The JISC Resource Discovery Landscape

A personal reflection on the JISC Information Environment and related activities

Andy Powell, UKOLN, University of Bath

May 2005

UKOLN is funded by 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.

1 Introduction

This document provides an opinion piece, looking at the way resource discovery technologies and services are being deployed across the UK HE/FE community. It considers some of the issues being raised by the JISC IE technical architecture, by our current approaches to technologies within that space and by the external trends we see happening around us.

The document represents only the views of the author, and should be treated as such. The author's credentials for providing such an opinion piece include: being one of the authors of the original DNER architecture study; advising JISC and the community about the JISC IE technical architecture and maintaining the documentation associated with it; being a member of the OAI-PMH technical committee, the DCMI Advisory Board and being current chair of the DC-Architecture Working Group; being a member of the OCLC Research Advisory Committee and the British Library eIS Technical Advisory Panel; being a member of the DLF Framework Working Group; and being a member of the UK eGovernment Metadata Technical Working Group. The author is also very familiar with the CIE activity, the People's Network Service being developed by the MLA and the NELH.

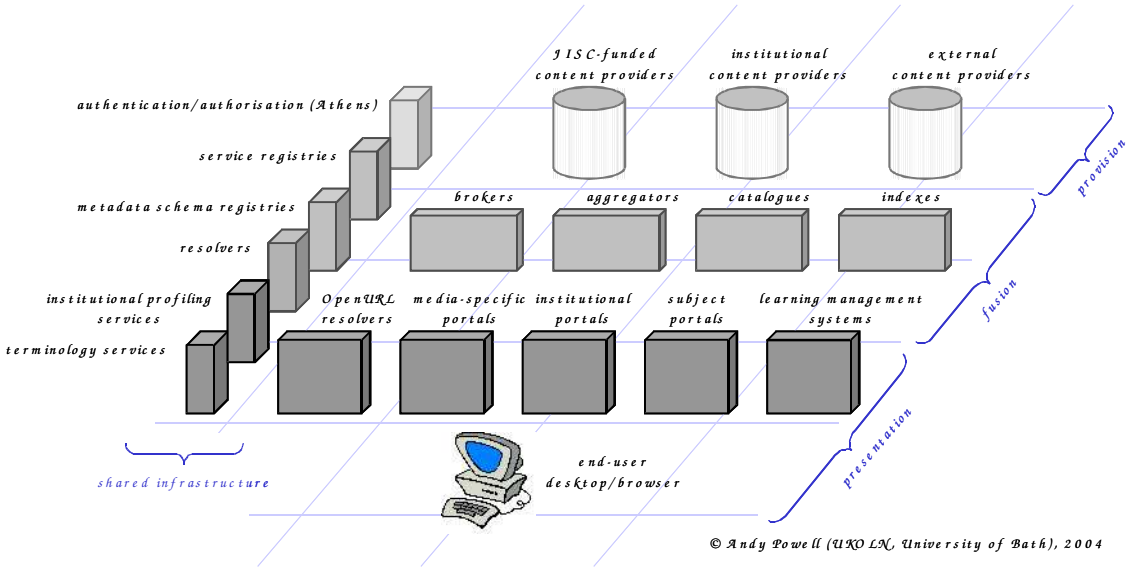

Figure 1 - The JISC IE architecture diagram

The JISC IE architecture diagram (see figure 1) will be used to frame this discussion. The JISC IE 'layers' (provision, fusion, presentation and shared infrastructure) will be analysed in turn and a summary of issues produced for each of them. Overall, the intention is

- to validate (or otherwise) the technical approaches adopted by the JISC IE, checking that they are in line with international trends in both the public and commercial sectors,
- to question the effectiveness of these standards and protocols,
- to review whether there are other technologies that might sensibly be adopted into the JISC IE, and
- to highlight those areas where JISC might usefully fund additional activities.

Service components within the JISC IE (portals, brokers, content providers, etc.) interoperate using a set of relatively mature, well defined, open international standards as described in the

JISC IE Standards document[1]. The standards provide support for cross searching (Z39.50 and SRW), metadata harvesting (OAI-PMH), alerting/news (RSS), metadata (Dublin Core and IEEE LOM), and context-sensitive linking (OpenURL). None of these standards is specific to the UK. By adopting international standards within the JISC IE, service components can benefit from a global approach that includes a wide variety of other national and international resource discovery initiatives. By adopting open standards, services can work with each other flexibly without needing to broker a potentially large number of bilateral agreements.

This document is intended to be read as a companion to the earlier "The JISC Information Environment and Google" discussion paper[2].

2 Summary of recommendations

1. The JISC community needs to work to ensure that the service oriented approaches being adopted by initiatives like ELF, the DLF framework, SAKAI, VIEWS, etc. will use the same conceptual frameworks and terminology as far as possible.
2. Engagement with the JISC IE by the commercial sector and other players is extremely valuable to the community and we should take care not to lose this important buy-in to our shared activities as we move forward with a more service-oriented approach.
3. The community needs to increase its investment in automated approaches to metadata creation and automated approaches to indexing and data-mining full-text and multimedia resources. We also need to remember that there will probably always be scenarios for which manually created metadata will be the most appropriate solution.
4. The JISC community needs to maintain good links with the Semantic Web Best Practice and Development Group and with other key players, particularly in the areas of metadata schema registries and terminology services.
5. Development of services that support community-driven approaches to building terminologies – so-called folksonomies – are worthy of consideration for JISC-funding. Evaluation of these kinds of approaches would also be useful to the community.
6. JISC should encourage the community to experiment with peer-to-peer (P2P) approaches (within single institutions, between a limited number of institutions and nationally) in order to gain some experience of their strengths and weaknesses.
7. The community needs to develop a typology of 'repositories' (eprint archives, institutional repositories, learning object repositories, content management systems, etc.) in order to understand their differences and similarities and in particular how to enter into appropriate dialogue with the commercial sector about the supply of software to deliver them.
8. The JISC community needs to collaborate internationally on the modelling of 'complex objects' and their packaging using standards such as METS, MPEG-21 DIDL and IMS C/P. Furthermore, the community needs to build an infrastructure that provides a

[1] JISC IE Standards <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/>

[2] The JISC Information Environment and Google: a discussion paper <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/ie-google/>

coherent view across disparate repositories in order to prevent individual service providers having to replicate significant pieces of knowledge engineering.

