INTERNATIONAL DIGITAL LIBRARIES – NSF 99-6

The IMesh Toolkit

An architecture and toolkit for distributed subject gateways

A proposal submitted to the JISC and the NSF under the International Digital Libraries initiative by the University of Bath (UK Office for Library and Information Networking), the University of Bristol (Institute for Learning and Research Technologies), and the Internet Scout Project, Computer Science Department, University of Wisconsin-Madison.

19 January 1999

1 Project summary

Background and objectives. Recent years have seen the emergence of the subject gateway approach to Internet resource discovery. In this model, databases of resource descriptions are built up through manual selection, 'cataloguing' and classification. This approach is now the subject of significant national initiatives in the UK (the Resource Discovery Network), the US (Project ISAAC), Australia, The Netherlands, Finland, and elsewhere. Recently, the leading gateway initiatives have come together in an informal consensus-making forum under the name IMesh (international mesh). The project partners are leading players in IMesh, and this proposal aims to advance the systems framework within which subject gateways and related services operate.

Various design and technical choices have been made in subject gateway services, however there are some common features: a metadata format (Dublin Core or IAFA templates, for example), a search and retrieve protocol (Z39.50, LDAP and Whois++, for example), and a mechanism for routing queries between gateways (the Common Indexing Protocol). There has been no common design approach. Subject gateway technology is now at a point where an architectural approach is necessary. Individually developed software tools cannot be expected to work together automatically; they can be 'glued' together using custom-developed code but this approach is not scalable or maintainable. What is required is an overall architecture which specifies individual components and how they communicate. Software with a well-defined architecture is known to be more maintainable, extensible and reusable. This project will define, implement and test such an architecture through the provision of an integrated toolkit for subject gateways and other metadata creators.

The strategic aims of this proposal are

- to develop an overall framework for subject gateways which levers individual development effort by supporting reuse of tools and metadata;
- to provide a more robust framework for interoperating between subject gateways and between those gateways and other network information services;
- to create a favorable environment for the development of systems and services by commercial and public sector players by providing a technology base-line.

These will be realized through the following specific objectives:

- to develop an architecture for subject gateways (and other metadata creation, management and access systems),
- to develop or identify existing APIs (programming interfaces) between the principal components of such an architecture to allow them to communicate,
- to integrate and distribute a set of tools for subject gateway management based on this architecture (some tools will be developed within the project, some through other work of the partners, and some will be third party tools),
- to develop an integrated development environment for metadata,
- to develop a metadata registry (which allows human or computer users to obtain information about metadata attributes, an increasingly necessary service in a distributed environment),
- to research issues of deployment and use, and
- to widely disseminate project lessons, approaches and software.

These are ambitious objectives. However, it should be noted that a major ambition is to lever other work the partners are engaged in, and related work elsewhere, and to add value to it by reusing it within a consistent framework.

Method. The project will work through desk research, software design and development, system development, and testing in operational services. The partners share development and operational goals, and have theoretical and practical interests in the work presented here.

Impact. The work proposed will have several impacts. It will provide some components and a development framework to software developers, improving the quality and sustainability of system development in this area and reducing entry costs. It will provide subject gateways with a modular and flexible approach to service development in a multi-protocol and multi-format environment, allowing them to hide more of the technical complexity of cross searching and merging of results from users. It will encourage more potential metadata providers to publish their resources, again, as entry costs are reduced. Recent developments have seen tremendous growth in subject gateway provision. The project will encourage that growth by providing a toolkit which promotes interworking, reuse of tools and metadata, and distributed working.

2 Table of Contents

3 Project Description

3.1 Introduction

This proposal comes from three initiatives with a strong UK and US presence, which are seen as international leaders in the subject gateway approach. The Internet Scout Project (ISP) [ISP] manages The Isaac Network [Isaac Network], a US based project that enables searching over distributed and heterogeneous resource collections. ILRT [ILRT] runs the SOSIG [SOSIG] and Biz/Ed eLib funded subject gateways and manage the ROADS (Resource Organisation and Discovery in Subject-based services) [ROADS] and DESIRE (Development of a European Service for Information on Research and Education) [DESIRE] projects. UKOLN [UKOLN] are partners in the ROADS, DESIRE, PRIDE (People and Resource Identification in Distributed Environments), BIBLINK [BIBLINK] and CHIC (Cooperative Hierarchical Indexing Coordination) [TF-CHIC] projects. UKOLN is also a joint provider of the Resource Discovery Network Centre (RDNC), a framework put in place by the JISC to coordinate the pioneering UK subject gateways, and to develop the network in various ways. The UK subject gateways are forming the Resource Discovery Network (RDN).

The Isaac Network and the RDN are taking comparable technical approaches based on a directory service that allows services to act autonomously but also participate in a cross-searching 'mesh' of servers. The Isaac Network uses LDAP [RFC 1777; RFC 2251], Dublin Core [Dublin Core; RFC 2413], and CIP [Allen & Mealling 1998]. RDNC services use ROADS, a set of tools that use Whois++ [RFC 1835], IAFA templates, and CIP. ROADS is also working to integrate LDAP and Z39.50 [ANSI/NISO Z39.50-1995; ISO 23950:1998] based services. Each service is developing a set of tools to support creation, administration, and use of resource descriptions in a managed environment.

The IMesh toolkit project will:

- develop a metadata management toolkit which integrates the Isaac Network and ROADS approaches, while also making data available in a Z39.50 environment
- explore end-user and service provision issues in international meshes.

The toolkit will allow for distributed searching over international networks using the dominant directory protocols in use today. This will result in a package that has been developed with both UK and US needs in mind, and will negate duplicate work being done simultaneously by both projects. Also, because both Isaac and RDN gateways are active in current "IMesh Guidelines" work [IMesh], which are international discussions including representatives from over 15 countries, the work done in this joint project can also take into account any similar work being done in other parts of the world.

An increasing number of subject gateway services, including the Resource Discovery Network members in the UK and the Isaac Network in the USA, are to be found on the Web. These services are significant in the future of resource discovery. In order to meet expectations a number of end-user and service provision issues must be addressed and the technology to support the findings must be developed. Areas in which current generation subject gateway technology is lacking include scalability, usability, interoperability, metadata sharing, authentication and charging. This project proposes an integrated approach to addressing key issues associated with moving subject gateway technology to the next level.

New and emerging technologies that were not available to current generation tools and which are expected to play a role in an integrated approach include: the Resource Description Framework (RDF) and its XML representation [RDF] which provide a standard way of representing and sharing metadata; recent developments with the Dublin Core as it moves to a more structured Version 2; new query routing technologies; the proposed W3C Query Language which will provide a standard way of querying XML/RDF information.

Existing projects including ROADS, the Isaac Network, DESIRE and ASF [ASF] have provided and are continuing to provide subject gateway toolkits and the tools that populate those toolkits. Each of these projects has made decisions regarding the metadata formats (Dublin Core, GILS, etc) and the query protocols (Z39.50, LDAP, Whois++) that they support.

