IRUS-UK

Institutional Repository Usage Statistics

IRUS-UK: the technical aspects

22 May 2013

Balviar Notay, Jisc
Ross Macintyre, Mimas
Paul Needham, Cranfield University
Angela Conyers, Evidence Base, BCU

IRUS-UK

- Funded by Jisc as part of UK RepositoryNet+
- Led by Mimas
- Primary Project Team Members:
  - Mimas
  - Cranfield University
  - EvidenceBase, BCU
- IRUS-UK: Institutional Repository Usage Statistics - UK

IRUS-UK: history

- Emerged as an outcome of PIRUS2
  - Publisher and Institution Repository Usage Statistics project
  - http://www.cranfieldlibrary.cranfield.ac.uk/pirus2/
  - Aimed to develop a global standard to enable the recording, reporting and consolidation of online usage statistics for individual journal articles hosted by Institutional Repositories, Publishers and others
  - Proved it was technically feasible, but thwarted by organisational and political issues
  - However, crisis/opportunity… we now knew it was possible to…

IRUS-UK: aim

- Enable UK IRs to share/expose usage statistics based on a global standard – COUNTER
  - Produced on the same basis as publishers
  - Filtered to remove robots and double clicks
  - Comparable
  - Reliable
  - Trustworthy
  - Authoritative

IRUS-UK: objectives

- Collect raw usage data from UK IRs for all item types within repositories
  - Downloads not record views
- Process those raw data into COUNTER-compliant statistics
- 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
  - demonstrate their value and place in the dissemination of scholarly outputs
- Offer opportunities for benchmarking
- Act as an intermediary between UK repositories and other agencies
  - e.g. global central clearinghouse, national shared services, Research Councils, SCONUL, OpenAIRE

IRUS-UK: Technical aspects

Files are being downloaded from repositories

1. How can IRUS-UK get information about those downloads from repositories?
   - How do we get the usage data from IRs to IRUS-UK?
2. Then what do we do with that information?
   - How do we process the raw usage data and convert to COUNTER-compliant statistics?
   - How do we display, share, re-expose those statistics?

IRUS-UK: Gathering usage data (1)

At top level, there are only two options

A. Repositories PUSH usage data to IRUS-UK
   - Think Google Analytics

B. IRUS-UK PULLS usage data from repositories
   - Think OAI-PMH

- Either way is technically possible
  - so which one have we gone for? And why?

IRUS-UK: Gathering usage data (2)

We've opted for the PUSH mechanism

- We receive notification of download events as and when they occur
- It makes life simple
  - A file is downloaded from a repository
  - Ping – a notification is sent to IRUS-UK server
    - as OpenURL key-value pair strings – I'll come back to that…
  - We handle all the processing from there
  - We're responsible for the COUNTER-compliance side of things
- It scales well
  - More IRs = bigger logs
  - But no other real issues
- We call it "the Tracker Protocol"

IRUS-UK: Gathering usage data (3)

Whereas the PULL mechanism, OAI-PMH

- Yes, it's a familiar protocol in a repository context, and it can be repurposed to expose usage events for harvesting – as OpenURL Context Objects - by IRUS-UK (and others)
- But then
  - For IRs: they would have to take responsibility for storing events locally, ensuring integrity and availability of the data, maintaining OAI crosswalks, incur annual audit charges as part of COUNTER-compliance
  - For IRUS-UK: it becomes difficult to manage the harvesting processes as the number of participating IRs grows

IRUS-UK: the Tracker Protocol spec (1)

The specification for this is quite brief and straightforward

- When a user clicks on a link to (i.e. downloads) a file from a Repository with the tracker protocol in operation, an OpenURL log entry is sent to a remote server for further processing.
- The OpenURL log entry should be based on a subset of the NISO OpenURL 1.0 standard *KEV ContextObject Format.* The OpenURL string must be URL encoded, with key-value pairs separated by &.

IRUS-UK: the Tracker Protocol spec (2)

Element	OpenURL Key	OpenURL Value (example)	Notes
OpenURL version	url_ver	Z39.88-2004	Identifies data as OpenURL 1.0. String constant: Z39.88-2004 (Mandatory)
Usage event datestamp	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)
Item OAI identifier	rft.artnum	oai:dspace.lib.cranfield.ac.uk:1826/936	(Mandatory)
MIMEtype	svc_format	application%2Fpdf	(Mandatory)
FileURL	svc_dat	https://dspace.lib.cranfield.ac.uk/bitstream/1826/936/4/Artificial_compressibility_Pt2-2005.pdf	(Mandatory)
Source repository	rfr_id	dspace.lib.cranfield.ac.uk	(Mandatory)

IRUS-UK: the Tracker Protocol spec (3)

Eliminating robots

- Information about (known) Robot downloads need not be transmitted to the remote server. Before attempting to transmit the OpenURL, a check should be carried out to eliminate robots as defined in the COUNTER official list, available as a series of user-agent regexes in xml and/or text files at:
  - List of Robots, xml file XML (20KB)
  - List of Robots, txt file TXT (2.5KB)
- Checking against these regexes should be case insensitive.

