The JISC Information Environment and Google

A discussion paper

Andy Powell
UKOLN, University of Bath
November 2004

1 Executive summary

1.1 This discussion paper compares the different approaches to resource discovery adopted by the JISC Information Environment (JISC IE) and by Google (and the other major Internet search engines)[1] and makes the following recommendations to JISC and the UK HE/FE community:

1.1.1 The JISC IE and Google are complementary rather than alternative approaches to resource discovery. The UK HE/FE community should be encouraged to develop and use resource discovery services that take advantage of the best of both approaches.

1.1.2 All JISC IE 'content providers' should make their content (or at least an abstract page[2] for their content) available in a form suitable for indexing by Google, in addition to supporting the JISC IE standards and protocols. The JISC development services should develop a set of best-practice guidelines for ensuring that Google effectively indexes JISC IE content and that it is possible to reliably and persistently link to that content.

1.1.3 JISC should enter into discussions with Google about allowing services within UK HE/FE to use the Google Web services and to access the Google PageRank score for all the resources it indexes.

1.1.4 JISC should enter into discussions with Google about special treatment for the high quality content in all UK institutional repositories (eprint archives, learning object repositories, etc.), along the same lines as the current DSpace/Google/OCLC discussions. These discussions should also aim to ensure that appropriate UK HE/FE content is made available through the Google Scholar service.

1.1.5 JISC should enter into discussions with Google, Amazon, OCLC and academic journal publishers to ensure that where Google indexes the abstracts and other Web pages for books and journal articles of interest to the academic community, those pages contain an OpenURL *button* (thus providing a link to institutional context-sensitive resolution services).

2 Introduction

2.1 The JISC Information Environment technical architecture (see figure 1) lays out a vision of the UK HE/FE resource discovery information landscape in terms of a high-level set of 'service components'. Each component is a fairly complex service (i.e. one that offers many facets) such as a 'portal', an 'aggregator', a 'content provider', etc.

2.2 The JISC IE encourages the providers of such service components to expose some of their functionality on the network in machine-oriented ways. This allows service components to interoperate with each other in order to provide a more seamless 'resource discovery' experience for the end-user. So, for example, rather than visit the Web site for each content provider of interest, the

[1] In most of this discussion paper, "Google" is used as shorthand for "Google (the search engine at www.google.com) and the other major Web search engines".

[2] The phrase 'abstract page' is used here to indicate a Web page that contains a description of (or some other surrogate for) the full content - an abstract of a journal article, a thumbnail of an image, a catalogue record for a book, etc.

end-user will be able to enter a single query into a JISC IE 'portal' of some kind[3], in order that a cross-search of multiple content providers can be performed.

2.3 The main types of interoperability envisaged by the JISC IE include distributed searching, metadata harvesting, news-feeds and context sensitive linking using protocols and standards such as Z39.50 (and SRW), the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), the OpenURL and RSS.

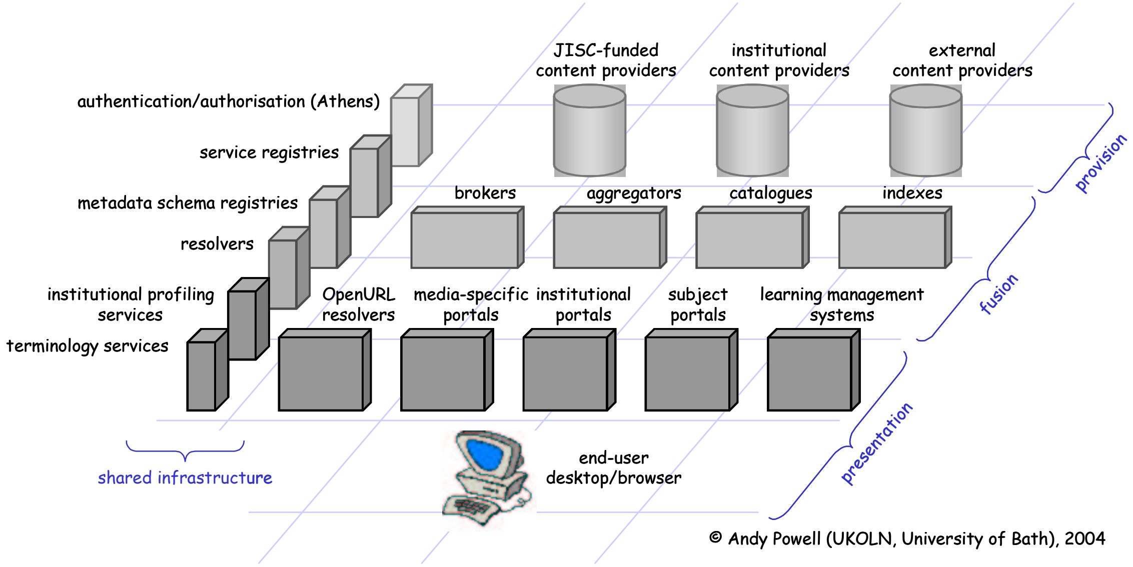
JISC-funded content providers
institutional content providers
external content providers
authentication/authorisation (Athens)
service registries
metadata schema registries
resolvers
institutional profiling services
terminology services
brokers
aggregators
catalogues
indexes
OpenURL resolvers
media-specific portals
institutional portals
subject portals
learning management systems
provision
fusion
presentation
shared infrastructure
end-user desktop/browser
© Andy Powell (UKOLN, University of Bath), 2004