In general, the tools developed within currently available toolkits are specific to the technical and architectural context of the toolkits. For example, it is not possible to take tools (such as a link checkers or language translators) from one toolkit and use them in another. This leads to duplicated effort for toolkit developers. There is now a need for an integrated approach in which modular components can be developed within the context of one toolkit and reused by other toolkits.

A further issue for current generation toolkits has been the introduction of cross-searching. The need for cross-searching across multiple protocols and multiple metadata formats has led to the development of a number of specific protocol to protocol gateways, usually with built in format conversion (DESIRE, CHIC). This approach is not scaleable since the introduction of a new query protocol, such as the proposed W3C Query Language, requires modification to all toolkits that wish to support it. This modification may be substantial since each toolkit may wish to be able to cross-search subject gateways that support the new standard and to be able to serve its own metadata via the new standard.

The next generation of subject gateways can be built with the benefit of hindsight, it is now clear that a number of query protocols will need to be supported and that new standards can be expected to emerge. It is also clear that if multiple toolkits continue to develop toolkit-specific tools then a lot of effort will be duplicated. What is needed is a framework that allows tools, including protocol specific gateways, to be slotted in. A framework would include an overall architecture, APIs that allow software developed by different parties to interoperate, and generic software components that are required by all subject gateways. Such a framework would allow tools such as protocol specific gateways to be slotted in.

This project will contribute to the development of a subject gateway framework. The architectural aspect of the framework will build on ongoing MIA (MODELS Information Architecture) work being carried out within MODELS [MODELS] and will also be influenced by the architectures of existing subject gateway toolkits. The CORBA-based InfoBus architecture of the Stanford Digital Libraries Project [Stanford Digital Libraries Project] also provides input into this work. The ROADS project has developed APIs to separate the front-end back ends of the ROADS toolkit, a similar approach has been taken in the development of ASF. Where appropriate existing APIs will be used or developed further.

The current generation of metadata management tools (including metadata editors, metadata format converters and metadata harvesters and indexers) tend to operate based on fixed, embedded representations of metadata formats. This causes problems when new metadata formats need to be supported. The problem occurs not only when new metadata standards, such as Dublin Core version 2, are released, but also when there is a requirement to support application specific metadata formats which may require extended attributes sets, or place restrictions on particular attributes. The concept of schema registries has been introduced as a way of managing this complexity [Workshop on Metadata Registries 1997]. A metadata registry would be able to provide both human-readable and machine-readable definitions of schemas, and of mappings between schemas. Metadata management tools would then we able to consult a metadata registry for metadata formats, enabling tools to support a wide range of metadata formats without needing to write new code for each one. The Resource Description Framework (RDF) and its representation in XML provide a basis for the provision of machine-readable schemas and mapping between them.

Work on schema registries is currently at an early stage. This project will build on work carried out in DESIRE and elsewhere and provide the tools required to operate a metadata registry service.

The project intends to facilitate the creation of metadata in a distributed environment. It will investigate the possibilities of enhancement and inheritance of existing metadata in order to reduce duplication of effort and encourage sharing of metadata between different communities. This will build on work undertaken in the BIBLINK project [Day et al. 1999] to use common metadata workspaces.

This proposal emphasizes the requirements of subject gateways. However, the approach discussed here is of wider applicability and the partners will be collectively and individually working to promote the benefits and wider use of results. In the UK there is particular interest in collaborating with the hybrid library projects (UKOLN is a partner in Agora, one of the hybrid projects), and in supporting the emergence of the DNER (the Distributed National Electronic Resource). It is anticipated that the toolkit will be useful for developers of hybrid libraries and other network information services.

3.2 Overall Approach

3.2.1 Scope

The scope of this project is the technology and end-user and service provider issues involved in the provision of subject gateway services (such as SOSIG in the UK and the Isaac Network in the USA).

The focus is on providing an integrated IMesh Architecture that will support the following activities:

- The development of an IMesh Framework which will specify the Application Program Interfaces (APIs) via which software components communicate.

- The integration of existing subject gateway tools to provide a unified, configurable and fully-featured toolkit from which new subject gateways can be built and which can be extended in order to provide new tools.
- The investigation of end-user and service provision issues.
- The development of new IMesh compliant tools to address end-user and service provision requirements. (Some IMesh compliant tools will be developed as a part of this work and third parties will be able to develop additional tools.)

3.2.2 Architecture

Individually developed software tools cannot be expected to work together automatically; they can be `glued' together using custom-developed code but this approach is not scalable or maintainable. What is required is an overall architecture which specifies individual components and how they communicate. Software with a well-defined (and well-documented) architecture is known to be more maintainable, extensible and reusable [Bass 98]. Subject gateway technology is at a point where an architectural approach is a necessity and such an approach is at the core of this proposal.

An IMesh architecture and corresponding Application Program Interfaces (APIs) will be developed. This will provide a basis for the integration of independently developed tools from projects such as ROADS, Isaac, DESIRE, and ASF. This approach will avoid replication of existing functionality in favour of modularizing and integrating existing tools into the IMesh framework. In areas where required tools do not exist - this is likely to be the case for Metadata Registries and RDF support especially - they will be developed within this project according to the IMesh APIs.

3.2.3 High-Level IMesh Architecture

A fully specified IMesh Architecture will be developed within the early stages of the project, however it is useful to have a high-level architecture for the purposes of this proposal. This architecture is shown in figure 1.

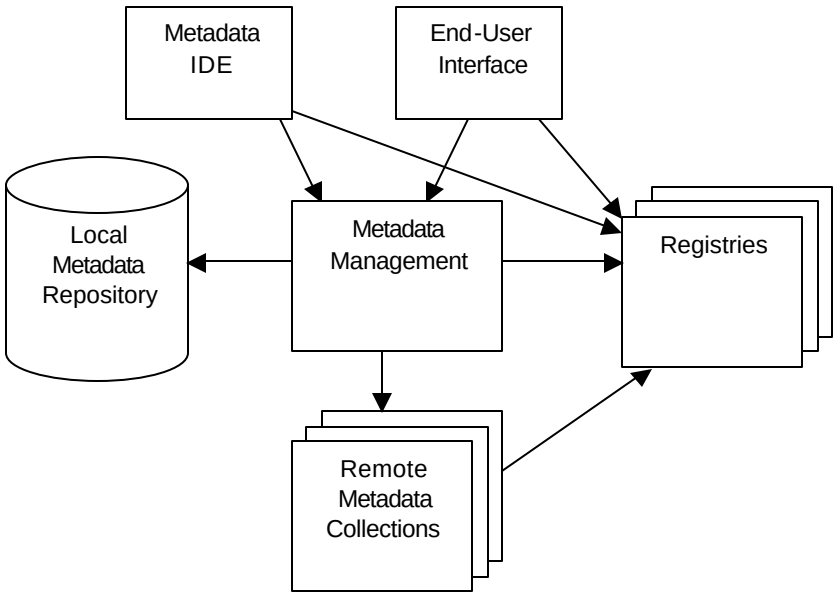

Fig 1: High-level overview of Cross-Searching Framework and Metadata IDE

The elements of this high-level architecture are:

Metadata Integrated Development Environment (IDE) - The Metadata IDE is the (web-based) application via which cataloguers enter metadata into an IMesh Subject Gateway.