IRUS-UK: the Tracker Protocol spec (4)

OpenURL Transmission

- Once the OpenURL has been constructed, it will need to be transmitted to a remote server where it will be stored and processed.
- If the transmission is successful the remote server will return a 200 OK code.
- If the transmission is *not* successful, e.g. a 4xx or 5xx code is returned, the OpenURL string should be queued for processing later, e.g. appended to a file held on the local server. A script, that can be scheduled to run periodically, should check if there are queued entries and, if there are, re-transmit them.

IRUS-UK: deploying the Tracker

- There are a (growing) number of software platforms offering Institutional Repository functionality, these include:
  - DSpace, Eprints, Fedora, intraLibrary (Intrallect), Digital Commons (Bepress), PURE Portal … and the list goes on
- Each of these platforms works differently under the hood meaning a single solution for deploying tracker functionality is not possible
- So, we've focused, initially, on DSpace and Eprints, which account for about two-thirds of UK repositories
  - Plugins/patches available for DSpace (1.8.x and 3.x) and Eprints (3.2+)

IRUS-UK: deploying the Tracker - DSpace

- Patches are available for DSpace versions1.8x and 3.x
  - Contact us for further detailed information
- Installation is reasonably straightforward:
  - Extract the files in the patch to a convenient directory on the server
  - Switch to the dspace-source top level directory and execute the patch command, e.g.
    `patch -p0 < /path/to/Atmire-Statistics-OAI-and-Harvester-3-0.patch`
  - Check and adjust a few config parameters
  - Rebuild with `mvn clean package` and `ant update`
  - Deploy the updated code and restart tomcat
- And that's it!

IRUS-UK: deploying the Tracker - Eprints

- The latest version of the Eprints add-on is available from http://files.eprints.org/816/
  - Works with version 3.2 or greater
- Installation is straightforward:
  - Extract the files in pirus-1.04.tar.gz to a convenient directory on the server
  - Copy pirus.pl to your repository's cfg.d/ directory
  - Copy PIRUS.pm to your Event plugin directory e.g. lib/plugins/EPrints/Plugin/Event/PIRUS.pm
  - Check pirus.pl to ensure the tracker location is set to http://www.jusp.mimas.ac.uk/counter/
- And that's it!

IRUS-UK: deploying the Tracker - Others

- We have to look at other repository software platforms on a case by case basis
- Fedora
  - Every Fedora repository is a one-off. But some general guidelines are available in Appendix O in the PIRUS2 Final Report, http://www.projectcounter.org/News/Pirus2_oct2011.pdf
- PURE Portals
  - We've opened discussions with Atira and we're hopeful that IRUS-UK Tracker functionality will be available for PURE portals …
- Other platforms
  - We would welcome dialogue with interested vendors & developers
  - Contact us!

IRUS-UK: processing data (1)

- Data gets stored in daily log files
- The Log for any given day is usually processed the following day
- We currently have a three step daily ingest process
- Step 1: Perl script parses the logs
  - Processes entries from recognised IRs
  - Sorts and filters entries following COUNTER rules to remove robot entries and double-clicks
  - Filters entries using additional IRUS-UK filters
    - *There's a lot of strange behaviour out there!*
  - Consolidates raw usage data for each item into daily statistics
  - Outputs to intermediate file

IRUS-UK: processing data (2)

- Step 2: Perl script processes the intermediate file
  - Intermediate file entries
    - 5 2 oai:usir.salford.ac.uk:9967 2013-03-18 1
    - 5 2 oai:usir.salford.ac.uk:9970 2013-03-18 6
    - 5 2 oai:usir.salford.ac.uk:9972 2013-03-18 2
  - Using the OAI identifier, it looks up each item against the Item Authority table in the IRUS DB to see if we already know about it
    - If we do, all well and good – there's an existing IRUS Item Identifier
    - If not, it adds a stub-entry to the Item Authority table – minting a new IRUS Item Identifier and adding the repository identifier, platform and OAI identifier
    - The rest of the metadata is set to 'unknown' at this stage
  - Finally, the script adds the download statistics associated with each IRUS Item Identifier to the Daily Statistics table