Figure 1 – The JISC IE technical architecture

2.4 This short discussion paper compares the JISC IE approach with that adopted by Google and makes some recommendations for how the community might align these activities.

3 Comparing Google and the JISC IE

3.1 Many institutions are now making use of the JISC IE standards and protocols through the adoption of "library" portal and OpenURL resolver products (for example, using the Ex Libris MetaLib and SFX products), through the deployment of institutional repositories (for example, using DSpace) and through the integration of search technologies and RSS channels into their institutional portals and learning management systems (for example, using uPortal and Sentient Discover).

3.2 However, there is also some evidence that Google is meeting many of the resource discovery needs of students, and to a lesser extent staff, within UK HE/FE and that end-users are voting with their feet (or mice!) in favour of these kinds of services. As Lorcan Dempsey has noted in his regular column for CILIP Update, "increasingly 'on web' means available in Google" and therefore content and service providers need to turn their "systems inside out, and try to make the functionality that was available only within monolithic search systems now available on the open web"[4].

3.3 Google operates by harvesting the full-text of a significant proportion of all available Web pages using software programs commonly referred to as Web robots or spiders[5]. The retrieved full-text is indexed and made available for searching through the Google home page, through toolbars embedded into

[3] The phrase "JISC IE 'portal'" is used here to refer to any application or service that uses the JISC IE standards and protocols to support resource discovery, whether it is delivered through an online Web presence or a locally installed desktop application or by some other means.

[4] Dempsey, Lorcan, *The Three Stages of Library Search* <http://www.cilip.org.uk/publications/updatemagazine/archive/archive2004/november/lorcan.htm>

[5] This process will be hereafter referred to as 'spidering'.

Web browsers, via machine-oriented Web services and, more recently, within a personal search tool that can be installed on the end-users desktop.

3.4 Clearly, Google can only spider Web pages that are publicly available – for example, content that is held within institutional intranets or behind publisher's access-controls is typically not available to search engine robots. However, as discussed in the next section, Google has recognised that it needs to get its robots more deeply into this 'invisible Web'.

3.5 Before going further, it is worth highlighting the key differences between a Google search and a JISC IE 'portal' cross-search.

3.6 When searching Google, the end-user's search terms are matched against the full-text of all the Web pages that have been spidered by the Google robots and a ranked list of pages (and/or images) is returned.

3.7 As noted above, this means that a significant proportion of the high quality content available to the UK HE/FE community does not feature in the Google search results (typically because that content is hidden behind the publisher's authentication controls).

3.8 The recall[6] of a Google search is typically quite high (at least for the content that Google has been able to index) because the term only has to appear once, anywhere, on the page in order for a 'hit' to occur. On the other hand, the precision[7] is often quite low. Fielded searching isn't supported (at least so far as that term is understood by the library community) – it isn't possible to use Google to do a *title* or *author* search for example[8].

3.9 Because precision is low, the ranking algorithms employed by Google are critical to its success. Like all search engine providers, Google is very protective of its ranking algorithms. However, what we do know is that they are partly based around a technique known as PageRank. In this technique, Web pages that have lots of other pages linking to them are assigned a higher rank – furthermore, if links to a page are from pages that are themselves highly ranked, those links are given additional weight. This relatively simple algorithm gives a surprisingly good method of ranking Google's search results and it appears to be a fairly commonly held view amongst end-users that the *right* results nearly always appear in the first page (or first few pages) of results from Google.

