JISC/NSF Digital Library Initiative
Final Project Report: Cross-Domain Resource Discovery Project ("Cheshire")

1. *Background*

This is the final project report for the JISC/NSF funded Cross-Domain Resource Discovery Project ("Cheshire"). This project addresses the need to develop and implement the advanced networking technologies relating to the integration of digital and internet-based services and digital content. To date, it has sought to do so by setting out a high-level systems component framework based on the Z39.50 information retrieval protocol; the development of a standards-based software system which is extensible for the accommodation of radically different architectural models; and the distribution and support of this system on an open source basis throughout the Joint Information Systems Committee (JISC) and US communities. For its three year life, the project has attempted to integrate a number of innovations in the access, analysis, and display of cross-domain resources including text, numeric, geo-spatial and bibliographic resources. It is currently used to serve over 100 datasets and is now used extensively in higher and further education institutions for use in national and international services.

This project was originally initiated to address the growing need to develop standards-compliant software supporting access to cross-domain resources which are distributed across locations, platforms, protocols, and projects. Our research agenda has been developed to enable users to work on a distributed basis to produce and distribute digital library materials. We feel this type of research is likely to be required if students, researchers, and teachers are to take full advantage of the infrastructure of information resources which have been funded by the JISC and various agencies in the United States. At our current stage, we are able to link a variety of resources and extend the capability of software and services which may allow educationalists to store and retrieve data objects for reuse in the portal environment.

The design of the system incorporates a client/server architecture based entirely on national and international standards for document description and information retrieval protocols, including SGML/XML, Z39.50, HTTP/CGI, with support for other protocols such as RDF, OAI, SOAP, etc., developed as part of the three year project. The system has been redesigned and implemented to user probabilistic information retrieval methods and to support Object-Relational Database Management Systems; it incorporates advanced retrieval algorithms which can form the basis of a sophisticated ontology engineering environment (which may be incorporated into managed learning environments).

The Cheshire software and tools link a wide variety of JISC resources and extend the capabilities of software and services characteristic of the digital library environment. In particular, the system allows users to retrieve and reuse data objects from the wide variety of protocols supported for use in learning systems such as portals, Virtual Learning Environments, etc.

1.1Why Cheshire is needed to support the national infrastructure

The Cheshire system has been designed to address the growing need to develop open-source, standards-compliant software which will support portal architectures catering for the widest variety of underlying databases and information retrieval protocols. The objective of the project has been to develop and implement a system which will support large-scale digital library services on a distributed basis, extending to non-bibliographic data.

As a result of research and development funded by the JISC/NSF grant, Cheshire now has the following capabilities. It will:

1. Support searches across multiple JISC services and datasets (digital libraries). This includes cross-protocol harvesting and serving (e.g. Z39.50, OAI, SOAP).
2. Support the indexing and delivery of non-bibliographic (e.g. full-text and numeric) databases as well as the linking of bibliographic references with corresponding full-text services.
3. Support advanced information retrieval algorithms to enable users effectively to discover information including cross-thesauri and trans-lingual information management.
4. Search support for unfamiliar metadata vocabularies and data filtering.
5. Allow users to request and deliver information from appropriate resources.
6. Supports a full distributed architecture, in which repositories can host and mediate their own data.

Cheshire's support for a distributed architecture means that:

1. Individuals, institutions, and projects are able to serve data; to allow that data to be easily used by national services; to give enhanced access to that data; and to manage that data at a greatly reduced cost.
2. Enables projects and services to be flexible in terms of the services they can include (e.g. it is easy for individuals, services, and projects to mix and match data resources for particular purposes).
3. Reduction in management costs for the national services and an emphasis for individuals on best-practice for creating, managing, and serving information.
4. The architecture is scalable to an indefinite extent with little or no loss of performance and functionality.

The Cheshire software and tools developed in response to JISC/NSF link a wide variety of JISC resources and extend the capabilities of software and services characteristic of the digital library environment. In particular, the system allows users to retrieve and reuse data objects from the wide variety of protocols support for use in learning systems such as portals, Virtual Learning Environments, etc. As far as we are aware, there are no other products or systems which have all the capabilities of the Cheshire system available in one package, either commercial or open source.

Compared with other information retrieval systems, Cheshire offers a much greater specification as well as cross-protocol support. As an example of its cross-protocol functionality, Cheshire's support for SRW ("Search and Retrieve Web Service" http://www.loc.gov/z3950/agency/zing/srw) is the most complete package available and the Cheshire project has been active in promoting and developing the protocol for wider use in the digital library environment.