9. The JISC community needs guidance about how best to expose the content in repositories to search engines like Google, whilst at the same time also investing in more structured disclosure approaches such as those based on metadata harvesting and cross searching.
10. The JISC community should work with selected content providers and end-users in order to undertake some appropriate research into the effectiveness of exposing full-text to Google and metadata to metasearch engines and the end-user benefits that such exposure brings.
11. The JISC community should work with the NISO Metasearch Initiative and appropriate elearning partners to evaluate the use of the A9 OpenSearch specification, SQI and other similar specifications as alternatives to Z39.50 and SRW/SRU.
12. JISC IE content providers need best-practice guidance for how to assign relatively persistent 'http' URIs to their resources and on when it is sensible to buy into alternative identification systems such as the DOI.
13. JISC should work with the providers of union catalogue services to investigate their use as points of contact with Google, both as places where metadata records can be exposed and as places where knowledge of physical and electronic holdings information can be disclosed. However, this work should not be undertaken unilaterally within the UK.
14. The community needs to see more development undertaken in the area of automated indexing and data mining of full-text and other content types. JISC should work to ensure that there are appropriate links in place where institutions are deploying full-text indexing techniques, e.g. in the provision of a university's Web-site search engine, with other institutional activities such as the development of eprint archives and/orinstitutional repositories.
15. In order to support seamless resource discovery approaches across the content of repositories and to support personalised 'views' of this content, JISC should investigate the benefits of developing a national federated architecture for repositories, in tandem with similar national initiatives elsewhere as appropriate. This would include agreeing common solutions to a variety of technical challenges such as the assignment and resolution of identifiers, the use of complex object packaging standards and the provision of format conversion tools.
16. The community needs to refine its performance measures for purely machine-oriented services such as those found in the fusion layer.
17. The community needs to balance the focus on 'portals' as Web-based services with a focus on the most effective mix of desktop and Web-based tools and services (both machine-oriented and human-oriented) that can be used to meet end-users' functional requirements.
18. The JISC community should contribute to the development of a global OpenURL resolver 'routing' service in order to encourage and streamline the deployment of OpenURLs on a very wide global scale.

19. The community should ensure that appropriate authentication, authorisation, and trust mechanisms are in place to support the potentially complex relationships between end-users, institutions, shared services, fusion layer services and content providers.
20. The community should attempt to reach agreements internationally about how to deploy distributed 'service registries' - including agreements on metadata standards and transport protocols. We also need to agree on the operational policies for service registries and the ownership and IPR issues associated with the metadata records being exchanged.
21. The JISC community needs to undertake more work in the area of mapping metadata schema and related services, looking particularly at the issues of mapping between Semantic Web and non-Semantic Web schemas. The JISC community also needs to consider setting up a registry of 'packaging profiles'.
22. The JISC community should continue to contribute to international discussions about the use of identifiers and the services associated with them.
23. As a community we need to refine our understanding about the best ways that our ontologies can be created and maintained and the kinds of services that we require on those ontologies. We also need to ensure that best-practice guidelines are developed for assigning identifiers to terms in the vocabularies (e.g. URIs) and for marking-up the vocabularies in machine-readable forms.
24. The JISC community should work towards building a licence registry (or registries) to encourage a consistent approach to the deployment and use of 'open access' licences.
25. The JISC community should encourage the development of automated metadata-creation tools and should deploy them as Web Services so that they can be embedded into presentation layer (and other) tools and services.
26. In order to deliver name authority services, JISC should work with various parties, including the BL, to determine if an 'authority list' of journal article author names exists or can be created, and if so to layer Web services in front of it. Alternatively, the JISC community could consider options for delivering a distributed 'name authority' service through a network of institutional (LDAP) servers.

3 General issues

This section addresses some general issues that affect all aspects of the JISC IE architecture.

3.1 Service Oriented Architectures

There is a growing tendency to adopt a 'service oriented' approach when considering architectures like the JISC IE. This is no bad thing, since it is very much in line with trends within the commercial software sector and within other international initiatives such as Sakai[3]. However, it should also be remembered that adopting a service-oriented approach is not a 'quick win'. In general, our communities are not familiar with this kind of approach and it will take time for them to become comfortable with it. Perhaps more importantly, even amongst those people who are well equipped to adopt this approach there are differences of

[3] Sakai <http://www.sakaiproject.org/>

opinion about the best conceptual model to use as a framework for forming our reference models.

The 'framework' needs to support a service-oriented approach to developing and delivering learning, research and management information systems within the academic community. Such an approach maximises the flexibility with which systems can be deployed, both in an institutional context and nationally. In particular, such an approach supports the integration of library management systems, virtual learning environments and management information systems. The framework allows the community to document its business requirements and business processes in a coherent way and to use these to derive a set of interoperable network services that conform to appropriate open standards. By documenting requirements, processes, services, protocol bindings and standards in the form of 'reference models' members of the community are better able to collaborate on the development of service components that meet their needs (both within the community and with commercial and other international partners). The 'framework' also functions as a strategic planning tool for the JISC and similar bodies internationally.

These are non-trivial issues to address. Currently different communities are addressing them in different ways, using different vocabularies and conceptual models. The current 'wall of bricks' offered by initiatives such as the E-Learning Framework and the Virtual Research Environment[4] is an initial step towards a services oriented approach that will be useful for the community. However, this approach suffers a little at the moment from a lack of clarity over how the abstract services are 'factored out' of the current landscape. Attempts to partition the wall of bricks by domain (e-learning, e-research, etc.) don't work well, partly because so many of the services are cross-domain in nature and partly because that approach does not help to show what broader functional requirement is being met.