IRUS-UK: processing data (3)

- Step 3: Perl script obtains the “unknown” metadata
  - Queries the DB to find the ‘known unknowns’
  - Using the OAI identifiers, issues OAI-PMH GetRecord calls to retrieve OAI_DC metadata
  - Parses the OAI records
  - Updates the metadata – Title, Author, Item Type, etc., in the Item Authority Table in the DB
  - Additionally maps the Item Type, as given by the source repository, to a smaller (more manageable list) of IRUS Item Types
- Step 4: A monthly Perl script
  - Consolidates the Daily Statistics into a Monthly Statistics table
    - The ‘traditional’ COUNTER granularity
  - So, as a service, we can easily work with statistics at either granularity

IRUS-UK: robots and unusual usage (1)

There's a lot of strange behaviour out there!

- Our starting point for eliminating robots and machine accesses was the COUNTER robots exclusion list
  - Holds regexes of User agents to exclude
- However, as we've taken on-board more repositories, it's become obvious: that list is not enough
- When LSE joined in January we identified further exclusions not in the COUNTER list
  - half a dozen user agents
  - a couple of IP ranges used by Baidu Spider
- And that turned out to be the 'tip of the iceberg'…

IRUS-UK: robots and unusual usage (2)

- With each new repository we find additional bizarre usage that really doesn't look like accesses by regular (human) users
- And we extend our filters, but
  - It's time consuming
  - Tedious
  - Stats periodically have to be restated
- It has become apparent that we need a much more sophisticated filtering system to eliminate (as much as possible) dodgy usage!

IRUS-UK: robots and unusual usage (3)

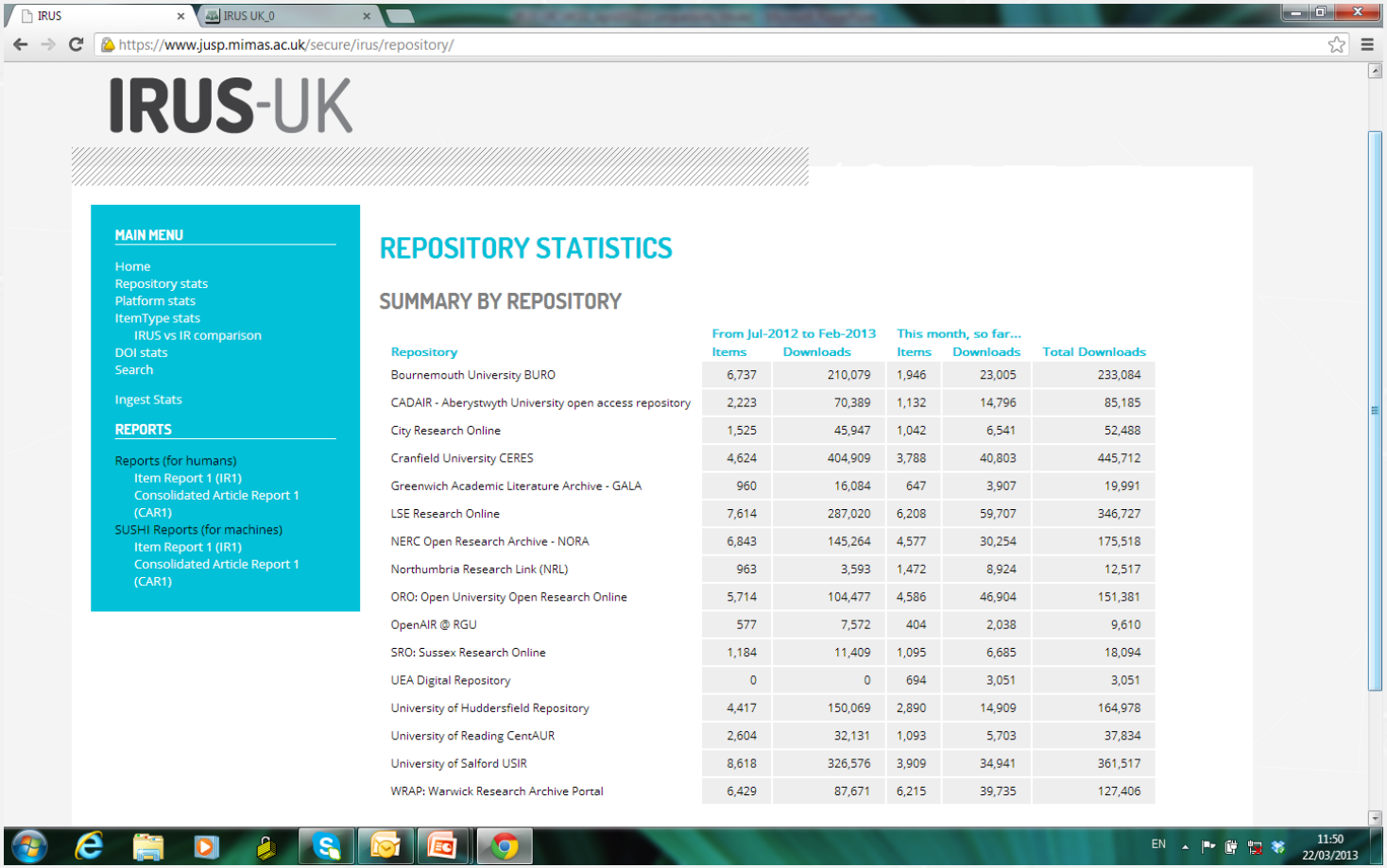
The practical solution looks to be an “adaptive filtering system”