1.2 How the Cheshire system supports the JISC Information Environment architecture

The system ("Cheshire") has been designed to support the infrastructural issues as set out in the DNER Architecture Document. This sets out an architecture based on the concept of portals querying or harvesting any number of information providers who may be serving information via OAI, Z39.50, or web service protocols not currently defined. This has brought about a need to for users to access these distributed information resources efficiently, even when there may be additional unanticipated protocols not anticipated in the DNER Architecture Document. In particular, there is a need to define some method of enabling managed learning systems to integrate with digital library services supporting a number of different protocols.

To address this situation, we have developed Cheshire's potential to act as a cross-protocol data harvester (for OAI, Z39.50, OGDI, etc.), providing both Z39.50 and HTTP interfaces to descriptions. As an outcome of the project, we have extended Cheshire's capabilities to support any number of protocols which might be used in the JISC information environment, particularly for teaching and learning architectures. These include Open Archives Initiative (OAI), Simple Object Access Protocol (SOAP), Search/Retrieve Web Service (SRW), Simple Digital Library Interoperability Protocol (SDLIP), and web service protocols such as Universal Description Discovery and Integration (UDDI), Web Service Definition Language (WSDL), etc. We intend to the Cheshire system to act as a transformation engine for managed learning systems, making existing resource descriptions to teachers an learners, currently served by any number of various protocols, available via Cheshire harvesters. This will create a large testbed of items for which descriptions already existing. It will result in an assimilation of an existing network of resources into managed learning environments.

The current stage of the system's development, as completed over the past three years, has focused on creating a distributed data object retrieval system, based on Z39.50 and SGML/XML, which offers advanced retrieval and discovery capabilities. This includes such features as:

1. The system may be extended to accommodate any number of other protocols, for example OAI, SOAP, ADLP, and SDLIP using the embedded scripting languages of TCL and Python.
2. Indexing and searching capabilities include the ability to extract geo-spatial and temporal information, extraction of single tags as pseudo-records, improved stemming and relevance ranking algorithms and the ability to dynamically adjust the indexing process based on processing instructions within the XML data.
3. Retrieval of individual known XML elements or strings from within a record on the fly.
4. Support for the Attribute Architecture specification for Z39.50 enabling use of multiple attribute sets at once, as well as more fundamental transactions such as sorting.
5. Virtual databases that provide integrated access to multiple datasets on a single server.
6. Storing XML in a preparsed form allowing for much faster access to complex record structures.
7. A Z39.50 client integrated within the open source Mozilla (Netscape) web browser.
8. The system was extended to support Mac OSX as well as Linux, with additional support for Windows platforms with the entire source tree available on an open source basis.
9. Many XML and SGML parsing issues were corrected, enabling the correct validation of even very complex DTDs and Schemas.

The outcomes of the JISC/NSF funding have resulted in a stable version of the Cheshire software which is intended to fulfill the development and redesign objectives of that funding, in particular the expansion of software support for Z39.50 and XML technologies. These capacities, now enabled, permit the dynamic implementation of different processing and retrieval methods, as appropriate to a given domain whether statistical, full-text, geospatial, or multimedia. The source code for the software is available for compilation on an open-source basis and has been progressively released throughout the grant period. We are currently compiling and releasing binaries for this version of the system which will enable support for Windows2000, Macintosh OS X, Linux, and Solaris 8 operating systems (the source code is available and may be independently compiled for these operating systems). This version has been set up as a "turn-key" search environment for Digital Library services.

1.2 Relationship to other projects and services

1.2.1 Use of Software and Tools

The Cheshire system is now being used by all the national services in the United Kingdom. Our model is to work within the framework of the services and collaborate with programmers to develop systems within their own areas of expertise. We believe this type of collaborative design process encourages cross-fertilization in the development of the software, access systems, and national services. Working with the national services, we have assembled a number of tailor-made implementation solutions based on established models. The project has been able to draw upon existing service infrastructures, already in place, which have organizational and support staff to maintain the software and its use over time.

One rewarding aspect of the project is to see how research and technical advances are incorporated into production services. These have been subject to formal evaluation and service review procedures taking place outside the project but which inform the development and success of the project's objectives. We have enjoyed working with the national services in disseminating the system more widely and investigating with them strategic support for the use of digital library activities, particularly in teaching and learning activities. In particular the work with Manchester Computing (MIMAS) has yielded valuable results which we hope to deploy with them (and UKOLN) for the JISC Information Environment Service Registry (IESR).

