PIRUS **Publisher and Institutional Repository Usage Statistics**

PIRUS2: Technical aspects

WP4 Software, standards and protocols

Paul Needham, Cranfield University
PIRUS2 Project Manager & WP4 leader

PIRUS2 End of Project Seminar - 23 February 2011

Funded by: JISC

PIRUS2: WP4 Software, standards, protocols

The technical challenge:

- A full-text journal article can be downloaded:
  - Directly from a Publisher web site
    - Or, via an Intermediary web site
  - from an Institutional Repository
  - from a Subject Repository
- In each case a usage event occurs
- Is it possible to capture those events from various sources and consolidate them to show the overall usage of that article???

PIRUS2: WP4 Software, standards, protocols

So, in PIRUS2, test if we can:

- Gather … usage data and statistics
  - From institutional repositories, publishers, etc.
  - For full-text article downloads (not record/abstract views)
- Consolidate …
  - In an article-level usage statistics demonstrator portal
  - Experiment and illustrate some possibilities
- Re-expose …
  - To authorized third parties

PIRUS2: WP4 Software, standards, protocols

Publisher/intermediary usage events

- Download events are logged as they occur
- Processed according to COUNTER rules:
  - Stripping out robot accesses
  - Eliminating double click entries
- COUNTER-compliant usage statistics are produced, reporting at the Journal level
- e.g. Journal Report 1 (JR1) Report: Number of Successful Full-Text Article Requests by Month and Journal
- Stats are shared with authorized parties via
  - MS-Excel/CSV files – manually downloaded
  - SUSHI – machine to machine

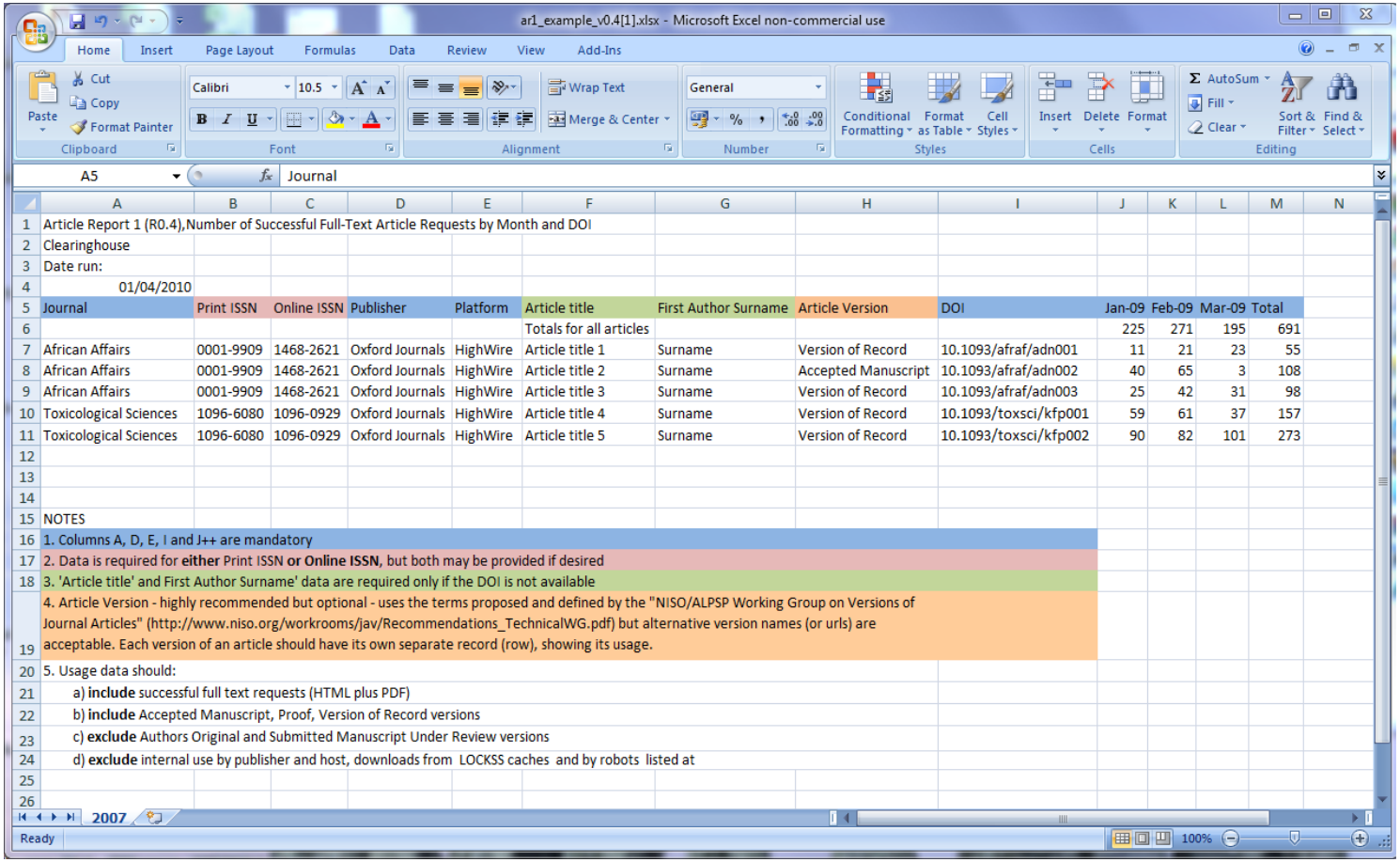
PIRUS2: WP4 Software, standards, protocols

