

Project Information			
<b>Project Acronym</b>	None		
<b>Project Title</b>	Shared OpenURL Data Infrastructure Investigation		
<b>Start Date</b>	1 December 2009	<b>End Date</b>	30 April 2009
<b>Lead Institution</b>	University of Edinburgh (EDINA)		
<b>Project Director</b>	Christine Rees		
<b>Project Manager &amp; contact details</b>	Tim Stickland <a href="mailto:t.stickland@ed.ac.uk">t.stickland@ed.ac.uk</a> ; 0131 650 3309		
<b>Partner Institutions</b>	None		
<b>Project Web URL</b>	<a href="http://edina.ac.uk/projects/Shared_OpenURL_Data_Infrastructure_Investigation_summary.html">http://edina.ac.uk/projects/Shared_OpenURL_Data_Infrastructure_Investigation_summary.html</a>		
<b>Programme Name (and number)</b>	JISC Repositories and Preservation Programme		
<b>Programme Manager</b>	Neil Jacobs		
Document Name			
<b>Document Title</b>	Project Final Report – Draft without Conclusions and Recommendations		
<b>Reporting Period</b>	NA		
<b>Author(s) &amp; project role</b>	Christine Rees (Project Director), Leah Halliday (Consultation Workpackage Leader), Tim Stickland (Project Manager) and Des Reid (Software Developer).		
<b>Date</b>	21 April 2009	<b>Filename</b>	OpenURLDataInfV1
<b>URL</b>			
<b>Access</b>	<input checked="" type="checkbox"/> Project and JISC internal		<input type="checkbox"/> General dissemination
Document History			
Version	Date	Comments	
V1	21 April 2009		
V1a	1 May 2009	CR comments	
V1b	8 May 2009	Additional info. From TS and LH; submitted as draft to Neil Jacobs, without conclusions, 18 May 2009	
V1c	6 July 2009	Draft Final report: conclusions/recommendations need team agreement	
V1d	9 July 2009	Conclusions/Recommendations finalised; Quality Check read through report by AB.	

# Shared OpenURL Data Infrastructure Investigation

EDINA (JISC National Data Centre)

Leah Halliday, Tim Stickland, Christine Rees and Des Reid

Copyright 2009 EDINA

Contact:

The Administrator

EDINA

Causewayside House

160 Causewayside

Edinburgh EH9 1PR

Tel: 0131 650 3302

Fax: 0131 650 3308



## Table of Contents

Acknowledgements .....	4
Background .....	1
Aim and Objectives .....	1
Methodology .....	1
Consultation .....	1
Technical work .....	2
The Consultation Exercise .....	2
Interviews .....	2
Staff in HE libraries .....	2
Resolver vendors .....	3
Questionnaire .....	4
User Scenarios – evaluated by means of private blog .....	4
Legal issues .....	6
Technical work .....	6
Requirements for successful sharing of OpenURL data .....	6
Using OpenURL Router data .....	8
Summary of the analysis .....	10
Conclusions and recommendations .....	11
Conclusions .....	11
Services based on aggregated OpenURL data .....	11
Libraries appear willing to share data, in principle .....	11
Availability of OpenURL resolver usage data .....	11
OpenURL router as a data source .....	12
Recommendations .....	12
Appendix 1: the Scenarios of Use and a précis of comments .....	14
(1) Comparative reports on outcomes .....	14
(2) Recommender service .....	14
(3) Ranking of papers based on ‘use’ (to benefit the HE sector) .....	14
(4) Collection development – national licences .....	14
(5) Emerging trends in UK scholarship .....	14
(6) Enhanced metadata for institutional services online 1 .....	14
(7) Enhanced metadata for institutional services online 2 .....	14
(8) Ranking of papers based on ‘use’ (to benefit users) .....	14
(9) Ranking sources within institution .....	14
Appendix 2: Report from online questionnaire .....	18
Appendix 3: Analysis of OpenURL Router logs .....	20
Preliminary investigations .....	20
Analysis based on frequency of requests from IP addresses .....	20
Sessions analysis .....	21
Appendix 4: What is an OpenURL resolver? .....	23
Appendix 5: What is the OpenURL router? .....	25

## **Acknowledgements**

This project was funded by JISC and was undertaken as part of the Repositories and Preservation Programme. The authors would like to thank all participants in the consultation process for their time and expertise.

## Background

“Many UK universities and colleges maintain OpenURL link servers<sup>1</sup>, which direct users to potential sources of the content they want. These redirect transactions might be logged and used in various ways. The MESUR<sup>2</sup> project has shown some of the possibilities that aggregations of such data might offer.”<sup>3</sup>

EDINA was invited by JISC to identify what OpenURL linking data is available within the UK and to scope potential uses and highlight any organisational and legal issues around use of the data. During JISC-funded project activity, as part of the JOIN-UP 5/99 programme, EDINA scoped, implemented and now runs as a service the UK OpenURL router<sup>4</sup>. This service is a potential source of UK OpenURL linking data. It has also provided EDINA with experience of and involvement with the OpenURL standard.

## Aim and Objectives

The aim of this project was to scope a UK architecture for the analysis of OpenURL linking data intended to provide a basis for ongoing services.

The objectives were:

1. Through desk research and consultation, to gain a good understanding of the technical, legal, and administrative challenges and opportunities related to sharing and using OpenURL link server data.
2. Through consultation with a few institutions, to identify the nature of the data that are currently collected within institutions, how they are used, whether they may be shared and on what condition, and the perceived benefits of sharing.
3. Using data from the OpenURL router, to establish the feasibility of identifying user behaviour across sessions or within sessions and analysing and comparing those behaviours.
4. To assess the relative value and complementary value of data from the OpenURL router and from OpenURL resolvers within institutions.
5. To explore potential uses of these data by developing and testing scenarios of use through consultation with various stakeholders in the UK Research and Higher Education community.
6. To identify any legal obstacles to use of the data and explore how the data may be used whilst respecting legal boundaries and constraints.
7. To identify organisational, policy and cultural issues related to aggregation of data from different institutions and explore, through consultation, the viability of gathering and analysing aggregated data.
8. To make recommendations for future gathering and use of OpenURL link server data based on: a review of the architecture proposed by the Los Alamos National Laboratory (LANL) MESUR project with reference to the findings of this project; and assessment of the type of service that may feasibly be provided with reference to the requirements and benefits, the ability and willingness of UK institutions to share data, and the willingness of other stakeholders to participate.

## Methodology

### Consultation

The project team sought to determine the context for such a service by exploring the feasibility of collecting and aggregating data as well as any legal, administrative or technical barriers to such. This

---

<sup>1</sup> See Appendix 4 for a description of an OpenURL resolver

<sup>2</sup> <http://www.mesur.org/MESUR.html>

<sup>3</sup> Neil Jacobs, JISC: project description provided to EDINA

<sup>4</sup> <http://openurl.ac.uk/doc/> : See Appendix 5 for a description of the OpenURL router

was undertaken through desk research and subsequently as part of the consultation exercise through liaison with staff in research libraries, representatives of vendors and with JISC Legal as follows:

- Interviews and an online questionnaire were used to determine what data are collected, how they are used, attitudes to sharing (and willingness to participate) and potential uses, benefits and risks as well as barriers to sharing, and the attitudes of vendors to such sharing.
- Nine scenarios of use for OpenURL usage log data were developed and tested for validity (i.e. whether they were considered to be useful, feasible and sufficiently attractive to warrant participation by contributing institutions) by publishing them on a private blog and inviting relevant stakeholders and experts (staff in libraries, researchers, and vendors) to comment on the user scenarios.
- Legal issues were explored through desk research and through consultation with JISC Legal.

## Technical work

The project team compared and contrasted the use of source data available from institutional OpenURL resolvers and those available from the OpenURL Router with a view to determining the extent to which OpenURL Router Data would be suitable for further work and in what instances shared institutional data would be required. From the outset it is apparent that:

- OpenURL Router data is a restricted sample. Whether a more representative sample could be achieved by aggregating institutional data was considered doubtful in the short/medium term, since the Router has around 90 registered resolvers. In the longer term this could have become a constraint.
- The OpenURL Router usage data are readily available at a very low cost and lower risk compared to aggregating institutional data.
- OpenURL Router data cannot be used for applications that require analysis of the resolver response to an OpenURL request.

OpenURL Router usage data were immediately available in large quantities and were used for initial investigations covering issues relevant to shared institutional data as well as those available from the Router. These included:

- Exploring the feasibility of identifying the actions of end users over extended periods of time, and the actions of users within a single session, and comparing the value of each of these approaches; this included considering the usefulness of IP addresses, cookies or other methods of discovering which requests originated from the same end user.

Finally, the team reviewed the architecture proposed by the LANL MESUR project in light of the project findings.

## The Consultation Exercise

### Interviews

Staff at six research libraries (five universities and a research council) were interviewed regarding collection and sharing of usage log data from their OpenURL resolver products. Representatives from the vendors of two of the market-leading resolvers were interviewed regarding their willingness to co-operate with an initiative to aggregate usage log data from resolvers in UK HEIs. The team corresponded with representatives from a third vendor but were unable to schedule an interview.