End-User Interface - The End-User Interface is the (web-based) application that allows users to query an IMesh Subject Gateway (and the subject gateways it cross-searches).

Metadata Management - The Metadata Management Subsystem is at the center of the IMesh architecture. It manages access to an optional local repository of metadata and access to remote subject gateways (including forward knowledge and native protocol s upport).

Registries - The IMesh framework will rely on a Metadata Registry for definitions of metadata formats, this will enable IMesh Subject Gateways to support new or variant formats without software modification.

Local Metadata Repository - Each subject gateway that wishes to catalogue resources (as opposed to just cross-searching existing collections) will have a local metadata repository, this will be accessed via the metadata management subsystem.

Remote Metadata Collections - Subject gateways may cross-search remote collections which may be IMesh compliant (i.e. built according to the IMesh architecture and APIs) or they may use a supported query protocol.

3.3 Plan of Work

This project will undertake work in two broad areas, researching the key issues associated with the development of an international mesh of subject gateways and developing the IMesh Toolkit. The product of this research will be a framework, as outlined above, and a software toolkit that will allow librarians and other information providers to create metadata records, store these records in collections, make these collections accessible over a variety of distributed search and retrieval protocols and allow the creation of search services that can query the metadata collections. The software tools will make it easier for organizations to create collections of metadata and apply metadata to existing collections of Internet resources. For example, a library may currently have a collection of World Wide Web "links" regarding a particular topic, such as history. Using the software tools, that library will be able to annotate the links, apply subject information from controlled vocabularies (for example, Library of Congress Subject Headings), and make their collection of history resources searchable by users on the Internet using specialized search clients (such as Z39.50 or Whois++) or searchable from Web-based search services. In addition, the software tools will make it possible for other institutions with similar collections to include the collection from our example as part of a larger, "brokered" collection. In this way, organizations can configure subject gateways that encompass multiple collections, providing the academic and research community with the ability to search multiple collections of high-quality resources with a single search command. Currently, making such a collection available to other institutions involves much technical work. In addition, the resulting collection is likely to be a "stand-alone" collection, accessible only if you (1) already know of the collection's existence and (2) have access to a client that implements the proper protocol to search the collection and retrieve the results. So, the results of this project should significantly lower the barriers involved in creating and sharing these collections of high-quality resource.

3.3.1 Research into End-User and Service Provision Issues

Development of the IMesh framework and tools will be informed by research into the key end-user and service provision issues associated with the development of international meshes of subject gateways.

Issues that currently require investigation include:

- **Metadata Sharing** - As the number of subject gateways increases it is inevitable that resources will be catalogued multiple times. It is not desirable to have multiple similar descriptions of the same resource but currently there is no tool support to prevent this. It is however desirable to have different metadata about the same resource, for instance SOSIG may catalogue a resource with a social science bias whereas EEVL [EEVL] may describe the same resource with an engineering bias. In this case it is necessary to look at ways of combining this information in a way that is meaningful to the user.
- **Resource Mirroring** - A mechanism is needed to allow subject gateways to direct users towards the closest (in network terms) copy of a resource. URNs/persistent URLs would need to be assigned to resources. Current subject gateways use URLs to refer to resources rather than permanent identifiers.
- **Query Routing** - If subject gateways are indiscriminately cross-searched by large numbers of other subject gateways then they are likely to become overloaded. To avoid such problems it is necessary to use query routing with technology such as CIP/centroids and Tagged Index Objects (TIOs) to send queries only to services *likely* to be able to provide a positive response. Query routing must avoid circular routing of queries and needs to develop to avoid sending queries to multiple mirrors of the same metadata. This is particularly relevant in the context of international subject gateway collaboration. For example, the US mirror of SOSIG, held at ISP, might usefully be a component of a distributed search mesh, but should not result in multiple 'hits' for the same record.

- **Service Directory** - As large numbers of subject gateways are developed, some of which may be mirrors of others and some of which may cross-search each other, it will become increasingly necessary for subject gateways and end-users to have access to directories of service descriptions.
- **Reducing Delay in Query Processing** - Cross-searching a large number of subject gateways is time-consuming. Mechanisms for both reducing the delay (parallel query processing, stateful connections) and minimizing the impact of delays on users (displaying results as they arrive rather than waiting for all results) need to be investigated. A further strategy to be considered is results caching where popular or recent query results are cached, avoiding the need to repeat the query when it is requested again.
- **User Interface Design** - As well as being able to provide underlying subject gateway technology it is also necessary to ensure that users can effectively make use of that technology. This is especially true when users are cross-searching (or cross-browsing) multiple subject gateways.
- **Authentication and Charging** - Metadata is a valuable resource; subject gateways are likely to want to know who is accessing their metadata records and may need the ability to restrict access in some cases, or even charge for access.

Ongoing work within the DESIRE project and elsewhere is expected to address some of these issues to a greater or lesser degree. It will be necessary to evaluate the current status of work at the beginning of this project. In order to direct effort towards issues that are considered important by the subject gateway community and which have not been fully investigated elsewhere, it will be necessary to carry out the following evaluations:

1. **Technology Evaluation** - an evaluation of the current status of research and tool development in the above areas.
2. **Evaluation of Subject Gateway Requirements** - an evaluation of issues considered important by RDN members in the UK and members of the Isaac Network in the USA will be carried out.

In addition to these initial evaluations it will be necessary to continually evaluate ongoing work and current user requirements throughout the project.

3.3.2 The IMesh Toolkit

The IMesh Toolkit will build on the work of ROADS, the Isaac Network, ASF and the DESIRE project to develop components that enable distributed searching over international networks and the creation and management of metadata in a variety of formats. It is anticipated that many of the components in the toolkit will already exist, some will be based on existing work with minimal modification and some will have to be developed from scratch. The Resource Description Framework will be used as the basis for the data-structures used within the toolkit. Emphasis will be placed on support for Dublin Core, initially version 1 but moving to version 2 as it is developed.