Publisher/intermediary usage events

- Additionally, some produce usage statistics reporting at the article level
  - E.g. PLoS ... Others are aware of increasing demand
  - Important that an Industry Standard should be applied (like COUNTER) to make figures comparable
- To get publisher data in for PIRUS2, we suggested a number of Article Reports:
  - AR1 Report: Number of Successful Full-Text Article Requests by Month and DOI
  - AR1j Report: Number of Successful Full-Text Article Requests per Journal by Month and DOI

PIRUS2: WP4 Software, standards, protocols

AR1 example:



ar1_example_v0.4[1].xlsx - Microsoft Excel non-commercial use

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Article Report 1 (R0.4),Number of Successful Full-Text Article Requests by Month and DOI													
2	Clearinghouse													
3	Date run:													
4	01/04/2010													
5	Journal	Print ISSN	Online ISSN	Publisher	Platform	Article title	First Author Surname	Article Version	DOI	Jan-09	Feb-09	Mar-09	Total	
6						Totals for all articles				225	271	195	691	
7	African Affairs	0001-9909	1468-2621	Oxford Journals	HighWire	Article title 1	Surname	Version of Record	10.1093/afraf/adn001	11	21	23	55	
8	African Affairs	0001-9909	1468-2621	Oxford Journals	HighWire	Article title 2	Surname	Accepted Manuscript	10.1093/afraf/adn002	40	65	3	108	
9	African Affairs	0001-9909	1468-2621	Oxford Journals	HighWire	Article title 3	Surname	Version of Record	10.1093/afraf/adn003	25	42	31	98	
10	Toxicological Sciences	1096-6080	1096-0929	Oxford Journals	HighWire	Article title 4	Surname	Version of Record	10.1093/toxsci/kfp001	59	61	37	157	
11	Toxicological Sciences	1096-6080	1096-0929	Oxford Journals	HighWire	Article title 5	Surname	Version of Record	10.1093/toxsci/kfp002	90	82	101	273	
12														
13														
14														
15	NOTES													
16	1. Columns A, D, E, I and J++ are mandatory													
17	2. Data is required for **either** Print ISSN **or Online ISSN**, but both may be provided if desired													
18	3. 'Article title' and First Author Surname' data are required only if the DOI is not available													
19	4. 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) but alternative version names (or urls) are acceptable. Each version of an article should have its own separate record (row), showing its usage.													
20	5. Usage data should:													
21	a) **include** successful full text requests (HTML plus PDF)													
22	b) **include** Accepted Manuscript, Proof, Version of Record versions													
23	c) **exclude** Authors Original and Submitted Manuscript Under Review versions													
24	d) **exclude** internal use by publisher and host, downloads from LOCKSS caches and by robots listed at													
25														
26														

2007

Ready

PIRUS2: WP4 Software, standards, protocols

The AR1:

