Using OpenURL Activity Data: Legal and policy issues – initial investigation

EDINA, February 2011

This is a brief report on the initial exploration of legal and policy issues associated with a project which seeks to create an aggregation of data from the UK OpenURL Router Service and possibly institutions' OpenURL Resolver services (Appendix 1 explains what the OpenURL Router Service is and what an OpenURL Resolver is). The project was preceded, in 2009, by an investigation into the viability and value of such an aggregation and exploration of the types of services that may usefully be based on such an aggregation [1]. To build on this work, our first steps in this project were to:

- Explore further the legal and policy issues associated with collection and processing of activity data through the UK OpenURL Router service and the aggregation of data collected through institutions' OpenURL Resolver services.
- Determine whether it is viable to pursue aggregation of these data.

Legal issues often appear to be complex and potentially delay progress as issues are clarified and decisions made on that clarification. At the outset in this project, we sought to clarify whether the data being collected by the Router and by Institutions' OpenURL Resolver services would be considered personal data and, if so, what impact this would have on the proposed aggregation of activity data which would be made available through an API as the basis of services built by third parties. We were advised by the University of Edinburgh's legal advisors and staff in the Records Management department who have expertise and experience in dealing with information law – in this instance Data Protection. Although we document here our approach and findings, we do not offer legal advice. Any service seeking to build a similar service should seek legal advice on their own behalf.

Legal and policy issues

Investigation during our earlier project suggested that activity data generated by OpenURL resolvers or the Router are unlikely to attract IPR (unless substantial investment was made in creating a structured database from those data, in which case such database may attract database rights) so, as we understand it, the main issue to address relates to data protection; if the activity data collected include anything that may be construed as personal information, processing of those data must comply with Data Protection legislation. Although OpenURL Router and Resolver activity data tend not to include the name, ID or email address of an individual, they usually include IP addresses. In fact those IP addresses are important to this project as data in an aggregation are often useful only if they give some indication of the activity of an individual user within a discrete session (i.e. if such activity maybe inferred from the data). The IP address links the user to the activity.

The Information Commissioner advises that IP addresses may constitute personal data in terms of the Data Protection legislation if an individual user can be identified by using a combination of that IP address and other information which is either (i) held by the organisation collecting the data; or (ii) publicly accessible. This applies even when personal data are anonymised after collection as they are, nevertheless, personal data when collected. Clearly, IP addresses do not always constitute personal information as they identify a computer not an individual and the IP address may be dynamic rather than static; the static IP address of a computer used by a single individual, when combined with other information about that person, would be personal data while the IP address of a computer in an Internet café would not. However, the Information Commissioner advises "When

[1] http://edina.ac.uk/projects/Shared_OpenURL_Data_Infrastructure_Investigation_summary.html

you cannot tell whether you are collecting information about a particular person, it is good practice to treat all the information collected as though it were personal data …"[2]

We have explored the steps that must be taken in order to collect the data required for the aggregation without exposing the collecting organisation to legal risk through breach of data protection legislation. We sought to review the status of IP addresses, whether their collection and processing must be disclosed and how such disclosure may be made.

We sought to further explore the legal and policy issues through dialogue with the university's legal advisor and staff within the University's Records Management Department who have expertise in data protection. We also included questions in our questionnaire (see below) to elicit the respondents' understanding of the privacy issues related to the activity data collected in their institutions.

Our advisors confirmed that IP addresses may constitute personal data if the organisation holds or can easily obtain information that links these addresses to an individual. Online information service providers are likely to have information about individuals that may be linked with IP addresses. An email, for example, indicates the email address of the sender, and often includes the IP address in the Internet header information. So, if a service provider receives an email from an individual (in whatever context) and that individual also uses a service being provided by that same service provider, which is collecting IP addresses, the latter may be deemed personal data, the processing of which must be consented to by the data subject. The implications of this are far reaching. It suggests that online information service providers can hardly avoid collecting personal data as IP addresses and other personal information are routinely communicated through use of the web and email. Clearly, in most instances, there is no intent to 'process' data (in the ordinary sense of the word). However, as the definition of 'processing' in the Data Protection legislation is wide enough to include deletion, this is difficult to avoid. The law has been overtaken by technology. Our aim is to stay within the spirit of the law by protecting the privacy of individuals whilst operating an online service and thus minimising the risk to the service provider.

IP addresses and the Router

The Router currently generates activity data, i.e. when a user clicks on a link to view a paper and their request is routed through the Router, the Router logs capture the IP address of the computer making the request along with the information in the request e.g. the OpenURL of the paper that has been requested. These are not currently collected and processed but the service envisaged here would involve anonymising IP addresses and creating an aggregation both of which constitute processing. We have been advised to err on the side of caution by seeking to ensure that collection and processing of activity data is disclosed to users. As the OpenURL Router service is middleware with no direct relationship with end users, we cannot expect end users to be aware that they are using an EDINA service and that EDINA is potentially processing their personal data (subject to the EDINA privacy policy). Instead, disclosure of this activity would have to be made by the institution whose end-users are routed through the Router, e.g. in the institution or library's privacy policy. There are currently 90 institutions registered with the Router.