The toolkit will comprise:

3.3.2.1 Cross-Searching Framework

This work provides the basis for developing a 'plug-and-play' approach to the development of distributed cross-searching services, providing support for the dominant search and retrieve and directory protocols in use today including LDAP, Whois++ and Z39.50. Rather than building a suite of protocol specific gateways, a framework will be developed into which various protocol-specific components can be slotted. This modular approach will reduce duplicated development across multiple search protocols and will provide an extensible framework into which future developments, for example the W3C search protocol, can be fitted. Research is necessary to determine the specific components necessary for such a framework, however the key components are likely to include:

- **Metadata Repository Component** - A repository for metadata will be created that can store metadata in a variety of formats. The repository will provide a common "database" usable by the other modules within the framework and by the Metadata IDE. The repository will likely use RDF to encapsulate the metadata.
- **Forward Knowledge Component** - Forward knowledge references are a key component of building a "mesh" of servers. The Toolkit will use the Common Indexing Protocol Version 3 (CIP V3) as the protocol for exchanging forward knowledge between servers. Currently the ROADS and Isaac software systems support CIP V3. CIP support will be added for servers utilizing the Z39.50 protocol. The CIP Tagged Index Object [Hedburg et al.1994] will initially be supported by the Toolkit. Research may need to be done to standardize the payload of the Tagged Index Object for use within the Toolkit, to resolve issues such as whether to stem terms, etc.

- **Registry Components** - Rather than having embedded schema information, tools within the framework will query a Metadata Registry for information about the element sets and metadata formats in use. Tools will also query a Collection/Service Directory to obtain information about available services. This information could be used to build a list of services for cross-searching for example.
- **Result Combination Component** - This component will combine results received from multiple servers to remove duplicates and rank results. Initially, simple ranking schemes, such as frequency counts of query terms will be supported. Later, more advanced methods of ranking will be incorporated.
- **Protocol Modules** - providing support for LDAP, Whois++ and Z39.50. The protocol modules will allow metadata records stored in the Metadata repository to be created, modified, queried and fetched. Both client and server modules will be created for each of the three supported protocols. Server modules will support the creation of subject gateways while client modules will be used in the creation of search interfaces and the development of the Metadata IDE.
- **Format converters** - providing conversion between the main metadata formats in use by the protocols above, including RDF, Dublin Core, ROADS/IAFA templates [Deutsch et al. 1994], MARC and GRS-1.

Each component of the Subject Gateway Cross-Searching Framework will have a well-defined API. This will allow individual components to be enhanced in the future without modifying others in the framework. Where possible, components developed elsewhere will be used.

The Framework will allow the creation of several types of services. For example, the toolkit will allow an administrator to create a subject gateway that consists of a number of individual collections of metadata. Collections may be local (stored in a local metadata repository) or remote.

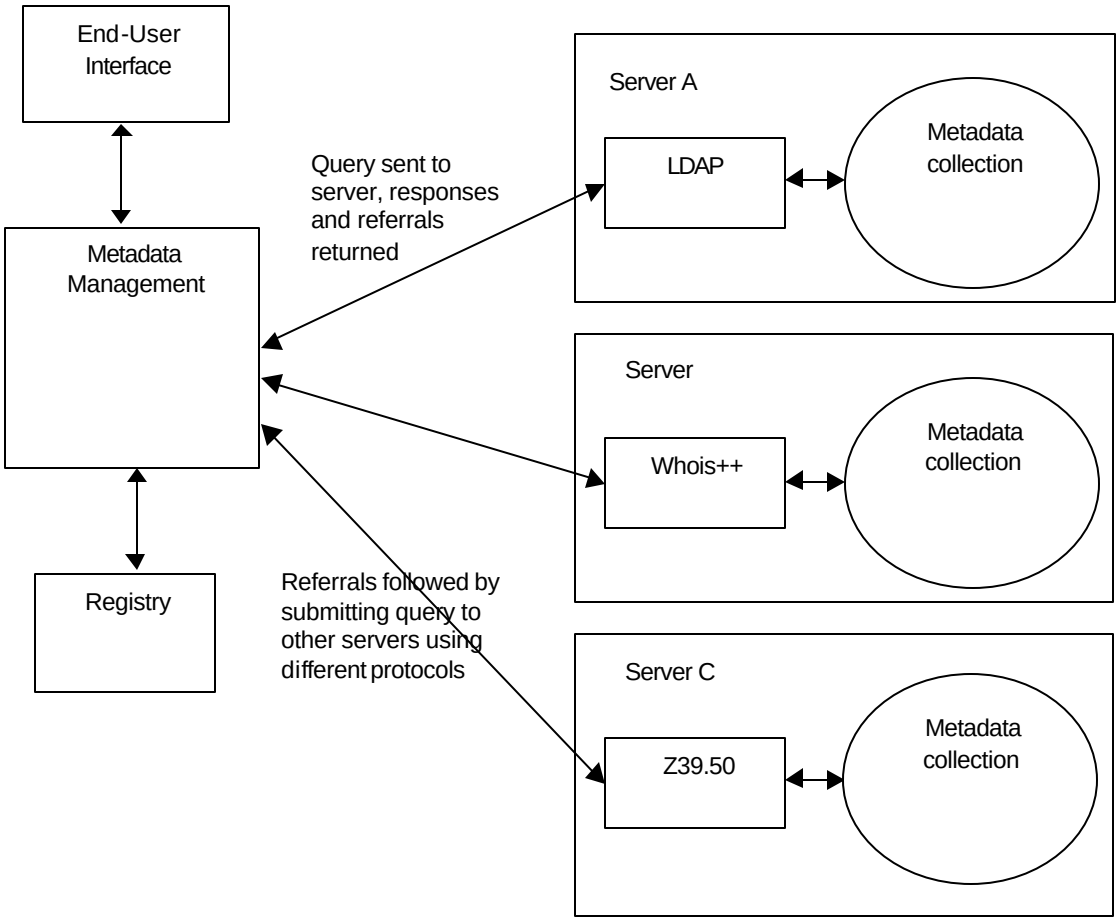

Fig 2: Multiprotocol client can query servers using different protocols

The toolkit will allow the creation of a local subject gateway that uses only metadata in a local collection. But the power of the toolkit lies in its ability to allow the creation of "virtual" subject gateways that consist solely of other metadata collections. One could configure a subject gateway that brokers dozens of collections, stored in different formats using different protocols, from around the world, combining the best collections of resources on a given subject area, and make it searchable as a single entity by users.

The toolkit will allow the creation of multi-protocol search services. Using the Toolkit components, one could create a network of metadata collections using different protocols. Using forward knowledge in the form of CIP V3 index objects, ROADS installations using Whois++ and Isaac Network installations using LDAP could be cross searched. An HTTP-CGI program will be able to use client protocol modules from the Toolkit to submit a query to an LDAP server that has index information from Whois++ servers. When a user submits a query, the search produces results (records from the local collection) and referrals to other servers. The other servers may be using any of the supported protocols. The HTTP-CGI program can "chase" referrals using the native protocol of each server.