The most ambitious use of the software to date has been for the Archives Hub and Manchester Computing. This service was established in 1999 with the aim of managing and serving archival finding aids throughout the United Kingdom for the Higher Education sector. In its current phase, the Archives Hub is becoming a distributed service with each of the repositories responsible for hosting their own data. This architecture is derived directly from the research findings of the JISC/NSF project, based on the use of Z39.50 SCAN to harvest indexes of distributed servers, as discussed above. In order to enable the use of the Cheshire software by non-specialists, we have provided an easy installation package which is in current use by archivists in universities throughout the country. A key outcome of the Archives Hub development will be the deployment of distributed technologies at a national level and their impact on the daily use by non-specialists.

1.2.2 Exploitation Plan and Dissemination

Rather than seeing this as a pure "research" project with limited local use, we have sought the widest possible use and dissemination of the system architecture as an outcome. Since Cheshire is being used by a number of national services, we are able to target a large number of potential audiences, including students and teachers, librarians, gateway and broker services, commercially provided portals, etc. To date the dissemination activities have provided general information about the project and its potential benefits; and information about research outcomes such as methods of metadata reuse. These have been disseminated in the following ways:

1. Papers on the design and development of the project and its components (see bibliography)
2. Papers documenting the project and significant results (see bibliography)
3. Technical reports made available via the project's web site(s)
4. The software and tools for the project are made available on an open source basis

As far as possible we have incorporated the architecture and methodology of the project more widely into existing services funded by the JISC. One objective is to to ensure that support and further development of the system can be sustained for use by individuals as well as by digital library services on an ongoing basis, thereby ensuring optimal take-up of the software and the tools described.

1.2.3 System Test, Evaluation, and Quality Control

The stakeholders in the development and dissemination of the project's objectives include the national services using the software as well as students and teachers. We have paid a particular focus on the needs of those individual users as well as by the national services and the suitability of the software for production.

The process of evaluation for this project is associated with the testing of the system to ensure that it adheres to national and international standards and supports the optimal methods for integrated information discovery across domains. As part of this, we have conducted interoperability tests with other servers (e.g. COPAC) and with a variety of datasets. We have, additionally, tested interoperability with portals as well as with other digital library systems and services. An additional component has been an investigation of the use of standardized metadata elements across domains and support of these using the new Z39.50 attribute architecture for cross-domain searching, semantic interoperability, etc.

The Cheshire project has undergone a number of summative evaluation procedures, both its use for national services and as an experimental prototype. For production services, the primary evaluation procedure has been conducted for services hosted by MIMAS (Manchester Computing), available via http://www.archiveshub.ac.uk/introduciton.shtml. This reflected a series of single-focus user trials with candidates from a variety of institutions in order to determine typical user-type transitions and modes of search behaviour using the system.

2. Methodology

Our initial strategy was to build an information access system that would provide an original methodology for information discovery and retrieval; and its record in exploiting the fundamental interconnections between diverse information resources which comprise the JISC Information Environment. These are:

1. Document databases which describe information about various topics ranging from news reports and library catalogue entries to full-text articles from academic journals
2. Numerical statistical databases which assemble facts about a wide variety of social, economic, and natural phenomena
3. Spatial and temporal databases associated with geographic information systems (GIS) which facilitate map-based display of geographic attributes.

The development work undertaken for this project focused on exploiting the fundamental interconnections between these three types of data in the digital library context. The original aim was to develop and implement a statistical association methods for mapping a query to appropriate metadata classifications in the various types of databases; while at the same time extending the system's capabilities and redesigning it with a new system architecture. These have been executed as follows:

Extension 1: ***Client Technologies.*** *The integration of the Z39.50 information retrieval protocol into the existing code base of a browser (Netscape/Mozilla).* We have achieved this through the release of a Z39.50 browser client, currently available as a download with an XPI (cross-platform installer) for ease of use. The new client is able to locate a server via a Z39.50 URL and allows the functionality of the Z39.50 protocol to be accessible via XPConnect, a library linking Javascript with underlying C++ code. Thus, the Z39.50 session objects are able to support and intelligently process Z39.50 v.3 functions such as SEARCH, SCAN, etc. The client also has an interface which allows for improved accessibility to Z39.50 servers. This interface is easy to use, but has added functionality. The interface is accessible via Mozilla's automatic install procedure, so that the client may be seamlessly layered on top of Netscape.