- We have commissioned a piece of work to come up with such a system
  - It’s a bit premature to go into detail, but it will involve various metrics and a scoring and weighting system
  - We’ll be trying it out in the coming weeks
  - COUNTER may adopt this system as part of the COUNTER PIRUS Code of Practice

IRUS-UK: Exposing statistics

- Through the User Interface - The Portal
- Various reports are available to Institutions
  - For humans, downloads as CSV/Excel spreadsheet files
    - Monthly and daily granularity
  - For machines, harvesting via SUSHI
    - Monthly granularity
- Usage statistics for incorporation into Repositories
  - Via an API/Web Service
    - Initial version in place
    - Further requirements to be determined

IRUS-UK: Repository Totals



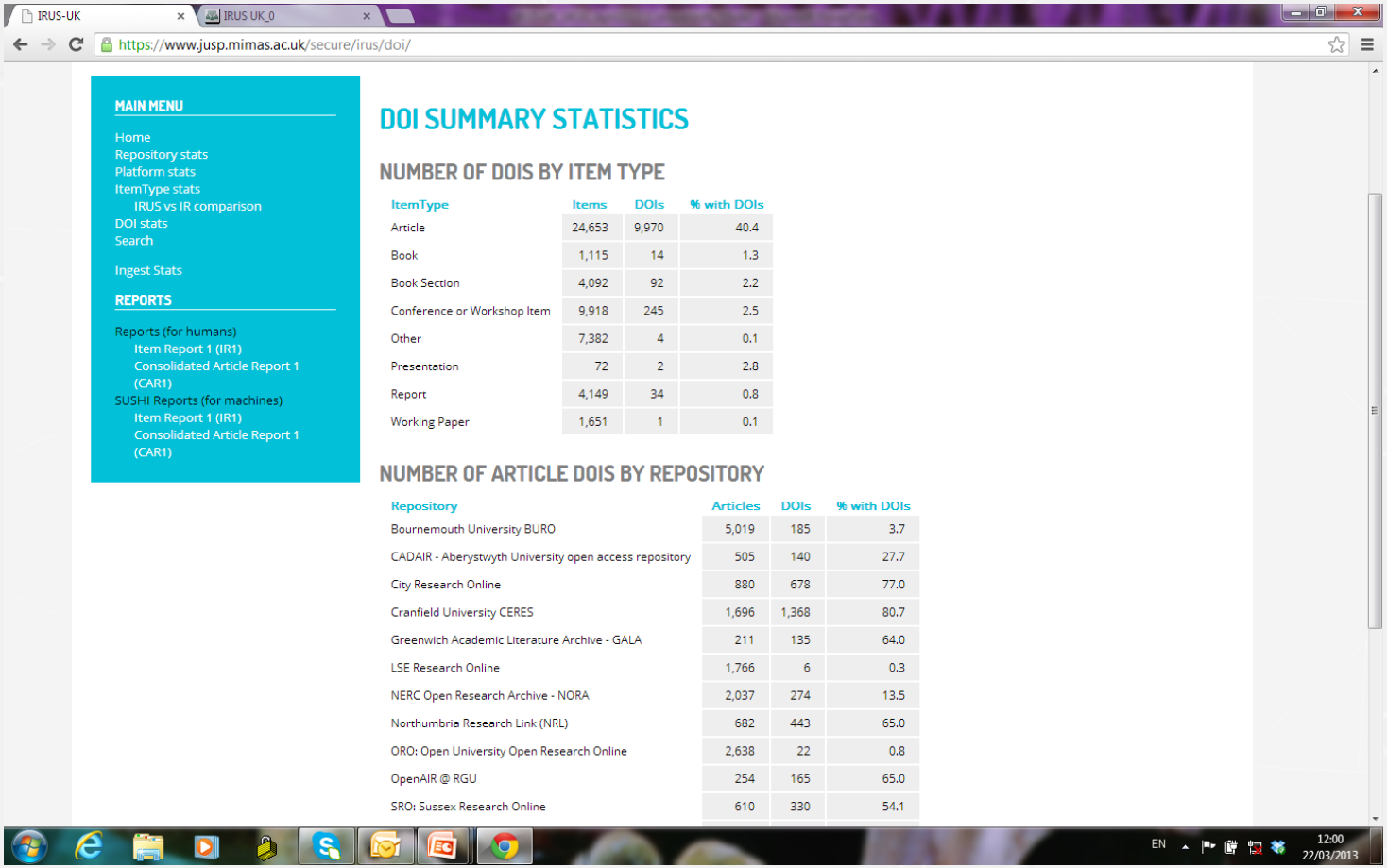
Repository	From Jul-2012 to Feb-2013 Items	From Jul-2012 to Feb-2013 Downloads	This month, so far... Items	This month, so far... Downloads	Total Downloads
Bournemouth University BURO	6,737	210,079	1,946	23,005	233,084
CADAIR - Aberystwyth University open access repository	2,223	70,389	1,132	14,796	85,185
City Research Online	1,525	45,947	1,042	6,541	52,488
Cranfield University CERES	4,624	404,909	3,788	40,803	445,712
Greenwich Academic Literature Archive - GALA	960	16,084	647	3,907	19,991
LSE Research Online	7,614	287,020	6,208	59,707	346,727
NERC Open Research Archive - NORA	6,843	145,264	4,577	30,254	175,518
Northumbria Research Link (NRL)	963	3,593	1,472	8,924	12,517
ORO: Open University Open Research Online	5,714	104,477	4,586	46,904	151,381
OpenAIR @ RGU	577	7,572	404	2,038	9,610
SRO: Sussex Research Online	1,184	11,409	1,095	6,685	18,094
UEA Digital Repository	0	0	694	3,051	3,051
University of Huddersfield Repository	4,417	150,069	2,890	14,909	164,978
University of Reading CentAUR	2,604	32,131	1,093	5,703	37,834
University of Salford USIR	8,618	326,576	3,909	34,941	361,517
WRAP: Warwick Research Archive Portal	6,429	87,671	6,215	39,735	127,406