[4] The E-Learning Framework <http://www.elframework.org/>

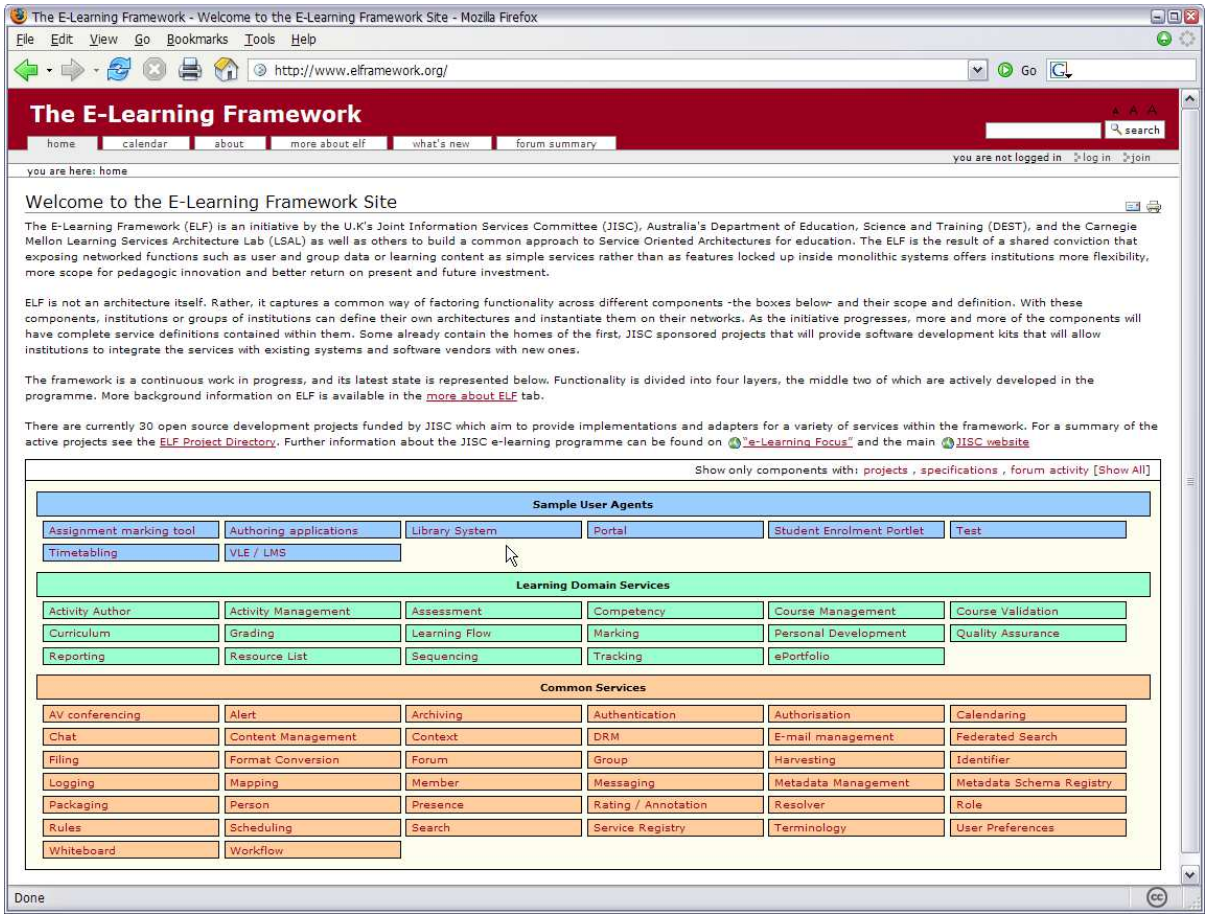

Figure 2 - the E-Learning Framework

The framework being developed by the DLF[5] is interesting in this context, since it provides some structure for the 'wall of bricks' based on an analysis of the 'business requirements' and 'business processes' that drive our service delivery needs. It also recognises that understanding the 'services' is not sufficient. We need to marry our abstract services with the primary 'business entities' (people, organisations, collections, items and so on) to which the services relate.

Ideally, the service oriented approaches being adopted by initiatives like ELF, the DLF framework, SAKAI, VIEWS[6], etc. will use the same conceptual frameworks and terminology and the community needs to work together to ensure that this happens as far as possible. Otherwise, there is a danger of the people engaged in our different service oriented approaches talking past each other.

It is also worth remembering that a set of 'abstract services' (which is a good example of term that is used differently by different communities currently, but it is being used here to mean a service definition that is not bound to a particular protocol or API) does not, on its own, provide true interoperability, other than at a fairly high 'abstract' level. Two service components that support the same 'abstract service' but do so by adopting different protocols or APIs do not interoperate. There is a very real danger that a service oriented approach that is flexible about what protocol and API 'bindings' should be used (for example that allows

[5] DLF Abstract Service Framework Working Group Wiki <http://wip.dublincore.org/dlfwiki/>

[6] VIEWS: Vendor Initiative for Enabling Web Services <http://www.niso.org/committees/VIEWS/VIEWS-info.html>

different services to make different choices about specific protocols) allows the community to feel like it is building interoperable services, while actually resulting in little or no useful interoperability in a practical sense.

Finally, it is worth noting that there is no fundamental problem in recasting the JISC IE architecture using a 'service oriented' approach. The JISC IE is inherently service oriented - though that term was never used during its development.

3.2 Maturity of the JISC IE

The JISC IE technical architecture is now reasonably mature, both in terms of the stability of the standards and protocols it adopts and in terms of the deployment of real service components on the network. It is now quite possible to point to all the boxes on the JISC IE architectural diagram and enumerate a list of one or more services that instantiate that particular box. The architecture is now easily recognisable by players in the commercial arena that have an interest in the 'discovery to delivery' space covered by the JISC IE technical architecture. This is extremely valuable to the community and we should take care not to lose this important buy-in to our shared activities as we move forward with a more service-oriented approach.

It is worth remembering that the JISC IE architecture has invented nothing. Rather, it has brought together a set of existing standards and protocols and uses them to deliver a coherent approach to resource discovery. There is no sense in which any of the standards and protocols adopted within the JISC IE technical architecture are UK-only, with the possible exception of ATHENS – though with the gradual move to a Shibboleth-based approach, even this peculiarity will be lost. This means that the commercial sector do not need to adopt a special "UK approach" in order to meet our needs. The same tools that are sold to us will be sold to other communities around the world.

3.3 Manual vs. Automated approaches

It is probably fair to say that the resource discovery approaches adopted in the JISC IE primarily focus around manually created metadata. This was discussed in more detail in the previous "The JISC Information Environment and Google" discussion paper. While these kinds of approaches remain valid in some contexts, it is clear that the community must continually question the cost-effectiveness of manual approaches against alternative automated approaches to resource discovery. In particular, the scalability of manual approaches to metadata creation in the context of the Internet is clearly a concern. This is made worse by a growing recognition that end-users do not necessarily do metadata well – because they are not cataloguers. This is particularly true for more complex metadata – such as that required to describe the pedagogic aspects of a learning resource. Creating metadata by hand is expensive, particularly if it is done well enough to support useful resource discovery services.

