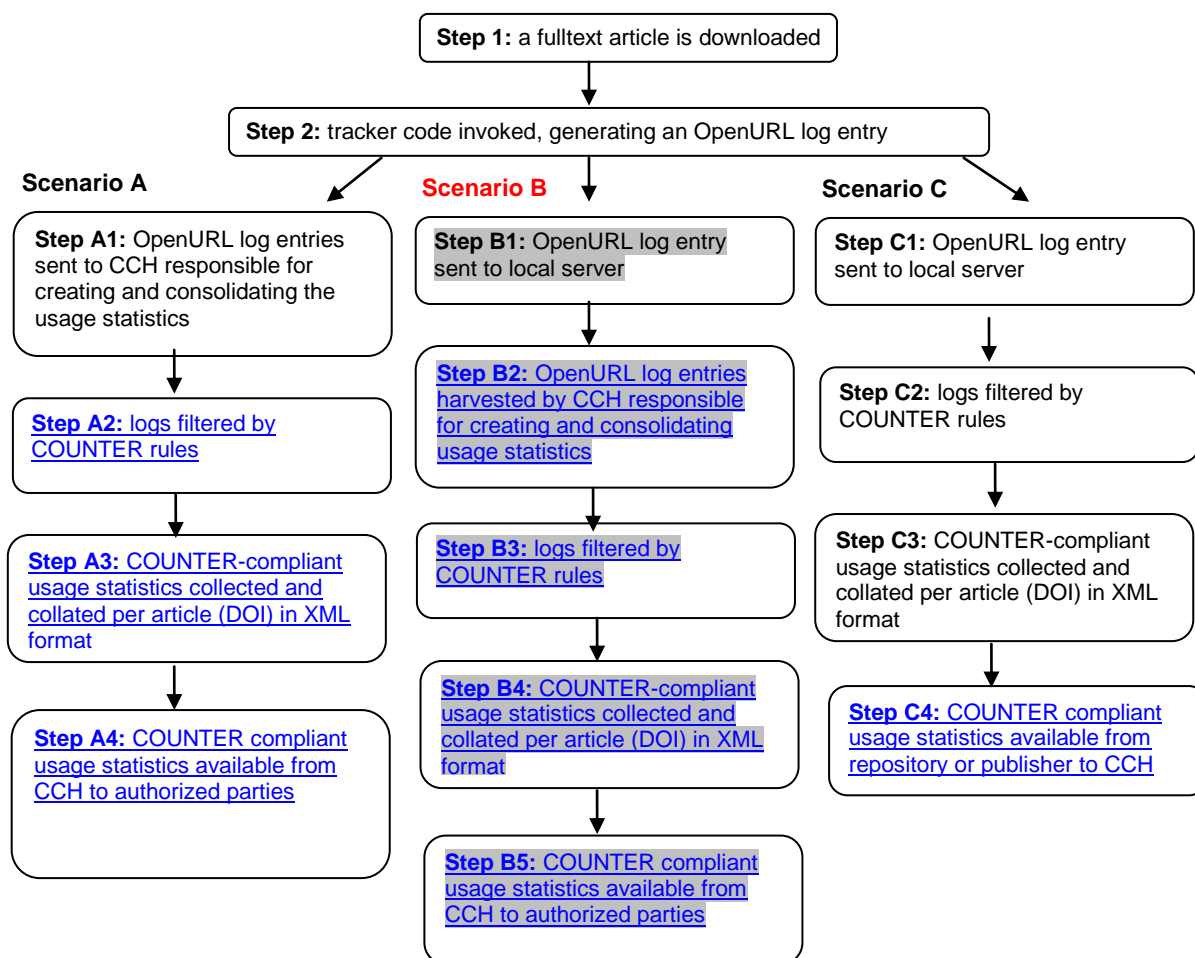
The CCH will have two broad roles. First, to collect, consolidate and process usage data from repositories and publishers. Second, to create and distribute usage reports to authorised parties (mainly publishers and repositories)

One recommendation of the original PIRUS project was that the CCH should be able to support three scenarios (A, B and C in Scheme 1, below) for the collection of usage data.

Scenarios A and B are likely to be prevalent among institutional repositories, while Scenario C will be prevalent among participating publishers and larger repositories. In the course of testing the repository usage data, however, the project team came to the view that we should drop Scenario B and adopt Scenario A (i.e. the tracker (push) approach rather than use the OAI (pull) approach) for repositories that cannot implement Scenario C (the great majority).

#### Scheme 1

Steps highlighted in blue text will take place in the CCH.



### c. PIRUS2 Central Clearing House: Economic Model

The tariffs outlined below are based on figures provided by a vendor with extensive experience in generating the existing COUNTER usage reports for a number of publishers. We have confirmed independently that the cost base for these tariffs is reasonable.

The proposed tariffs for both Scenario A and Scenario C are provided. Please bear in mind that these figures are necessarily averages, designed to provide a general picture of the likely tariffs for different sizes of publisher. If the project proceeds to implementation there will be a formal tender process to select the organizations that would be involved in its management and the generic tariffs listed below would be negotiated individually with each publisher.

