

E-INFRASTRUCTURE STRATEGY

REPORT OF THE WORKING GROUP ON SEARCH AND NAVIGATION

Members of the Working Group

Michael Jubb (Chair) – *Research Information Network*

Keith Adlam – *British Geological Survey*

Richard Boulderstone – *British Library*

Rachel Bruce – *JISC*

Peter Burnhill – *EDINA*

Stéphane Goldstein – *Research Information Network*

Adam Hodgkin – *Exact Editions*

Balviar Notay – *JISC*

David Pearson – *University of London*

David Smith – *CABI Publishing*

Report drawn up by

Sheridan Brown – *Key Perspectives Ltd*

March 2006

1. Summary

Search and navigation services provide the key to unlocking the wealth of literature and data that exists in digital form. As such they are fundamental to achieving the Government's goal of providing high quality and effective information mechanisms for the research community. This report presents the views of the working group on search and navigation, convened by the RIN on behalf of the OST-led steering group which is responsible for developing a national e-Infrastructure strategy. The report has three key sections:

- In section three we present a concise overview of the current provision for search and navigation, giving examples of the different types of resource available both in the body of the report and in tabular form at Appendix 1. In addition we highlight a number of issues that are central to any consideration of search and navigation, such as user behaviour, metadata standards, privacy, the general direction of the market and public/private sector interaction.
- Building on this foundation, section four outlines our view of the search and navigation resources that, ideally, will be developed during the next 10 years. We perceive a number of important developments, including: a big increase in the scale and scope of digitised content; the widespread, effective use of standardised metadata; the development of a wide range of new search and navigation tools and services; and a regulatory framework that deals with the issues of privacy and intellectual property rights.
- Finally, in section five we recommend four programmes of activity designed to reduce the gap between the current situation and our vision of the ideal situation. These focus upon user behaviour, finding aids for the nation's physical information collections, semantic interoperability and, lastly, information search and navigation tools and technologies. We also flag up the importance of metadata and the need for an appropriate regulatory framework.

2. Introduction

The Government's Ten Year Framework for Science and Innovation provides the basis for the development of a national e-infrastructure strategy, the ultimate goal of which is to provide high quality and effective information mechanisms for the research community. The development of the strategy is being overseen by an OST-led steering group. The steering group has appointed working groups to consider six key activity strands. This is the report of the working group on search and navigation which presents an overview of the current level of provision, a vision of the "theoretical ideal" situation in ten years time, and finally ideas on how best to stimulate development in the direction of this ideal situation.

Search and navigation services may be defined as interactive electronic means used by people to discover and utilise a wide variety of information sources relevant to their needs. The subject has been considered in its widest sense, encompassing bibliographic resources, directory services, generic services and the underpinning framework. It is important to note at the outset that the working group identified three key distinctions that apply to search and navigation resources: **literature**; **data** (of all types); and the **facilities** that serve to enable people to discover such information. The working group recognises several additional distinctions which should be flagged up at this point:

- This report focuses on *national* issues but recognises the importance of the *international* context.
- The transient nature of the current range of search and navigation resources: some are *long established*, others in their *formative* stages and others whose longevity is defined by *public funding timelines*.
- There is a key distinction between resources that are *freely available* and those to which access is restricted by *toll barriers*.
- There is a difference between services which *point* to full text and data resources and those which provide *direct access* to such resources.
- Finally, different search and navigation services operate at different levels of granularity, ranging from low level granularity at the resource discovery level to high levels of granularity when searching within literature and data information sources.

3. An overview of the current provision for search and navigation

The wide scope and variety of issues, resources, and activities that naturally reside within the ambit of “search and navigation” provision presents a significant challenge: how best to reflect that variety in a concise and coherent fashion. In recognition of this we have distilled six main themes, designed to encapsulate the nature and breadth of activity in this field.

3.1 Main themes

In order to be concise, we have focussed upon established resources that exemplify the types of resources that fit in each of our main themes. The aim is not to be comprehensive in coverage, but rather to provide an overview of the range of resources that exist today. These resources are outlined below and more detail is presented in the accompanying table [Appendix 1].

3.1.1 Scholarly and technical information resources

This theme is sub-divided into three parts, described below:

First, **libraries and archives** are major repositories and points of access for scholarly and technical literature. Such institutions create their own catalogue records to describe and enable access to their holdings of books, journals, e-resources and other materials for study and research as well as retrospectively digitise manual records. On a national level, union catalogues such as COPAC cover books and journals held by members of the Consortium of Research Libraries (CURL).

Second, **abstracting and indexing services** provide information about, and increasingly access to, a significant proportion of scholarly and technical information. The majority of these services are provided by the commercial sector. Historically abstracting and indexing services have provided value through manual indexing, creating rich and consistent metadata, though we believe it likely that this process will increasingly be done automatically as a matter of course. Services tend to be discipline-specific though two notable examples strive towards comprehensive coverage: Web of Knowledge published by Thomson Scientific and, more recently, Scopus published by Elsevier.

Third, access to **full-text content** has always been very important to researchers and controlling the rights to unique full-text content confers upon primary publishers a privileged and very strong market position. The majority of scholarly publishers provide electronic access to full-text content though at present most of this content lies behind toll barriers. The largest single source of scholarly full-text currently available is Science Direct, produced by Elsevier. It should be noted that a small but increasing proportion of full text content is becoming available on an open access basis. Most publishers now permit researchers to self-archive their published papers, with certain limitations. A good example of a community-based approach is the Open Content Alliance, whose aim is to help build a permanent archive of multilingual digitised text and multimedia content.

3.1.2 Aggregation services (gateways and portals)

Aggregation services, which include portals and specialist directories, provide users with a single point of access to multiple resources. A good example of a **general** gateway is the Resource Discovery Network (RDN), a collaboration of over 70 educational and research organisations

which offers a free national catalogue of internet resources for the learning, teaching and research community. A **grid portal** is a problem-solving environment which allows researchers to use distributed grid applications with conventional desktop tools. Examples include the National Grid Service Portal, myGrid and GROWL. **Subject** or **domain-specific portals** facilitate access to resources in particular subject areas or of particular kinds. The EDINA Go-Geo! Portal, for instance, performs simultaneous searching of a wide range of geospatial metadata catalogues. Finally, **specialist directories** exist to provide straightforward access to information that researchers may otherwise find difficult to assemble. One example is the ARCHON Directory, funded by the National Archives.

