September 2010

PIRUS2: Creating a Common Standard for Measuring Online Usage of Individual Articles

Peter Shepherd

Project COUNTER, pt_shepherd@hotmail.com

Paul Needham

Cranfield University, UK

Follow this and additional works at: <https://docs.lib.purdue.edu/atg>

Part of the [Library and Information Science Commons](#)

Recommended Citation

Shepherd, Peter and Needham, Paul (2010) "PIRUS2: Creating a Common Standard for Measuring Online Usage of Individual Articles," *Against the Grain*: Vol. 22: Iss. 4, Article 10.

DOI: <https://doi.org/10.7771/2380-176X.5598>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

PIRUS2 (Publisher and Institutional Repository Usage Statistics): Creating a Common Standard for Measuring Online Usage of Individual Articles

by **Paul Needham** (Cranfield University, UK)

and **Peter T. Shepherd** (Director, Project COUNTER, Edinburgh, UK) <pt_shepherd@hotmail.com> www.projectCounter.org

Introduction

One of the more significant developments since scholarly articles have been published online has been the growing role of institutional and subject repositories as hosts for these articles. The publishers of journals, though still the most important hosts, no longer have a monopoly of the distribution of these articles that they enjoyed in the print world. This trend has been given considerable further momentum by the Open Access movement, which encourages the free availability of the outputs of scholarly research, especially where that research has been publicly funded.

A reader searching online for a particular article may now find it in a number of different locations:

- the main journal publisher Website (e.g., **Elsevier's ScienceDirect**)
- a content aggregator site (e.g., **ProQuest**)
- a subject repository (e.g., **PubMed Central**)
- the author's local institutional repository (e.g., **Oxford University Research Archive - ORA**)

It is not the purpose of this article to argue the pros and cons of this highly distributed system for the publication of scholarly articles, still less to present the case for or against Open Access publishing. Rather, we accept that

these trends are now well established and that any system for recording and reporting online usage of articles must take them into account. This makes the task of counting usage at a global level rather challenging. For a start, it will no longer suffice to record and report usage at the journal level: the journal as a package of articles is used by publishers, but not by repositories, which are organised on the basis of individual items. Then we have to consider the status of different versions of articles and which versions should be counted. Clearly, the accepted version of an article, or the published version of record has higher status than the author's initial draft, but does this mean that usage of the latter should not be counted at all, or does it mean that such usage should be weighted differently? These and other issues become highly pertinent in this increasingly heterogeneous publishing environment, and the aim of the **PIRUS (Publisher and Institutional Repository Usage Statistics)** project is to address them.

COUNTER as a Basis for Individual Article Usage Statistics

Currently the only widely implemented global standard for measuring online usage of scholarly information has been set by **COUNTER**, but until now the most granular level at which **COUNTER** requires reporting of usage is the individual journal. Demand for usage statistics at the individual article level

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. 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 evaluation of research outputs.
- 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** protocol facilitates the automated consolidation of large volumes of usage data from different sources.

Aims and Objectives of PIRUS2

The aim of **PIRUS2** is to specify **COUNTER**-consistent standards and protocols (as well as an infrastructure and an economic model) for the recording, reporting, and consolidation of online usage of individual articles hosted by repositories, publishers, and other entities.

In order to achieve this overall aim, the project will seek 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 repositories.

continued on page 28

MESUR: A Survey ... from page 24

Proceedings of the National Academy of Sciences, 102(46), 16569-16572. DOI: 10.1073/pnas.0507655102.

Johan Bollen, Marko A. Rodriguez, and Herbert Van de Sompel (2006). Journal status. *Scientometrics*, 69(3), DOI: 10.1007/s11192-006-0176-z.

C. T. Bergstrom (2007). Eigenfactor: Measuring the value and prestige of scholarly journals. *C&RL News*, 68(5).

Loet Leydesdorff (2007). Betweenness centrality as an indicator of the interdisciplinarity of scientific journals. *Journal of the American Society for Information Science and Technology*, 58(9), 1303—1319.