3.10 In short, the PageRank algorithm uses links between Web pages as a simple, open and highly democratic method of determining the 'quality' of a Web page, though it must be said that this is largely measured in terms of popularity rather than any academic or educational notion of quality.

3.11 Of course, all ranking algorithms are open to abuse, particularly the practice of spamming search engines (using embedded keywords and fake links) to achieve a higher than expected ranking in search results. Google must constantly update its algorithms in order to try and stay one step ahead of the spammers.

3.12 It is also worth noting that end-users are often not encouraged to form deep-links direct to the high quality academic content made available by publishers, either because the URIs for that content are not persistent or because publisher prefers links to be made to their home page. This means that even in cases where that content is available for indexing by Google, the PageRank assigned to it will remain quite low, thus hiding high quality material towards the end of long lists of search results.

3.13 Searches initiated from JISC IE portals typically do not make use of full-text. Instead, the queries are passed across the network in parallel to multiple search 'targets' where the terms are typically matched against metadata records about the content held by a particular content provider. Results are returned in the form of lists of metadata records.

[6] The 'recall' of a resource discovery service is an indication of the percentage of the items in a body of material that would satisfy a request that are actually found by a search.

[7] The 'precision' of a resource discovery service is an indication of the percentage of hits found by a search that satisfy the request that generated the query.

[8] Google does support 'advanced' searches which allow the end user to restrict search results by language, date, format, links, etc. but it does not support 'fielded' searching in the sense of 'metadata fields' such as title and author.

3.14 Although, in most cases, an ordered list of results is returned, the ranking algorithms adopted by each target are different and limited to knowledge about the metadata records and content within a single content provider. The portal must merge together the lists of metadata records returned by each target to form a single list of search results.

3.15 Some features of this approach are worth noting. Firstly, sending queries across the network and waiting for results to be returned is a time consuming process, typically much slower than querying the same database held locally. In cases where multiple targets are being searched, the performance of the overall search is limited by the performance of the worst performing target.

3.16 Secondly, the quality of search results is highly dependent on the quality and consistency of the metadata records in all the collections being searched. Although precision is typically good (fielded searches are possible for example), in some cases the recall may be limited because the end-user has used different search terms than those anticipated by the metadata creator (i.e. the cataloguer).

3.17 Thirdly, the task of merging together several individually ranked lists to provide a single overall ranked list of results is a very difficult one. Many of the current portal systems choose not to do this – instead, they simply present the lists to the end-user separately, often in order of the speed they were returned across the network or alphabetically by content provider name[9]. Unlike the case of Google described above, there is no widely accepted mechanism for determining the relevant 'quality' of individual results across multiple content providers.

3.18 There are also various hybrid solutions, where metadata records are harvested from multiple content providers into centralised search services. This is most obviously typified by the use of the OAI-PMH, a protocol commonly adopted by eprint archives and learning object repositories for making copies of all their metadata records available for harvesting.

3.19 Although a metadata harvesting approach gets over the performance problems of distributed searching and allows a more consistent approach to ranking the metadata records from multiple collections, it does nothing to overcome the problems caused by mismatches between the search terms chosen by the end-user and the terms used in the metadata records created by the cataloguer. Furthermore, because ranking algorithms can only be based on the available metadata records, they tend to be rather primitive in nature. In response to this, some OAI-PMH-based search services (e.g. ePrints UK) are attempting to gather both metadata records and full-text in order to allow the end-user to carry out both fielded, metadata-based searches and full-text-based searches at the same time. This approach has the potential to deliver better recall without significantly compromising the precision of the results, provided that the full-text is available for indexing.

3.20 The resource discovery approaches adopted by the JISC IE and by Google are complementary rather than alternatives. The Google Web index is available for searching via a Web service and can thus be integrated into JISC IE cross-searching approaches[10]. However, it must be noted that Google limits the use of its Web services to 'personal use' and attempts to enforce this by restricting the number of accesses from any given service to a daily maximum. Usage of the Google Web services from JISC IE 'portals' may therefore break the Google terms of use.

3.21 If JISC could reach agreements with Google such that the services being developed within UK HE/FE were allowed to access the Google Web services on an unlimited, non-personal use basis, then those services could take a more innovative approach to resource discovery building on the best features of both approaches. In addition, if UK HE/FE services had access to the Google PageRank score for any particular item of content, then it might be possible to perform ranking across multiple content providers more effectively.