Extension 2: ***Metadata Reuse.*** *The development of tools which make it easier for users to discover information, even though they may be unfamiliar with the classification, categorization, and indexing schemes characteristic of the databases being searched.* To achieve this, we investigated and implemented methods of reusing pre-categorized and pre-classified records for improving cross-domain searching, information management, and resource discovery. We exploited Cheshire's ability to harvest and index data using an advanced "clustering" technique which will enable common terms to be interlinked automatically and retrieved quickly.

Extension 3: ***The creation of an infrastructure of Z39.50 databases and harvesters which may be used by the national services and which may be integrated into commercially available portals or Virtual Learning Environments (VLEs).*** Cheshire exploits the Z39.50 information retrieval protocol to link any information resource to any other information resource without necessarily any direct interchange between the two. As part of the project, we developed additional support for use of Cheshire as a cross-protocol data harvester and transformation engine, which may be used to turn existing Z39.50 resource descriptions to other standardized metadata, e.g. EAD, MARC, IMS/SCORM/IEEE LOM. The objective is to exploit this feature in a way which will permit services and datasets to interoperate in new and effective ways.

This work underpinned a number of related research and development projects based on the Cheshire infrastructure, as follows:

1. *The Bandalino suite of research projects relating to user interfaces for search, text data mining, and empirical computational linguistics, and automating web site evaluation.* Within this context, Cheshire is being used as a search interface for heterogeneous web intranets, such as those found at large universities, corporations, and government sites. It is being used to organize and group search results over large intranets into coherent structures.
2. *The Biothreat Reduction Program at the Los Alamos National Laboratory.* Within this context, Cheshire is being used to search for and synthesize information about textual descriptions supporting biomedical research, in particular for cross-domain cataloguing, statistical analysis and strain/species identification based on forensic and attribution information. The Cheshire infrastructure is being used to support the development and deployment of statistical approaches to natural language processing, which will identify entities and relations between them in bioscience texts. This will in turn facilitate more effective search and synthesis.
3. *The Metadata Research Programme, which uses Cheshire to explore information retrieval in a networked environment.* Within this context, the Cheshire software is being used to design, build, and experiment with front-end prototypes, strategic search commands, entry vocabulary modules, and multi-database navigation.
4. *The "Search Support for Unfamiliar Metadata Vocabularies".* Within this context, Cheshire is being used to develop prototypes to assist with searching across various classification, categorizing, and indexing (metadata) schemes.
5. *The Seamless Searching of Numeric and Textual Resources Programme.* Within this context, the Cheshire software is being used by the Institute for Museum and Library Services in a development project to improve access to written material and numerical data on the same topic when searching very different types of databases and numerical data.

6. *Translingual Information Management Programme.* Within this context, the Cheshire system is being used to prototype new methods of cross-lingual searching, information management, and resources of language engineering.
7. *The Cheshire Record and User Management, which uses Cheshire to provide a web based interface to the creation and maintenance of SGML/XML records within a Cheshire database.* By using the standard web authentication system to validate users in conjunction with a Cheshire based user databases, very sophisticated levels of distributed management are possible.

The methodology behind the Cheshire development programme has, in many ways, been driven by its use within these types of research and development programmes. Our strategy has been to develop Cheshire capabilities for use within R&D projects, work with researchers to formalize the software capabilities to meet their needs, and then release a production version of the Cheshire software with these capabilities which may be used in a production environment for the national services. In this respect, we have adopted a collaborative methodology, involving other researchers and services. These capabilities are covered in greater detail in Section 3.

2.1 Project Workplan and Deliverables

We have executed the project workpackages in accordance with the Cheshire Project plan:

1. The base client and server building was executed within the first 18 months, leading to a release of the browser-client and an incrementally developed server side implementation.
2. The metadata and format handling phase was executed within the next 10 months, which led to the use of the SCAN service of Z39.50 servers to build combined indexes containing information "harvested" from individual servers. This is now in place and is used in production services such as the Archives Hub and MerseyLibraries.org.
3. The Entry Vocabulary Module support was executed also during this 10 month period. This aspect of the project enabled intuitive mapping from natural languages to controlled vocabularies. It is now used in production services such as the Archives Hub and MerseyLibraries.org (under the guise of "Subject Resolver").
4. Support for Geographic Information Retrieval (GIR) applications was executed near the end of the project, which resulted in the capability of extracting geo-spatial and temporal information.
5. The support for non-Z39.50 protocols has been added within the last 12 months, including support for SOAP, ADLP, OAI, SDLIP, and web service protocols. Cross-protocol search and retrieve is now available as part of the Resource Discovery Network (which provides Z39.50 access to harvested OAI data) and MerseyLibraries.org (which supports the cross-searching of Z39.50 and SOAP protocols to include Amazon and Google results alongside bibliographic results).
6. We have not only embedded Cheshire not only in testbed applications such as those cited above (e.g. Bandolino), but have also used Cheshire to underpin a number of national services in a production environment.