Bollen J., Van de Sompel H., Hagberg A., Bettencourt L., Chute R., et al. 2009 Clickstream Data Yields High-Resolution Maps of Science. *PLoS ONE* 4(3): e4803. DOI:10.1371/journal.pone.0004803.

Bollen J., Van de Sompel H., Hagberg A., Bettencourt L., Chute R., 2009 A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE* 4(6): e6022. DOI:10.1371/journal.pone.0006022.

Johan Bollen, Marko Rodriguez, Herbert Van de Sompel, Lyudmilla Balakireva, and Aric Hagberg. The Largest Scholarly Semantic Network...Ever (poster). In *Proceedings of the 16th International World Wide Web conference*, May 2007. Note: Best Poster Award!

Marko. A. Rodriguez, Johan Bollen and Herbert Van de Sompel. A Practical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage. In *Proceedings of the Joint Conference on Digital Libraries*, Vancouver, June 2007.

Johan Bollen, Herbert Van de Sompel and Marko A. Rodriguez. Towards usage-based impact metrics: first results from the MESUR project, *JCDL 2008*, Pittsburgh, PA, June 2008. (arXiv:0804.3791v1, best paper finalist). 🐼

- Develop a prototype article-level Publisher/Repository usage statistics service comprising a technical demonstrator and a set of business model recommendations for a central clearing house.
- Define a core set of standard useful statistical reports that repositories could/should produce for internal and external consumption.

Benefits of PIRUS2

The work of PIRUS2 will ensure that usage data are available for journal articles wherever held (publisher sites, repositories, aggregators), whilst going further than Web analytics software and more able to meaningfully address the consistency of the usage data and the resultant quality of the reports.

Repositories will benefit from a technical point of view as PIRUS2 will provide them with access to new functionality to produce standardised usage reports from their data.

Digital repositories systems will be more integral to research and closely aligned to research workflows and requirements, as the project addresses production of authoritative usage data.

The authoritative status of PIRUS2 usage statistics will serve to enhance trust across repositories; furthermore, the data will provide a firm evidence base for repositories to take firm steps to defining clear policies to support their goals.

Which Article Versions to Count?

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

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

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

Peer Review Status

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

Repositories Host More than Journal Articles

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

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

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

Metadata elements typically used when cataloguing articles in repositories include:

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

All repositories include Title, Author and Resource type metadata. Research carried out for PIRUS confirms that many repositories add citations identifying the published versions of articles in their records.

More than a purely Technical Challenge

The original PIRUS project (<http://www.jisc.ac.uk/whatwedo/programmes/pals3/pirus.aspx>), demonstrated 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. If this is to be translated into a new, implementable COUNTER standard and protocol, further research and development will be required, specifically in the following areas:

- **Technical:** further tests, with a wider range of repositories and a larger volume of data, will be required to ensure that the proposed protocols and tracker codes are scalable/extensible and work in the major repository environments.
- **Organizational:** the nature and mission of the central clearing house/houses proposed by PIRUS have to be developed, and candidate organizations identified and tested
- **Economic:** assess the costs for repositories and publishers of generating the required usage reports, as well as the costs

of any central clearing house/houses; investigate how these costs could be allocated between stakeholders

- **Political:** the broad support of all the major stakeholder groups (repositories, publishers, authors) will be required. Subject repositories, such as PubMed Central,

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

Table 1: Stages in the Publication of an Article

continued on page 30

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

PROGRESS ON PIRUS2 Standards and Protocols

A new **COUNTER** report, Article Report 1: Number of Successful Full-Test Article Requests by Month and DOI (AR1), has been developed. This provides a standard, **COUNTER**-compliant format for publishers and repositories for the submission of usage statistics at the individual article level. A specification for Article Report 1 is available on the **PIRUS2** Website in XML and MS-Excel formats at <http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/tiki-index.php?page=Project+Plan+and+Progress>.