#### *Staff in HE libraries*

Four of the library staff work in libraries that use the three market leading products (two use SFX, one uses Serials Solutions 360 Link and a fourth uses Innovative's WebBridge). The other two interviewees use commercial products with a small share of the UK HE library market.

Four of the interviewees (three of the universities) collect usage log data although in most instances, practical use of those data is still being explored or has yet to be achieved.

Staff at all of the university libraries would, in principle, be willing to share data so that it may be aggregated as the basis for services. Most would wish that the data be anonymised, some would wish that the institution also remain anonymous and some that it be used only on a non-commercial basis. The university libraries would, generally, not have resource available to assist with any sharing initiative so would participate only if they could do so with minimum effort. Some indicated that the aggregator would have to obtain the data directly from the resolver vendor (with the library's permission). One systems librarian remarked that the volume of data would be enormous. Librarians providing samples to the project team reiterated this comment and the issue of volume was raised by participants in the evaluation of the Scenarios. EDINA technical staff commented on the blog that a National Data Centre is equipped to handle large volumes of data. The viability of doing so in this instance is not clear at this stage.

Generally, the interviewees struggled to imagine what kind of services might be built on such aggregated data. Two interviewees made suggestions which were reflected in the Use Cases developed as part of this project (Scenario 1 – comparative reports on outcomes and Scenario 9 – Ranking sources within institutions).

Two of the interviewees stated explicitly that institutional benefits (as well as community benefits) would be required as incentive to participate. Others, when asked how data may usefully be employed, simply described institutional benefits (some that do not require aggregation as they could be based on usage logs from the local resolver).

### *Resolver vendors*

During the course of the project, ExLibris announced the imminent release of its new article recommender service bX which is based on OpenURL usage log data and derives from the work by Herbert Van de Sompel and Johan Bollen at Los Alamos National Laboratory in the US.

Thus, ExLibris clearly has invested in and has an interest in this field. As the only vendor with such a service, it's not possible to report on the interview whilst maintaining the anonymity of that vendor. The interviewee, Jenny Walker has given permission for us to name her in this report.

SFX is the leading OpenURL resolver in UK HE. The bX service is based on data from instances of SFX throughout the world. We asked Jenny Walker if ExLibris would co-operate with an initiative to aggregate UK data with a view to building services on those data. She indicated that while ExLibris would not wish to block such an initiative, it would also wish to consider the impact of such an initiative on bX.

As far as the team can tell from interviewing vendors and librarians, SFX is the only commercial resolver being used in UK HE that has implemented OAI harvesting of usage log data. As part of the bX development, OAI-PMH capability has been implemented for all SFX sites but will be activated only on licensing of bX. The usage logs are in a proprietary format but before log data are harvested, the data are published in an OpenURL context-object format. This facility would be available to those licensing bX.

Interviews with two vendors and with librarians suggest also that SFX collects and stores far more data than the other resolvers. In SFX these data are archived but they are brought online again to be harvested for bX.

The other vendor interviewed does not have a facility for harvesting usage log data and would not consider developing the system to facilitate an initiative such as the one explored by this project. User institutions may output data in a CSV file from the vendor's system for use by an aggregator. Nevertheless, the vendor wondered whether such data would be 'sufficiently compatible' to be useful to a national aggregation service. Following the interview, project staff (with the help of the vendor) obtained and inspected sample data from a UK university using this resolver. It seems that the resolver collects no citation data. Only data that identify the referring service, the target resource and e.g. time and date were included in the sample. Without citation data, none of the services described in the scenarios would be possible.

## Questionnaire

A full report from this part of the questionnaire responses is appended (Appendix 2). There were 31 usable responses. All of the scenarios were interesting to at least a third of respondents. Two were interesting to more than half of the respondents. These were that data would be used to generate reports on trends in scholarship and to compare the institution's data with aggregated data from multiple institutions in the UK. The latter concurs with comments on the private blog suggesting that direct benefits to institutions would be important. The former is a scenario that received no comment on the blog other than a reference by Herbert Van de Sompel to papers on this topic.

Most of the respondents were librarians responsible for e-serials or e-resources or are systems librarians. Others were either managers responsible for e.g. collection development or e-resource assistants. The number of e-journals held at the respondents' institutions varied from 100 to 60,000 with a median and a mean of around 21,000.

Twenty questionnaire respondents would share their data, five would not and six skipped the question. Eight respondents commented on their willingness to share data. Comments included that the decision to share would depend on the value to be derived from doing so; that data from different resolvers may be incompatible; that data from their own resolver was so limited that they doubted these data could be the basis for such a service; and that they were concerned about the effort required to participate.

Eighteen respondents said that their resolver was registered with the UK Router service and eight said that theirs was not. The remaining respondents skipped the question. It is possible that they either do not know what the router is or that they do not know whether their resolver is registered. Unfortunately, the questionnaire did not offer 'don't know' as a response to this question.

Regarding legal, technical, administrative and policy barriers to sharing, many respondents did not know of any, a few named the data protection act and the requirement for anonymity and one said that 'some ejournal licence agreements prohibit data sharing'.

## User Scenarios – evaluated by means of private blog

The project plan indicated that, based on interviews and the online questionnaire, the team would generate a series of use scenarios describing services that may be built onto a database of aggregated OpenURL usage data and would evaluate those scenarios by means of one or two focus groups. As the team conducted the interviews, it became clear that it would be difficult to populate even one focus group on this topic. The interviewees were sufficiently interested to discuss the topic on the telephone but very few would commit to travel to a focus group. The team decided, instead to use online tools to evaluate the user scenarios. It created nine scenarios, published these on a private blog using Wordpress, and invited participants to comment. (Appendix A lists the Scenarios).

The blog was open for a limited period of 10 days. The team believed that opening the blog for a longer period would simply delay response rather than elicit more input and, in fact, after the blog opened, there were no comments for a couple of days. The facilitator prompted participants with individual messages and dialogue began on day three.

The blog was intended as a focus for dialogue rather than as a site to be visited and commented in a single visit. The facilitator sought to encourage dialogue through email, e.g. when TR posted a comment that she thought would interest DB, she would email DB inviting him to respond. This had limited success but most participants simply visited once and posted their comments. Through email correspondence later, one participant told the team that whilst continually juggling priorities in terms of urgency and importance, his focus, when visiting the blog was on those scenarios where he could best contribute.

The team had planned to follow the blog consultation with a conference call to complete the evaluation exercise but, again, on the scheduled date, only one participant confirmed her availability. The team decided not to reschedule the call as it seemed that participants had made their points clear

on the blog; it was not clear that a conference call would have made good use of the time of the invited participants. Instead, the facilitator sought to fill any gaps from the blog through email contact with the participants.

Eleven invited participants registered with Wordpress and were authorised to access the blog. Eight of those commented on the blog. Two EDINA staff members also commented on the blog in order to respond to and query the comments of participants.

Most of the participants remain anonymous. They included an Electronic Serials Librarian, two Managers responsible for development of library systems at large research universities in the UK, a Systems Librarian at a medium-sized university, the manager responsible for systems development in the library of a research council, an information professional working on a related JISC project, and the vendor of one of the market leading resolvers. It is difficult to describe the role of Herbert Van de Sompel (MESUR, Los Alamos National Laboratory) without identifying him. He agreed to be identified as author of his contribution.

It seemed likely that some of the Scenarios of Use developed by the team would be considered more plausible and/or useful than others but the team did not wish to pre-judge these. During this evaluation exercise, none of the scenarios was clearly favoured by more than one or two participants.

Several scenarios were considered potentially useful with various caveats:

- Scenario 1 (Comparative Reports on Outcomes) would be useful as long as Universities could specify their own comparator institutions. (A service would have to take care to maintain anonymity of institutions so perhaps, in this case, a minimum number of comparator institutions should be specified by a participant institution which would then receive details about its own statistics compared with the aggregate of the e.g. eight institutions which it has specified).
- Scenario 2 (Recommender Service): participants suggested that it may be useful but would probably be more so if it were part of a global initiative rather than a national initiative.

Several of the scenarios were considered to be potentially useful services but participants believed that they should more effectively be based on data from other sources e.g. from COUNTER or CrossRef. These included:

- Scenario 3 (Ranking of Papers Based on Use - to benefit the sector)
- Scenario 7 (Enhanced Metadata for Institutional Services Online 2)
- Scenario 8 (Ranking of papers based on use – to benefit users).

Participants considered a further 2 scenarios to be interesting:

- Scenario 6 (Enhanced Metadata for Institutional Services Online 1)
- Scenario 7 (Enhanced Metadata for Institutional Services Online 2)

However, they doubted that data aggregated from OpenURL requests would be of sufficient quality to drive these services. Scenario 7 was considered useful by one participant but he suggested that it may be too expensive to implement across the various regimes in different institutions.

Two of the scenarios attracted little comment other than references to related work (MESUR):

- Scenario 4 Collection development – national licences
- Scenario 5 Emerging trends in UK scholarship

When the blog closed, the project team asked participants why they had not commented on these scenarios. Those that responded explained that they did not have relevant experience and expertise in these areas or (with regard to Scenario 5) that it was not clear, within the scenario, how the information would be presented. This reflects the importance of securing input to the evaluation process from people who have expertise in all of the Scenarios as well as providing useful feedback for future development of 'Scenarios of Use'. On reflection, in email, one participant indicated that he would consider Scenario 4 to be useful.