The community needs to spend more time investing in automated approaches to metadata creation and automated approaches to indexing and data-mining full-text and multimedia resources. However, we need to recognise that this is an area where there is still much real research to be done and where commercial tools are not cheap. In particular, it should be noted that although a significant amount of work has been done over the last 20 years or so on information retrieval and text indexing algorithms, much less work has been done on

automated metadata generation. We also need to remember that there will probably always be scenarios for which manually created metadata will be the most appropriate solution.

3.4 Semantic Web

There are currently few overlaps between the JISC IE and the W3C's Semantic Web activity[7], even though the two share some common aims. The exceptions to this are the use of the RDF Schema language to describe metadata schemas in the JISC IE Metadata Schema Registry and the recommended use of RSS version 1.0 (the RDF variant of RSS) to expose alerting/news channels.

Generally speaking, most of the interesting work going on in the area of ontology creation and management is happing in the Semantic Web context, notably the recent SKOS work[8], though there are many interesting challenges still to address – for example, how best to identify terms within an ontology. However, the fairly recent announcements about VDEX[9] (within IMS) and the Z39.19 'thesaurus standard'[10] (by NISO) means that there is also ongoing activity in this area outside of that group.

The JISC community needs to maintain good links with the Semantic Web Best Practice and Development Group[11] and with other key players, particularly in the areas of metadata schema registries and terminology services.

3.5 Community-led approaches

Given the scalability issues with manual metadata creation, particularly in those cases where cataloguers are explicitly funded to create metadata records using a fairly centralised model, it seems sensible to look to community-led activities to do some of the manual work for us. Examples of such approaches include the Open Directory Project[12], the del.icio.us shared book-marking service[13] and the recently announced Connotea service from Nature Publishing[14].

The critical factor in making such services seems to be keeping things simple (in terms of what information the end-user has to supply) and making sure that the resulting service offers something of real value to the end-user concerned. Of course, the creation of most terminologies is community driven to some extent. What seems to characterise these newer approaches is their informality, the use of Web technologies to provide a simple user-interface, and a relatively low level of editorial control.

---

[7] W3C Semantic Web <http://www.w3.org/2001/sw/>

[8] SKOS <http://www.w3.org/2004/02/skos/>

[9] Vocabulary Definition Exchange <http://www.imsglobal.org/vdex/>

[10] NISO Ballot Announcement: Z39.19-200X <http://www.niso.org/standards/balloting.html>

[11] Semantic Web Best Practices and Deployment Working Group <http://www.w3.org/2001/sw/BestPractices/>

[12] Open Directory Project <http://dmoz.org/>

[13] del.icio.us social bookmarks <http://del.icio.us/>

[14] Connotea beta <http://www.connotea.org/>

Evaluation of these community-driven approaches to building terminologies – so-called folksonomies[15] – and the development of tools and services that support their creation and maintenance are worthy of consideration for JISC-funding.

3.6 Peer-to-peer approaches

The JISC IE has tended to focus on approaches to resource discovery based around cross searching and harvesting metadata from the fairly 'formalised' repositories of content offered by institutions, commercial providers and other organisations. This has led to a somewhat rigid client-server architecture - for example, portal-to-content provider or broker-to-content provider.

An alternative model is to move towards a less rigid peer-to-peer architecture, where there is a lowering of the barrier to being considered a 'content provider' and where services discover each other in a somewhat more hap-hazard way. In such a model, every desktop machine has the potential to become both a 'content provider' and a 'portal'. Clearly there are non-technical issues to do with setting 'acceptable-use' and other policies, enabling long-term preservation of materials and so on in such a de-centralised architecture.

There are some examples where a peer-to-peer approach is being adopted. For example, the Mellon-funded LionShare initiative at Penn State University and elsewhere[16]. JISC should encourage the community to experiment with these kinds of approaches (within single institutions, between a limited number of institutions and nationally) in order to gain some experience of their strengths and weaknesses.

4 Provision

This section considers some of the issues facing service components in the provision layer of the JISC IE architectural diagram. Readers are referred to the longer and more detailed Digital Repositories Review study recently completed by Heery and Anderson[17] for a more thorough treatment of some of the services in this layer.

4.1 When is a repository not a repository?

The study referred to above provides us with a reasonably narrow definition of repository:

- *content is deposited in a repository, whether by the content creator, owner or third party on their behalf*
- *the repository architecture manages content as well as metadata*
- *the repository offers a minimum set of basic services e.g. put, get, search, access control*
- *the repository must be sustainable and trusted, well-supported and well-managed*

However, real-world usage of that term tends to be quite broad, particularly in the elearning sector. For example, the term repository can be applied to collections of metadata records only, i.e. the kind of service that has tended to be called a 'catalogue' until recently. As was

[15] Wikipedia entry for 'folksonomy' <http://en.wikipedia.org/wiki/Folksonomy>

[16] LionShare <http://lionshare.its.psu.edu/main/>

[17] Digital Repositories Review <http://www.jisc.ac.uk/uploaded_documents/rep-review-final-20050220.pdf>

the case with the word 'portal', such broad and unrestrained usage tends to devalue the term to the point where it becomes more or less useless.

This is further compounded by confusion within the institutional context about the differences between an 'institutional repository', an 'eprint archive', a 'learning object repository', a 'content management system' and a variety of other terms in common usage. Institutions are in danger of investing in an infrastructure that includes a number of components, each of which does more or less the same job, but with a different content type.

The community needs to understand the differences and similarities between these components and in particular how to enter into appropriate dialogue with the commercial sector about the supply of software to deliver them. For example, a software vendor selling a 'content management system' may not understand that their product can also be used as a 'repository' and may therefore not respond appropriately to an invitation to tender.

4.2 Handling complex objects

The problems of exposing simple item-level metadata in the form of simple DC are now reasonably well understood. Such an approach forces 'service providers', services that harvest the simple metadata records, to work quite hard in order to provide a coherent view to end-users of the simple metadata they have to deal with. As described in the final technical report of the ePrints UK project[18], in some cases, for example in the case of reliably linking to the full-text of an eprint, the task becomes almost impossible. The community needs to agree how to model the complexities of the objects it wishes to expose metadata about rather better than it does now. Furthermore, we need to agree how to expose complex objects, containing both the metadata and the content of the item itself, structured according to the models agreed above, using formats such as MPEG-21 DIDL, METS and IMS Content Packaging Specification.

In the case of eprints, such modelling needs to build on the work done within the Functional Requirements for Bibliographic Records work undertaken by IFLA[19]. However, we also need models for all the other kinds of content that we expect to be made available in 'repositories', particularly learning objects and datasets.