In view of the wide range of repository softwares currently implemented and the different ways they operate, not all repositories will be able to provide Article Report 1, and it is impractical to propose a single approach that will work in all situations. For example, currently, when a full-text article is downloaded:

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

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

Diagram 1: Proposed Approaches to Recording and Reporting Usage Statistics for Repositories

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

Repository Software Applications

Although a few years old now, the Budapest Open Access Initiative (BOAI) report "A Guide to Institutional Repository Software v 3.0" (<http://www.soros.org/openaccess/software/>) provides an excellent introduction to repository software applications. It details and compares nine softwares available under an Open Source licence, namely: **Archimede**, **ARNO**, **CDSware**, **DSpace**, **Eprints**, **Fedora**, **i-Tor**, **MyCoRe**, and **OPUS**. It includes a useful System Feature & Functionality Table (http://www.soros.org/openaccess/pdf/OSI_Guide_to_IR_Software_Table_v3.pdf) providing summary comparison of the nine applications.

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

The Table below shows that, globally, four out of five (80%) of listed IRs are based on just five software applications:

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

Two-thirds of all repositories appear to be based on just two applications, **DSpace** and **Eprints**, while **Fedora**-based repositories appear to be under represented in the ROAR listings.

As part of the **PIRUS2** project plugins have been developed for three of the major repository software applications (**DSpace**, **Eprints**, and **Fedora**), and these are in the process of being tested in a range of repositories using these applications.

Repository Test Usage Data

Institutional Repositories are supplying usage data via:

- Diagram 1: Scenario A – push: tracker code sends an OpenURL log entry to a central clearinghouse
- Diagram 1: Scenario B – pull: the central clearinghouse will harvest usage data from IRs using OpenURL context objects via OAI-PMH
- Usage data are exposed as: (A) OpenURL Key-Value Pair Strings; (B) OpenURL Context Objects.

The OpenURL approach was first suggested by **MESUR** (<http://www.mesur.org/MESUR.html>) and taken forward in Europe under "Knowledge Exchange" – an initiative involving DEFF, DFG, JISC, and SURF foundation (<http://wiki.surfoundation.nl/display/standards/OpenURL+Context+Objects>).

Usage data must be filtered according to **COUNTER** rules to eliminate Robots and Double clicks and processed into monthly statistics.

continued on page 32

Creating a Common Standard ... from page 30

At this stage, the **PIRUS2** team consensus is that it is not yet appropriate for repositories to attempt to supply **COUNTER**-compliant AR1 reports. The AR1 standard is still being developed. Technically, it is challenging to incorporate **SUSHI** into the wide range of repository softwares, and there are issues, even among publishers, about the size of **SUSHI** reports, lack of compression, etc. Business-model-wise, it would incur costs/time/effort for each and every IR to undergo regular **COUNTER** audit for compliance.

Publisher Test Usage Data

Ultimately, publishers will supply AR1 usage statistics reports via **SUSHI**. However, the AR1 Report is not yet an agreed **COUNTER** standard, and **SUSHI** implementations are technically demanding both on the server and client sides, so — for the purposes of the tests — **PIRUS2** has agreed to accept data in MS Excel format. Test usage data is now being obtained from the following **COUNTER**-compliant publishers: **ACS Publications**, **Emerald**, **IOP Publishing**, **Nature Publishing Group**, **NEJM**, **OUP**, **Springer**, and **Wiley**.

So far test usage data for 450,000 individual articles from 5,500 journals has been collected and is being processed.

User Interface

A skeletal user interface is in place; its development and testing is ongoing.

Central Clearing House

We face two main challenges in attempting to create a Central Clearing House (CCH) to consolidate individual article usage statistics at

a global level. The first is primarily technical. Not only will the CCH have to receive and manage usage data from a range of publishers, but is also has to deal with the diversity of repository softwares and implementations that are in use.

The second challenge is in persuading repositories, publishers, and other organizations to participate in and support such a CCH service. Meeting this challenge will require us to demonstrate not only the benefits of providing global usage statistics at the individual article level but also that this can be done cost-effectively and reliably.