3.3.2.2 Metadata Integrated Development Environment

A Metadata IDE enables users to: automatically generate metadata for a resource; edit existing metadata and create new metadata with access to documentation relevant to the metadata format (e.g. Dublin Core) that they are working with; convert between metadata formats; save metadata (the actual storage of metadata will be handled by other IMesh Framework components) and export metadata for use in other applications. The Metadata IDE will rely on human-readable and machine-readable definitions of metadata formats.

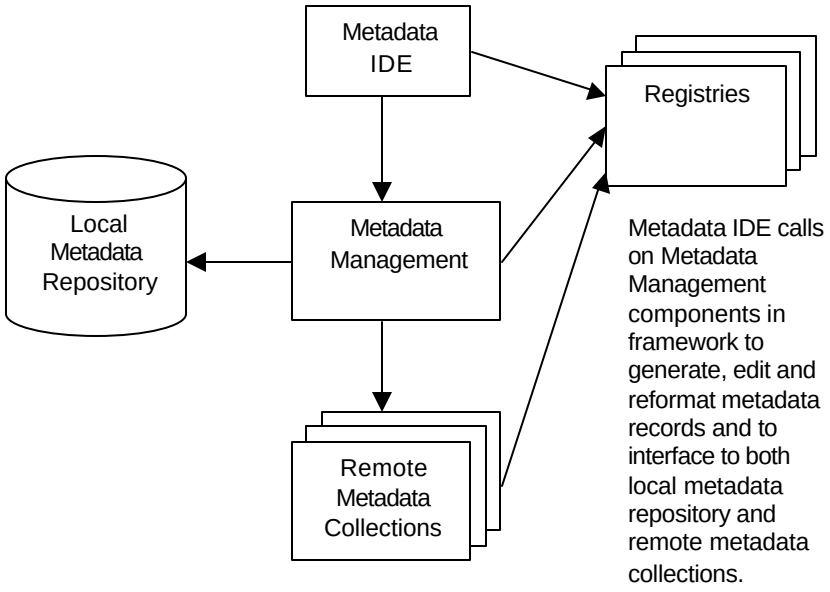

Fig 3: Metadata IDE

The Metadata IDE will extend the Framework, adding the following components:

- Editing Component - The Metadata IDE will allow the contents of metadata records to be edited and updated on servers utilizing any of the supported protocols.
- Auto-generation Component - Software will be developed that can extract metadata embedded within Web resources (including embedded Dublin Core in HTML META tags or RDF/XML). When the user enters a URL, the metadata will be extracted and formatted for inclusion in the metadata editor.
- Re-formatting Component – Using information from the Metadata Registry, the Metadata IDE will allow users to reformat metadata records from one supported metadata format into another. For example, a user will be able to view a ROADS/IAFA Template record and reformat it into Dublin Core or MARC format.

Where possible the Metadata IDE will make use of the Cross-Searching Framework components described above. Again, each component will have a well-defined API. This will allow individual components of the IDE to be re-used in a variety of ways. The UKOLN metadata creation tool DC-dot will provide some background for this piece of work.

3.3.2.3 Registry Development

The components developed in the two parts of the IMesh Toolkit described above will not have information about metadata formats hard-coded into them. Instead they will make use of a Metadata Registry to obtain information about the metadata formats in use. The development of a registry will build on work done in the DESIRE project and elsewhere. Schema Registries are likely to include machine-readable schema definitions (using, for example, the RDF schema definition language), human readable information about the schemas available and application profiles. Application profiles detail the specific use of schemas and formats as used by a particular service. In addition, framework components will be able to obtain information about available services from a Collection/Service Directory. The project will develop the tools necessary to build and maintain both Schema Registries and Collection/Service Directories, for example RDF Schema editors.

3.4 Project Benefits

This project will develop a software toolkit that will allow subject gateways to build and disclose their metadata repositories in an open and collaborative way. By supporting a range of protocols and metadata formats, it will be possible to integrate subject gateways with other resource discovery services such as those provided by the library, museum, archival, education and government communities. This will allow subject gateways to search across the range of services offered by these communities and, in turn, allow services offered within those communities to provide access to subject gateways.

To date most subject gateways have been set up to serve large groups of users, for example all the members of the UK or US academic communities. However, other types of gateways are also being established, for example those set up to serve all the members of a particular organization, perhaps across multiple subject areas. These smaller gateways will benefit greatly by being able to re-use metadata records already created by the larger subject specific gateways. The framework developed in this project will enable such sharing of metadata records through the Metadata Integrated Development Environment. Sharing will be possible across the range of protocols and formats supported by the framework. Similarly, the larger subject specific gateways will also benefit by being able to re-use records created by the smaller services. Such records may be enhanced by the subject gateways and then be exported back to the originating service.

3.5 Dissemination

Dissemination is integral to project success. It is also integral to the mission of the partner organizations who collectively have significant and varied experience of event management, scholarly publication, current awareness and promotional activity, consensus and standards work, and production information services. The partners have high visibility and are extensively networked into international resource discovery and digital library initiatives. Project outcomes will be disseminated during the normal course of this work, but in addition we propose the following specific activities:

- A project Web-site - public documentation, partner details, related work.
- Academic articles - scholarly articles on the architecture will be submitted to a computer science journal and an information science journal in the middle of Phase 1.
- Professional press - articles will be prepared for D-Lib Magazine and Ariadne (published by UKOLN) at the beginning of the project, and at the end of Phase 1.
- Partners regularly attend relevant conferences and will present the project where appropriate.
- The partners will be collaborating on IMesh guidelines workshops. These will be used to promote the work of the project and gain input.
- The partners will promote the results of the work through relevant standardization channels within W3C, IETF, Dublin Core, and elsewhere.
- The partners will organize a project workshop at the end of Phase 1 in association with a relevant event.

Software developed within this project will be open-source, released under the GNU GPL [GNU] (or similar) license. This will enable other developers to study, change and improve the software provided that modifications are made public in order to benefit the whole community.

3.6 Evaluation

The project partners have experience of developing evaluation strategies for the variety of projects in which they have been involved. The project will commission evaluation of the project from an expert third party and, in consultation with them, will draw up performance indicators to measure effectiveness. We see the evaluation as taking a variety of forms with particular emphasis on user feedback. It will consider the overall management of

the project, identify areas for improvement in the dissemination strategy as well as reviewing the effectiveness of the toolkit. In order to provide timely feedback the evaluation will be initiated in Phase 1 of the project.

3.7 Relationship to Other Work

UKOLN and ILRT have been involved in a number of successful projects over the last four years in the areas of metadata and resource discovery. As part of the ROADS project (together with the University of Loughborough) we have provided a user-oriented resource discovery system for the existing UK subject gateways and have promoted the use of common metadata formats and cataloguing guidelines. ROADS has implemented the means for subject services to offer cross-searching and query routing. Together with European partners in DESIRE we have extended the functionality of ROADS and moved towards a toolkit approach, incorporating robot generation of metadata, additional indexing services and automatic classification, and introduced a multi-lingual approach to resource description [Dempsey 1996] . Other work on the tool-kit is planned in DESIRE Phase2.