#### a) PIRUS2 CCH - model for allocation of costs to publishers - scenario A

##### Assumptions

- a. the basis for the tariffs for publishers will be a combination of total revenues and article download activity
- b. A **Large Publisher** is defined as one with i) annual revenues in excess of \$50 million and ii) more than 100 million full-text article downloads per month (Assume 150 million per month for the calculation)
- c. A **Medium Publisher** is one with i) annual revenues of between \$5 and \$50 million and ii) between 20 and 40 million full-text article downloads per month (Assume 30 million per month for the calculation)
- d. A **Small Publisher** is one with i) annual revenues of less than \$5 million per annum and ii) less



than 500k downloads per month (Assume <500k per month for the calculation)

Annual tariffs are built up as follows:

- Annual Membership Fee ( \$38,700 for a Large Publisher; \$9,700 for a Medium Publisher; \$1,700 for a Small Publisher)
- Reporting Services Costs ( Setup cost : \$58,500; annual infrastructure and operational costs: \$103,200)

	<b>Large Publisher</b>	<b>Medium Publ.</b>	<b>Small Publ.</b>
Membership fee	\$38,700	\$9,700	\$1,700
Reporting services			
Y1	\$808	\$808	\$808
Y2 +.....	\$515	\$515	\$515
Transaction-based fee costs	\$630,000	\$264,000	\$4,800
<b>Total</b>			
<b>Y1</b>	<b>\$669,508</b>	<b>\$274,508</b>	<b>\$7,308</b>
<b>Y2+.....</b>	<b>\$669,215</b>	<b>\$274,215</b>	<b>\$7,015</b>

**b) PIRUS2 CCH - model for allocation of costs to publishers - scenario C**

In Scenario C the Publisher, or the Publisher’s own vendor, creates the usage reports, which are harvested by the CCH for consolidation

**Note:**

- Publishers will be charged only an annual flat fee, based on annual revenues, by the CCH for the Scenario C service
- Publishers who do not wish to have their usage reports harvested by the CCH for consolidation, but instead wish to harvest usage data from the CCH for consolidation into their own usage data would be charged an annual fee of approximately 50% of the Scenario C tariffs below

Assumptions

- A **Large Publisher** is defined as one with annual revenues in excess of \$100 million
- A **Medium Publisher** is one with annual revenues of between \$5 and \$10 million
- A **Small Publisher** is one with annual revenues of less than \$1 million per annum

Annual tariffs are built up as follows:

	<b>Large Publisher</b>	<b>Medium Publisher</b>	<b>Small Publisher</b>
<b>Annual tariff</b>			

Annual fee	\$66,900	\$16,900	\$2,900
<b>Total</b>	<b>\$66,900</b>	<b>\$16,900</b>	<b>\$2,900</b>

#### **d. Publisher Responses to Questions**

18 COUNTER compliant publishers were included in the Survey. 12 provided responses and their responses to each question are summarised below:

##### **Question 1: Which scenario would you prefer to implement?**

Scenario A 2  
Scenario C 10  
Other Scenario (please describe):

1. Publisher gathers and consolidates usage data on its own

##### Publisher Comments

1. To be perfectly honest, we are not 100% sure what value the CCH actually would add. Since we are already collecting COUNTER compliant usage statistics on an article level, it would be easy for us to 'publish' COUNTER compliant article level statistics on our own servers (just as we do with the COUNTER journal and book reports).

##### **Question 2: Into which Publisher category does your organization fit?**

Large 2  
Medium 7  
Small 3

##### Publisher Comments:

##### **Question 3: Do you find the overall level of the proposed tariffs reasonable?**

Yes 2  
No 10

##### Publisher Comments

1. Is it worth investigating 5 tiers in order to spread the costs across a broader range of company sizes? There seem to be some large step-changes with just three.
2. The costs for scenario C are preferable, particularly because this scenario would allow our vendor to supply statistics that would be consistent with the COUNTER-reports we supply directly to customers. However we feel that in both scenarios – A and C, the tariffs are too high, particularly because we are not convinced of the benefits of article-level metrics to our business and to our customers
3. The amount would be quite substantial in our case. And, as mentioned, we don't see a big 'value add' by the CCH (yet).

##### **Question 4: Do you find the proposed model for allocation of costs reasonable?**

Yes 4

No 8

Publisher Comments:

1. We already spend a considerable sum on creating the COUNTER usage reports and would find it hard to justify additional expenditure on usage statistics
2. Who would find individual article usage data useful? A low priority for us at the moment.

**Question 5: Have you any other comments on the proposed role for the CCH, or on the tariffs associated with it?**

1. As noted previously, I think the costs are high and will significantly limit involvement of smaller publishers and repositories.
2. To summarize, we feel that:
  - We do not yet have enough information to convince us of the benefits to us or our customers of article-based usage reporting
  - We feel both preferred scenarios, A and C are too expensive
  - The publisher bandings do not seem consistent
3. Assuming we want to consolidate the usage data into reports for our web site, would repositories that have our articles be able to request our usage data from CCH? I think that would not be something we would want.

**e. Conclusions**

1. It is clear that the great majority of the publishers still find the proposed tariffs too high. In follow-up phone-calls it became clear that there is a great reluctance to incur **any** extra costs to support a PIRUS2 service at this time.
2. Most publishers prefer Scenario C, where they themselves generate the usage statistics for consolidation by the CCH, but still appear to have problems with the relatively modest costs involved
3. There is clear evidence of growing author demand for usage statistics for their own articles and at some point publishers will see a competitive advantage to providing this.

## Conclusions

PIRUS2+ has demonstrated that:

- While it is technically feasible to create consolidate and report usage at the individual article level based on usage data from a range of sources based on different platforms, there are considerable economic barriers to publishers participating in such a service at this time
- It is feasible to set up an IRUS service in the short term
- While it is feasible to set a standard for adoption by publishers and repositories for the recording and reporting of individual article usage statistics, there is not enough support to set up a comprehensive CCH in the short term. Instead, it would make sense to set up an IRUS service for repositories, while encouraging publishers to adopt the PIRUS standard for their own, article-level usage statistics