IRUS-UK: Item Types Totals

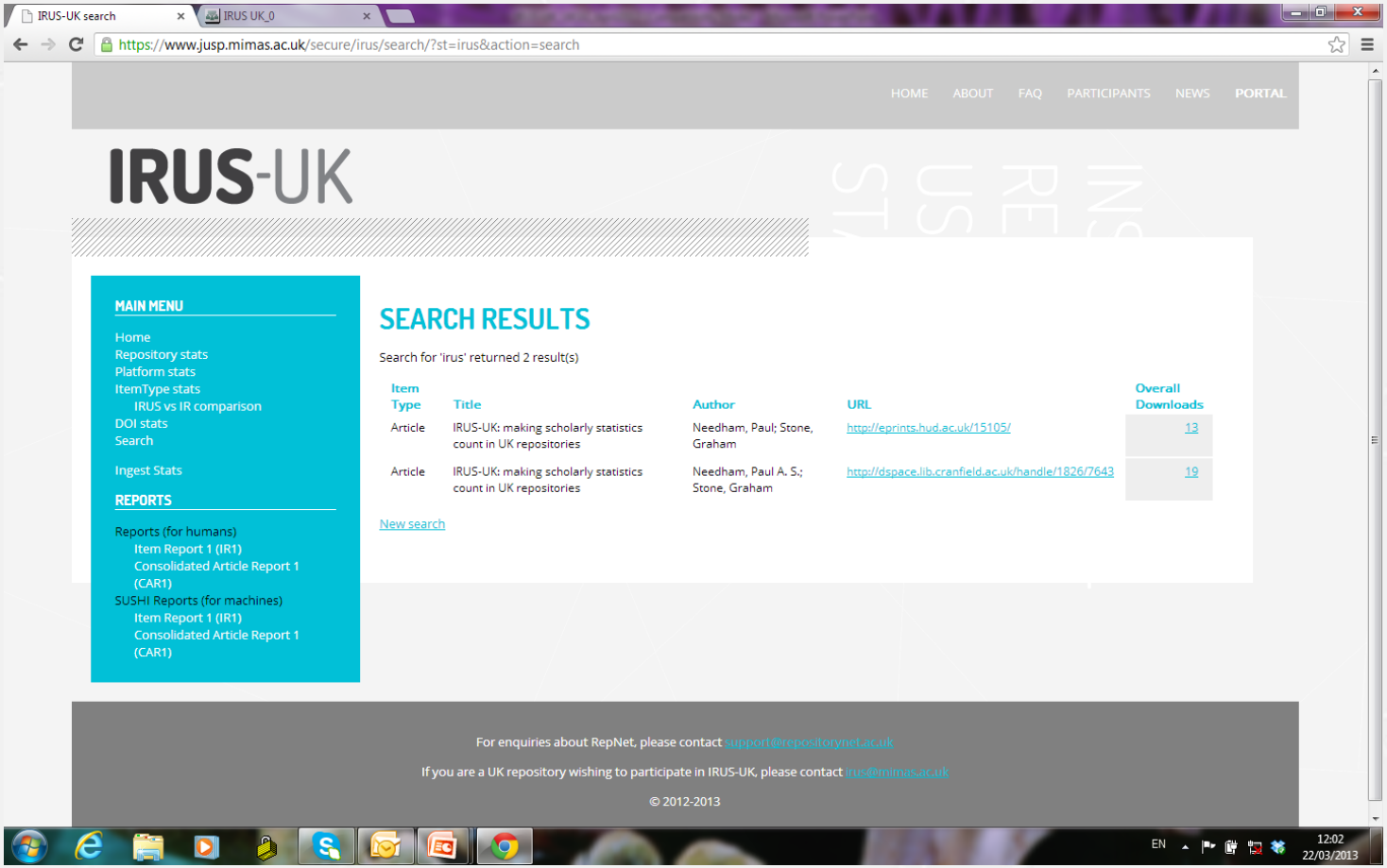
IRUS ItemType	From Jul-2012 to Feb-2013 Items	From Jul-2012 to Feb-2013 Downloads	This month, so far... Items	This month, so far... Downloads	Total Downloads
Article	21,172	688,930	15,222	128,025	816,955
Book	962	36,164	579	6,728	42,892
Book Section	3,399	127,259	2,177	23,716	150,975
Conference or Workshop Item	8,216	193,161	5,238	28,805	221,966
Dataset	15	447	11	71	518
Exam Paper	4,025	25,187	1,148	2,903	28,090
Image	826	3,868	222	468	4,336
Learning Object	261	8,333	115	1,754	10,087
Manual	3	62	2	4	66
Other	6,614	276,751	4,929	53,169	329,920
Patent	27	139	16	28	167
Presentation	63	926	40	193	1,119
Report	4,034	84,061	2,446	16,165	100,226
Thesis or dissertation	7,291	379,000	6,735	63,435	442,435
Unclassified	8	9	1	1	10
Unknown	764	3,731	249	1,360	5,091
Website	1,759	15,889	1,163	4,056	19,945
Working Paper	1,593	59,273	1,405	11,022	70,295