## Legal issues

Little has been written on the legal issues associated with aggregating usage data. The JISC Usage Statistics Review Project<sup>5</sup> explored legal issues and reported that the recording and processing of IP addresses and the use of cookies can breach privacy laws. This is determined by national laws and EU regulations and the strength of regulation varies across countries. In the UK, the issue of whether IP addresses constitute personal data for the purposes of Data Protection legislation is currently debated (J. Campbell, JISC Legal, personal communication). Personal data are data that relate to a "living individual who can be identified".

In isolation, an IP address is unlikely to be deemed personal data because it identifies a computer rather than an individual. With reference to a complaint about Yahoo!'s disclosure of information about a journalist to the Chinese authorities, the Hong Kong Privacy Commissioner stated that "an IP address per se does not meet the definition of 'personal data'." Thus, if IP addresses are collected to analyse aggregate patterns of website use, they are not necessarily personal data and are not subject to the Data protection Act (DPA)<sup>6</sup>.

However, the Information Commissioner's view is that an IP address may fall within the definition of personal data under the DPA where it can be linked to an individual user, perhaps through other information held, or from information that is publicly accessible<sup>7</sup>. (This concurs with the view of an independent EU Working Party on Data Protection<sup>8</sup>). It applies even when data are anonymised as they are, nevertheless, personal data when collected. As the services being explored in this project depend on the collection of IP addresses in conjunction with 'requester' entities in the OpenURL request, the IP addresses may be construed as personal data. This becomes even more likely where IP addresses are 'static' i.e. relate to a specific computer and thus are more likely to be traceable to a specific individual. As the status of IP addresses has not been tested in court, it is considered prudent to consider them personal data<sup>8</sup> with reference to the DPA. Thus, in order to stay legal, it would be necessary that the referring service disclose this intended collection and use of IP addresses to its users. This would usually be included in the Privacy Policy for that service<sup>9</sup>.

## Technical work

### Requirements for successful sharing of OpenURL data

The requirements described here assume a mechanism for aggregation of OpenURL usage data similar to that proposed by the LANL group<sup>10</sup>.

<sup>5</sup> <http://www.jisc.ac.uk/publications/publications/usagestatisticsreviewreport.aspx>

<sup>6</sup> Out-Law.com (Pinsent Masons) (2009) IP Addresses and the Data Protection Act, <http://www.out-law.com/page-8060> (accessed 10 July 2009).

<sup>7</sup> See Information Commissioner (2003) *Data Protection Good Practice: Collecting Personal Information Using Websites*, [http://www.ico.gov.uk/upload/documents/library/data\\_protection/practical\\_application/collecting\\_personal\\_information\\_from\\_websites\\_v1.0.pdf](http://www.ico.gov.uk/upload/documents/library/data_protection/practical_application/collecting_personal_information_from_websites_v1.0.pdf) (accessed 10 July 2009).

<sup>8</sup> Article 29 Data Protection Working Party (2000) Privacy on the Internet – an Integrated EU Approach to Data Protection, [http://ec.europa.eu/justice\\_home/fsj/privacy/docs/wpdocs/2000/wp37en.pdf](http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2000/wp37en.pdf) (accessed 10 July 2009)

Article 29 Data Protection Working Party (2007) Opinion 4/2007 on the Concept of Personal Data, [http://ec.europa.eu/justice\\_home/fsj/privacy/docs/wpdocs/2007/wp136\\_en.pdf](http://ec.europa.eu/justice_home/fsj/privacy/docs/wpdocs/2007/wp136_en.pdf) (accessed 10 July 2009).

<sup>9</sup> See, for example, University of Edinburgh Library (2009) Privacy Policy, <http://www.ed.ac.uk/about/privacy-policy> (accessed 10 July 2009).

<sup>10</sup> see Bollen and Van de Sompel, "An Architecture for the Aggregation and Analysis of Scholarly Usage Data", JCDL 2006, <http://arxiv.org/abs/cs/0605113v1>

1. *Institutional resolvers must collect usage data*  
The usage data must be sourced from institutional resolvers, so this is a fundamental requirement. Our investigations indicate that some usage data is collected by all major resolver products.
2. *Usage data must include "who", "what" and "when"*  
The usage data must include some information identifying (anonymised) users, the items being requested and date/time of requests. If these data are absent the applications suggested by Bollen and Van de Sompel are not possible.  
  
SFX (the resolver offered by ExLibris) is known to log these data. Other resolvers appear to offer only summaries of usage with no information at all related to individual requests; we do not know what data are logged for the purpose of producing these summaries.
3. *Institutions must be willing to share usage data*  
Cooperation from institutions is a prerequisite for aggregating OpenURL usage data; minimally, access to a harvesting mechanism would have to be enabled, the need for some local development work should be envisaged.  
  
Librarians were generally positive regarding the prospect of sharing usage data, but would be reluctant to invest time or money to do so as resource in libraries is scarce and whilst interesting, they did not consider services built on aggregated data to be sufficiently interesting that they could justify investing in participation. It is not possible to estimate the extent of effort required unless vendors make usage data accessible (see requirement 4).
4. *Institutions must be able to share usage data*  
The usage data must be accessible, which requires a mechanism for harvesting or downloading. Our investigations indicate that these data are generally *not* available to institutions. SFX stores the data in a proprietary format, and they are not accessible to institutions unless they subscribe to bX (the commercial service based on shared OpenURL data, offered by ExLibris), in which case a harvesting mechanism is enabled. It is not clear whether other resolvers log usage data, but it is known they offer no mechanisms for accessing the data.
5. *Usage data must be available in standard form*  
The mechanism proposed by Bollen and Van de Sompel would require a standard format for shared usage data; they suggest OpenURL XML ContextObject format. Given the failure to satisfy requirement 4, we did not investigate this requirement. In the event that harvesting or downloading became possible, but formats could not be standardized, we would suggest the aggregating body works around this requirement by transforming the data.
6. *Usage data must be aggregated*  
It would be necessary to establish procedures for "publishing" usage data and have an infrastructure for gathering and storing it to a central location. This can be designed and costed when mechanisms for collecting and harvesting usage data (requirements 2 and 4) are established.
7. *Aggregated use data must be more useful than individual institutional data sets.*  
An aggregated data set is only useful if it permits actions that cannot be performed (or performed as well) with individual institutional data sets. Such an evaluation depends on usage data being collected from a range of resolvers, to determine the extent and quality of information (see requirement 2).
8. *Must gain expertise in techniques for applications*  
Some of the applications suggested by Bollen and Van de Sompel require specialist skills and knowledge. Before a judgment can be made on this usage data would need to be collectable from a range of resolvers, and the feasibility of various applications considered. We envisage that a National Data Centre could do this work in collaboration with academic specialists.

At the outset of this project, it was envisaged that usage data would be generally available for librarians to share if they wished to do so, and (implicitly) that obstacles may be connected with legal issues, resourcing within institutions, or the level of interest in the services that might be offered based

on the aggregated data.

In fact the basic requirement, that usage data should be available, seems not to be satisfied in most cases. Furthermore, while libraries would, in principle, be willing to share, they are not willing to invest effort to do so. As collection of 'session' data may fall within data protection legislation, I would be necessary to secure permission for such collection from the institutions registered with the router and to secure their commitment to inform their users that such data will be collected. .

### Using OpenURL Router data

The OpenURL Router service provides an alternative source of OpenURL data. This service acts as a registry of institutional OpenURL resolvers, and provides an HTTP redirect function: any OpenURL request may be addressed to the OpenURL Router, and it will be redirected to the appropriate OpenURL Resolver for the end user<sup>11</sup>.

The OpenURL Router has 90 registered resolvers, which gives it the advantage of providing large quantities of data. From 1 April 2008 - 1 April 2009 the Router handled 25,282,750 requests. The nature of the router is such that a proportion of these requests do not contain OpenURL data, and the first step is to eliminate these. Of those requests, 4,894,534 did contain OpenURL data. The requests of particular interest are those that contain enough information to enable us to identify a particular journal article<sup>12</sup>, of which there are 1,701,025.

### Applicability to Scenarios of Use

These usage data are not appropriate for several of the possible Scenarios of Use (Appendix 1), for various reasons.

Since Router data describe only requests to resolvers, and do not reveal anything about the outcome of a request<sup>13</sup>, scenarios 1 (Comparative reports on outcomes) and 9 (Ranking sources within an institution) are ruled out. It is also not clear if scenarios 3 and 8 (Ranking of papers based on use), 4 (Collection development), 5 (Emerging trends in UK scholarship) can be addressed, when data relate to articles *sought* by users, and not those necessarily *read* by them. The discrepancies between seeking and reading papers must to some extent distort measures of impact and availability.

The OpenURL Router is also known to be used by a subset (though probably a large one) of institutions with resolvers, and is certainly used more heavily by some referring services than others. It would be naive to assume that the data are a statistically representative sample. This will certainly have an impact on scenarios 3 and 8 (Ranking of papers based on use), 4 (Collection development) and 5 (Emerging trends in UK scholarship).