ILRT are providing subject gateways for social science information (SOSIG) and business information (Biz/Ed), both using ROADS software. Training and awareness sessions are held throughout the UK aimed at end-users and Internet cataloguers.

Our work on these projects has given our organisations a detailed knowledge of a range of search technologies and metadata formats. We have experience of using a variety of existing tools integrating these with software developed as part of the projects. We have been closely involved in standards making activities such as Dublin Core [Dempsey & Weibel 1996] and the W3C Resource Description Framework [Brickley et al. 1998] .

UKOLN has led the MODELS initiative supported by eLib and the British Library. MODELS provides a forum for UK library and information communities to consider architectural solutions and move towards a shared view of the applications framework needed to manage the range of distributed information services being offered today. Other projects of relevance are PRIDE (EC funded) which will provide distributed user and service/collection information which can be integrated into wider information systems, and BIBLINK which aims to improve the flow of metadata between electronic publishers and national bibliographic agencies.

The Internet Scout Project is an NSF-sponsored organization charged with promoting the progress of research and education by improving the Internet's information infrastructure through the advancement of its resource-discovery tools. The Internet Scout Project has focused their efforts on making resources more accessible to end-users through current awareness publications, such as the Scout Report and subject-specific reports, and selective dissemination of information services, like Net-happenings and SCAN. The other main area of endeavor is developing tools for resource discovery and retrieval and information infrastructure evolution. Scout Report Signpost is one such development. Signpost is a catalog of Internet resources organized and indexed according to existing standards, such as Library of Congress Subject Headings and Library of Congress Classification, and developing standards, such as the Dublin Core. The Isaac Network, the Scout Project's newest initiative, grew out of a desire to link the Signpost with other similar "subject gateways" so that end-users could easily search several selective collections of resources with a single query.

The current proposal will build on the outcomes of these projects, and will bring a more accessible toolkit to the subject gateway information provider. It will develop those aspects of the applications framework which are in their initial stages such as metadata registries and metadata sharing.

3.8 Management Plan

The Internet Scout Project, UKOLN, and ILRT are natural partners in an effort to advance the state of distributed resource discovery on the Internet. Each group has years of experience in improving the ease of discovery of quality Internet resources for the higher education community in the US, the UK, and beyond. Each group has both past and current resource discovery research projects to their credit, and each has operational experience in making newly developed services available to the community in a timely manner. The combined years of experience in Internet resource discovery and cross-searching protocols is for the three organizations totals twelve years.

On another level, the three projects are natural partners because their philosophies on how to proceed are similar, even though these philosophies were developed independently on opposite sides of the Atlantic. In fact, while the principals were discussing the potential for collaboration, it was discovered that the impetus for the development of the Isaac Network was based on the same philosophy of the initial ROADS proposal, even though that proposal had never been read by the founders of Isaac. The current differences between ROADS and Isaac lie in the protocols that are in use, not in any different view of how distributed searching can be

accomplished on the Internet. The fact that different protocols are in use is simply a matter of what was available at the time of implementation.

Because the Internet Scout Project, UKOLN, and ILRT are natural partners and share the same philosophy, it is felt that management of the collaborative project will be smooth, as described below. This successful collaboration will result in successful research results.

The ROADS project had three partners - ILRT, Loughborough and UKOLN. To preserve continuity and exploit expertise within this grouping, UKOLN will sub-contract approximately 0.2 FTE over the lifetime of the project to Loughborough. It is anticipated that Martin Hamilton will either supply or coordinate the supply of this effort.

3.8.1 Partitioning of Research Activities

The research into and development of the IMesh Toolkit call for a breadth of knowledge in both the Computer Science and Library Science disciplines. One of the primary reasons that ILRT, UKOLN and ISP choose to collaborate on this project is that each brings needed expertise to the project. The development of an overall architecture for the IMesh software at an early stage of the project will provide a shared view on which to base collaboration. A modular approach resulting in well-specified and well-documented APIs will enable software development work to be shared across the collaborators, each contributing effort according to expertise.

ILRT and UKOLN have experience using the Whois++ protocol for resource discovery and the development of subject gateways. ISP has experience using the LDAP protocol for resource discovery. ILRT and UKOLN have experience of Z39.50. All of the organizations have been involved in the development and application of metadata standards to Internet resources. Therefore collaboration among these groups is the most efficient path to an ambitious undertaking such as the IMesh Toolkit. The resulting partnership will draw upon each group's strengths, as well as each group's existing code base and staff expertise. This will allow us to more quickly produce portions of the toolkit by tapping protocol expertise already in place instead of duplicating knowledge and effort both in the US and UK.

3.8.2 Overall Coordination of Activities

Each organization in the partnership has designated a single individual as a point of contact for technical issues and a single individual for overall management issues. While there will be multiple people included in the decision making process within each group, it is important that there be a clear understanding between groups as to who has final decision-making power and who to contact for ultimate clarification of any given issue.

The deliverables of the project have been outlined in detail in advance, and approved by all participants. In this way the possibility of confusion between participants regarding the goals of the project has been minimized as we proceed.

A schedule of work has been determined, including interim goals and deliverables for the first 18 months of the project. A detailed schedule of the second 18 months of the project will be assembled at the halfway mark after initial results have been produced. The schedule of work includes specifics about which group is responsible for each deliverable and when.

Communication will take place via email, telephone, and scheduled face to fact meetings (as described below). In addition, team members will experiment with the feasibility and usefulness of video-conferencing facilities on the Internet, and with shared white-board applications for describing and diagramming ideas. Partners already have experience of using BSCW and other collaborative tools to support joint project working. The feasibility and usefulness of these facilities will be summarized in the team's regular project reports in order to inform others in the community who may be considering their use for other collaborative projects.

3.8.3 Coordination and Evaluation of Progress

The technical and managerial point person on each team will be responsible for monitoring the progress of the project and for regular reporting of this progress to other team members. Since the proposed project and budget do not include separate staffing for the administration and reporting of the team's work, written reports will be kept to a minimum and will be brief and to the point. The team has agreed that brief, quarterly summaries of the status of the work will be sufficient for keeping all team members informed, since the overall number of people involved in the project is relatively small. The quarterly reports may be as simple as a one-page listing of pending work and the status of that work.

The quarterly reports will be followed-up with conference calls including all point persons, when deemed necessary by one or more members of the team to clarify any issues.