3.1.3 Repositories of digital content

Repositories can contain many different types of digital information, including literature, data and visual information objects, and can be subject or institutionally based. There is keen interest in and funding support for the development of repositories particularly in the higher education sector, not least because of their potential impact in facilitating the dissemination of research outputs. Examples of successful subject-based repositories of scholarly literature are arXiv, regarded as a key resource by physicists, and PubMed Central, an archive of life sciences scholarly literature. Other services, such as the Arts and Humanities Data Service or the British Geological Survey, focus on data generated by or of use to researchers.

3.1.4 Commercial suppliers of published content or services

There exists a stratum of organisations that currently fit in the middle of the distribution channel between publishers and institutional purchasers. Many of these intermediaries offer aggregation services bundling such as full-text scholarly literature with proprietary search software. Others offer electronic publishing services to publishers who choose to outsource this aspect of their business; Highwire Press is a good example.

3.1.5 Generic services and products, including search engines

There is a growing class of services and products that are becoming “generic” in the sense that they boast significant infrastructure and scalability and are very widely used. Typical examples include Google, Yahoo and Microsoft Network (MSN) – all of which are based in the USA. Such services are useful and simple to use but the breadth of their coverage gives rise to questions about the precision and relevance offered by these discovery tools. Thus while they are, and will continue to be, used by researchers, and it is likely that their usefulness to researchers will increase as the services develop further, we do not believe that they can be relied on as central pillars of the service infrastructure to meet the Government’s wish to “deliver an effective system of high quality and effective information mechanisms for the [UK] research community.”

3.1.6 Underpinning activities, projects and infrastructure

The e-infrastructure currently in place in the UK is thought to be of good quality in comparison with international competitors, and it continues to evolve. SuperJANET 5 will enhance the capability of the UK’s optical network and the JISC has a programme to establish a network of digital repositories as part of its Information Environment. Other important underpinning elements of the current infrastructure for search and navigation include: open URL resolvers; the national open URL router; licensing structures; persistent identifiers; metadata standards; burgeoning middleware and machine-to-machine services.

3.2 Central issues

In judging the effectiveness of the current provision of search and navigation services for the research community in the UK, and in considering the ways by which it could and should be developed, we wish to highlight a number of issues that cut across the current pattern of provision and will need to be addressed as part of the process of planning for the future.

3.2.1 User behaviour

It is one thing to strive to provide excellent infrastructure, services and tools but it another matter entirely to persuade users to make best use of these facilities. At present students and researchers are known to have widely adopted resource discovery habits that rank ease of use and convenience higher than quality of results. This means that Google, for example, is used far more extensively than other resources which may be more appropriate for the task. This is a cultural and behavioural issue of critical importance, one that is addressed in section 5.1.

We have also noted the growth of peer to peer communication in relation to information discovery and the development of folksonomies which, many believe, are an important step towards an effective semantic web. User tagging, an activity which currently applies particularly to open access scholarly literature, has the potential to become an important feature of the search and navigation environment. Users (rather than authors) add tags, which may be classificatory or evaluative in nature. The process is anarchic – or democratic – and proponents claim it to be a version of the semantic web. This active involvement by users in classifying and evaluating information may become an important aspect of search and navigation. An example is Connotea, promoted by Nature Publishing Group.

We are already seeing advances in the development of folksonomies and tools such as wikis, blogs and mashups. While these are not discovery tools as such, they are sources of information which in themselves are becoming gateways. Witness, for instance, the extent to which Wikipedia is already becoming a vast, even if imperfect, means of finding information

3.2.2 Metadata standards

Metadata is structured information that describes the key characteristics of information objects. Resource discovery is critically dependent upon metadata, which itself is essential for effective machine-to-machine interaction. In order for search and navigation to move forward it is necessary to encourage information providers to adhere to metadata standards, for new, richer metadata standards to evolve, and crucially for all information providers to expose their metadata effectively so that information can be indexed and searched.

3.2.3 Privacy

There are increasing concerns about the possibility of the misuse of data about individuals on two distinct levels: first, researchers' search and navigation behaviour, where an electronic trail of users' searching activity is recorded by computers; and second, personal information which individuals may provide in confidence to one or more information providers. If an information system is to be truly effective it must be perceived by individual users to be transparent and, above all, trusted. In the UK and the rest of the European Union privacy and human rights legislation needs to be taken into account when designing systems that rely upon data that may be regarded as private.

3.2.4 Intellectual Property Rights (IPR)

Current IPR arrangements derive essentially from a pre-digital age and do not reflect the ways in which information – especially information created and used by the research community – is now created, disseminated and used. IPR arrangements thus constitute a significant barrier to the development of effective search and navigation, complicating the route between discovery

and access to source information. These issues need to be addressed both at national level (initially through the Gowers review) and at international level (through the EU Information Society programme and through WIPO).

3.2.5 The general direction of the market

A large proportion of the search and navigation arena for researchers as for other major sectors is served by commercial organisations. By their nature these organisations need to adapt to their changing environment if they are to survive. Key examples of current pressures and opportunities are given below as pointers to the general direction in which the market is moving.

- The past few years has seen consolidation among major publishers. Important abstracting and indexing services have been bought and mainly subsumed into existing (larger) portfolios and systems, reducing the variety available in the search and navigation market.
- Information providers are engaging in forward and backwards integration with, for example, subscription agents becoming aggregators of data and full text content, and publishers of scholarly literature becoming creators of abstracting and indexing services.
- The relationship between author and end users is changing, becoming more direct. This is putting pressure on major players in the value chain to adapt.
- The importance of literature and data which is not published via traditional channels is increasing. Examples include the growth of folksonomies, peer-to-peer communication and wikis. Indeed, researchers and their institutions are becoming “publishers” of information in their own right.
- Electronic information is rapidly becoming the dominant form, increasing the relevance of effective search and navigation tools.
- Many publishers have a “long tail” of information assets which, with the critical mass of online traffic now using the web, are becoming easier to exploit.