IRUS-UK: Item Type <->IR: Item Type

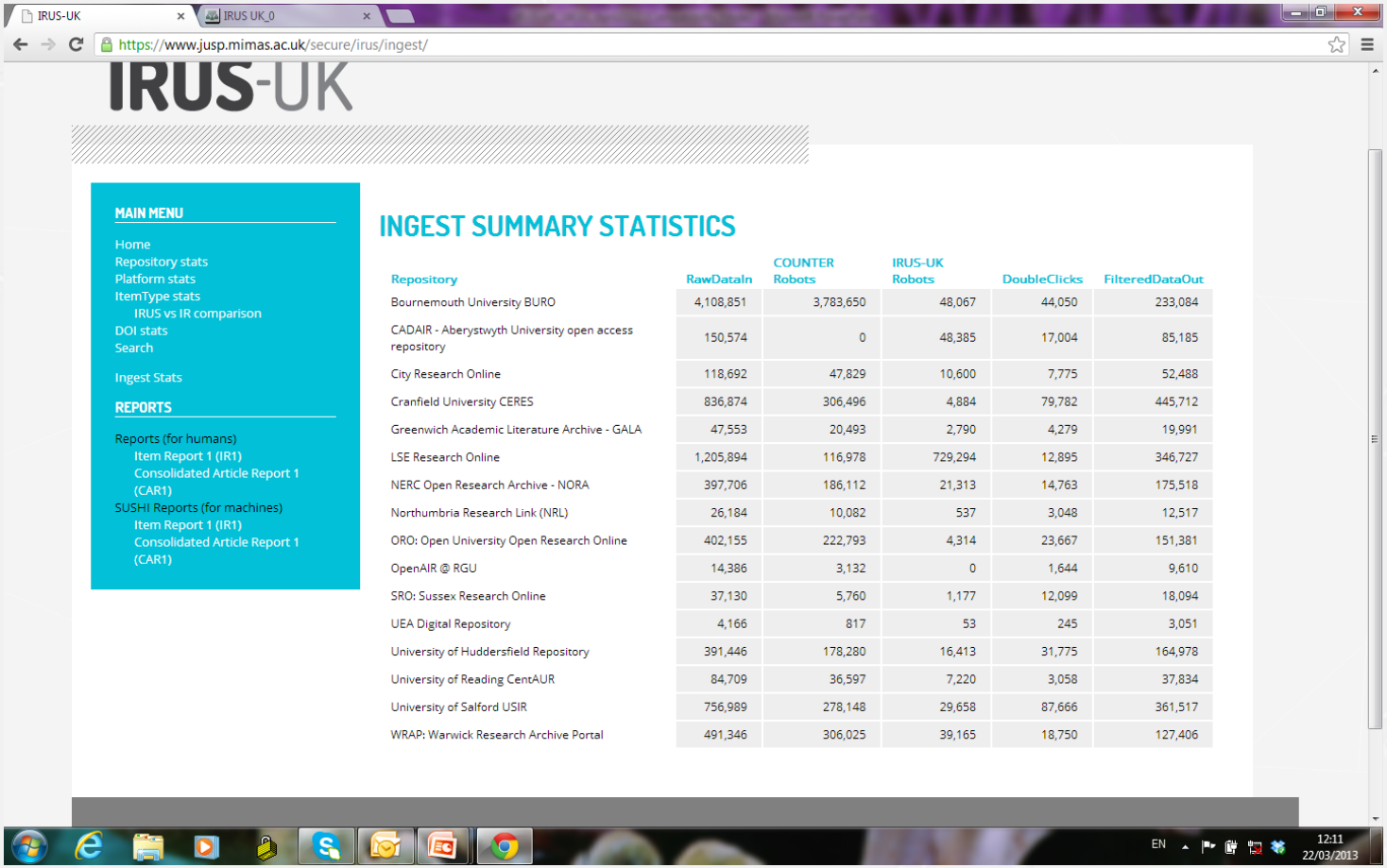
IRUS-UK: DOI Summary Stats



IRUS-UK: Title/Author Search

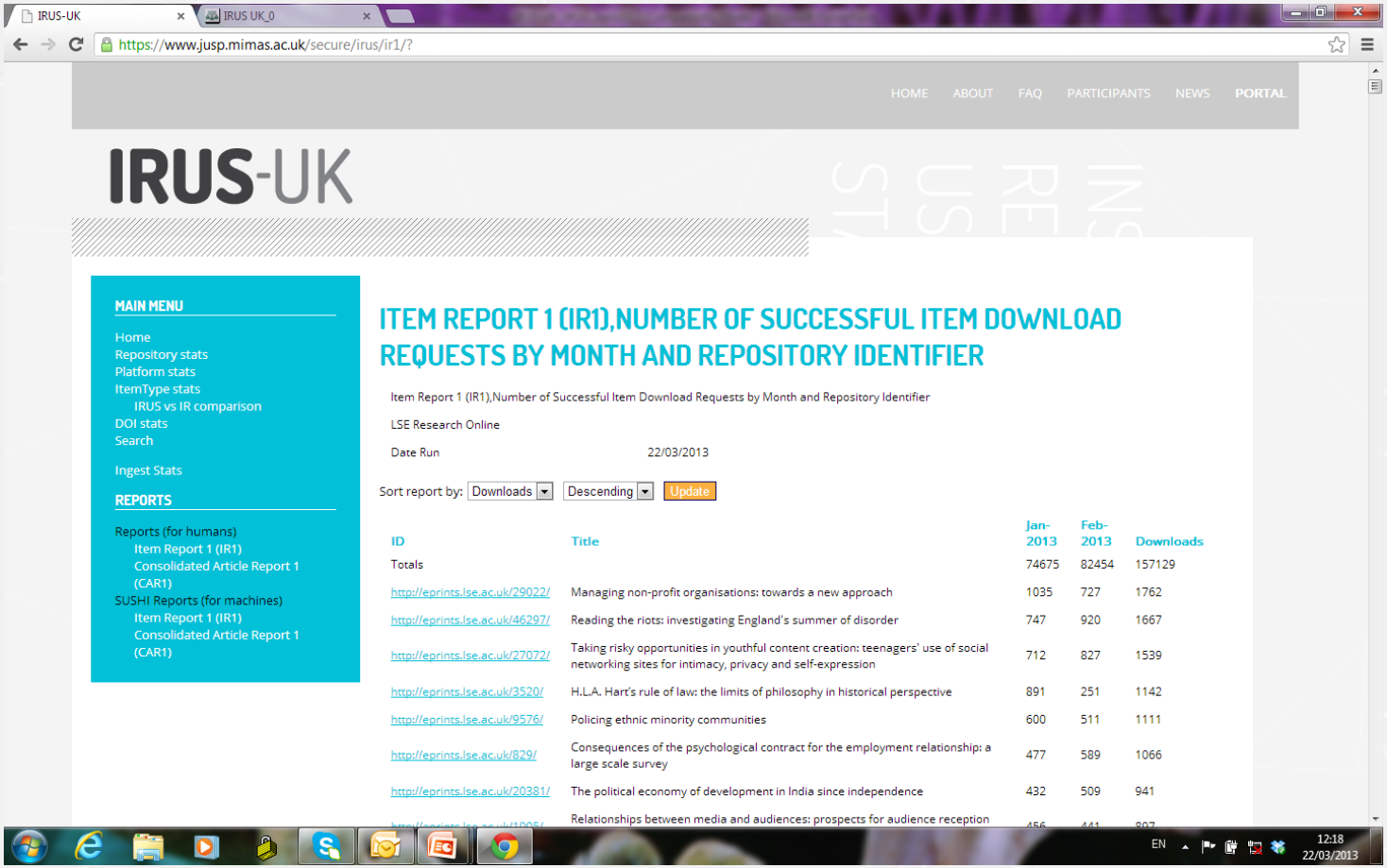


IRUS-UK: Ingest Summary Stats



Repository	RawDataIn	COUNTER Robots	IRUS-UK Robots	DoubleClicks	FilteredDataOut
Bournemouth University BURO	4,108,851	3,783,650	48,067	44,050	233,084
CADAIR - Aberystwyth University open access repository	150,574	0	48,385	17,004	85,185
City Research Online	118,692	47,829	10,600	7,775	52,488
Cranfield University CERES	836,874	306,496	4,884	79,782	445,712
Greenwich Academic Literature Archive - GALA	47,553	20,493	2,790	4,279	19,991
LSE Research Online	1,205,894	116,978	729,294	12,895	346,727
NERC Open Research Archive - NORA	397,706	186,112	21,313	14,763	175,518
Northumbria Research Link (NRL)	26,184	10,082	537	3,048	12,517