- This is the report format used for data supplied to us by our participating publishers
- For the project, we've been accepting MS-Excel files
  - The AR1 standard is still being developed, not yet an agreed **COUNTER** standard
  - In real world, data would be gathered using SUSHI
- SUSHI
  - Standardized Usage Statistics Harvesting Initiative Protocol
  - a SOAP-based web service, i.e. Machine to machine
  - used to expose COUNTER Release 3 compliant usage statistics to institutions and consortia

PIRUS2: WP4 Software, standards, protocols

Gathering publisher sample data:

- Received from ACS, Emerald, IOP, Nature, Oxford Journals, Springer and Wiley
- Processed and loaded into the PIRUS2 demonstrator database
  - Using a mix of manual processing and scripting
- Statistics for
  - 581,556 Articles
  - 537 Journals
  - 93,729,498 downloads across 2,743,839 records

PIRUS2: WP4 Software, standards, protocols

Institutional Repository usage events

- Many different IR softwares
  - Open Source, e.g. CDSware, **DSpace, Eprints, Fedora**, i-Tor, MyCoRe, OPUS
  - Proprietary, e.g. Digital Commons (BePress), Digitool (Ex Libris)
- IRs commonly catalogue: **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.
- Great variations in the way they work
  - Different programming languages and platforms
  - Different methods of logging download events
  - No common standard applied where stats are exposed
  - Figures quoted often include robot accesses

PIRUS2: WP4 Software, standards, protocols

Institutional Repository usage events

- Key is to get usage data out in a standard manner
- We considered three scenarios for gathering:
  - (A) 'tracker' code – a 'push' method
    - a server-side 'Google Analytics' for full-text article downloads
    - When a download occurs a message is transmitted to a central remote server
  - (B) OAI-PMH harvesting – a 'pull' method
    - Open Archives Initiative Protocol for Metadata Harvesting
    - protocol familiar to repositories
    - Used to expose and share metadata for items in IRs
  - (C) SUSHI

PIRUS2: WP4 Software, standards, protocols

Institutional Repository usage events

- Quickly put Scenario C (SUSHI) to one side
  - AR reports still under development, not yet standard
  - Technology complex and unfamiliar to repositories
  - Auditing cost implications of producing ready-made COUNTER-compliant reports
- Turned our attention to Scenarios A & B
- Usage data are shared in an OpenURL format
  - An approach first suggested by MESUR. Taken forward in Europe under 'Knowledge Exchange' initiative see: http://wiki.surffoundation.nl/display/standards/OpenURL+Context+Objects
  - Scenario A (tracker) OpenURL Key-Value Pair Strings. (URLs)
  - Scenario B (OAI-PMH) OpenURL Context Objects. (XML)

PIRUS2: WP4 Software, standards, protocols

Institutional Repository usage events

- With so many IR softwares about, as a project, we couldn't carry out development on all of them
- Decided to focus on DSpace, GNU Eprints and Fedora
  - All open source
  - Underlying software for around two-thirds of IRs
- PIRUS2 Repository software plug-ins/extensions
  - DSpace – developed by @mire
  - Eprints – developed by Tim Brody, Southampton University
  - Fedora – developed by Ben O'Steen, Oxford University
  - Links and downloads on PIRUS2 project web site

PIRUS2: WP4 Software, standards, protocols

Institutional Repository usage events

- DSpace
  - Patches for v1.6.2 for both the tracker and OAI-PMH
    - At Cranfield, tested both approaches and shown both to work
    - Adopted the tracker approach for wider testing
      - Privacy of data and simplicity
      - Looking to future auditing implications, simpler and cheaper option
- Eprints and Fedora
  - plug-ins using the tracker approach
- Links and downloads on PIRUS2 project web site

PIRUS2: WP4 Software, standards, protocols

Gathering Institutional Repository usage data

- 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)

PIRUS2: WP4 Software, standards, protocols

Gathering Institutional Repository usage data

- PIRUS2 server logs receiving 19MB data a week
- Usage data in those logs must be:
  - filtered according to COUNTER rules to eliminate Robots and Double clicks
  - Processed into monthly statistics
- Using a series of scripts, processing and loading data into the PIRUS2 demonstrator database
- So far, good … but what about consolidating the IR usage statistics with publisher statistics?