3.3 Estimate of current expenditure

While we give some indications of current expenditure in the Appendix, we have concluded that beyond headline figures for projects funded by public sector organisations there is no reliable way to adduce or even estimate overall levels of current expenditure. In the public sector the search and navigation element of investment programmes is not normally disaggregated either at the level of work or of cost. It would, therefore, be very difficult to isolate expenditure that relates solely to search and navigation. The private sector may well keep detailed records for the costs involved in developing and maintaining search and navigation services, but such information is rarely available in the public domain.

3.4 Summary of plans for future provision by key providers

Similarly, we have concluded that it is not possible to adduce the plans for future search and navigation provision by key providers, especially those in the commercial sector, beyond an assumption that the biggest players such as Google will continue to strive to become entrenched in all aspects of the information landscape. The extent to which they are successful will have an impact on the threats or opportunities presented to smaller providers.

3.5 International comparison with leading research countries

Budgets and related activities are rarely disaggregated to the level of search and navigation so figures are few and far between, but there is a perception that the UK is relatively well advanced in its provision of ICT infrastructure and the availability of search and navigation resources to researchers - though public investment is proportionately far less than in the USA. The UK is an active participant in the European Union's work to establish an ICT policy for the coming years, an initiative branded i2010.

3.6 Public/private sector interaction

The current pattern of search and navigation services for the research community involves a significant degree of interaction between the public and private sectors, and this will continue into the future. But while there is little doubt that the private sector will continue to play a major role in the development and provision of search and navigation services, the precise nature and scope of interaction between the two sectors is uncertain, and it remains important to recognise that their strategic goals may be complementary, but they are not necessarily congruent.

- There are examples of beneficial interaction between public and private sectors, such as library catalogues interacting with Google.
- The private sector cannot guarantee comprehensive coverage. Many specialist fields of study are not particularly attractive to commercial publishers, sometimes because of the relatively small size of these niche research communities.
- The private sector is sometimes reluctant or unable to guarantee the long term preservation of information. In an environment where libraries are increasingly licensing access to electronic information rather than buying print media that can be archived, interaction and partnership between the public and private sectors will be of increasing importance in ensuring not only preservation, but the development of search and navigation services through which researchers can find and gain access to information.

4. An outline of a theoretical ideal set of search and navigation resources in the next 10 years

In order for an effective system of search and navigation resources for researchers to develop over the next 10 years there are a number of factors that need to be in place: more digitised content; better tools; the technology to generate standardised metadata automatically; and the development of appropriate ontologies and mechanisms for searching across ontologies. Our conclusions on a theoretical ideal set of search and navigation services are presented in a list format below, following the headings adopted for section 3.

4.1.1 Scholarly and technical information resources

- More digitised content together with the deployment and exposure of rich metadata standards will mean that a far greater proportion of the nation's information currently stored within libraries, archives and museums will be widely available for discovery and research.
- Abstracting and indexing services will routinely use automatic metadata extraction.
- Information objects of all kinds – including novel content types such as video – will benefit from the application of standardised metadata to enhance the opportunity to discover and then search within collections of information.
- There will be enhanced provision of linked full text content, data, references and commentaries, free at the point of use.

Focusing on the viewpoint of the individual researcher and their typical workflow patterns, we expect that:

- Articles will be linked much more directly and reliably with author emails and home pages, so that natural human connections will be seamlessly integrated and available.
- Lineages will be generated automatically on a variety of specifiable variables, such as citation lineage for articles and authors and influence lineages for topics
- Commentary and blog space will become interconnected with research space so that download and citation counts can always be augmented by looking at who is saying what about the work and when.
- Personalisation will be built into systems and their architectures, taking due care not to insulate researchers from the entirety of the search and navigation resources available.

4.1.2 Aggregation services (gateways and portals)

- Smarter search and navigation services will reduce the need for aggregation services in general, though there will be a need for such services to serve specialist fields of research.
- Search and navigation tools will be in place to meet the information needs of researchers working on and across the boundaries between traditional disciplines and subject areas.
- Services will be widely available to facilitate searching across language barriers.
- The search and navigation environment will be enhanced by widespread semantic interoperability; in particular people will be able to search and navigate on the bases of space, time, person and concept.

- There will be effective tools to support the tracing of provenance of information, from data creation through to all its subsequent manifestations

4.1.3 Repositories of digital content

- A wide range of repositories will be in place – some offering a wide range of content, others being content-specific especially when they are guided by their core mission (the Ordnance Survey providing cartographic information, for example). Data will be captured systematically as part of researchers' workflow process. They will adhere to metadata standards and expose their metadata to facilitate the effective functioning of machine-to-machine services.
- There will be a set of services – building on current work on processes to support cross-domain knowledge generation - to enable researchers effectively to search and navigate across repository boundaries

4.1.4 Commercial suppliers of published content or services

- Commercial suppliers of published content or services will continue to influence the development of search and navigation services through, for example, technical developments and licensing models. Their activities will, however, be affected by public and private sector developments in the information infrastructure as well as the related developments listed in this section.

4.1.5 Generic services and products, including search engines

- The biggest players will continue to be pervasive, augmenting their blanket coverage/indexing with developments designed to offer niche services, some of which may be directed to the research community. The scale of their infrastructure and technological resources is likely to insulate them from environmental changes for the foreseeable future.
- The major players will begin to provide more personalised search and navigation services deploying, for example, intelligent search agents.
- These big players will provide the context for the search and navigation environment within which smaller players from the public and private sectors will seek to identify niche markets where they are well positioned to provide value added search and navigation services to the research community.

4.1.6 Underpinning activities, projects and infrastructure

- We will see the development of a semantic grid, designed to provide meaning to data and to generate an appropriate semantic context for the meaningful interpretation of data.
- The infrastructure will have moved towards a ubiquitous system where computing devices and communications technology combine to enable researchers to use search and navigation services whenever and wherever they need to.

4.2 Central issues

4.2.1 User behaviour

- We shall have advanced towards an environment in which researchers use the search and navigation resources most appropriate to their information needs, and have been trained to acquire the skills to use those resources effectively.

- The motivations and changing needs of researchers in relation to search and navigation services will be clearly understood. To achieve this, researchers will be fully engaged in the process of developing new services through tried and tested processes such as user behaviour research, concept and prototype testing, and observation.
- Alongside more formal search and navigation services, researchers will be using folksonomies and social tagging, and tools such as wikis, blogs and mashups.