Our advisor suggested that EDINA post notification on the page where institutions register to use the Router which tells the institution that EDINA collects activity data and intends to include them in an anonymised form in such an aggregation and tells those institutions that they should advise

[2] Information Commissioner's Office (2010) *Personal Information Online: Code of Practice,* http://www.ico.gov.uk/upload/documents/library/data_protection/detailed_specialist_guides/personal_information_online_cop.pdf *(accessed 14 February 2011).*

their users of this activity (e.g. through their privacy policy). It would be important to offer the option for institutions to opt out of the aggregation. The 90 institutions that have already registered should be advised of the aggregation by email and also given the option to opt out. Data from those registered institutions which opt out will be excluded from the aggregation. (Identifying data related to a specific institution and deleting it constitutes processing of data but, again, this would be done to avoid inclusion in an aggregation and to protect the wishes of the institution with regard to data that may be construed as personal data. Thus, we seek to minimise the risk that privacy is breached). Clearly, data generated before this process of disclosure is undertaken cannot be included in the aggregation.

To determine institutions' willingness to participate in an aggregation of data from their resolvers or to engage in advising their users about aggregation of data from the Router, EDINA created an online questionnaire to investigate:

- whether libraries are currently collecting activity data through their OpenURL resolvers;
- whether and how they disclose this collection (and possibly processing) of data to their users;
- which vendor's resolver those libraries use; what data are being collected; and
- whether libraries subscribed to the UK OpenURL Router service would be willing to advise their users that the Router collects activity data (the questions are attached as appendix 1).

The questionnaire was publicised through the JISCmail email discussion lists lis-e-resources and lis-link, it was tweeted by Andy McGregor, the JISC Programme Manager and the questions were sent directly to those who participated in an earlier, related study (with the option to respond by email or to complete the questionnaire online). We anticipated that the response rate would be low as it was difficult, in 2009, to find people who were actively collecting activity data through their OpenURL resolvers. Twenty six people responded to the survey but only 15 of these responses were useable. (We deleted those that contained no information and those that contained only the name of the institution and the title of the respondent.)

Questionnaire responses

A more detailed account of the questionnaire responses is given in Appendix 2. In brief, most of the respondents to the questionnaire are using either SFX (8) or WebBridge (5). We hoped that among the respondents would be staff in libraries who would be willing to work with us to explore the process of contributing to an aggregation – i.e. that some of them would contribute their data during the second phase of this project. However, although most of those collecting activity data said that they would be willing to contribute to an aggregation or discuss the possibility with their colleagues, only three of those provided contact details. Nevertheless, these responses are encouraging as a significant majority of those collecting data are willing to consider contributing to an aggregation.

Most respondents collecting activity data believe that it doesn't include personal data although a few indicated that the data include IP addresses. (It is not currently clear whether and how these data may be useful if they do not include IP addresses. This requires further clarification). None of the respondents currently advises users that they collect these data but several would do so or would investigate the necessity to do so if they were to contribute the data to an aggregation. Similarly, those respondents who are registered with the Router would either be willing to advise their users that they do so or would be willing to discuss this possibility with their colleagues.

Next steps

We were concerned, when we began this project that the status of IP addresses as personal data may thwart the project; a requirement to disclose collection of those data to end users may act as a barrier to progress. We have been advised that collection and processing of these data should be disclosed to end users. As those users are routed through the Router by virtue of their institution's registration, it seems most appropriate that the institution disclose to users that these data are collected and processed. This requirement adds to the work of this project. Before proceeding, it was important for us to determine whether institutions would be willing to disclose this to their users. It seems that they would consider doing so.

The response to our online questionnaire was limited, as anticipated. Nevertheless, it was encouraging as respondents are willing to consider contributing data from their OpenURL resolver services to an aggregation and those who are registered with the Router are willing to consider disclosing to their users (e.g. through a privacy policy) that the Router collects and processes activity data. On the basis of this, we conclude that it is worthwhile pursing aggregation of data.

Our next steps will be:

(1) to approach institutions which we believe are collecting activity data that are sufficiently rich to be useful as part of an aggregation (i.e. that include citation data as a minimum) and invite them to work with us to explore the process of creating an aggregation. SFX users are collecting such data. We are currently seeking to determine whether those institutions using WebBridge are collecting such data.

(2) to approach those institutions that are registered with the Router and ask them to consider disclosing to their users that the Router is collecting and aggregating activity data from the Router, anonymising it and making the anonymised data available through an API for third-parties to build services based on those data.

Appendix 1: What is an OpenURL resolver and what is the UK OpenURL Router?