One of the problems with the current suite of packaging specifications is the lack of any semantic markup within the package, in order to indicate what underlying conceptual model has been used. Unless we can get very widespread agreement (i.e. international agreement) about the packaging models that we are going to use, there is a danger that we will not achieve very much interoperability between services that expose and use packages based on different models.

The JISC community needs to collaborate internationally on the modelling of 'complex objects' and their packaging using standards such as METS, MPEG-21 DIDL and IMS C/P. Furthermore, the community needs to build an infrastructure that provides a coherent view across disparate repositories in order to prevent individual service providers having to replicate significant pieces of knowledge engineering in order to understand the heterogeneous repositories space.

---

[18] ePrints UK - Technical Documents <http://www.rdn.ac.uk/projects/eprints-uk/docs/technical/>

[19] FRBR <http://www.ifla.org/VII/s13/wgfrbr/wgfrbr.htm>

4.3 Metasearch vs. full-text indexing

Content providers are currently faced with choices about whether to expose metadata about the items that they make available for harvesting or searching and also whether to expose the full-text (or other forms of multimedia content) of items for indexing by other services.

This is not an either/or choice, since it is possible to expose metadata and full-text and to support both a search interface and a harvesting interface! However, in practice services may feel reluctant to invest in all these possibilities. They may therefore feel forced to make choices about which options they support and which they do not. Of course, in practice some metadata will be required to support the management of resources, so content providers may find that they need to create and manage metadata internally, even if they choose not to expose it to support resource discovery.

In the age of Google Scholar, the option of exposing full-text selectively to Google (and other search engine) robots may appear to offer the fastest and most cost effective payback for both commercial and non-commercial content providers, in terms of bringing paying and non-paying customers to their content.

However, as the investment in metasearch engines, such as Ex Libris Metalib, continues to grow within institutions, pressure on content providers to support a machine interface to their metadata records is also likely to increase.

The JISC community needs guidance about how best to expose the content in repositories to search engines like Google, whilst at the same time also investing in more structured disclosure approaches such as those based on metadata harvesting and cross searching.

It might be helpful for both the community and commercial content providers who sell into our community, for JISC to work with selected content providers and end-users in order to undertake some appropriate research into the effectiveness of exposing full-text to Google and metadata to metasearch engines and the end-user benefits that such exposure brings.

4.4 Complex vs. simple search interfaces

Over the last few years we have seen a general trend towards a simplification of search interfaces from complex standards such as Z39.50, through simpler revisions of it in the form of SRW and SRU leading ultimately to very simple proposed standards such as the Amazon A9 OpenSearch specification[20]. On the way we have seen the development of SOAP-based search interfaces that are offered by Google and Amazon. While these appear to be popular, it is probably possible to argue that the use of SOAP for these interfaces makes them more complicated than they need to be.

The experience of library portal software vendors is that the majority of targets still have to be searched using custom HTTP and XML interfaces rather than by using one of the standard interfaces listed above.

The development of the A9 OpenSearch interface, building on a simple set of HTTP CGI parameters and the RSS specification, is interesting in this context since it offers something that may well be simple enough for the vast majority of 'target' services to support. This approach is currently being considered by the NISO Metasearch initiative as a simple alternative to more complex approaches like Z39.50.

[20] A9 OpenSearch <http://opensearch.a9.com/>

There are also related initiatives within the e-learning community, notably the Simple Query Interface (SQI)[21].

The JISC community should work with the NISO Metasearch Initiative and appropriate elearning partners to evaluate the use of the A9 OpenSearch specification and SQI as alternatives to Z39.50 and SRW/SRU.

4.5 Identifiers for stuff

The importance of content providers assigning persistent, stable identifiers to the stuff they are exposing (both objects and metadata) cannot be overstated. Of all the aspects of the JISC IE technical architecture, this is probably the one that has been least vigorously emphasised, yet it is probably the thing that has most direct impact on the value of an 'information environment'. Without a reliable and persistent way of identifying and linking to the resources they discover, end-users will find it difficult to cite those resources in the context of their research and learning activities. For example, it will be difficult for a lecturer to build an online reading list.

Assigning persistent identifiers to stuff is a non-trivial task that requires forward planning in terms of how Web servers are delivered, how information resources are organised and a commitment to not changing underlying technologies in ways that have a drastic impact on how people can link to resources.

However, there are few clear-cut 'right' answers in this area. It can be argued that all the identifiers used in the JISC IE should be URIs[22] (because the URI is the only global identifier framework currently available that has any chance of being persistent). Furthermore, it can also be argued that the chosen forms of URI should conform to registered URI schemes (because this is the only way to achieve unique identifiers). However, even this level of guidance is met with disagreement in some forums.

Assigning identifiers is even harder in cases where the resource being identified is not exposed as an discrete thing on the network - for example where the resource is a person, an organisation, a 'collection', a concept and so on. Assigning URIs (even 'http' URIs) to these kinds of things is perfectly feasible and acceptable but there is little available guidance for how to do it currently.

For most content providers, the 'http' URI (commonly called the URL) will be used as an identifier for the majority of their resources. While identifier approaches based on HTTP redirects (such as the PURL) have some benefits in terms of persistence and are valuable in some contexts (for example, when identifying metadata terms), their use for information resources and other first-class objects should be treated cautiously, since the HTTP redirect may interfere with ranking algorithms such as Google PageRank.

Overall, content providers need best-practice guidance for how to assign relatively persistent 'http' URIs to their resources and on when it is sensible to buy into alternative identification systems such as the DOI[23].

---

[21] SQI Specification < http://nm.wu-wien.ac.at/e-learning/interoperability/sqi/sqi.pdf>

[22] URIs, URLs, and URNs: Clarifications and Recommendations 1.0 <http://www.w3.org/TR/uri-clarification/>

[23] Digital Object Identifier <http://www.doi.org/>

5 Fusion

This section considers some of the issues facing service components in the fusion layer of the JISC IE architectural diagram.

5.1 Union catalogues and Google

Issues around how content providers should best interact with Google and similar indexing services has already been touched on above. The various union catalogues available in the JISC IE (COPAC, the ePrints UK database, SunCat, etc.) are a natural place to consider in terms of both 'discovery' and 'delivery' in this context.

Firstly, union catalogues present a natural place for disclosing information to 'discovery' services such as Google. Secondly, they offer a place through which users can be brought back (or at least a source of information for services through which users can be brought back), before passing them to local 'delivery' services such as their institutional OpenURL resolver.