3.8.4 Project Staff

ISP staff and commitments are listed below. Susan Calcari, Director, is the management contact. Michael Roszkowski, Researcher, Operations and Research Coordinator, is the technical contact. UKOLN and ILRT may recruit additional staff, but the following people will have some involvement. Andy Powell, Technical Development and Systems Group Coordinator, will be the UKOLN technical contact. Rachel Heery, Metadata Group coordinator, will be the management contact. Tracy Gardner, Technical development and research officer, will take a lead on architecture development. Lorcan Dempsey, Director, will provide contact with the RDNC and strategic input. At ILRT, Dan Brickley will be the technical contact, and Nicky Ferguson, Director, will be the management contact. UKOLN expects to subcontract some work to Martin Hamilton, one of the principal developers of ROADS, based at Loughborough University.

The project plan, to be produced in Month 2, will include committee and reporting structures, ensuring clear lines of accountability and decision-making.

3.8.5 Anticipated Travel Requirements

It is anticipated that a maximum of four face to face meetings per year will be required to assure good communication between team members. These meetings will be held in conjunction with other meetings or conferences whenever possible. For example, they may be held in conjunction with an Internet Engineering Task Force meeting, a DESIRE or TERENA conference, or one of the proposed IMesh Guidelines Workshops.

3.9 Deliverables

All deliverables will be produced collaboratively by the project partners. Where a lead partner is identified below, they will be responsible for the overall coordination and quality assurance of the deliverable and for its production in a timely manner. It is anticipated that all partners will contribute effort to each of the deliverables. Internal deliverables will be made available to partners on an internal project Web-site. External deliverables will be made available in line with the dissemination plan outlined above.

The project will be split into two phases, each lasting 18 months. All the internal deliverables listed below will be delivered during the first phase of the project according to timescales listed below or specified in the Project Plan. Initial versions of all the external deliverables will be made during the first phase according to times cales specified in the Project Plan. It is anticipated that all the external deliverables will be updated on an ongoing basis during the life of the project.

3.9.1 Internal Deliverables

Project Plan – A document providing details of the project's internal and external deliverables, delivery dates, lead and contributing partners, work schedules, committee and reporting structures and coordination plans for the first phase of the project. This deliverable will be made 2 months after the project start. An updated Project Plan will be made available to funding bodies at the start of the second phase of the project.
Lead partner: UKOLN

Evaluation Strategy – A document detailing the evaluation strategy for the project as outlined above. This deliverable will be made 5 months after the project start.
Lead partner: ISP

Technology Review – A report providing an evaluation of the current status of research and tool development in relation to metadata management, subject gateway maintenance and cross-searching and a review of existing architectures, drawing on the knowledge and expertise of the project partners in these field.

Subject Gateway Requirements - An evaluation of issues considered important by RDN members in the UK and members of the Isaac Network in the USA.

3.9.2 External Deliverables

IMesh Toolkit Recommendations - Recommendations on end-user and service provider issues investigated during the project.
Lead partner: ISP

IMesh Toolkit Architectural Overview - A document describing the architecture that provides the basis for work within this project. This document should enable readers to understand the operation of the IMesh subject gateway framework at a high level. It should also act as a starting point for developers wishing to develop new software tools to slot in to the framework. The document may reference existing architectures that provide a basis for the IMesh architecture.
Lead partner: UKOLN

IMesh Toolkit Framework APIs - The APIs developed or adopted by this project should be made available in an appropriate format (such as IDL); appropriate supporting documentation should also be provided. The APIs should enable third parties to develop IMesh-compliant components that can be slotted into the IMesh framework to replace or add functionality to existing components.
Lead partner: ISP

IMesh Toolkit Software Distribution - A set of generic tools implementing the IMesh framework packaged with specific tools that have been developed within the project and externally developed tools where they exist. The tools should be accompanied by appropriate documentation and installation instructions. The Metadata IDE and the Metadata Registry are important components of the toolkit that do not exist in a sufficiently advanced form outside of this project; these components are identified as separate deliverables. The software distribution should enable the cross-searching of Isaac (Dublin Core/LDAP) and ROADS (ROADS template types/Whois++) subject gateways and should enable third party components supporting other query protocols (such as Z39.50 and ODBC) to be slotted in.
Lead partner: ILRT

IMesh Toolkit Metadata IDE - A Metadata IDE that can be used to manage subject gateway metadata. The Metadata IDE will use the IMesh Metadata Registry to obtain details of metadata formats. Storage of metadata will be via the IMesh framework. Tools such as metadata extractors and registry based format converters will be shared with the other components of the IMesh framework. The Metadata IDE will be accompanied by both End-User documentation and software documentation.
Lead partner: UKOLN

IMesh Toolkit Metadata Registry - A set of tools to support human and machine-readable aspects of metadata registries. A prototype metadata registry service will be implemented in order to support development of other framework components.
Lead partner: UKOLN

IMesh Toolkit User Guide – A documentation set describing the Framework Architecture and APIs and the installation, configuration and use of all of the components in the IMesh Toolkit Software Distribution. The User Guide will be intended for use by the resource operators of subject gateways and other metadata repositories and for those setting up cross-searching services.
Lead partner: ISP

4 References Cited

[Allen & Mealling 1998] Allen, J. and Mealling, M., 1998. "The architecture of the Common Indexing Protocol (CIP)". Internet-Draft. <http://search.ietf.org/internet-drafts/draft-ietf-find-cip-arch-02.txt>

[ANSI/NISO Z39.50-1995] ANSI/NISO Z39.50-1995. *Information retrieval application service definition and protocol specification for open systems interconnection*. Bethesda, Md.: National Information Standards Organization.

[ASF] Advanced Search Facility (ASF). <http://www.asf.gils.net/index.html>

[Bass 1998] Bass, L., Clements, P., and Kazman, R., 1998. *Software architecture in practice*. Reading, Mass.: Addison-Wesley.

[BIBLINK] BIBLINK: Linking Publishers and National Bibliographic Services. <http://hosted.ukoln.ac.uk/biblink/>

[Brickley et al. 1998] Brickley, D., Guha, R.V. and Layman, A., eds., 1998. Resource Description Framework (RDF) Schema Specification. <http://www.w3.org/TR/WD-rdf-schema/>

[Day et al. 1999] Day, M., Heery, R. and Powell, A., 1999. "National bibliographic records in the digital information environment: metadata, links and standards". *Journal of Documentation*, Vol. 55, no. 1, pp. 16-32.

[Dempsey 1996] Dempsey, L., 1996. "ROADS to Desire: some UK and other European metadata and resource discovery projects". D-Lib Magazine, July/August. <http://www.dlib.org/dlib/july96/07dempsey.html>

[Dempsey & Weibel 1996] Dempsey, L and Weibel, S., 1996. "The Warwick Metadata Workshop: a framework for the deployment of resource description". D-Lib Magazine, July/August. <http://www.dlib.org/dlib/july96/07weibel.html>

[Dempsey et al. 1998] Dempsey, L., Russell, R. and Murray, R., 1998. "The emergence of distributed library services: a European perspective". Journal of the American Society for Information Science, Vol. 49, No. 10, pp. 942-951.