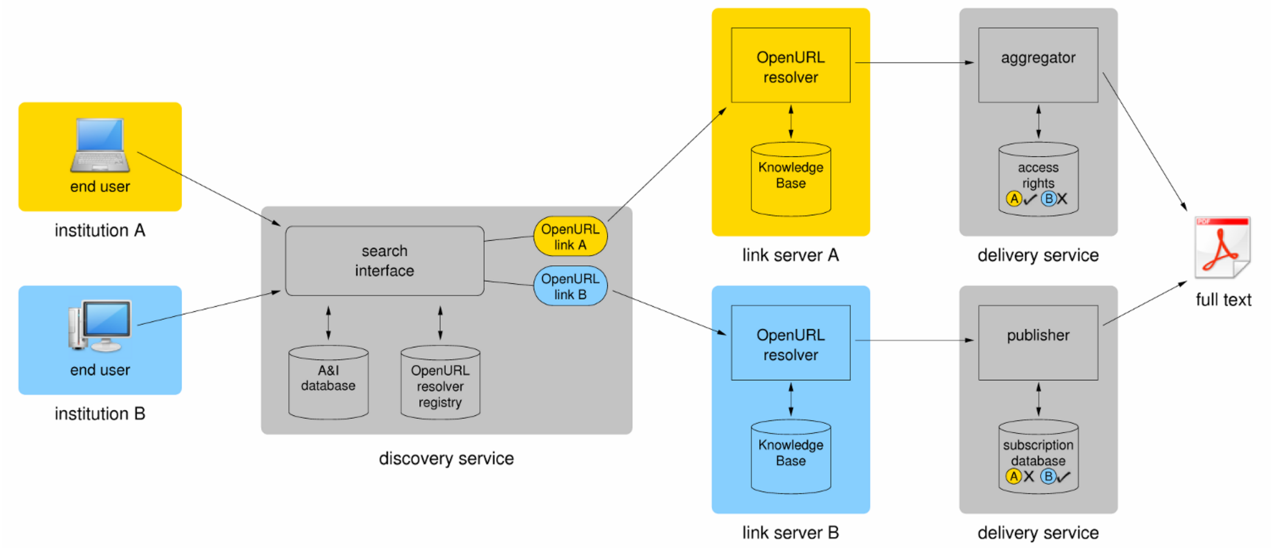
OpenURL Resolver

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.

OpenURL Router

The OpenURL Router works by offering a central registry of institutions' OpenURL resolvers. Any UK HE or FE institution may registers 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 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, 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, 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 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.

All OpenURL requests made by end users via the Router at 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 requiests 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/.

Appendix 2: Questionnaire responses

Which resolvers are used, and who is collecting what data

The OpenURL resolvers used by respondents: 5 are using WebBridge, 8 SFX, 1 Sirsi, and 1 360 Article Linker.

One respondent indicated that although registered with the Router, the institution does not actively use it. However, they would be interested in aggregated data so this would be an incentive to look at how they may contribute data.

Finding institutions willing to explore contribution to an aggregation

We hoped that among the respondents would be staff in libraries who would be willing to work with us to explore the process of contributing to an aggregation – i.e. that some of them would contribute their data during the second phase of this project. Nine respondents are collecting activity data through their OpenURL resolvers and eight of those said that they would either be willing to contribute to an aggregation (with incentives and/or support; two) or would be willing to discuss this possibility with their colleagues (six). However, of the eight people who provided their contact details, only three are collecting activity data: two are SFX users and the other is using WebBridge (Two of the eight respondents who provided contact details do not know if they are collecting activity data. One of these is using 360 Article Linker and the other SFX). Nevertheless, these responses are encouraging as a significant majority of those collecting data are willing to consider contributing to an aggregation.

Personal data; IP addresses

Of the 8 who responded to the question 'Do the data contain any information that may identify an individual user' 7 answered 'no' and 1 ‘yes’. When asked if those data include IP addresses, 3 answered 'yes', 4 answered 'no' and 1 'I don't know'. When asked whether they tell their users that they are collecting activity data, 7 said 'no' and 1 'I don't know'. If they were to contribute to an aggregation of activity data, 4 respondents do not know whether they would advise their users that they do so, 3 would do so (although one of them noted that this would be unnecessary 'if they were not identifiable'.), and one would ‘explore their obligations’. It is not clear from the questionnaire responses how data from a discrete session are identified if neither IP addresses nor any other personal information are collected. It would be important that this be clarified if institutions are to contribute data to an aggregation.

Which resolvers are collecting activity data?

Our scoping study in 2009 suggested that the only resolver product collecting data that is sufficiently rich to be useful as the basis of services (e.g. because it includes citation data) was SFX. The information gathered during this study does not change that conclusion.

Respondents who are registered with the Router

Of the 15 respondents providing useable responses, 13 indicated whether or not they are registered with the UK OpenURL Router service: 9 are registered and 4 are not. All who are registered with the Router would be willing to advise their users that the Router collects and processes activity data (2) or would be willing to discuss this possibility with their colleagues (7).

Additional points made by respondents

One respondent reiterated a point made in our earlier study, namely that their institution would be more likely to participate if participation among UK HEIs were high. A high participation level would increase the value of an aggregation. With other services this participants has also found that high participation provides a community of people who can help to iron out any difficulties that may arise (although there was no reference to difficulties that may arise in this context).