It is better to encourage the union catalogue services to do this than to get the individual contributing catalogues to do it, since on the discovery side of the equation, the union catalogue can de-duplicate a significant number of records before exposing them to Google (thus not polluting the Google indexes with multiple records about the same resource). Similarly, the knowledge about holdings collated in union catalogues, provides an ideal source of knowledge (a 'knowledge-base') on which to base 'appropriate copy' decisions.

JISC should work with the providers of union catalogue services to investigate their use as points of contact with Google, both as places where metadata records can be exposed and as places where knowledge of holdings-information can be disclosed. However, this work should not be undertaken unilaterally within the UK. Consideration should be given to working with other union catalogue providers internationally, for example OCLC.

5.2 Indexing and data mining

As discussed above, the community needs to see more development undertaken in the area of automated indexing and data mining of full-text and other content types. These are non-trivial issues in which much research still has to be done. It is therefore arguable whether such R&D activity falls within JISC's remit or elsewhere.

However, JISC should work to ensure that there are appropriate links in place where institutions are deploying full-text indexing techniques, e.g. in the provision of a university's Web-site search engine, with other institutional activities such as the development of eprint archives and/orinstitutional repositories.

5.3 Hiding the complexity of the provision layer

As the number of institutional and other repositories grows we will see a growing complexity in the landscape that portals and other presentation layer services have to deal with. This complexity is not just in terms of the number of 'targets' with which services have to interact, but in terms of the complexity and variety of the objects being disclosed by those services.

Services in the fusion layer, particularly aggregators, can significantly help hide that complexity from presentation layer services. Approaches such as the aDORe architecture

currently being deployed within LANL[24] can be used to transform a complex repository landscape into a much more coherent resource. The end-user needs to interact with multiple repositories (whether the repositories are within one institution or many), to search across multiple resource types, to view sub-sets of repositories, to link between deposited resources such as journal articles and data-sets. By ensuring a common approach, the JISC IE can offer such a unified view of repository content. Ensuring that identifiers are assigned in a consistent way, that those identifiers can be resolved, that complex objects are described in a consistent way, and that format transformations are provided will contribute to a coherent 'fusion' infrastructure that will help to support the development of a significant range of 'presentation' layer services, including metadata and full-text indexing resource discovery services and preservation services along the lines of LOCKSS[25].

5.4 Performance measurement

Services positioned in the fusion layer are conspicuous by their invisibility. Sitting between service components in the presentation layer and content providers in the provision layer, successful fusion services need present no visible user-interface to end-users.

Yet being invisible in the context of JISC-funded services is not an enviable position, since the primary mechanisms that we have for assessing the performance of services involves counting the numbers of eyeballs on Web pages.

Take for example the Resource Discovery Network 'catalogues'. In theory, these could function very effectively by becoming an invisible source of metadata records, surfaced only through third-party services. RDN records could be made freely available to institutional library catalogues, to library portals, to commercial 'portals' like Web of Knowledge, to Learning Management Systems like Blackboard and WebCT, to reading list services like Sentient Discover and so on. Such an approach might mean that more people would see and benefit from RDN records. However, at the same time it would reduce the number of hits on RDN Web sites to near zero. There is little incentive for the RDN to take this approach, since it would be difficult for them to demonstrate the benefits it brings to end-users in any measurable way.

There is a related problem in terms of how end-users are expected to perceive an 'invisible' service. Fusion layer services rely on end-users understanding the 'value proposition' of the aggregations of resources and metadata being offered. It is difficult to badge the quality of an invisible component.

Therefore, for a variety of reasons, services in the fusion layer are unlikely to ever disappear completely from the end-users horizons. The trick for these services is to balance the needs of funders in delivering measurable service performance, the needs of end-users in being able to understand the qualities of the service, with the over-arching desirability of disappearing completely from view.

The community needs to refine its performance measures for purely machine-oriented services such as those found in the fusion layer.

---

[24] aDORe: a modular, standards-based Digital Object Repository <http://arxiv.org/abs/cs/0502028>

[25] LOCKSS <http://lockss.stanford.edu/>

6 Presentation

6.1 Portals, portals, portals

JISC has tended to treat portals as the primary service component within the presentation layer - hence the development of an explicit 'portals programme' for example. However, it must be remembered that presentation layer functionality can also be delivered by tools sitting on the end-users desktop – and in many cases this provides a much more 'usable' interface than can be provided by offering the same functionality through a Web-based 'portal' tool.

The most common examples of desktop-based tools are in the RSS 'alerting' application area, where there are now a number of tools, such as FeedReader[26] that allow users to access their favourite RSS channel on their desktop. Even more recently, we are seeing an interest in and use of Podcasting technologies[27], almost all of which are based on the end-users desktop (in order to provide close coupling with the end-users iPod or similar device).

We are also seeing a growing trend, particularly in the context of the Mozilla Firefox browser[28], for the browser to become an application framework. So, for example, there are now many plug-ins and extensions to Firefox that offer additional functionality (often associated with particular content types), over and above the basic browser functionality offered by Firefox itself. The advantage of this approach is that the functionality is delivered directly from the end-users desktop (i.e. closely embedded into the end-users working environment).

The community needs to balance the focus on 'portals' as Web-based services with a focus on the most effective mix of desktop and Web-based tools and services (both machine-oriented and human-oriented) that can be used to meet the end-users functional requirements.

6.2 Portals and portlets

It remains to be seen if the much hyped notion of a portal framework integrating a set of remote portlets using standards such as WSRP really comes to pass. So far it seems fair to say that portal/portlet specifications remain too complex and unstable to have had a real impact. It may also be fair to say that where we are seeing some impact, it isn't in the form of truly open, cross-platform solutions. Instead, we are seeing a focus on Java-only solutions. This situation may change of course, and it is to be hoped that it does.

If the situation does change, we are likely to see an explosion of 'portlets' being offered within an institutional context, with a similar growth in Web services sitting behind them. This is likely to require the deployment of institutional service registries (see below) since the majority of institutional portlets and Web services are unlikely to be offered for public consumption.

It is also important to remember that a growth in the use of 'portlets' does not, on its own, provide a growth in interoperability between systems at the level of exchanging data and

---

[26] FeedReader <http://www.feedreader.com/>

[27] Wikipedia entry for 'podcasting' <http://en.wikipedia.org/wiki/Podcasting>

[28] Firefox <http://www.mozilla.org/products/firefox/>

information. Rather, it simply provides a standard way in which we can share our user-interfaces between different human-oriented systems. This is no bad thing of course, but we need to take care not to mislead the community into thinking that by adopting portlet technologies they are facilitating an open exchange of interoperable data and metadata.