ORO: Open University Open Research Online	402,155	222,793	4,314	23,667	151,381
OpenAIR @ RGU	14,386	3,132	0	1,644	9,610
SRO: Sussex Research Online	37,130	5,760	1,177	12,099	18,094
UEA Digital Repository	4,166	817	53	245	3,051
University of Huddersfield Repository	391,446	178,280	16,413	31,775	164,978
University of Reading CentAUR	84,709	36,597	7,220	3,058	37,834
University of Salford USIR	756,989	278,148	29,658	87,666	361,517
WRAP: Warwick Research Archive Portal	491,346	306,025	39,165	18,750	127,406

IRUS-UK: IR1 Report LSE Jan-Feb 2013



IRUS-UK: CAR1 Report Jan-Feb 2013

IRUS-UK: community engagement

- Growing number of repositories sending data to IRUS-UK
- Currently 24 participants:
  - Bath Spa, Bournemouth, City, UEA, Glasgow School of Art, Greenwich, Huddersfield, Kent, Lancaster, LSE, Middlesex, NERC, Northampton, Northumbria, Open, Reading, Salford, Sussex, , Warwick (Eprints)
  - Aberystwyth, Cranfield, Imperial, RGU, St Andrews (DSpace)
  - Others in the pipeline

IRUS-UK: how to join

- If you are a UK repository:
  - Contact us at irus.mimas.ac.uk to register your interest
  - Answer a few questions on the type of repository you have and the version you are running
  - Get advice from us on what work will be involved depending on your repository type and version
  - Implement any changes advised and then see your usage data instantly in IRUS-UK with no more work from you

“The set up was quick and painless, which is always a delight!”

“Consistent collection of statistics without me having to do it!”

Contacts & Information

- If you are a UK repository wishing to participate in IRUS-UK, please contact
  - irus@mimas.ac.uk
- For general enquiries, please contact
  - support@repositorynet.ac.uk
- Project web site:
  - http://www.irus.mimas.ac.uk/
- Thank you!