Functions to be fulfilled by Central Clearing House

It has been agreed that the CCH will have to perform the following basic functions:

1. Receive and store the following categories of data:
  - a. Open URL logfiles from repositories
  - b. **COUNTER**-compliant usage statistics from repositories, publishers, and other organizations
2. Harvest Open URL logfiles from repositories, publishers, and other organizations
3. Collect and collate usage statistics by individual article (DOI)
4. Store usage statistics by individual article for a specified period
5. Control access to the stored usage data

Capabilities required of the Central Clearing House

1. Conversion of logfiles to **COUNTER**-compliant usage statistics
2. Collection, collation, and storage of usage statistics

3. Collection, collation, and storage of relevant metadata
4. Creation and management of a Registry of Participating Repositories
5. Management of access control
6. Billing of costs to participating entities

Organizational options for Central Clearing House

Broadly speaking, there are two organizational options:

1. A global organization that would be responsible for carrying out all the functions listed above
2. A network of national/regional organizations that would carry out the functions listed above in their own nation/region

Organizationally, the favoured option is to go for a global organization, as this will make it easier to implement and adhere to standards, and we are now exploring this. International standards organizations already exist in STM publishing and have shown that it is possible to collect and collect large volumes of publication-related data on a global basis. It may well be that no single organization has, or wishes to develop, all the capabilities required, but one can imagine a partnership between organizations with complementary capabilities to create a global service.

Project Timetable and Further Information

Work on **PIRUS2** commenced in October 2009 and the project is scheduled for completion in December 2010. Further information on **PIRUS2** may be found on the project Website at <http://www.cranfieldlibrary.cranfield.ac.uk/pirus2>. 🐼

Consistent Squeeze

by **Gary Geer** (Collection Development Librarian, University of South Carolina)
<geer@sc.edu>

“Do you have a consistent squeeze?” asked my boss, **Alexander (Sandy) Gilchrist**. I didn’t understand what he was asking me. My task, sometime back in the 1980s, was to figure out how many books we had in certain subject areas. In the days before you could ask and get an answer to this question from a computer, we had to have a method to quickly and accurately count the number of titles on a particular subject. Most card catalog users were familiar with the author, title, and subject sections of the card catalog. The part of the catalog they never saw was called the shelflist. The shelflist was the whole card catalog, but arranged in call number order. It was located in the technical services area of the Library, and not generally accessible to our users. Since the **Library of Congress** call number system is a subject classification system, books with call numbers in the range E 441 to E 665, for

example, all deal with U.S. History, Slavery, and The Civil War. To get a reasonably accurate count of the number of titles in a subject area, we held the cards in that call number range straight in the catalog drawer, measured the width of that group of cards, and then would multiply by the number of cards per inch. To know how many cards there were per inch, you had to be able to squeeze the cards with a consistent amount of pressure while you measured. If you had a weak squeeze, the number of cards per inch might be 50, a strong squeeze and your average might be 75, so your squeeze could make a big difference in your count. It took some practice to get your squeeze consistent and to figure your cards per inch average. I don’t remember what my squeeze equaled in cards per inch. I suspect it’s a bit less today.

This is a skill they just don’t teach in library school these days. 🐼

Rumors from page 12

if you ask me. I have been reading an incredibly interesting book called *Hamlet’s BlackBerry: A Practical Philosophy for Building a Good Life in the Digital Age* (HarperCollins, 2010) by **William Powers** about this phenomenon. **Powers** wrote an earlier essay called “Hamlet’s BlackBerry: Why Paper is Eternal” in 2005/2007. I think we should have a book discussion group online about this. Anybody interested? <http://www.willampowers.com/about-me>

And did you know that **Elaine Robbins** (see above) is the new editor of *The Charleston Report (TCR)*? www.charlestonco.com

Speaking of *TCR*, the brainy **Laura Barfield**, Systems Librarian at Trident Technical College <laura.barfield@tridenttech.edu>, the last editor of *TCR*, just won an IMLS planning grant in her spare time. The project is called “Lowcountry Foodways.” [As] rapid

continued on page 38