4.2.2 Metadata standards

- Organisations that create information that they wish to be made widely available will be providing metadata that adheres to a recognised standard. In future it is very likely that metadata will be extracted automatically, helping ensure not only adherence to standards but also providing greater levels of richness than one sees today.
- When either individuals or institutions deposit information objects in repositories, they will be required to assign their own metadata tags (either manually or automatically).

4.2.3 Privacy

- There will be a regulatory framework, backed by statute if necessary, to deal with the issue of privacy in a digital world and to promote adherence to the regulations, to police and prosecute transgressors, and to protect the rights of researchers in relation to privacy.

4.2.4 Intellectual Property Rights (IPR)

- A machine-to-machine rights management system will be in place in order to mitigate the barrier effects of IPR issues in the search and navigation environment. The implementation of machine-readable licence registries will enable a process of automatic authentication, smoothing the search and navigation process.

4.2.5 The general direction of the market and levels of expenditure

- We cannot of course second guess the market, but our view is that the current major private sector providers of search and navigation services will thrive as the quantity of digitised content grows massively. Among the smaller providers, those which are nimble and adaptable will benefit from the rapidly changing market environment. In terms of expenditure, as described in section 3.3, there is currently no way to provide reliable estimates.

4.3 International comparison with leading research countries

- The UK library and archive sector compares favourably with some international competitors in the extent to which collaboration or public investment has delivered unified finding tools and related infrastructure to facilitate resource discovery for researchers. But the UK still lacks a national union catalogue of the content available through major libraries, and it will lose its competitive edge both in library and information services and in the support of research if there is not continued investment in the development of search and navigation services. A number of countries (e.g. USA, Australia) have government-supported programmes to develop digital information content and associated search and navigation services, thereby demonstrating their understanding of the link between such activity and national research excellence.

4.4 Public/private sector interaction

- Both the public and private sectors will continue to invest in search and navigation resources and the infrastructure that underpins them. There will be increasing interaction between the two sectors, leveraging the different strengths both can offer.

5. Moving towards the ideal situation

Rather than try to put a price on individual project proposals, we have identified potential *programmes* of investment and development that reflect the key themes identified in the preceding section.

5.1 User behaviour

Many of our recommendations focus on the technical aspects of developing an efficient search and navigation environment, but unless it is leveraged effectively by researchers the potential benefits will not be realised. As well as providing the means, it is important to influence user behaviour in order to optimise the research process. We identified in section 3.2.1 the cultural phenomenon whereby people – from casual users to professional researchers – habitually turn to search engines such as Google, Yahoo and MSN. These are useful tools but they are far from being the only ones available and, depending on the subject area, different tools and resources may be more appropriate.

Service providers need to understand in greater detail the basis for researchers' current search and navigation-related behaviour, and to determine the extent to which the historically fragmented nature of information and data provision inhibits optimum discovery behaviour. They also need to ensure that researchers are fully engaged in the development of new services and in decision-making about them; and that we avoid the danger of provider-driven rather than demand-driven services. In so doing they will build effectively upon work that has already been done in this field. Once this groundwork has been done, there will be a need for technically-aware, full-time trainers based within universities, colleges and other research institutions, whose job will be to educate and train researchers to exploit search and navigation resources effectively in pursuit of their research goals.

Some of this training will be aimed at bringing more of the current generation of researchers up to speed with developments in information technology. But it is equally important that school children should be taught from an early age not only to use search and navigation resources effectively but critically to assess the validity and relevance of the information sources they use. More specialised training in research techniques and skills can then build on this foundation.

Recommendation: *Invest in a programme of education and training for researchers and information professionals and encourage their engagement in the development of specialist services. There is also a case for considering the introduction of search and navigation techniques to school children.*

Benefit: The main benefit will be to enhance the effectiveness of the research, by, enabling researchers to make best use of the investment that is being made in ICT infrastructure as well as search and navigation resources.

Risk: The main risk is that the human tendency to take the path of least resistance will persist (which often translates to using generic search engines rather than more focused resources), meaning the investment described above will not be leveraged to its full potential.

5.2 Finding aids for the nation’s physical information collections

In a world where access to electronic information resources is becoming the norm for researchers, the working group envisages it will be increasingly important to deliver the full range of available research material digitally. At the present time, vast quantities of material remain accessible only in traditional physical formats, discoverable only via clumsy or labour-intensive mechanisms.

Recommendation: *To bring these physical collections to the research community through a process of creating digital catalogues and other “finding aids”, as well as digitising content where it is feasible to do so either in collaboration with private sector partners, where appropriate, or through public funding initiatives.*

Benefit: Researchers and the research process will benefit from increased remote access to information resources that are currently only available in physical form at particular geographic locations.

Risk: There is a risk that the cost of the process will outweigh potential benefits if the resources prove to be poorly utilised by the research community.

5.3 Semantic interoperability and related technologies

There is currently progress being made across a number of distinct but related fields such as semantic interoperability, shared services, concept mapping, distributed registries and knowledge organisation systems. Not only do these endeavours need to be supported since they are fundamental to improving the nature and scope of search and navigation, but the advances need to be corralled and harnessed in the pursuit of better search and navigation services. Services should facilitate searching on a variety of bases, not least space, time, person and concept. For search and navigation to step boldly into the future there is a need for investment in developing processes that enable machines to understand concepts and meaning.

Recommendation: *Coordinate and support a programme of technology development focussing on semantic interoperability in its widest sense.*

Benefit: Semantic interoperability has the potential to significantly improve search and navigation processes.

Risk: It will be difficult to achieve on a wide scale.

5.4 Information search and navigation tools and technologies

The next ten years will be characterised by a massive increase in the availability of data – a so called “data deluge”. The development of innovative search and navigation tools and technologies will be central to researchers’ ability to make the best of the possibilities and opportunities presented by this wealth of data. The programme of development should include the following areas: data mining; text mining; moving picture, sound and content-based image retrieval; sound transcription and indexing techniques; tools to trace data provenance and lineages. This will bring together skills in disciplines such as computer science with traditional library and information science techniques.