[9] Note that some portal software uses sophisticated 'post-coordinated' techniques for grouping sub-sets of the returned results together by topic or other criteria. In some cases this can provide a very effective view of the returned results.

[10] Note that 'indexes' appear in the fusion layer of the JISC IE architecture diagram above.

4 Googling the invisible

4.1 As mentioned earlier, Google and the other big Web search engines are now making increased efforts to index those parts of the Web that they have previously been unable to reach. Some examples of this kind of activity are listed here.

4.2 **Google and academic publishers.** At the UKSG 2004 conference earlier this year, the presentations by both Cathy Gordon of Google and Geoffrey Bilder of Ingenta[11] hinted very strongly that Google were talking to many of the major academic publishers about how they can index their content (or abstract pages that describe their content). The issues associated with this activity are non-trivial since Google presumably want assurances that they are not indexing 'blind alleys' – i.e. that end-users will not simply be presented with a pay-per-view page when they click on a Google search result. Google seem to be asserting that, at the very least, the publisher makes an abstract page for the book or article freely available to all end-users.

4.3 **DSpace/Google/OCLC.** In April this year, Google, MIT and OCLC announced a joint initiative to ensure that the contents of institutional repositories based on the DSpace software is available to Google for indexing. The pilot project currently covers 17 institutions, including Cranfield University in the UK. While there is no theoretical requirement that this initiative should be restricted to those repositories based on the DSpace software, it currently is limited in that way (which possibly demonstrates the benefits of having a strong software 'brand'). Although it is difficult to be sure of the technical details of this collaboration, it seems unlikely that it is based on any formal harvesting of metadata records – rather, it is more likely to be based on spidering and indexing the full-text of the eprints using the repository's human-oriented Web interface.

4.4 **Yahoo/OAIster**. In a similar move to the one above, Yahoo and OAIster (a global OAI search-engine) have teamed up in order to provide Yahoo's search engine with access to all repositories worldwide that support the OAI-PMH.

4.5 **OpenWorldCat.** In another collaboration with Google, OCLC have initiated a pilot project to expose a sub-set of metadata records from the OCLC WorldCat service[12] to Google for indexing. This activity allows end-users to discover library catalogue records through Google. Clicking on the search result for a book takes the end-user through to an OCLC HTML page for the catalogue record. The page contains links to various services related to the book, including a link to the OCLC member libraries that hold a copy of it in the local area of the end-user. In this way, the end-user is moved fairly seamlessly from discovery to delivery. It would also be possible for the page to contain an OpenURL *button*, supporting onward linking to a context-sensitive OpenURL resolver within an end-user's institution. While it would be technically possible for UK HE/FE libraries and union catalogue services like COPAC to similarly expose their catalogue records for indexing by Google, this approach might be self-defeating, since it would pollute the Google index with multiple catalogue records for each book[13]. It would be better for the UK HE/FE community to collaborate with OCLC and Google to ensure that there is sufficient coverage (but little duplication) in the catalogue records exposed to Google and that appropriate links (using the OpenURL) are embedded within those catalogue records to ensure that, having discovered a catalogue record, end-users can navigate to an appropriate book 'delivery' service that meets their needs.

4.6 **Google Scholar.** The potential impact of these kinds of initiatives is clearly demonstrated by the recent announcement of the Google Scholar service. This beta service "*enables you to search specifically for scholarly literature, including peer-reviewed papers, theses, books, preprints, abstracts and technical reports from all broad areas of research*"[14]. It applies Google's spidering,

[11] Bilder, Geoffrey, *Beyond Google: The Web is Changing Behind Your Back* <http://www.uksg.org/presentations4/bilder.pdf>

[12] OCLC WorldCat is a union catalogue of OCLC member libraries.

[13] Note that individual libraries should certainly consider exposing catalogue records for any unique or rare items that they hold in order that Google can index them.

[14] Google Scholar <http://scholar.google.com/>

full-text indexing and ranking techniques to a focussed collection of high-quality academic content, augmented by a level of citation analysis. The service appears to include material from institutional Web sites, eprint archives and less formal online journals as well as peer-reviewed literature and books. There is therefore a clear opportunity for UK HE/FE to expose its high-quality institutional content to Google Scholar for indexing, as well as benefiting from the service as end-users.