PIRUS2: WP4 Software, standards, protocols

Consolidating usage data from Publishers and IRs

- How do we match events from publishers and IRs?
- The key is the DOI
  - Unique identifier available for the majority of articles
  - Most (but not all) publishers allocate DOIs to articles
  - Most (but not all) IRs add DOI metadata to some (but not all) records pertaining to articles
- Where both Publisher and IR stats have a DOI, the match is easy and certain
- But where we don't have a DOI? What then?
- Two scenarios:
  - Scenario 1: Published article has a DOI, but IR hasn't catalogued the DOI
  - Scenario 2: Published article doesn't have a DOI, so IR can't add a DOI

PIRUS2: WP4 Software, standards, protocols

Consolidating usage data from Publishers and IRs

- Scenario 1: Published article has a DOI
  - In all cases, from IRs, we have the Article title and first author surname
  - Use those to query and retrieve the DOI from:
    - The CrossRef database
    - Our own database
  - Tried and tested this … and it works (most of the time)!
- Scenario 2: Published article doesn't have a DOI
  - We're stuck with relying on the Article title and first author surname
  - And any other supplementary metadata available, like Journal, Volume, Issue, ISSN, local identifiers
  - Relies somewhat on fuzzy matching
  - Doable, but needs further examination and consideration…

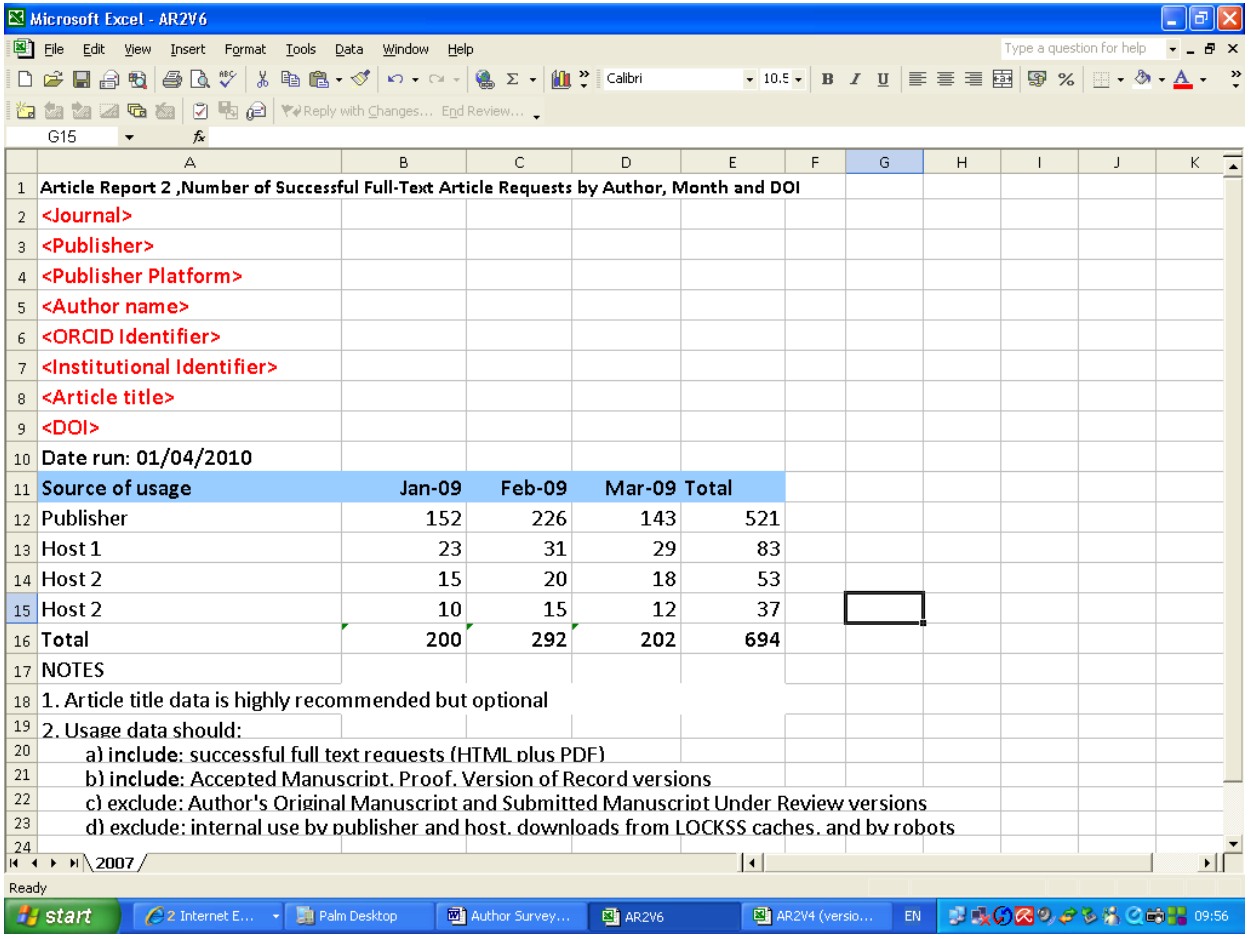
PIRUS2: WP4 Software, standards, protocols