Recommendation: *Coordinate and support a multidisciplinary programme of development of search and navigation tools and technologies.*

Benefit: The programme will bring together relevant work being done in different disciplines, harnessing advances to ensure the development of new, more effective search and navigation tools and technologies is done as efficiently as possible.

Risk: It can be difficult to manage potentially synergistic programmes such as this effectively.

5.5 International standards and metadata

There is little doubt that the widespread adoption of metadata standards is central to the future development of search and navigation services. There will be many different standards according to the type of data being tagged, but these will be automatically mapped for search and navigation purposes. National and international standards will continue to be created and adapted for all types of data including moving pictures, sound and images. At this point the working group is not in a position to put forward a recommendation for additional activity in this area, other than to reiterate the importance of data providers adopting standardised metadata.

5.6 Regulatory Framework

Finally we wish to flag up the need for a regulatory framework designed to deal with issues such as IPR and privacy. Since such a framework will affect all the topics covered by the six working groups, we propose that this be discussed further by the Steering Group.

Appendix 1

SEARCH AND NAVIGATION

Tabular information on selected current provision and future plans of service providers

Colour coding for table headings:

	Scholarly and technical literature resources
	Aggregation tools (for data and information)
	Source data services and repositories
	Generic services and products
	Commercial suppliers of published content/services
	Underpinning framework

D,L,R,A function codes – D: discover, L: locate, R: request, A: access

Name	Description and Function	Scale and Maturity	Est. Current Expenditure	Future Plans	Developments in other countries	Is target online or offline?
Scholarly and technical literature resources – libraries & archives						
Academic libraries	Catalogue records Created by all libraries to describe and enable access to their holdings of books, journals, e-resources and other materials for study and research. Typically provided through online web-accessible systems. D-L-R-A	Ca. 100 million records in HE and national libraries, in aggregate?	£30 million (est) spent across HE sector in creating and maintaining catalogues and catalogue records	Ongoing provision of local catalogue services is likely to be an integral part of library operations for the foreseeable future		
COPAC	Union catalogue of books and journals held in major research libraries in membership of CURL (Consortium of Research Libraries). Freely available over the web. D-L-A	32 million records	c.£150K p.a. in staffing and infrastructure costs, split between JISC and CURL	Strategic review of the future development of COPAC is under active consideration by both CURL and JISC	COPAC is one of a number of major union catalogue databases available globally; for the English-speaking world, the largest is developed in USA by OCLC (61M records)	
SUNCAT	National Union Catalogue for Serials for all research libraries in the UK; open access to researchers; additional, authorised services to librarians in Contributing Libraries. Steering Committee has representatives from BL and larger research libraries, building on initial RSLP sponsorship. D-L-A	Phase 2 of project: researcher interface launched in 2005; over 4 million item records, drawn from the holdings of the UK national libraries and 30+ largest university and specialist libraries, and the records of the ISSN Register and the CONSER database.	approx. £700kpa, including set-up, from the JISC	To extend coverage to about 60+ libraries, including key civic libraries In Phase 2 (by end of 2006), and perhaps beyond in Phase 3 (2007 -). To extend to electronic subscription information, using ONIX for Serials (Serials Online Holdings), Subject of project activity. To provide focus for journals, including other subject, regional and national union catalogues, see http://www.suncat.ac.uk/unioncatalogues/	UK is one of the last European countries to establish a national Union catalogue for serials. Co-operation through participation with ISSN –IC.	
ZETOC	Provides consolidated access to tables of contents of journals held by the British Library. Freely available to the academic community,	Covers 20,000 current journals and 16,000 annual sets of conference	Primarily funded by JISC	Strategic review of future development being undertaken by JISC, summer 2006. Various possible development paths,		

	available on subscription to others. D-L-R-A	proceedings		covering both additional content or amalgamation with other services		
Archives Hub	Consolidated searching tool covering the archive collections of UK HE institutions D-L-A	Covers 140 HE institutions	Funded by JISC, on behalf of CURL	One of a number of archive discovery tools/union catalogues in the UK; the obvious path for future development is to join them up	RLG are currently developing an American union catalogue of archive collections (ArchiveGrid)	
Scholarly and technical literature resources – A&I services						
CAB Abstracts	Commercial Subscription Bibliographic database covering the applied life sciences (emphasis on agriculture, environment and animal sciences. Major coverage of non English language materials and 'grey' literature. D-L-R-increasingly A To deliver access to research in the applied life sciences. Historically a location tool but like all bibliographic dBs these days, the emphasis is on maximising one click access to full text etc	7.7 million records. 1913 onwards.	£5 - 10 million	Full text hosting to accompany the dB (solutions for those who lack the resources to get their material online). Looking at indexing and access to additional content types.		
ISI Web Of Knowledge	Commercial - Subscription Multifaceted platform for bibliographic data with citation linking and direct access to multiple discrete data objects (patents, full-text articles, chemical structures, web sites). D-L-R-D (via open URL tech) Like Scopus, it aims to be a one stop solution for location and access to the gamut of scientific research outputs. A flagship product. Also enables niche databases to be searched and integrated within it Cover Arts and Humanities and Social Sciences	28 million (2004) but note - the cited references info does not form part of this number. Cited references approx 20 million per annum. 51 years old. Index goes back just over 100 years.	Unknown (Thomson Scientific had a Cap-Ex of \$98 million in 2004)	Looking to apply the same technologies and thinking to an assessment of institutional repositories. No doubt plenty of other things as well.		