The scenarios to which Router usage data are likely to be useful are 2 (Recommender service) and 6 and 7 (Enhanced metadata for institutional services).

### Feasibility of enhanced metadata service

Scenarios 6 and 7 require the aggregation of bibliographic metadata in large quantities. There is no doubt that the Router logs provide very large quantities of bibliographic metadata, which might be compiled to form a citation database. It would in effect be a catalogue built from user contributed data.

---

<sup>11</sup> with various fall back options if the end user cannot be recognized, or is from an institution with no registered OpenURL resolver

<sup>12</sup> for the purposes of this investigation we used a simple assumption that an article can be identified when ISSN, volume, issue and starting page numbers are present; in all likelihood, further research would provide additional techniques for identifying articles (using a range of slightly more ambiguous identifiers such as journal title, article title, author names etc.), so increasing the sample size further

<sup>13</sup> *i.e.* was the resolver able to offer the end user links to full text

We would not expect such a database to be remotely as comprehensive as a properly compiled bibliographic index, either in breadth of coverage or quality of metadata; but it might be suitable for applications focused on citations commonly requested by users. The obvious advantage, compared to a conventional bibliographic index, would be very low cost.

The success of a service for enhancing metadata would depend on its ability to “fill in the gaps” in citations. Citations sometimes include few metadata, and sometimes incorrect metadata. For example a citation containing an ISSN but without a journal title could have the latter added; or a citation with an ISSN and journal title could be checked to ensure the values match.

A proper evaluation of the extent to which specific functions could be provided, that satisfy actual demand, is beyond the scope of this study. It was possible to conduct a brief survey of the metadata included in OpenURL requests, which gives some indication of what functionality might be possible.

Many citations do not include both ISSN and journal names. Mapping an ISSN to a journal name or vice versa could be a useful function.

Based on our sample data, ISSN is provided in 95.4% and title in 38.3% of requests. There is some ambiguity in title values, due to poorly constructed requests that confuse journal and article titles, but it should be possible to eliminate this by accepting only frequently occurring pairings of journal titles and ISSN. This statistical approach assumes that journals tend to be requested many times.

A quick survey shows that 667 journals were requested 1,000 times or more. This is a small proportion of all available journals, but accounts for 99.97% of all requests. The top 6,180 journals have at least 100 requests each, and account for more than 99.99% of all requests. This supports our assertion that using “user contributed” citations would be a poor way of creating a comprehensive catalogue, but would provide the basis of a service that could satisfy actual end user demands.

Other citation metadata are provided in large numbers, most commonly:

1. First author surnames are provided in 34.1% of requests, with initials in 19.5% and first names in 6.4% of requests.
2. Volume numbers in 37.7%, issue numbers in 36.1%, pages numbers in 38.4% of requests
3. Journal article titles in 35.2% of requests
4. Date information (at least year) in 95.0% of requests
5. ISBN in 10.5% and book title in 8.6% of requests

These are all valuable metadata that have the potential to enhance citations.

### **Feasibility of recommender service**

The OpenURL Router usage data represent a large sample of end user requests, and the possible lack of statistical representativeness is not necessarily a major issue; where a recommendation can be made, it would be based on genuine associations between end users' requests for articles. If a journal or subject area is under-represented in the available data, then recommendations for those articles would be unlikely to be made; this is unfortunate, but it is not a problem in the same way that an inaccurately low impact metric (scenario 3 and 8) would be.

The key issue for implementing a successful recommender service, based on Router usage, is whether it is possible to associate each requests with a particular individual user. Bollen and Van de Sompel describe this as the “who” data, *i.e.* data that describe the person requesting an item either explicitly (user identifier) or analogously (an attribute such as IP address which may be related to a identity). “Who” data are the most difficult data to extract from OpenURL logs<sup>14</sup> in a usable form, and the focus of our investigation.

Ideally the “who” data would permit user behaviour to be tracked over extended periods, possibly separately by weeks or months. Another possibility is to record *session* identity, which tracks

---

<sup>14</sup> The other data include “when”, the date and time of a request, and “what”, the journal article or other item being requested; these are explicitly logged by the Router

behaviour during a single period of activity. The aim is to obtain samples of requests made by individuals, and these samples would be larger if user identity could be established. On the other hand, it may be possible to establish session identity with much greater certainty than user identity, which may mean that using session identity yields a greater number of samples.

There are various possible approaches to establishing user or session identities:

- *“Requester” entities in OpenURL requests*  
It is possible to include metadata describing the end user in an OpenURL request using the optional *requester* entity. In practice this entity is seldom, if ever, used by referring services.
- *OpenURL Router cookies*  
Browser cookies could be used by the OpenURL Router to enable us to associate requests with particular end users<sup>15</sup>. This would require the development of a privacy policy and negotiation with administrators of institutional resolvers. Individuals accepting cookies on their computers would have to consent to collection and manipulation of their usage data.
- *IP address*  
IP addresses may be used to distinguish between some end users' machines. The use of proxies (such as web caches and firewalls) is a problem since many (possibly all) users from an institution may share a single IP address. It may be possible to differentiate between proxies and desktop machines by various methods:
  - the number of requests from each IP address may reveal proxies, which can be expected to appear much more frequently in the logs
  - host names (from reverse DNS lookup) containing words such as “cache” or “firewall” are a good indication that a machine is a proxy
  - the HTTP\_X\_FORWARDED\_FOR header may be added to HTTP requests by proxies, to pass on the IP address of the end user's machine; this field may be excluded<sup>16</sup>, and in any case is not authoritative (it can arbitrarily be set to any value and cannot reasonably be verified), and so relying on this header would probably be a last resort

Requester entities and cookies would provide good methods for establishing user identity, if implemented. Implementation of requester entities would require “buy in” from a wide range of referring service providers; experience suggests that this approach is not viable. A cookie mechanism might be added to the OpenURL Router, if the privacy implications are acceptable to resolver administrators: this should be considered in future if user identity is considered necessary.

IP addresses are an intriguing possibility, since these are available from existing OpenURL Router usage logs. Any use of IP addresses assumes that we can differentiate proxies from desktop machines. IP addresses might yield user identity or session identity, depending on the assumptions we are willing to make. If we are willing to assume that a single IP address is a desktop machine used by a single person, we can treat IP addresses as user identifiers. Alternatively, we can try an approach of looking for “clusters” of requests made within a short period of time from a single IP address, and assume this is a single person: this uses the IP address as a session identifier.

Clearly, as we are required to consider addresses ‘personal data’ with reference to Data Protection legislation users’ consent must be obtained before they are collected (see Recommendations).

### *Summary of the analysis*

The OpenURL Router logs provide a substantial quantity of usage data, including requests for journal articles that may be grouped into “sessions”, i.e. where it may reasonably be assumed that a set of articles was requested by the same individual. These are the sort of data that could form the basis of a “recommender” service. It is likely that additional research could refine the methods used for

<sup>15</sup> Strictly speaking, cookies identify individual installations of user agents. This is likely to correspond to an individual end user, but this may not be the case if computers are shared. Errors may be minimized by limiting the duration of the cookie, but this would decrease the sample size (number of items known to have been requested by an individual end user). These issues would require investigation prior to implementation of a cookie mechanism.

<sup>16</sup> HTTP\_X\_FORWARDED\_FOR may be suppressed for security, or simply may not be enabled

identifying a “session”, and improve the quality and possibly the quantity of data by more careful consideration of the metadata required to unambiguously identify a journal article.

We wish to emphasise that this preliminary investigation did not attempt to produce mappings that would enable recommendations; we sought only to establish whether useful data, including bibliographic metadata and analogues for user identity, could be extracted from the OpenURL Router usage logs.

Full details of the log analysis undertaken are given in Appendix 3.

## Conclusions and recommendations

### Conclusions

The aim of the project was stated as:

*“to scope a UK architecture for the analysis of OpenURL linking data intended to provide a basis for ongoing services”*

The project team set out to gain a good understanding of the technical, legal, and administrative challenges and opportunities related to sharing and using OpenURL link server data and to assess the relative and complementary value of data from the OpenURL router and from OpenURL resolvers within institutions by gathering and inspecting those data. We also sought to explore potential uses of these data through consultation and through manipulating the sample data available. Our conclusions are organised by four themes: (1) the level of interest and viability of services based on aggregated OpenURL data; (2) libraries’ willingness to share data; (3) the availability of OpenURL resolver usage data; and (4) the value of the OpenURL Router as a source of data on which useful services may be built.

#### *Services based on aggregated OpenURL data*

- The Evaluation of the nine Scenarios of Use described by EDINA indicated that whilst some were of interest to some of those consulted, no single scenario was of interest to all or even to a majority.
- Participants in the consultation exercise were most interested in scenarios 2 (Recommender service) and 6 and 7 (Enhanced metadata for institutional services)

It seems that some of the services described, e.g. a recommender service, would be more useful were it based on a global aggregate of data rather than a national aggregate.

#### *Libraries appear willing to share data, in principle*

The consultation exercise revealed that libraries are, in principle, willing to share data but they do not have resource to commit and would wish to be reassured regarding how those data would be used. Thus, they would be willing to share:

- If there were clear benefits to them in sharing.
- If it were easy to contribute data or if data could be obtained directly from the resolver vendor.
- If anonymity were guaranteed – for both users and institutions.
- If the data were to be used only for non-commercial purposes.
- If use of the data adhered with data protection legislation
- As long as there was no requirement to provide data on an exclusive basis, i.e. they wish to retain control of their data and how they are used.