5 An 'academic Google'?

5.1 Google are now marketing their Search Engine Appliance to the UK HE/FE sector, primarily for use within institutional intranets. It might therefore be possible to purchase the Google software for use on a national basis, using it to spider and index everything of interest to UK HE/FE in the context of the JISC IE. Alternatively, it may be possible to negotiate a deal with Google such that they host a special instance of their search engine, targeted specifically at the needs of the UK HE/FE community.

5.2 This kind of approach has been adopted by the BBC, which has worked in the past with Google and now works with Yahoo to index both their own content and a set of external Web sites that they consider to be of value to the end-users of their services.

5.3 However, any attempt to apply such an approach in the context of the JISC IE will face significant 'collection development' issues – the problem of defining the subset of all information globally that is relevant to end-users within UK HE/FE. One possible course of action would be to use (or build on) the RDN[15] catalogue of high-quality Internet resources and the collections and Web sites listed in the JISC IE Service Registry as a set of starting points for the Google robots. However, this is unlikely to be sufficiently comprehensive to meet all end-user needs. As a result, and because there is no single Web site associated with the JISC IE, a national 'academic Google' may find it difficult or impossible to convince end-users to move away from using the standard Google search engine.

5.4 Furthermore, because the pricing model for the Google Search Engine Appliance tends to be based on the number of links being indexed, adopting such an approach on such a large scale may be prohibitively expensive.

5.5 Of course, building a UK-specific Web search engine does not need to be based on the Google software. There are many 'products' that could be used, ranging from homegrown or open source solutions to high-end knowledge management systems like Autonomy. But the real problem is that users want the content they are interested in to be surfaced in the service or services they already use to undertake resource discovery. Expecting users to change their resource discovery habits and move to using a 'special' UK academic search engine is probably unrealistic.

6 Is the JISC IE still relevant?

6.1 This paper might be interpreted as painting a pessimistic view of the future of the JISC IE. However, there are significant areas of overlap in the Google and JISC IE approaches and there are some very successful applications of the JISC IE standards and protocols so it would be prudent not to jump to too hasty a conclusion. More importantly, there is a significant amount of high quality academic content which is not yet available through Google but which is currently available in forms suitable for discovery via JISC IE 'portals'.

6.2 At the same time, there are now various digital library activities around the world that are complimentary with the IE (in some cases building directly on the work funded by JISC). There is a continuing, and global, groundswell of activity around the more 'structured' approaches used within the JISC IE.

6.3 A good example of this is the NISO Metasearch Initiative, an activity led largely by library system vendors. This activity is specifying the standards by which library 'portals' can more effectively search the high-quality content being made available by publishers. The JISC IE technical architecture, and work more generally going on within the context of the JISC IE, has provided a very solid basis for some of this activity.

[15] Resource Discovery Network <http://www.rdn.ac.uk/>

6.4 Other examples of global activities that are interoperable with the JISC IE include the 'open access' movement and the adoption of the OAI-PMH across the education and cultural heritage sectors worldwide, the IMS Digital Repositories Interoperability Specification (which promotes the use of many of the same standards used within the JISC IE for use within learning management systems[16]) and the fairly rapid and widespread take-up of RSS and the OpenURL.

6.5 The Google and JISC IE approaches tend to have strengths in different areas, not just because of the differences in coverage but also because of the different search functionality that is available. Google is less good for performing academic searches that are intended to be comprehensive of a subject area. Nor can it easily be used to highlight the *seminal* academic texts for a particular topic. JISC IE 'portals' are more likely to be comprehensive of the academic content in a particular subject area, but may tend to miss some of the more informal content currently.

6.6 The underlying fielded searches in the JISC IE approach can also be used to support browse interfaces, with points in the browse hierarchy being implemented as 'saved searches'. It would not be easy to use 'saved searches' against Google in this way.

6.7 Finally, it is worth remembering that JISC IE 'portals' tend to be built around particular communities of end-users (subject-based, media-based, institutionally-based or whatever). This is one of their strengths because it allows their user-interfaces and functionality to be specifically targeted to meet a particular community's set of needs. It also means that the collection of resources presented to those communities can be built and managed in a coherent way.

6.8 With particular reference to the UK HE/FE sector, developing portals around particular communities means that the pedagogic needs of end-users can be taken into consideration when building these kinds of services. For example, a 'subject portal' may be specifically designed to support the end-user's educational objectives by allowing them to learn about and understand the range of literature that is available in a given subject area.