6.3 OpenURL 'link servers'

The growing deployment of OpenURL resolvers (or 'link servers') within UK HE/FE institutions is very positive and suggests that the use of OpenURLs is likely to have a significant beneficial impact on resource discovery services generally. However, there remain one or two issues with the widespread adoption of OpenURLs which the UK academic community is well placed to help resolve.

The OpenURL is just a special kind of URL, which includes as its first part a BASEURL - typically the URL of an institutional OpenURL resolver. For any two readers of this document who are members of different institutions, the 'correct' BASEURL will to be different (in order to link to the appropriate OpenURL resolver). It is therefore difficult to use OpenURLs to cite journal articles from a static document such as this one because the chosen BASEURL will not work correctly for all the intended readers.

The OpenURL Router[29], hosted by EDINA, is intended to get round this problem. By using the 'openurl.ac.uk' BASEURL, the creator of an OpenURL can be more confident that it will work for the majority of end-users because the router maintains knowledge about which BASEURL to use for any given end-user and can issue an HTTP 'redirect' to the correct OpenURL resolver.

Unfortunately, this service only works in the context of the UK (or to be even more specific, in the context of UK HE and FE). It would be highly desirable to devise a global mechanism that offers this kind of functionality, allowing OpenURLs to be deployed more easily and reliably on an international scale. This would potentially encourage global services such as Amazon or Google Scholar to embed OpenURLs into their user-interfaces and/or Web services.

Such a mechanism would almost certainly need to be widely distributed in order to offer the required level of reliability and performance. Therefore, the technology used would not be the same as for the current OpenURL Router service. Whilst this is a non-trivial issue to resolve, without it we may not see the more widespread take-up of OpenURLs beyond the current set of 'bibliographic' players. By developing a URI compliant form of OpenURL that is independent of a particular BASEURL, and a global resolution mechanism on the back of it, we will enable very widespread uptake of the OpenURL standard by services that may currently view it with some suspicion. The JISC community is well placed to contribute to the development of such standards.

This activity would also position the OpenURL as a true 'identifier' for bibliographic and other works, rather than it being simply a 'locator'. This would be very valuable in the context of being able to describe such resources as part of the Semantic Web, since in the Semantic Web all metadata is associated with the resource being described via its URI.

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

7 Shared Infrastructure

7.1 Authentication/authorisation/accounting

Discussions about the future of ATHENS and the gradual transition of the community to a Shibboleth-based approach have tended to happen outside direct discussion of the JISC IE technical architecture.

Therefore, this paper will not discuss this area in any detail. However, one issue for potential consideration is how well Shibboleth fits into the multi-level resource discovery hierarchy adopted by the JISC IE (namely, end-user->browser->portal->broker->content provider).

The community should ensure that appropriate authentication, authorisation, and trust mechanisms are in place to support the sometimes complex relationships between end-users, institutions, shared services, fusion layer services and content providers.

7.2 Distributed service registries

The current JISC IE Service Registry[30] (a database of descriptions of available collections and services) has been developed primarily as a single centralised service for the UK HE/FE community. It is becoming clear that such registries are likely to need to be distributed, both between multiple organisations within the JISC community and between other organisations in the UK and world-wide.

In order for such a distributed service registry approach to work on a global scale we need to reach agreements between the key players about what metadata standards we want to use and about what transport protocols we are going to use to share records between co-operating registries.

We also need to agree on the operational policies for the registries and the ownership and IPR issues associated with the metadata records being exchanged. For example, there may be significant problems in determining when two collection descriptions from two different service registries describe the same collection.

The Web Services standards, WSDL and UDDI, offer one possible way forward in this space. But it is not clear that these standards offer the best solution for our needs, given that they were developed primarily with e-commerce service description in mind. Alternative solutions, perhaps those offered by the use of the Dublin Core Collection Description metadata application profile[31] and the OAI Protocol for Metadata Harvesting, may offer a solution that is more appropriate for our community.

The JISC community needs to initiate discussions with key 'digital library' and elearning communities about how best to move towards solutions in this area.

7.3 Metadata schema registries

The JISC IE formally endorses the use of simple Dublin Core metadata and IEEE LOM (in the form of the UK LOM Core and various related application profiles). In practice, of

[30] JISC IE Service Registry <http://www.iesr.ac.uk/>

[31] DCMI Collection Description Working Group < http://dublincore.org/groups/collections/>

course, the community uses a wide array of metadata standards including various extensions to RSS, MARC records in libraries, SCORM and so on.

Dublin Core metadata is very closely aligned with the Resource Description Framework, the metadata model that underpins the Semantic Web. IEEE LOM is not, and uses a different underlying model. Other XML-based standards are likely to use different underlying models. One of the challenges for the JISC IE Metadata Schema Registry[32] is how far it is possible to provide a coherent view across metadata schemas that are based on different underlying models, and in particular how far it is possible to map concepts between such differing models. It is important to be able to do this, since one of the intended benefits of metadata schema registries is to support a 'mix-and-match' approach to the usage of properties from different metadata schemas.

What we see in practice is a mix of 'semantic Web' and non-'semantic Web' metadata standards being used along-side each other. The problem is made worse by the growing use of XML packaging standards such as METS. Packaging standards and metadata schemas are different, but they are often used in close proximity. It is therefore sensible to consider whether a metadata schema registry is the natural place to maintain knowledge about the packaging conventions (e.g. METS profiles[33]) being adopted by the community. Alternatively, it might be sensible to set up a separate, but related, 'packaging profile' registry.

The JISC community needs to undertake more work in the area of mapping metadata schema and related services, looking particularly at the issues of mapping between Semantic Web and non-Semantic Web schemas. JISC also needs to consider setting up a registry of 'packaging profiles'.

7.4 Identifier services and resolvers

The problems of assigning identifiers to resources in ways that result in a persistent binding between the identifier and the object have been discussed above. Identifier resolution services such as those offered at purl.org, dx.doi.org and hdl.handle.net are very important infrastructural services but it is not clear that we, as a community, need to be building any new similar services.

However, there may be a requirement for 'identifier' lookup services - i.e. services that allow people to see what identifiers have already been assigned and, in particular, to see which resources have already had identifiers assigned to them. Such lookup services will reduce the chances of resources being assigned multiple identifiers, and will help aggregators, brokers and presentation layer services in determining when they are dealing with descriptions about the same resource.