#### *Availability of OpenURL resolver usage data*

The availability of OpenURL resolver usage data for aggregation depends on a number of factors, including: the data logged by the resolver; the accessibility of those data; and the formats in which they may be exported from the resolver service. EDINA inspected logs and reports from institutions using three different OpenURL resolvers. These suggested that aggregation of data from institutions across the UK may be limited more by the lack of ready access to those data and/or lack of relevant

data fields than by the lack of a standard format. SFX collects sufficient data to drive a recommender service but OAI-PMH is enabled only for institutions subscribing to the bX service.

Unlike data from the Router service, those from OpenURL resolvers may potentially indicate the outcome of the request i.e. whether the user was offered a link to full text or some other service (library holdings, ILL, etc.)

### *OpenURL router as a data source*

The data available through the national OpenURL router service can provide data that could form the basis of a “recommender” service.

These data are readily available, but cannot record the outcome of a request nor are they representative of the whole UK research community as data would relate only to those institutions registered with the resolver, only to services covered by the OpenURL resolvers of those institutions, and, among those, only those services that use the router.

## **Recommendations**

Clearly, any further work should be based on a clear perceived need for services based on the data, rather than on the availability of aggregated data. No single service described in the user scenarios developed by the project team was of great interest to those consulted. This may reflect a limitation in the range of scenarios rather than indicating that there is no useful service which the OpenURL usage data might support. The project team recommend a number of further activities:

**(1)** We propose that further work be undertaken to explore how the OpenURL usage data may be released for use by third parties. There may well be scenarios of use that were not identified by the project team or the users consulted. Making the data freely available would allow others to explore applications and mash-ups that we did not consider during the project.

We would begin by making the OpenURL Router data available thus providing the basis for a model to release such usage data by institutions.

Before doing so, we would seek agreement from the institutions that use the OpenURL router service. During the consultation exercise, librarians from some institutions indicated that they believe that these data belong to the institution and thus their aggregation and use should be subject to agreement with that institution. There are no intellectual property rights in these data, so the main issue is one of privacy, primarily whether the IP address identifies an individual user although some institutions are also concerned that the institution remain anonymous wherever the data are used. As indicated above (see Legal Issues), at this stage, we should proceed on the assumption that IP addresses are personal data and thus, should secure permission from the institutions using the OpenURL Router before we collect it and should secure their commitment to inform their users that data will be collected and used in the ways specified. With a view to making this as easy as possible, EDINA would provide sample words to indicate how the data will and will not be used (e.g. ‘We will not use your IP address to identify you in any way.’) Nevertheless, EDINA staff fear that this requirement may deter some institutions from registering with the Router; registration is currently very quick and easy.

The data made available in this way may be a UK contribution to a global aggregate. For example, an organisation with data collected from institutions across the globe but with only partial representation in each country may use the UK data to enrich its UK representation. In that way the services run on that data set may be run on a richer foundation while those services based on UK use (e.g. Collection Development for National negotiation) may be run on the UK subset of that global aggregate.

**(2)** We propose further dialogue with vendors to explore the potential for release of suitable usage data. Our contact with vendors suggested that before they could release institutional data their resolvers may require development to gather and export those usage data. While we established good dialogue with two vendors during the project, contact with the appropriate vendor staff in a third company was established only towards the end of the project. There is scope for further talks with resolver vendors to determine whether and how more complete and accessible log data could be

provided. We assume that although some vendors lack the scale of transactions required to drive a recommender service, they would be interested in contributing to an open solution as a way of maintaining the competitive position of their own products (i.e. by lowering the competitive advantage of larger products with recommender services). Thus these vendors may be interested in contributing data to such an open solution.

**(3)** We propose that a prototype be built using the OpenURL router data for scenarios 2 (Recommender service) and 6 and 7 (Enhanced metadata for institutional services). This would:

- Act as a catalyst for thinking about other types of use of OpenURL usage data and as a means to engage with institutions on usefulness of a service based on it
- Allow us to test the initial outcomes from the technical investigation and evaluation.
- This development should seek to include volunteer services which would try out the prototype and provide feedback and should include:
- A review of other recommender services already available;
- Exploration of the analysis required to make a recommendation i.e. a ranked list.
- Further investigation to determine/define a 'user session' based on usage data – building on the initial definition developed during this project.
- Further investigation of metadata analysis: for example, the project erred on the conservative side when identifying an article.
- Further investigation of any legal issues e.g. the possibility that an IP address, whether anonymised or not, constitutes personal information.

## Appendix 1: the Scenarios of Use and a précis of comments

This Appendix documents the Scenarios of Use which were published on a private blog for comment by invited participants. They were:

- (1) *Comparative reports on outcomes*
- (2) *Recommender service*
- (3) *Ranking of papers based on 'use' (to benefit the HE sector)*
- (4) *Collection development – national licences*
- (5) *Emerging trends in UK scholarship*
- (6) *Enhanced metadata for institutional services online 1*
- (7) *Enhanced metadata for institutional services online 2*
- (8) *Ranking of papers based on 'use' (to benefit users)*
- (9) *Ranking sources within institution*

The title of each scenario is followed by the text of the scenario as it appeared on the blog which is followed by the comments of participants. This scenario text is italicised here to differentiate it from the précis of participants' comments.

### (1) *Comparative reports on outcomes*

*The University of West Scotland participates in the Higher Education Data Aggregation Service (HEDAS). Participation required very little from staff at the University. It took less than 30 minutes to configure the resolver so that usage data could be harvested by HEDAS. In return, the University receives reports that compare usage of the local resolver with usage of resolvers throughout the country. It can be difficult to explain to managers in Central Administration, the difficulties of linking resources together and providing a reliable, streamlined service to staff and students. The comparative reports from HEDAS show that UWS does this as well as any other university in the country.*

Two participants indicated that comparative reports on outcomes would be useful only if they provided figures that compared their own university's performance with that of comparable universities rather than with the whole sector. . They would like to specify those comparable organisations Even then, these would form only one part of a picture which would most usefully be completed with figures from other sources, e.g. COUNTER statistics. A library systems vendor suggested that such reports should be based on various groupings, e.g. library type or CURL members.

### (2) *Recommender service*

*Alice is studying Comparative Religion at the University of Trumpton. She's writing an essay contrasting Taoism and contemporary Western thought. Comparative Religion doesn't benefit from many specialist databases so Alice often starts with Google and the University Portal. She starts by searching for sources on Taoism and the West. The results list looks interesting; it includes some materials that she has already read and some that are new to her. She selects a couple and moves them into her folder. A sidebar offers recommendations based on the selections of other readers – 'those who chose this also chose that'. She's curious to see that the recommendations include a paper on quantum physics. What's that got to do with Taoism and the West? She follows her curiosity and discovers systems theory. Wow, lucky find!*

*Alice can flag the recommendations that she found to be most helpful. The recommender system is adaptive, and on subsequent visits she finds that the recommendations that are offered reflect her interests with increasingly accuracy.*

The recommender service was one of two scenarios that attracted most comment. While some contributors consider this to be a valuable service, there was some question about its feasibility and affordability as well as some clear ideas about the parameters of such a service. One participant suggested that rather than creating a service based on UK data, the UK community could derive greater benefit by joining a global initiative such as ExLibris' bX. Herbert Van de Sompel agreed adding that 'a global span of usage data allows delivering recommendations that are aligned with global research (reading) trends and ... once one has a global dataset, it also is possible to restrict recommendations to particular subsets. Best of both worlds!'

### *(3) Ranking of papers based on 'use' (to benefit the HE sector)*

*The UK has replaced the Research Assessment Exercise with a metrics-based evaluation of research output Thomson. ISI provides authoritative citation data on global use. These are more valuable in some disciplines (e.g. Chemistry and Physics) than others (Law). Furthermore, as 80% of publications are cited only 5 times and a considerable proportion are never cited \*, it is useful to supplement the ISI data with data from various UK-based sources. Among the measures used to evaluate the output of a department is the volume of traffic linking to the published papers of researchers from the department. HEDAS is a source of these data.*

*\*Oxford Workshop on the Use of Metrics for Research Assessment,  
<http://www.wolfson.ox.ac.uk/metricsworkshop/>*

While some participants considered this to be a potentially valuable service, one suggested that it would be more appropriate to base such a service on statistics of use from target services e.g. COUNTER statistics. One respondent identified two other projects that are currently working to develop alternative metrics in this field. Two respondents commented that linking does not equate to but merely implies usage. Thus, 'ranking based on usage should be subject to appropriate and widely understood rules.' Herbert Van de Sompel suggested that 'agreements on the kinds of expressions of interest (which intentional data, e.g. article download, view of abstract, email of abstract, download of citation, etc.) to include for the computation of "usage" based metrics' would be required and possibly also agreement on the fraction they contribute to usage-based score'. He added that a lot can be done with a representative sample of data and referred to a paper published from MESUR (<http://arxiv.org/abs/0902.2183v1>).