Exposing consolidated usage statistics

- We've successfully consolidated Publisher and IR statistics in the PIRUS2 demonstrator
- What to expose and to whom are 'political' issues
- But, as examples, we're able to output:
  - Article Report 2 (AR2): Number of Successful Full-Text Article Requests by Author, Month and DOI
    - A report for a single article
  - Article Report 2a (AR2a): Number of Successful Full-Text Article Requests by Author, Month and DOI
    - A report for all articles available within the system for a particular author

PIRUS2: WP4 Software, standards, protocols

AR2 example:



Microsoft Excel - AR2V6

	A	B	C	D	E
1	Article Report 2 ,Number of Successful Full-Text Article Requests by Author, Month and DOI				
2	<Journal>				
3	<Publisher>				
4	<Publisher Platform>				
5	<Author name>				
6	<ORCID Identifier>				
7	<Institutional Identifier>				
8	<Article title>				
9	<DOI>				
10	Date run: 01/04/2010				
11	Source of usage	Jan-09	Feb-09	Mar-09	Total
12	Publisher	152	226	143	521
13	Host 1	23	31	29	83
14	Host 2	15	20	18	53
15	Host 2	10	15	12	37
16	Total	200	292	202	694
17	NOTES				
18	1. Article title data is highly recommended but optional				
19	2. Usage data should:				
20	a) include: successful full text requests (HTML plus PDF)				
21	b) include: Accepted Manuscript, Proof, Version of Record versions				
22	c) exclude: Author's Original Manuscript and Submitted Manuscript Under Review versions				
23	d) exclude: internal use by publisher and host, downloads from LOCKSS caches, and by robots				
24					

2007

Ready

PIRUS2: WP4 Software, standards, protocols

Exposing consolidated usage statistics

- And we can generate reports for IRs
  - Giving them back COUNTER-compliant usage statistics for items in their repository
  - Containing DOIs they can use to update and enhance their own records
- Regarding formats
  - Currently working with MS-Excel/CSV files
  - Going forward SUSHI and other formats can be exposed as appropriate

PIRUS2: WP4 Software, standards, protocols

Issues we wanted to address but couldn't (yet)

- Article versions
  - There are definitions in both the publisher and repository worlds
  - But, in practice, we found that neither are applied consistently enough to be used as yet

NISO/ALPSP Definitions	VERSIONS Definitions
Authors Original (AO)	Draft
Submitted Manuscript Under Review (SMUR)	Submitted Version
Accepted Manuscript (AM)	Accepted Version
Proof (P)	
Version of Record (VoR)	Published Version
Corrected Version of Record (CVoR)	Updated Version
Enhanced Version of Record (EVoR)	Updated Version

  - No standard as to which metadata element should be used
- Peer-review status
  - Again, not consistently enough available
  - No standard as to which metadata element should be used

PIRUS2: WP4 Software, standards, protocols

The PIRUS2 demonstrator portal

- Built using MySQL, PHP and Perl scripts
- Labour of love – particularly writing the scripts to get repository data in.
- Behind an authorization/authentication challenge to allay privacy concerns from various contributors – so, not available to the general public
- But, this afternoon, we'll have a session giving you a peek inside…

PIRUS2: WP4 Software, standards, protocols

Thank you for listening!

For more information:

http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/