The JISC community should continue to contribute to international discussions about the use of identifiers and the services associated with them.

[32] JISC IE Metadata Schema Registry < http://www.ukoln.ac.uk/projects/iemsr/>

[33] METS Profiles < http://www.loc.gov/standards/mets/mets-profiles.html>

7.5 Terminology and terminology services

Building and using controlled vocabularies and thesauri is difficult and it gets harder as solutions are attempted on a global scale and/or across very broad 'topic' areas. However, the use of controlled vocabularies and thesauri is the key to building useful subject-based services, since it is primarily through the adoption of one or more vocabularies, and the use of good mappings between those vocabularies, that it is possible to build coherent search and browse interfaces through any given collection of metadata records.

The Higher Education Academy community seems to have been quite successful in building up controlled vocabularies in the areas of pedagogy and policy/strategy and there are perhaps some lessons from this that the community more generally can learn. Recent attempts to revitalise discussion around the list of UK Educational Levels have continued to move this discussion forward, though we are not yet close to resolving all the problems.

However, it is in the area of subject classification that we still don't seem to be able to move towards any kind of widespread agreement. For example, as a community we are not yet in a position to be able to make recommendations for how subject terms are assigned in HE/FE institutional repositories. The recent announcement by OCLC that the top three levels of Dewey[34] are now freely available for use may help to get the community over some of the licensing hurdles that appear to have stifled progress to date. However, it is not clear that the top levels of Dewey (or even all of Dewey for that matter) provides a useful subject vocabulary for the community as a whole.

Despite this lack of clarity about our current and future requirements it is likely that different parts of the community will want to maintain and use different vocabularies in the areas of subject, resource type, audience/educational level, spatial coverage, temporal coverage and so on. The community will require ways of dealing with this complex vocabulary space and may well require infrastructural services that map between and across terms in these vocabularies.

As a community we need to refine our understanding about the best ways that our ontologies can be created and maintained (see, for example, the references to 'folksonomies' above) and the kinds of services that we require on those ontologies.

Without knowing exactly what form these services might take, we can help in their future development by ensuing that best-practice guidelines are developed for assigning identifiers to terms in the vocabularies (e.g. URIs) and for marking-up the vocabularies in machine-readable forms (e.g. RDFS[35], OWL[36] and VDEX). The SKOS Core RDF Vocabulary[37], mentioned in the semantic Web section above, looks to be a significant contender in this space.

The JISC community should work towards reaching agreements about how to mark-up community-developed vocabularies in a machine-readable form and how to assign URIs to vocabulary terms.

---

[34] Dewey Decimal Classification < http://www.oclc.org/dewey/>

[35] RDF Schema Language < http://www.w3.org/TR/rdf-schema/>

[36] OWL Web Ontology Language < http://www.w3.org/TR/owl-features/>

[37] SKOS Core RDF Vocabulary < http://www.w3.org/2004/02/skos/core/>

7.6 Other services

The list of infrastructural services shown on the JISC IE architecture diagram above was not intended to be exhaustive. As the community and landscape evolves we can expect to see new shared services being developed, both within and without the JISC community.

7.6.1 Licence registry

The development of the Creative Commons licences[38], particularly the recent work to build UK-specific versions of those licences, is encouraging 'open access' content providers to be more explicit about the licences under which resources are being made available. We are therefore likely to see growing use of machine-readable licences (e.g. using the ODRL standard), either directly attached to resources or attached indirectly via a metadata record that includes the licence URI.

In the early phases of this process the community is likely to benefit from easily available knowledge about what licences are in use within the community. This will encourage the adoption of the same (or similar) licences and will help to prevent duplicated effort in developing new licences where existing ones are already available.

In the longer term, there are likely to be problems in dealing with licences that are no longer available on the Web. Therefore, a persistent repository of information about past and present licences may be of value.

The JISC community should work towards building a licence registry (or registries) that help meet these needs.

7.6.2 Automated metadata-creation tools

The community needs to move towards services that automatically create metadata whenever possible. Examples of such tools include automatic subject classification tools (for example the automatic Dewey classification services offered by OCLC and the LCSH tools in the iVia Infomine software[39]), format assignment tools and so on.

The JISC community should encourage these kinds of tools to be offered on the network as infrastructural Web services so that presentation layer (and other) services can make use of them.

7.6.3 Name authority tools

"Author search" is one example of a resource discovery approach that is best met by metadata-based solutions rather than by full-text indexing since it is difficult to reliably extract the author's name from the full-text. However, the requirement can only be met in full by assigning names to people and organisations according to an agreed set of rules. Typically this is achieved by selecting a controlled form of name from a name authority file. Offering a machine interface to a name authority Web service, in order that presentation layer and other services can make use of it, will help to achieve an improved level of conformity with the way names are assigned.

---

[38] Creative Commons <http://www.creativecommons.org/>

[39] Infomine iVia Software < http://infomine.ucr.edu/iVia/>

However, the experience of the ePrints UK project, which worked with OCLC to deliver a name authority file for eprint authors and contributors, is that we need a suitable database of names to underpin such a service (e.g. the names of UK journal article authors in the context of creating metadata for eprint archives).

JISC should work with various parties, including the BL, to determine if such a list of names is available or whether it could be created and maintained centrally, and if so to layer Web services in front of it.

An alternative approach is to push the name authority problem down into institutions. By using a common technology approach, encouraging institutions to control their part of the name authority space, and by providing a single, coherent, point of access to that distributed network of authorities, the community might be able to build and maintain it's own distributed name authority file. It is not clear what technology would be used to build such a system, but LDAP would be the usual choice for 'person'-related directory services. This approach might fit well with existing institutional directory services and with requirements for supporting Shibboleth in the future.

The JISC community should consider options for delivering a distributed 'name authority' services through a network of institutional LDAP servers.

8 Conclusions

This report has considered each of the main areas of the JISC IE architecture diagram in turn (provision, fusion, presentation and shared infrastructure) and has listed some possible areas of activity that might be taken forward by JISC and the community. These are summarised in the executive summary section above.

This report has stopped short of making firm recommendations about which areas are funded by JISC and which are funded by other means. Nonetheless, it is hoped that this report is found to be useful by the JISC in making decisions about future areas of activity.

Acknowledgements

The author would like to thank Rachel Bruce (JISC) and Rachel Heery (UKOLN, University of Bath) for commenting on previous versions of this document.