### *(4) Collection development – national licences*

*JISC Collections and NESLI2 are among the services negotiating licences on behalf of the UK HE/FE sector. HEDAS informs their Collection Development plans by identifying gaps between what users request and what is available through national licence schemes.*

There was no discussion of this scenario although Herbert Van de Sompel referred to two papers that discuss use of usage data for collection development (See <http://www.dlib.org/dlib/may03/bollen/05bollen.html> and <http://www.dlib.org/dlib/june02/bollen/06bollen.html>).

### *(5) Emerging trends in UK scholarship*

*Sunil has been offered scholarships to do his PhD at two different universities. Both departments have excellent research reputations as do the professors who would supervise his work. He wants to make the right choice and is basing his decision on many factors. He's*

*been to the departments to meet his potential colleagues and is gathering information about their research. He's keen to work with the most progressive researchers. Data on emergent trends in UK scholarship, available from the Higher Education Data Aggregation Service (HEDAS) help with this. He can see the directions of growth in his field and compare these with the current output of the Departments that offered him scholarships.*

Project MESUR has explored this type of scenario. Herbert Van de Sompel was the only participant to comment. He explained that although usage data precede citation data, they still pertain to published literature and so are subject to publication delays. He identified various 'true' indicators related to research funding and conference presentations but conceded that these lack structure and so it would be difficult to derive meaningful trend analysis from them. Thus, he said that, pragmatically, linking servers provide a useful source of data.

#### *(6) Enhanced metadata for institutional services online 1*

*Jo is depositing an eprint in her institutional repository. The paper has already been published in the European Journal of Ecopsychology. The repository requires her to complete some mandatory metadata fields. She keys in the title of the paper, the title of the journal, the year of publication, and her own name. There's an option to 'complete the record?'. She selects that and the metadata record is, miraculously completed with the ISSN of the journal, the volume, issue and page numbers, and an OpenURL to provide electronic access to the published version. (The repository accesses a bibliographic database compiled from OpenURL logs. The database is clearly not comprehensive as it is created from user-contributed data but it is often useful for completing incomplete records.)*

This was one of the Scenarios that received most comments. Those participants working in libraries considered this to be a valuable service, one on condition that it facilitates bulk upload. Another questioned the need to develop it given that this facility is available with some commercial repositories. (The benefit of a community service is that it would be available to all universities regardless of which, if any, resolver they licence). Herbert Van de Sompel expressed concern about the quality of metadata that would be available for such a service as metadata in OpenURLs are typically insufficient for 'decent bibliographic citation'.

There was tangential discussion about whether or not researchers who have published papers in journals would be motivated to deposit these in repositories.

#### *(7) Enhanced metadata for institutional services online 2*

*Jason is creating a reading list for his students. He has a collection of journal articles that he wants to include but the data for many of them are incomplete. He goes into the VLE and keys in the titles of the papers, years of publication and first-named author. Then he clicks 'Complete the records'. Further detail for each paper is provided including, essentially, a link to the electronic copy. In one instance, he'd mistyped the date – this is corrected. Jason doesn't know how this happens but it helps him provide a useful list to his students without a lot of boring additional work on his part. (The repository accesses a bibliographic database compiled from OpenURL logs. The database is clearly not comprehensive as it is created from user-contributed data but it is often useful for completing incomplete records.)*

One participant considered this to be a useful function but suggested that it may be very difficult to support generically across institutions with different reading list regimes. He also wondered whether incorrect amendments may be made. Another participant suggested that the Talis product Aspire fulfils this function and a third the Cross-Ref database would be a better source of data for such a service.

*(8) Ranking of papers based on 'use' (to benefit users)*

*Ted is applying for his first lectureship. As a researcher he has excelled in his field. His publication record demonstrates this. Nevertheless, he knows that competition for the post will be fierce. He has a sense, through communicating with peers in the field, that his work is well respected but this is difficult to evidence as citation data take time to filter through. He is delighted to discover the HEDAS service. This provides evidence that, within UK HE at least, his papers are popular relative to others within his field. HEDAS cannot report readers but the numbers of readers linking to his papers compare well with the numbers linking to other papers in the field.*

Two participants suggested that other sources of data may be more useful to inform the service described than aggregated link server data. EDINA would consider the benefit of using OpenURL usage logs as the basis for this service to be that they reflect what users want to access rather than what they succeed in accessing (as would be case if statistics from targets were used).

*(9) Ranking sources within institution*

*The University of West Scotland has electronic journals in a wide range of sources. It thinks it has 20,000 discrete titles. It's difficult to be sure because the content of packages fluctuates and changes, sometimes month by month. While the core journals in any package are stable, the lesser known titles come and go. These titles are often included in many packages. At any time, UWS can be subscribed to as many as 9 instances of a single title. It could use its OpenURL resolver to direct users to a 'preferred' copy but it doesn't because user preferences vary. Individual users get accustomed to the interface of a favourite aggregator. The library configures the resolver so that users are presented with a list of options (as many as 9). While those users with a clear preference have no difficulty choosing between the options, many other users are simply confused by choice. What's the difference and which one is best? A new service helps those users by ranking the options in order of popularity. Users can see which is most the most popular source and which the least popular.*

Three participants commented that this service may be useful but one of them already has the service as a function of her link resolver and a second suspected that the effort required to link this service into institutional and commercial resolver systems so that it may be embedded would outweigh any benefit.

## Appendix 2: Report from online questionnaire

An online questionnaire was published on 3 February 2009. Responses were invited on both lis-link and lis-e-resources. Over a period of 5 weeks until 13 March 2009, 42 responses were collected of which 31 were usable. (Eleven respondents provided so little information that their responses were of no use in this survey, for example, some gave information about their role and library but nothing about OpenURL data usage logs.)

### Respondent's roles

Most of the respondents were either librarians responsible for electronic resources or electronic journals/serials or systems librarians. These accounted for 24 responses. The remaining 7 included e-strategy managers, a Librarian and a Deputy Librarian, three Heads/Managers of Collections, Purchasing, or Procurement, an Information Assistant and an Electronic Journals and Database Administrator.

### Number of journals held by respondents' institutions

The number of journal titles held by respondents varied from 100 to 60,000. The median was 21,000 journal titles. The mean was 21,669.

### Resolvers used at respondents' institutions

The four resolver products most commonly used by respondents were SFX, Innovative WebBridge, Ebsco LinkSource and Serials Solutions 360 Link with SFX and WebBridge in the lead by a large margin. Thirteen respondents use SFX, 8 use Innovative WebBridge, 4 use Ebsco LinkSource and 3 use Serials Solutions 360 Linkfinder.

### Collecting usage log data

Twenty seven respondents said that their resolver logs usage, three said that theirs does not and one skipped this question.

Only 13 respondents use the usage logs collected by their resolver, 16 do not, the remaining two skipped this question.

### How usage log data may be usefully employed

Respondents were asked to identify, from a list of five suggestions, how usage data may be usefully employed. The two most popular options, with 16 votes each, were 'to generate reports on trends in scholarship' and 'to compare your institution's data with aggregated data from multiple institutions within the UK'. The remaining three options were selected by 9, 12 and 13 people as follows:

- To provide a 'recommender' service making suggestions such as 'other readers who chose this article, also chose that article'(9)
- To rank articles based on numbers accessing them. (12)
- To provide a service that identifies the most popular source for a journal that is common to many aggregated services. (13)

Respondents used the comment box in this question to suggest additional uses of data. These all relate to local uses of local data. Examples include that usage data could indicate which of their journals are well used and which are not or that usage data could indicate which of the journals to which the library is not subscribed do users seek and which of the library's subscriptions remain unused. It's interesting that one of the most popular uses, 'to provide reports on trends in scholarship' was the scenario that attracted no comment on the blog. The facilitator asked blog participants why they hadn't commented on this scenario.

### Sharing usage log data

20 people said they would be willing to share their data in order to derive more value from it. Five said they would not and 6 skipped the question.

Eight people commented on this question four of them to say that their decision to share data would depend on the value to be derived from doing so. One of them suspected that data from different resolvers may be incompatible. Another was concerned about the effort required to participate as they currently have difficulty finding time to analyse the data from their own resolver. One respondent

said that sharing would be contingent on addressing concerns related to privacy and anonymity, another that s/he would participate only if the data were restricted to non-commercial use, and another said that sharing log data would conflict with the current disclaimer [presumably, this refers to a disclaimer to library users].

### **Identifying legal, technical, policy and administrative constraints to sharing**

Eleven people responded to the question 'Do you know of any legal, technical, administrative or policy issues that would impose constraints on sharing of data by your institution?' Two said they didn't know. A further 8 said that there would be no constraints (although one with the qualification that the data should be anonymised and another that use should comply with legislation including the Data Protection Act). One respondent said that 'some ejournal licence agreements prohibit data sharing'.

### **Registration with OpenURL Router service**

Eighteen respondents said that their resolver was registered with the Router service, 8 said that theirs was not. The remaining respondents skipped the question.