PubMed	<p>Non-commercial - free at point of use. US Public Funded (Federal funding via National Library of Medicine) Bibliographic database (with increasing full-text linkage).</p> <p>D-L(sometimes)-R-A(increasingly)</p> <p>Aiming to get the researcher from a query to the papers (and additional data) they need as rapidly as possible.</p>	<p>16 million records back to 1950 (600k per annum).</p> <p>1879 onwards (1950 electronic onwards) online 1971.</p>	\$15 -20 million est	<p>Artificial Intelligence indexing terms system</p> <p>Addn dataset info</p> <p>In next 10 years crawling websites (publishers sites to grab additional info that could be useful)</p>		
Aggregation tools – general gateways and portals						
Resource Discovery Network (RDN)	<p>The Resource Discovery Network is the UK's free national gateway to Internet resources for the learning, teaching and research community. In contrast to search engines, the RDN gathers resources which are carefully selected by subject specialists in our partner institutions. You can search and browse through the resources, and be confident that your results will connect you to Web sites relevant to learning, teaching and research in your subject area.</p>	<p>The RDN funded by the JISC and is a collaboration of over seventy educational and research organisations, including the Natural History Museum and the British Library, and builds upon the foundations of the subject gateway activity carried out under the JISC's eLib Programme.</p>		<p>The service currently links to more than 100,000 resources via a series of subject-based information gateways (or hubs). The RDN is primarily aimed at Internet users in UK further and higher education but is freely available to all.</p>	<p>Total number of direct catalogue access for last year: 37,287,306</p>	
Aggregation tools – subject portals						
NERC DataGrid Portal	<p>NERC Metadata gateway, which is based on Z39.50 technology and in principle links together all the NERC data repositories at the discovery metadata level.</p> <p>Funded by NERC and the National e-Science programme.</p>	<p>They expect to retire it within the next couple of months in favour of a discovery service based on OAI-PMH, which obviously where possible will exploit OAI repositories, but if necessary we'll handle legacy z servers for a short while.</p>	<p>As part of the NERC DataGrid project, they are spending approximately 1.5 million over five years developing a completely new way of handling search and navigation for atmospheric and oceanographic data, which reaches down to individual geographically orientated features.</p>	<p>NERC expects us to role out the NDG in two years time, and then role it out over all disciplines in the coming years. We expect at that point to be on a roadmap towards a system which is secure and standards compliant, which allows secure (and where appropriate, chargeable) navigation seamlessly between all NERC's data holdings in such a way that data products can be overlaid and visualised using standard commercially available tools.</p> <p>In two years time we expect only to have in place such a system with secure (but not chargeable) access to NERC's primary atmospheric and oceanographic data</p>	<p>They are working on discovery interoperability with the Australian National Grid and the U.S. National Centre for Atmospheric Research and the U.S. Programme for Climate Model Diagnosis and They expect to start similar work with German colleagues shortly.</p>	

				<p>holdings, and to be outlining how we can take the technology to the wider discipline coverage of NERC.</p> <p>The technology will be open, and freely extensible, and based on OGC protocols; however, these will need to be extended, as they currently do not support everything we need.</p>		
EDINA Go-Geo! Portal	<p>Using an ANSI standard, Z39.50-1995, the portal undertakes simultaneous searching across a number of geospatial metadata catalogues found in UK tertiary education. It also cross searches a number of external catalogues including the national Gigateway service (see below) and its network of geospatial metadata catalogue services. A key feature is the ability for users to find other related resources, such as books, photographs, projects, maps, for their geographic area of interest. These resources are discovered by cross searching the JISC Information Environment and other online information services. The focus of the portal is therefore on where a resource about and less on the what it is about, which is the focus of other JISC portals.</p>	<p>Although still viewed 'a service in transition' by JISC it has actually been available to the UK academic community for 2.5 years and is well used and widely promoted on other web sites.</p>	<p>Funding comes from JISC for development work and from EDINA for 'service' operation, total approx £95K.</p>	<p>It has an active development programme which includes investigating linkages to geo-spatial data repositories and interoperability both in terms of metadata standards and with other portals through the deployment of web services.</p> <p>EDINA expects that like the NERC DataGrid portal its technology will be open, and freely extensible, and based on OGC protocols. The goal is for the Go-Geo! portal to move beyond telling users where they can find geospatial data, to permit interaction, integration and visualisation of retrieved and derived data.</p> <p>It is a core component of the UK academic spatial data infrastructure.</p>	<p>In the next phase the plan is to begin interoperating with geo-portals in other countries.</p>	
Gigateway	<p>Operated by the UK Association for Geographic Information and funded by the Office of the Deputy Prime Minister via NIMSA (National Interest Mapping Services Agreement), Gigateway provides metadata on geospatial data set. The service comprises a gateway which cross-searches, using Z39.50, catalogues hosted by government and research agencies.</p>	<p>Launched in 1999 (askG!raffe) it is a mature service whose future is now in some doubt because of questions over the future of NIMSA. Currently funding is only available to until end of Aug 2006.</p>	<p>The AGI budget for the service for 2004-5 was £316,000. This does not include the cost of setting up and operating the third party catalogues or of people creating and depositing metadata.</p> <p>There is a widely held view that it should be publicly funded for the national good.</p>	<p>Unclear.</p> <p>It is recognised that Gigateway, or something like it, will be required if the UK is to meet its commitments to INSPIRE - a Directive Of The European Parliament and Of The Council for establishing an infrastructure for spatial information in the European Community.</p>	<p>Has links with similar initiatives in other European countries and the US.</p>	