[Dempsey et al. 1999] Dempsey, L., Russell, R. and Murray, R., 1999. "A utopian place of criticism? Brokering access to network information". *Journal of Documentation*, Vol. 55, no. 1, pp. 33-70.

[DESIRE] Development of a European Service for Information on Research and Education <http://www.desire.org/>

[Deutsch et al. 1994] Deutsch, P., Emtage, A., Koster, M. and Stumpf, M., 1994. "Publishing information on the Internet with Anonymous FTP". Internet-Draft. <http://info.webcrawler.com/mak/projects/iafa/iafa.txt>

[Dublin Core] Dublin Core metadata. <http://purl.oclc.org/dc/>

[EEVL] Edinburgh Engineering Virtual Library. <http://www.eevl.ac.uk/>

[GNU] GNU's not Unix! <http://www.gnu.org/>

[Hedberg et al. 1998] Hedberg, R., Greenblatt, B., Moats, R. and Wahl, M., 1998. "A Tagged Index Object for use in the Common Indexing Protocol". Internet Draft. <http://search.ietf.org/internet-drafts/draft-ietf-find-cip-tagged-07.txt>

[ILRT] Institute of Learning and Research Technology, University of Bristol. <http://www.ilrt.bris.ac.uk/>

[IMesh] IMesh: International Collaboration on Internet Subject Gateways. <http://www.ilrt.bris.ac.uk/discovery/imesh/>

[Isaac Network] The Isaac Network: Information Seeker's Avenue to Authoritative Content. <http://scout.cs.wisc.edu/scout/research/index.html>

[ISO 23950:1998] ISO 23950:1998. *Information and documentation - Information retrieval (Z39.50) - Application service definition and protocol specification*. Geneva: International Organisation for Standardisation.

[ISP] Internet Scout Project, University of Wisconsin-Madison. <http://scout.cs.wisc.edu/scout/>

[ISP 1999] Internet Scout Project, 1999. *Project Isaac Architecture Overview for Collaborators*. <http://scout.cs.wisc.edu/scout/research/rfp/IsaacCall.html>

[Kirriemuir et al. 1998] Kirriemuir, J., Brickley, D., Welsh, S., Knight, J. and Hamilton, M., 1998. "Cross-searching subject gateways: the query routing and forward knowledge approach". *D-Lib Magazine*, January. <http://www.dlib.org/dlib/january98/01kirriemuir.html>

[MODELS] MOving to Distributed Environments for Library Services. <http://www.ukoln.ac.uk/dlis/models/>

[RDF] Swick, R., Miller, E., Schloss, B. and Singer, D., 1999. *Resource Description Framework (RDF)*. <http://www.w3.org/RDF/>

[RFC 1777] Yeong, W., Howes, T. and Kille, S., 1995. "RFC 1777: Lightweight Directory Access Protocol". <ftp://ftp.isi.edu/in-notes/rfc1777.txt>

[RFC 1835] Deutsch, P., Schoultz, R., Faltstrom, P. and Weider, C., 1995. "RFC 1835: Architecture of the WHOIS++ service". <ftp://ftp.isi.edu/in-notes/rfc1835.txt>

[RFC 2251] Wahl, M., Howes, T. and Kille, S., 1997. "RFC 2251: Lightweight Directory Access Protocol (v3)". <ftp://ftp.isi.edu/in-notes/rfc2251.txt>

[RFC 2413] Weibel, S., Kunze, J., Lagoze C. and Wolf, M., 1998. "RFC 2413: Dublin Core metadata for resource discovery". <ftp://ftp.isi.edu/in-notes/rfc2413.txt>

[ROADS] Resource Organisation And Discovery in Subject-based services. <http://www.ilrt.bris.ac.uk/roads/>

[Roszkowshi & Lukas 1998] Roszkowski, M. and Lukas, C., 1998. "A distributed architecture for resource discovery using metadata". *D-Lib Magazine*, June 1998. <http://www.dlib.org/dlib/june98/scout/06roszkowski.html >

[SOSIG] Social Science Information Gateway. <http://www.sosig.ac.uk/>

[Stanford Digital Libraries Project] The Stanford Digital Libraries Project. <http://www-diglib.stanfo rd.edu/>

[TF-CHIC] TERENA Task Force on Cooperative Hierarchical Indexing Coordination (CHIC). <http://www.terena.nl/task-forces/tf-chic/>

[UKOLN] UKOLN: The UK Office for Library and Information Networking. <http://www.ukoln.ac.uk/>

[Workshop on Metadata Registries 1997] Joint Workshop on Metadata Registries, University of California, Berkeley, California, 7-11 July 1997. <http://www.lbl.gov/~olken/EPA/Workshop/>

5 Biographical Sketches

5.1 Institute for Learning and Research Technology, University of Bristol

The Institute for Learning and Research Technology (ILRT) at the University of Bristol is host to more than twenty-five national and international projects at the forefront of learning and research technology and is the largest group of its kind in the UK. The mission of the Institute is to be a centre of excellence in the development and use of new technology in teaching, learning and research. In pursuit of this mission, the main objective of the Institute is to initiate research projects in the use and development of technology-based methods in teaching, learning and research and to provide national and international services, consultancy and support using these methods.

5.2 Internet Scout Project, Computer Sciences Dept., University of Wisconsin-Madison

The Internet Scout Project (ISP) is an NSF-sponsored organization charged with promoting the progress of research and education by improving the Internet's information infrastructure through the advancement of its resource-discovery tools. ISP has focused their efforts on making resources more accessible to end-users through current awareness publications, such as the Scout Report and subject-specific reports, and selective dissemination of information services, like Net-happenings and SCAN. The other main area of endeavor is developing tools for resource discovery and retrieval and information infrastructure evolution. Scout Report Signpost is one such development. Signpost is a catalog of Internet resources organized and indexed according to existing standards, such as Library of Congress Subject Headings and Library of Congress Classification, and developing standards, such as the Dublin Core. The Isaac Network, the Scout Project's newest initiative, grew out of a desire to link the Signpost with other similar "subject gateways" so that end-users could easily search several selective collections of resources with a single query. The Internet Scout Project is a part of the Computer Sciences Department at the University of Wisconsin-Madison, in Madison, Wisconsin.

5.3 UK Office for Library and Information Networking, University of Bath

UKOLN, the UK Office for Library and Information Networking, is a national centre for support in network information management in the library and information communities. It provides awareness, research and information services and has the following goals:

- to promote the awareness of emergent issues at technical, service and policy levels,
- to provide a focal point for research, development and performance measurement,
- to influence policy makers and service providers in the interests of the communities it serves,
- to demonstrate high-quality information services.

UKOLN is funded by the British Library Research and Innovation Centre, the Joint Information Systems Committee of the Higher Education Funding Councils, as well as by project funding from the JISC's Electronic Libraries Programme and the European Union. UKOLN also receives support from the University of Bath where it is based.