## Appendix 3: Analysis of OpenURL Router logs

### *Preliminary investigations*

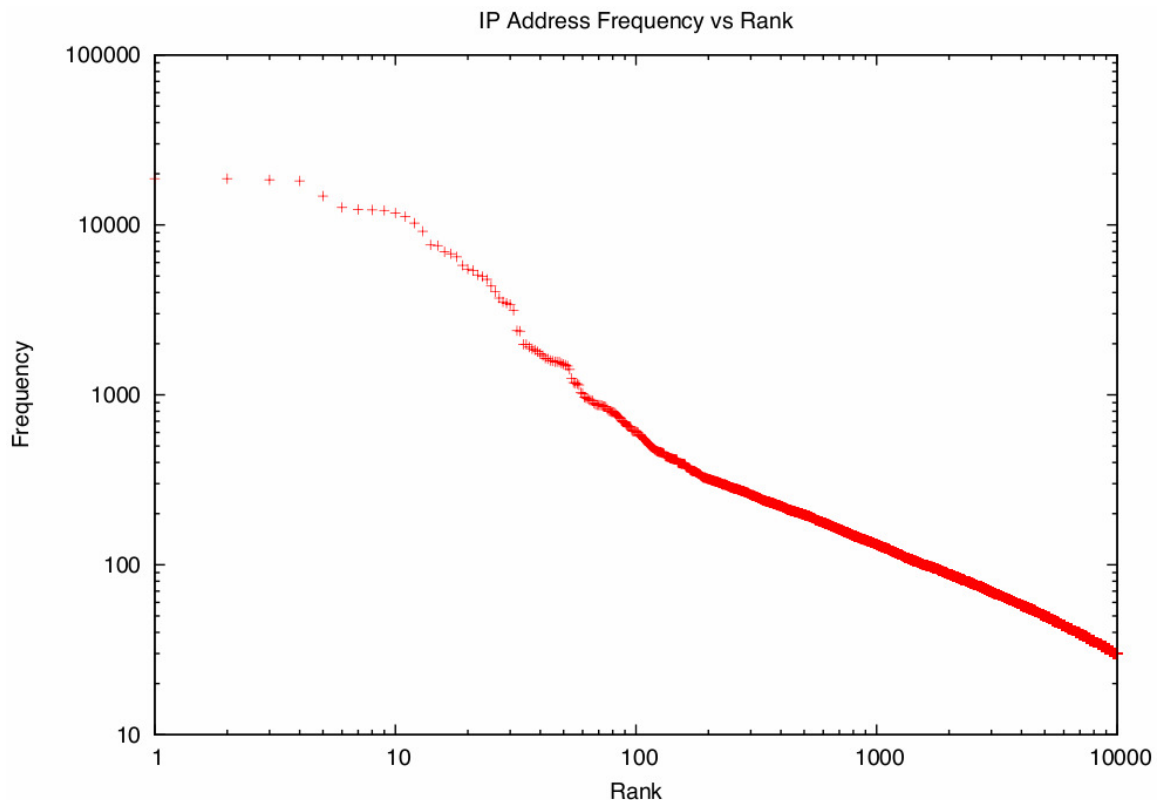
Within the time frame of this project it was only possible to conduct a preliminary analysis of the OpenURL Router logs. We used a sample from 1 April 2008 to 1 April 2009, a total of 4,894,534 OpenURL requests.

We were able to confirm immediately that none of these requests contained a requester entity (*i.e.* end user information supplied by the referring service). The use of IP addresses is the remaining option for identifying users or sessions.

### *Analysis based on frequency of requests from IP addresses*

Bollen and Van De Sompel proposed a statistical solution to the problem of identifying proxies. The distribution of the frequency of requests from 'genuine' end user machines should obey a Power Law. Thus a log plot of frequency of occurrence of IP address against rank should show a straight line. Proxies or other non-human users (such as search engine bots) with very high traffic should show themselves as departures from this straight line. This should enable us to eliminate the vast majority of such requests.

1,701,025 of the OpenURL requests in our sample contained metadata sufficient to unambiguously identify a journal article. These requests are the ones of interest for the various applications that have been proposed, and so we restricted our analysis to this subsample, which we refer to as "article requests".



**Figure 1:** IP addresses ranked according to frequency (number of requests observed from each address). Linearity begins at roughly a rank of 190, suggesting IP addresses ranked below this are predominantly end user machines.

We ranked IP addresses according to frequency (number of requests observed from each IP address). Figure 1 shows that linearity begins at roughly a rank of 190. The flattened zone at the start of the graph is explained by the presence of set of IP addresses belonging to load balanced proxy servers. Inspecting reverse DNS lookups of the first 189 indicates that many of the addresses preceding this point are indeed proxies. Inspecting reverse DNS lookups of a sample of addresses below the rank of 190 suggests that few are not genuine end user machines. Clearly this is never going to be a certainty, but it tends to confirm the Van De Sompel / Bollen theory, and the elimination

of requests from those first 189 IP addresses should eliminate the majority of proxy / cache noise from our sample.

If we accept the assumption that the 190 highest ranked IP addresses are proxies, and others are end user machines, then we are left with a sample of 1,296,061 article requests that originate from end user machines.

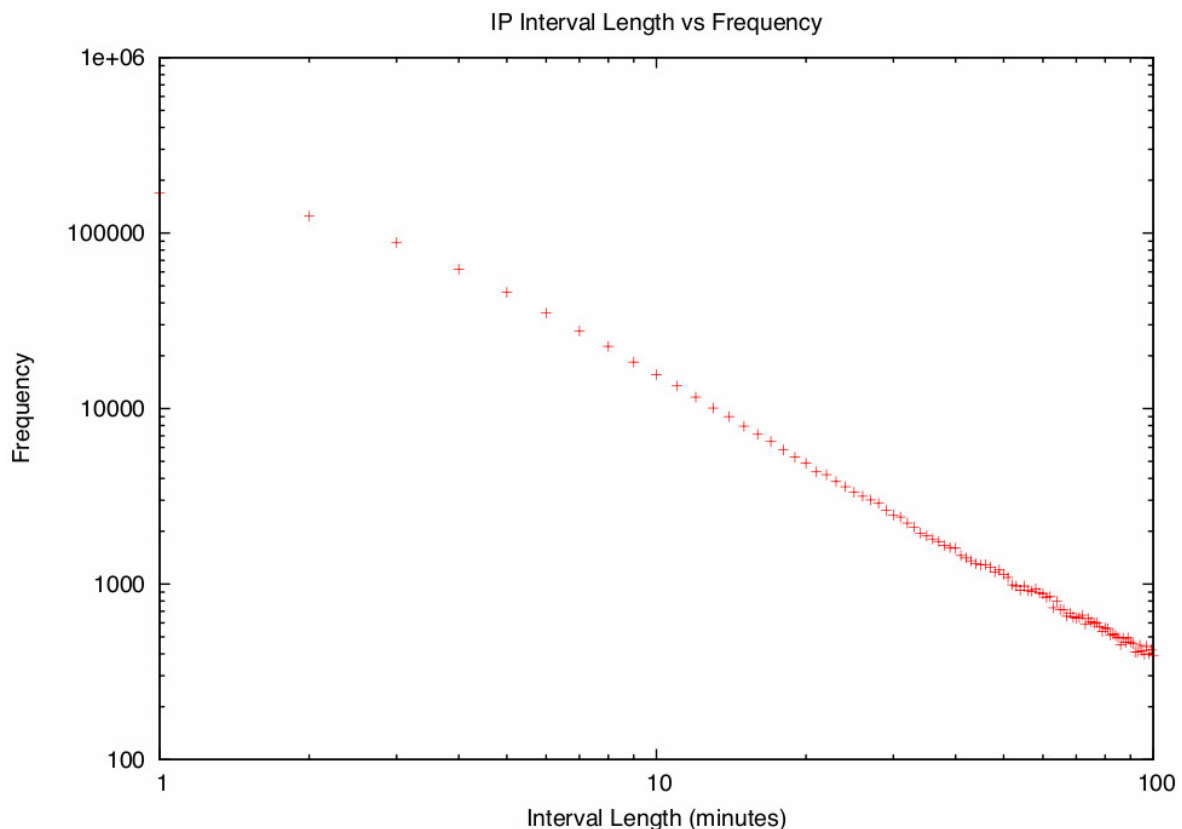
One possible approach is to assume each of the remaining IP addresses represents an individual person, i.e. that the machine is not shared. This treats the IP address as a user identifier. We are not in a position to judge whether this is a safe assumption, but we can propose an alternative approach that would avoid it; this involves identifying separate sessions of activity.

### Sessions analysis

An appropriate cluster analysis of requests should be able to identify sessions. It would be reasonable to assume that, in the majority of cases, two requests from the same IP address from a non-proxy machine that occur within one minute of one another would be from the same user. However, two such requests separated by one hour are not so clearly related. Somewhere between one minute and one hour lies a reasonable cut-off point at which it could be assumed that requests no longer belong to the same 'session'. Establishing a sensible place for such a cut-off point should be an aim of the log analysis.

For this analysis, we used the sample of 1,296,061 article requests originating from IP addresses that are assumed to be end user machines. This sample was separated into sets of requests from each individual IP addresses.

The first step is to establish a “reasonable cut-off point” for the interval between two requests, at which it can be assumed that they belong to separate “sessions”. We looked at all the request intervals in our data, and plotted their frequency against length in minutes (see Figure 2). The resolution of our log data is to the nearest minute, so intervals started at 0 minutes (meaning 0 to 59 seconds). A 1 minute interval means 60 to 119 seconds and so on.