Aggregation tools – Grid services portals						
GROWL Project	<p>GROWL is a client interface to a Grid programming environment, and aims to provide scientists of various disciplines, including Social Science, who have a limited interest in and knowledge of computing with an easy vehicle to improve their access to computational power. The GROWL toolkit will provide client and wrappers to existing VRE resources and services developed in e-Science projects, integration of new common services (e.g. Condor, NetSolve and SRB) into GROWL, and produce clients for the National Grid Service.</p> <p>Funded by JISC under the VRE programme and EPSRC</p>	<p>GROWL is a modular programming toolkit. Currently it has modules for Authentication, File upload/ download, 3rd party File transfer, Globus job submission, Condor job submission, SRB file management. New modules are being added and the server side re-designed as a full multi-threaded C++ gateway application. Security is handled using Grid certificates.</p>	£150,00 (Feb 05 to Oct 06)	<p>GROWL is being extended in the ESRC-funded CQeSS project (Collaboratory for Quantitative e-Social Science). It is also written into a proposal to ESRC for a wider e-Infrastructure and is in the work plan for the NW-GRID middleware deployment.</p>	<p>A "Lightweight Grid" workshop is being held in April to compare with other solutions such as RealityGrid, AHE, GAT, GridSite, gLite, WSRF::Lite etc.</p>	
National Grid Service Portal	<p>NGS Portal is an end-user interface for applications on the four NGS nodes, with a possible extension to the affiliates. It uses the StringBeans Java framework and has HSR-168 compliant portlets for: Authentication, MyProxy certificate management, Globus job submission, Job monitoring, GridFTP file transfer, SRB file management and a Discussion Forum.</p>	<p>NGS Portal is being used in production mode from a BladeCentre at Daresbury. Additional tools are being developed and deployed based on user feedback and requirements. It is used in the NGS training course.</p>	Currently 0.5 FTE effort for development and maintenance	<p>Written into GOSC proposal recently funded by JCSR. Portlet interfaces to SRB, UDDI, OGSA-DAI, PayPal services have been demonstrated as prototypes. Shibboleth interface being added via the JISC-funded ShibGrid and SheBangs projects.</p>	<p>Developers at Daresbury have regular workshops/ meetings with international counterparts (GridSphere, GridPort, Sakai). A full report was produced in 2003. A workshop comparing JSR-168 frameworks was held in March 2005.</p>	
myGrid	<p>myGrid provides a language and software tools to facilitate the easy use of workflow and distributed compute technologies for the life science community. myGrid also supports the e-science life cycle, capturing experimental provenance and providing semantic web technologies to help in workflow construction.</p> <p>myGrid was originally an EPSRC pilot e-science project, this was followed by an</p>	<p>myGrid has followed two development tracks – core development and research prototyping. The core myGrid components, including the flagship workflow component, Taverna, offer stable releases and support an</p>	£1.2 million over three years	<p>myGrid has recently become an OMII-UK node. As part of OMII-UK, myGrid will be made into production-grade software and joins with the OGSA-DAI data access toolkit from Edinburgh and the original OMII from Southampton.</p>	<p>myGrid has a user community spread throughout the world. We also have collaborators from other countries, for example, BioMoby, and the Virginia Bioinformatics Institute (see the</p>	

	EPSRC platform grant and it is now funded by the OMII (Open Middleware Infrastructure Institute)	active user community. To date, there have been 13600 downloads			website http://www.mygrid.org.uk/index.php?module=pagemaster&PAGE_user_op=view_page&PAGE_id=70&MMN_position=82:82 for more details) Also, the LinK-Up - e-Science sisters programme provided an opportunity to collaborate with other Grid projects in the states (See Link-Up website http://www.mygrid.org.uk/linkup/)	
Aggregation tools – specialist directories						
ARCHON Directory	The ARCHON Directory includes contact details for record repositories in the United Kingdom and also for institutions elsewhere in the world which have substantial collections of manuscripts noted under the indexes to the National Register of Archives. Funded by the National Archives					
Source data services and repositories – specialist browsing tools						
Ensembl	Ensembl is a joint project between EMBL - European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute (WTSI) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust .	Release of data and analysis into the public domain immediately. Open, collaborative software development: Ensembl imposes no restrictions on access to, or use of, the data provided and the software				

		used to analyse and present it. Collaboration on agreed standards for distribution. Timely development.				
GenBank This is US funded so should this be here? Also I'm nor sure if this does a cross search?	DNA Sequence Database (repository) GenBank is maintained by the National Center for Biotechnology Information (NCBI), a part of the National Library of Medicine, National Institutes of Health. Submitters to GenBank currently contribute over 3 million new DNA sequences per month to the database. More information about GenBank may be found on the NCBI Web site at http://www.ncbi.nlm.nih.gov	DNA sequence database has exceeded 100 gigabases.				
DOAR	Initially called OpenDOAR, DOAR aims to be an authoritative, comprehensive and quality assured list of Open Access Repositories that contain research information. The project team is undertaking a survey of the available repositories to determine the scope and scale of the developing repository network. They intend to create a classification structure of such repositories, with a comprehensive set of metadata to describe each one. The list is designed to provide a bridge between repository administrators and third-party service providers, allowing m2m interfaces for service provision. In this we hope that the list will facilitate the development of innovative services and repository metadata. Project is partnership between University of Nottingham, UK and Lund University, Sweden. See: www.opendoar.org	This is a project, funding for which lasts into the summer of 2006. It has a public, end-user interface. The initial list covers approx. 340 repositories. Each has been checked personally, to eliminate dead-links, out-of-scope repositories, etc. Version 2 will depend on the completion of the survey and categorisation work. Service providers have been contacted and have expressed interest in working with OpenDOAR.	circa 2FTE - £65k pa plus costs. For the future duration of the project, both partner sites have full-time researchers involved. University of Nottingham has a full-time technical post funded.	Completion of initial project plan - survey, categorisation, liaison with service providers, release of Version 2 list.	Partnership is international, between the UK and Sweden. Funding comes from the UAS-based OSI, SPARCEurope, and the UK-based CURL and JISC Response and support from repository administrators, and service providers has been international.	
Source data services and repositories – full text content						

Scopus	Commercial - Subscription Bibliographic and citation linking database + Full text linkage. D-L-R-A Heavy full text linkage via Science Direct (Elsevier's Full-Text package) in fact at this time, you have to subscribe to Science Direct to get Scopus.	Approx 30 million records Also approx 10 million records are citation enhanced Also the 'web' results from Scirus. 18 months old Back file to approx 1965.	Unknown	Unknown - probably looking to global ubiquity of this product ahead of any further enhancements to the service		
Science Direct	Elsevier's Full Text delivery Platform (see Scopus) Comprises Journals (over 2000 titles) and Books (unknown number) (at the moment). D-L-R-A To push the entirety of Elsevier's Scientific Full text content to scholars via an institutional subscription.	2000+ journals Backfile of 6.75 million articles (full-text) Major Books and Reference works collection also available. Approx 5 years (? Check).	Unknown	Science Direct +Scopus and Scirus appear to form the backbone of the Elsevier content delivery strategy for the next 5 - 10 years.		
Generic services & products – search engines						
Google	Web Search Engine (websites, maps, books, video, images, misc other data objects). D-L-R and now starting to think about actual delivery "Google's mission is to organize the world's information and make it universally accessible and useful"	Effectively the complete visible web (and an increasing amount of the hidden web) index measures in the billions. <10 years.	Cap Ex est at US\$1 billion in 2006	Clearly heading beyond indexing to being a provider (and holder) of access to discrete digital items (eg Google video, various Google software initiatives) Aim to be the place to go. Google believes that machines can bear the brunt of indexing, sorting and presenting disparate datatypes. And they invest accordingly.		
Google Scholar	Web Search engine dedicated to "Scholarly Content" Current consists of an index based on a set of PubMed data, plus other data sources from multiple publishers. Note: As of Mar 2006 Elsevier NOT in GS. Also note, Definition of "Scholarly" open to interpretation and also not explicit. No documentation on	Difficult to estimate. Smaller than PubMed and certainly not consistently updated at this time. Note: GS is a beta.	Unknown - as of late 2005 Google had allocated 2 full time engineers to the project, but was looking to focus more on this area.	Unknown		Online