In addition to the original workpackage objectives stated above, we have developed a number of new support tools for cross-protocols, as follows:

1. Cheshire/Python integration, including Python ZOOM; appropriate SOAP toolkits (ZSI); Creation of SOAP object and message bindings; Mapping from ZeeRex files to autogenerate SRW configurations; Result and authentication handling in a stateful server environment; Redesign for integration into JISC projects and services.
2. The development of the system's scripting capabilities, including Changes to the display code for distributed resolving; MetaClusters for distributed cluster harvesting; Scripts for automating updates; ZeeRex (creation, standardization, processing, automation, dissemination); Postgres database for maintaining harvested state; Postgres databases for user information, including result sets; Scripts and databases for harvesting; More efficient indexing routines.

3. Activities

This section surveys the activities of the project during its three year period. Further detailed information is included in the annual reports, available on http://sca.lib.liv.ac.uk/cheshire or http://cheshire.berkeley.edu

3.1 The development and implementation of an open source Z39.50 browser client (Cheshire Extension 1)

Before starting to work on this extension, we conducted an initial assessment exercise to assess the design objectives in light of new software technologies which evolved since the proposal was submitted in 1999. Although the project proposal implied a Java based system for the client, we subsequently decided to utilize the Mozilla framework being developed on an open source basis. By extending the browser (Netscape/Mozilla) to handle the Z39.50 protocol, we assessed that the client functionality would be familiar to the general user and, additionally, we could take advantage of the modular structure of the Netscape/Mozilla framework. This has ensured that every time Netscape/Mozilla is updated, the changes needed to keep the Z39.50 component sychronized are minimized. Because Netscape/Mozilla also implements XPI (the Cross-Platform Installer), it has also meant that new components or user interface modifications could be installed automatically. The new architecture has meant that, with a single XPI installation, users are able to get a much speedier response in comparison to a Java client (which would comparatively have a slower start up time). The Z39.50 browser client is entirely standards based: It supports the Document

Object Model level 1 (DOM1), HTML version 4.0, ECMAscript (standardized javascript), Resource Discovery Framework (RDF), eXtensible Markup Language (XML), Cascading Style Sheets (CSS), and so forth. We regard adherence to such standards as essential to ensure that the application remains usable with different servers and data types.

The construction of the client base was done using the following phases:

1. ***Z39.50 URL Phase.*** **This was to ensure that the client is able to locate a server via a Z39.50 URL. This capability is used to retrieve a single document as well as search the server and display the results. The URL scheme was integrated into Netscape/Mozilla's existing network code structure, reusing existing functions and objects. (The URL scheme was not that proposed in RFC2056, as this did not incorporate user authentication, searching, or other Z39.50 commands.) The retrieve data is then able to be handled by the HTML, text, or other mime type handlers already in Netscape/Mozilla. The work consisted of preparing the draft specifications for the URL scheme, and the start of development work for the same.**
2. ***Z39.50 Scripting Phase*****. This was to engineer a client which is scriptable via Javascript and XUL. This required the Z39.50 functions to be accessible via XPConnect (Netscape/Mozilla's library to link Javascript and the underlying C++ code). As such, it needed to be compatible with XPCOM (Netscape/Mozilla's cross-platform component object model). This has ensure that Z39.50 session objects are available to the javascript components that can call CONNECT, DISCONNECT, SEARCH, SCAN, and so forth, to enable the client intelligently to process information available via the Z39.50 information retrieval protocol.**
3. ***Interface Building Stage.*** **This phase built upon the Z39.50 scripting phase, outlined above. Using these scripting capabilities and XUL, we built and tested a new interface that allows for imprved accessibility to the target servers. The interface is familiar to the end user, but has additional Z39.50 functionality. This interface is accessible via Netscape/Mozilla's automatic install procedures, so that the client may be seamlessly layered on top of Netscape.**

Although the client and server are designed to work together in a seamless fashion, they must also interoperate with any other Z39.50 compliant system providing such systems support SCAN, EXPLAIN, SEARCH, etc., to a minimum version 3 standard.