**Figure 2:** The frequency of time intervals between requests received from individual IP addresses (excluding those believed to be proxies). Requests separated by short intervals are more common than those separated by longer intervals. The pattern does not suggest a “cut off point” that would separate sessions of user activity.

We would expect shorter intervals to occur more frequently than long intervals. The data reveal a steady decline in frequency as the length of the interval increases. A sudden dip might indicate a most common session abandonment time, or a suitable “cut off” point beneath which we might assume all requests are part of a single session of user activity. However, there are only the smallest of dips at around 50 and 70 minutes, which are inconclusive.

Plotting these data over an extended period of time shows the steady decline in interval length continues, followed by rise as we approach the 24 hour mark, then an oscillation on further 24 hour cycles. This suggests only that machines are used for similar periods each day.

### **Data available from sessions**

In the absence of an objective measurement of session lengths, we chose a period of 15 minutes as a cut off point. This is the timeout period that has been used for many years by EDINA discovery services to terminate idle user sessions, and which corresponds to typical user activity.

Using an arbitrary method to identify separate sessions is not very satisfactory, however we wished to examine the possible implications of basing an analysis on OpenURL usage data on separate short sessions of user activity. This is an important issue, because this approach avoids making the assumptions that IP addresses can serve as user identifiers; however the risk is that by using data derived only from short sessions, we will only be able to generate small samples of articles that can be associated with each other.

Using this technique, a “session” is identified from being two or more requests from the same IP address with no more than 15 minutes separating them. From the 809,233 unique article<sup>17</sup> requests we were able to identify 246,195 individual sessions, which contained a total of 612,410 unique articles.

In other words, 76% of all articles requested occur within a session during which at least one other article was requested. This suggests that an analysis based on sessions will provide a good sample of articles which can be associated with each other, on the basis of very reasonable assumptions that suggest they were requested by the same end user.

---

<sup>17</sup> For the purposes of this study, article uniqueness was based on distinct values of ISSN, volume, issue and starting page numbers. Articles lacking these metadata were discarded from the sample. These criteria might be refined (e.g. by investigating ways in which journal title could be used as an identifier) to provided a larger sample.

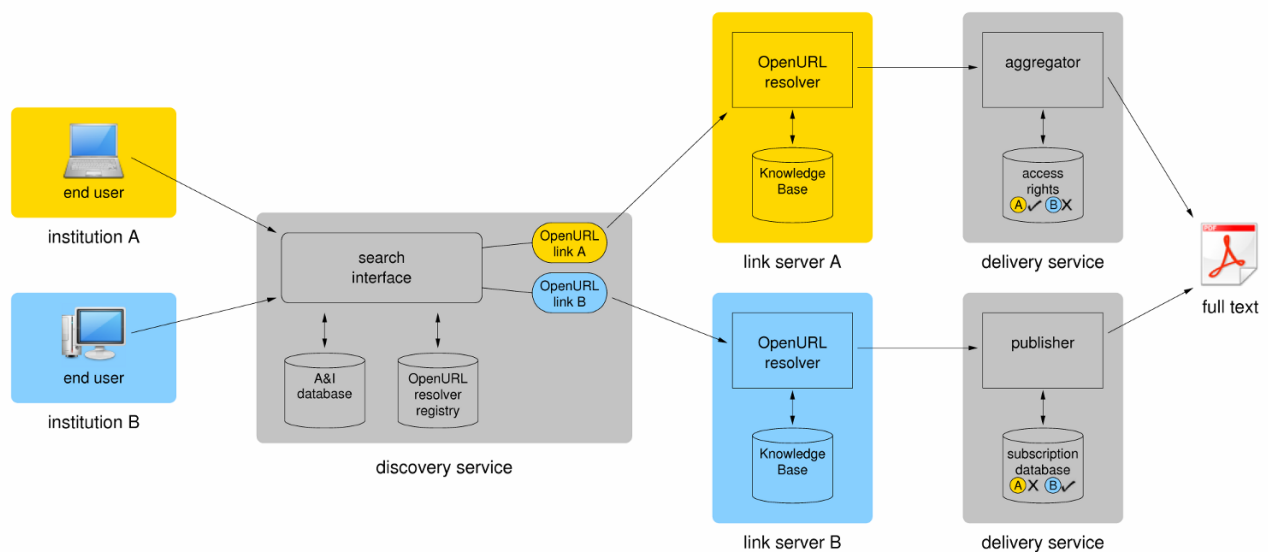
## Appendix 4: What is an OpenURL resolver?

The OpenURL framework for open reference linking in the networked information environment was developed by Herbert Van de Sompel at Ghent University in a project conducted between 1998 and 2000 (see Van de Sompel and Bein-Arie 2001). The OpenURL framework was designed to address the 'appropriate copy' problem. This is the fact that any single journal article (or other information object) for which descriptive metadata is available (typically, conventional citation details such as author, year, journal, volume, issue and page) may be available from many different service providers. In an electronic environment, the user desires a seamless link from a reference to the paper it describes. If the user is affiliated with a library, they may be authorized to access the paper from one or more services to which their institution subscribes. If a link directs them to a different service, access may be denied or payment may be requested.

The framework was declared a standard by the US National Standards Organisation NISO in 2004 (ANSI/NISO 2004). The NISO standard does not specify the function performed by an OpenURL Resolver, so the "appropriate copy" problem for which OpenURL was originally devised is in fact only one possible service; however at the current time the term "OpenURL Resolver" is commonly regarded as synonymous for a service that addresses the appropriate copy problem by attempting to locate articles that are freely accessible to the end user.

The OpenURL is generated by a referring service, which can be any source of a journal reference. The OpenURL must contain information about the article, and optionally may contain information about the referring service, the end user making the request, and the type of service being requested from the resolver. Of the optional elements, information regarding the referring service is the only one commonly included (though generally it is unlikely to be useful for anything other than logging). Information regarding the end user and service being requested is generally redundant, since for any institutional resolver that addresses the appropriate copy problem, the end user identity and desired function are implicit.

The OpenURL resolver parses the OpenURL to obtain information about the article, and consults a database (often called a "knowledgebase" or "KB") of journal full text providers and subscriptions to determine the best available source for the end user. As well as onward links to full text services, local library holdings records and other services (inter library loan, google searches) may be offered. Andy Powell provided a good description of the conventional OpenURL resolver designed to address the "appropriate copy problem in Ariadne, 2001: <http://www.ariadne.ac.uk/issue28/resolver/intro.html>.



### OpenURL linking as a solution to the appropriate copy problem.

The above diagram shows each end user, one in Institution A and the other in institution B, discovers a journal article in a discovery service, and wishes to access it. Their access rights differ, and each

must use a different route to access the full text. The discovery service provides context sensitive links, so each user is directed to an OpenURL resolver that will provide links to full text based on the subscriptions taken by their institutions.

## Appendix 5: What is the OpenURL router?

The OpenURL Router works by offering a central registry of institutions' OpenURL resolvers. Any UK HE or FE institution may register details of its resolver, free of charge. When the resolver has been registered, the OpenURL Router enables service providers to link users from that institution to their local resolver.

The aim of the OpenURL Router is

1. to enable OpenURL linking without the need for service providers to have prior knowledge of the resolver applications installed at their end users' institutions,
2. to avoid the need for librarians to register their resolver with all of the services to which they subscribe, and
3. to enable OpenURL linking from within services that do not require user authorisation: in these services, no information is held that identifies users or institutions, so without the OpenURL Router it would not be possible to link to institutions' OpenURL Resolvers without considerable overheads.

Service providers can use [openurl.ac.uk](http://openurl.ac.uk) as the "base URL" for all of their OpenURL links for UK HE and FE customers. Service providers can also make a "lookup" request to [openurl.ac.uk](http://openurl.ac.uk), and receive an XML response with details of a user's resolver. If an OpenURL aware service provider does not choose to use [openurl.ac.uk](http://openurl.ac.uk), their customers will of course still be obliged to configure links to their resolvers. In this case an institution can simply configure the links using [openurl.ac.uk](http://openurl.ac.uk) as the base URL. The OpenURL links in that service will direct users to their local resolver via the OpenURL Router. This will be transparent to users, but in the event of the resolver being changed the institution would then only need to make a single change to their configuration registered with [openurl.ac.uk](http://openurl.ac.uk).

All OpenURL requests made by end users via the Router at [openurl.ac.uk](http://openurl.ac.uk) are logged, and (subject to the metadata included by the referring service) provide a record of the article that user was attempting to find via their local resolver. The Router redirects these requests to the appropriate local resolver for each user, so though the Router logs each request, the ultimate outcome (whether the end user obtained a copy of the article, and from where) is unknown to the Router.

The OpenURL Router supports various types of requests other than links direct to local resolvers. These include the "lookup" requests (registry searches), requests for the preferred button image to be used for each resolver, and various browser redirects that may be required to identify an end user. These requests are all logged, but they are not OpenURL requests and do not contain bibliographic metadata. These requests are excluded from the analysis described in this report.

More details of the OpenURL Router are available at <http://openurl.ac.uk/doc/>.