	actual coverage. D-L (Some R and A if Library has signed up to program)					
Yahoo	Web Search Engine (websites, maps, video, images, misc other data objects). D-L-R and via partners programme more able to deliver now.	Effectively the complete visible web (and an increasing amount of the hidden web) index measures in the billions. About 10 years.	Cap Ex est at US\$400-500 million in 2006	Seem to be investing heavily in various "folksonomy" start-ups. Yahoo seem to believe that machines AND humans (in Yahoo's case end users) can organise data better than machines alone.		
MSN (Microsoft)	Web Search Engine (Part of Microsoft Corp - global software conglomerate). D-L From operating systems to the internet (and now home entertainment) Microsoft want's to be the glue that connects user devices together. Search is one aspect of their work.	95% of the planet' s computers run a Microsoft Application. 25 years (MSN the web portal and search engine about 10 years)	No hard data - assume to be similar to Yahoo? Or more like Google (if you bundle in the development streams for their various products).	http://www.live.com/ Looks as though Microsoft intend to cross link their OS (Vista the latest version due out in 2006 in theory) and their office suite of applications to derive some species of online services product where users can tap in to a multitude of resources (for a payment no doubt)		
Commercial suppliers of published content/services						
Amazon	They sell books (and other products)! Big index of published material and the 'Amazon identifier' has become a very sophisticated ISBN type identifier for online services. D-L-R-A Find the book, buy the book. review the book, have other books suggested to you, see what others bought. A good paradigm for context driven improvements to search and navigation issues.	15% of the total book market in the US goes through Amazon.com. similar numbers starting to be seen for Amazon.co.uk and other national variants. The single largest online bookstore. 10 years old.		Unknown - but assume circa \$ 50-100 million	Upgrade your book purchase to include an online version accessible globally via amazon. Also seriously investigating the issues involved in deconstructing the book into granular sections for purchase. Also looking at the mechanisms for successful delivery of online books across multiple electronic platforms.	

Underpinning framework – R&D activities						
Text Mining	The National Centre for Text Mining (NaCTeM) provides text mining services to the UK academic community. The Centre provides software tools and services which will allow researchers to apply text mining techniques to problems in their areas of interest. (Initially Biomedical domain)	Developing a pilot service.	£979,00 (2004 – 2007)	Possible CSR funds (£500,00) under the e-Infrastructure strand will involve text mining to extend beyond Biomedical domain and to broaden benefits of the Grid. CSR funds (£600,00) for AHRC projects. Some projects may involve text mining.	Inspired by NaCTeM, other counties are considering setting up such centres e.g. Germany. Other research projects:	
Data Mining	Pervasive as part of the above services					
Underpinning framework – projects						
HILT	High Level Thesaurus: focussing - on terminology and thesauri requirements at the collection level, but also bearing in mind the need to extend this in due course to the needs of item level retrieval.	The project is charged with creating an M2M demonstrator that will: Offer web-services access via the (SOAP-based) SRW protocol , but be designed so that a possible extension offering other protocols (Z39.50 , or SRU , for example) at a later date could be an option. Use SKOS-Core as the 'mark-up' for sending out terminology sets and classification data responses but be designed so that adding other formats such as MARC and Zthes would be an option at a later date.	Current funding: 2005/06 £59,000	Future developments dependent on outcomes of the current phase and terminologies study being conducted at UKOLN and other work in this area.		
JISC Information Environment	The IESR has been developed to provide a registry of information about electronic resources that are of value to teachers,	Demonstrator project. The aim is to create a reliable source of	£ 424,306 total November 2001 to April	The funding of the current phase of the IESR will end on 31st July 2006. The project team intend to submit a proposal for	USA - The OCKHAM Digital Library Services Registry:	

<p>Service Registry (IESR)</p>	<p>researchers and learners. The IESR project is part of JISC's Shared Services Programme.</p>	<p>information that other applications, such as portals, can freely access through machine-to-machine protocols, in order to help their end users discover resources of assistance to them.</p> <p>The IESR contains information about the resources themselves, technical details about how to access the resources, and contact details for the resource providers. For resource providers the IESR will hold a master description of their electronic resources, to which other potential users of the resources may be directed.</p> <p>The IESR currently holds this information for a selected set of electronic resources within the JISC Information Environment, provided by:</p> <p>Arts and Humanities Data Service (AHDS) Edina MIMAS Resource Discovery Network UK Data Archive UK Mirror Service (now JISC National Mirror Service)</p>	<p>2006</p>	<p>continuing funding to enable the future development and maintenance of the IESR. It is expected that the focus of the proposal will be on the steps required to move the IESR from a project to a service-in-development. Important elements of the next phase</p> <p>include: increasing the usefulness of the IESR by encouraging the creation and maintenance of content; demonstrating use of the IESR through collaboration with other JISC shared services, repository and portal projects; continuing to disseminate information about the registry.</p>	<p>http://www.ockham.org/ Funded by National Science Foundation. They used the IESR metadata schema.</p>	
<p>GeoXwalk</p>	<p>geoXwalk is JISC funded middleware implementing a digital gazetteer service and server for the UK academic Higher and Further Education community. The rationale</p>	<p>It currently is part of the Go-Geo! Portal functionality as mentioned above. Funded by the JISC</p>	<p>Current funding 2005/06 £40,000</p>	<p>The JISC is investigating how to move this project from development to service.</p>		

	<p>behind the project is that there is currently no unified entry point to assist in geographic searching within the existing academic network as each information provider/service adopts different geographic coding conventions (some use postcodes, others placenames, some grid references etc.). geoXwalk is designed to make geographic searching transparent by 'crosswalking' these different geographies.</p>	<p>Shared Services programme.</p>				
--	---	-----------------------------------	--	--	--	--