6.9 Taking all this into account, there seems to be little doubt (at least in the opinion of the author) that the JISC IE is still very relevant (both within the UK and globally) but that end-users within UK HE/FE will be best served if resource discovery tools and services can be built that make the most effective use of the complimentary approaches offered by Google and the JISC IE. The challenge for JISC and the community is how best to take advantage of and combine these approaches.

7 Conclusions and recommendations

7.1 The JISC IE and Google are complementary rather than alternative approaches to resource discovery. While it is not possible to be sure about how resource discovery services will develop in the future, it does seem clear that neither approach in isolation will fully meet the requirements of all resource discovery scenarios. The UK HE/FE community should be encouraged to develop resource discovery solutions that use the best of both approaches.

7.2 All JISC IE service components that make content available with the intention that it may be freely used by anyone, should make that content available in a form suitable for indexing by Google (in addition to supporting the JISC IE standards and protocols). In practice this means designing robot-friendly Web sites, having persistent URIs and domain names, ensuring important text is text and not images, etc. The JISC development services should develop a set of best-practice guidelines for ensuring effective indexing by Google.

7.3 All JISC IE service components that make content available but where there are restrictions in place controlling who has access rights to that content, should make an abstract (or metadata record or some other surrogate) for that content available in a form suitable for indexing by Google (in addition to supporting the JISC IE standards and protocols). The JISC development services should develop a set of best-practice guidelines for ensuring effective indexing by Google.

[16] This paper has focused on the relationship between Google and the JISC IE. However, much of what is said here can also be applied to the E-Learning Framework and the Virtual Research Environment activities being undertaken by JISC, since the services related to resource discovery are 'common' across those three areas of work.

7.4 The OCLC 'collections grid'[17] provides a useful mechanism for considering the classes of content that need to be exposed to Google. JISC will need to work with publishers in order to encourage them to expose Web pages about their licensed published material in some way. Institutions should consider exposing Web pages about the items in the unique or rare collections that they hold: 'special collections' in libraries; e-theses and eprint archives; databases of digitised slides, posters, and other materials; research data repositories; departmental Web sites; learning object repositories; etc. These issues need to be considered irrespective of whether the content is hosted within the institution or by some external service (a JISC-funded national repository for example). Similarly, academic libraries could expose Web pages about their general book and journal holdings, but the community needs to avoid a many to many situation in which lots of libraries expose Web pages about the same content. Avoiding this will require collaboration between JISC, the UK HE/FE community, OCLC, Google, journal publishers and Amazon. Finally, although Google is good at surfacing content that is already on the Web, this may be usefully augmented in some way by building on the descriptions of high-quality Internet resources made available by the RDN (though the author makes no firm technical proposals for how such a collaboration might take place).

7.5 These cross-party discussions should also attempt to ensure that the Web pages for books, journal articles and other resources of interest to the academic community being indexed by Google contain an OpenURL *button* if appropriate (thus providing a link to institutional context-sensitive resolution services). The most effective way of doing this will probably be to take advantage of the UK 'OpenURL router' service[18].

7.6 All JISC IE service components should issue guidance for how to generate persistent hypertext links to the resources that they make available and should encourage end-users to take advantage of such links. This will improve the effectiveness of search engine ranking algorithms, like the Google PageRank algorithm, that are based on numbers of links to a given resource.

7.7 JISC should enter into discussions with Google about allowing services within UK HE/FE to use its Web services in an unrestricted way and to access the Google PageRank score for all the resources it indexes.

7.8 JISC should enter into discussions with Google about special treatment for the high quality resources in all UK institutional repositories (such as those listed above) along the same lines as the current DSpace/Google/OCLC discussions. These discussions should also aim to ensure that appropriate UK HE/FE content is made available through the Google Scholar service.

Acknowledgements

The author would like to thank Lorcan Dempsey, Brian Kelly, Rachel Heery, Paul Miller and Rachel Bruce for commenting on previous versions of this paper (though this should not be taken as implying that they agree with all the views expressed here). Any errors or omissions remain the sole responsibility of the author.

UKOLN is funded by the MLA: The Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

[17] Dempsey, Lorcan, *Libraries, digital libraries and digital library research* (slide 17) <http://www.ecdl2004.org/presentations/dempsey/>

[18] The OpenURL Router <http://openurl.ac.uk/doc/>