3.2Enhanced Support for Metadata and Vocabulary (Cheshire Extension 2)

A primary aim of the project was to enable the enhanced retrieval of unfamiliar metadata across domains, e.g. constructing linkages between natural languages expressions of topical information and controlled vocabularies for geospatial, textual, and statistical. To this end, we developed methods of using Z39.50 to automatically "cluster" together topics which may be semantically related for digital library projects; and have incorporated this technology in a number of national services some cross-domain. The techniques of "Classification Clustering" use natural language parsing software to identify phrases in the language of the users of bibliographic databases, taken from the titles and abstracts in the literature to be searched, and then apply statistical association techniques to associate these words and phrases with the metadata terms of the target.

Through this way, we sought to develop a research-oriented method of providing access to subject headings, no matter how unfamiliar they may be to the end user, by automating the process of association between natural language and their subject headings. This capability appears to have been effective in enabling users to map their query to the controlled vocabularies (subject headings) used in descriptive metadata; it may be used to cross-search different thesauri and automate associations between them and the user's inquiry.

We are currently using this to facilitate automatic subject retrieval across any number of thesauri supported by a number of distributed datasets. The initial findings suggest that this functionality may facilitate access to metadata describing geospatial datasets. Specifically, methods of mapping geographic place names in text (natural language) to probable geographic coordinates; for mapping geographic coordinates to sets of nearby named places at different levels of geographic or political detail and of different place name types (e.g. city, country, state or province, country). This would require further development of techniques and standards for authority control of events, for time-lines, and for the generation and display of *ad hoc* time lines of events relating to any given theme.

3.3Distributed Search Support ("Meta-Search" capability) (Extension 3)

One of the most important research advances of the project is the development of an architecture enabling efficient searching across hundreds or thousands of distributed network nodes using the Z39.50 SCAN service. This capability will be required for the DNER and similar digital library projects to function effectively.

Our initial work in Cross-Domain Resource Discovery concentrated on using the facilities of the Z39.50 protocol to implement what we are calling a "Meta-Search" capability using existing Z39.50 servers and resources, extending also to non-Z39.50 protocols such as SOAP and OAI. Many existing attempts at distributed search and resource discovery

have relied on the broadcasting of search requests to all servers making up the distributed resources to be searched. There are a number of practical problems with this approach. One of the chief drawbacks is that all systems must be searched before the user or the search controller can determine which systems are most likely to provide the results that th user is seeking.

Instead of using such broadcast searches, we are using the SCAN service of Z39.50 servers to build combined indexes containing information "harvested" from individual servers. The SCAN service permits Z39.50 requests directly to server indexes and returns results containing index information including the words or keys in the index along with their frequency of occurrence information for the database. With this information, indexes combining information from many servers and databases can be combined and statistical ranking methods can be used to rank those servers and databases according to the probability that they contain relevant information for a given user query.

The Z39.50 SCAN service is included in all the Cheshire servers. We have implemented a special indexing mode for the Cheshire system, which will take a list of servers and use the Explain and SCAN services to build combined "Meta-Indexes" for those servers. This functionality is included in all Cheshire servers, making it easy for any Cheshire server to function as a Meta-Search server for some group of other servers. This facility can be recursively executed, so that the hierarchies of Meta-Search servers can be constructed. This has meant the construction of topically-oriented Meta-Search servers as well as "global" servers that summarize index information from each of the lower layers in the search hierarchy). This architecture is operating in production for the Archives Hub service, MerseyLibraries.org, etc.

The support for SRW (Search/Retrieve Web Service) and SRU has extended this capability beyond the Z39.50 protocol to other protocols, which provides a common usage framework for SOAP, OAI, Z39.50, WSDL, and other protocols which may be use in the future. Cheshire's support for SRW/SRU is part of the Information Environment Service Registry (IESR) and MerseyLibraries.org, to name two implementations.

4. Outputs

The five major outputs resulting from the three-year project are:

1. The system capabilities have been extended to allow users to access different domains and information resources (text and document retrieval, numeric databases, and geographic information systems) through the support for *transverse searching* (in which data found in a text database can be used to find related data in a numeric or geo-spatial database). This functionality may be used for across, retrieving, and rendering the different types of metadata characteristic of geo-spatial, textual, and statistical in nature. This, in turn, will allow users to conduct searches in each type of database by accepting a query in the users' own terms and then suggest specialized categorization terms to search for in the information resource through statistical associative techniques.
2. The project has extended development of these associative techniques to provide support for "subdomain" vocabularies, e.g. association dictionaries which will lead searchers to the appropriate term or cluster of subject access terms that are likely to satisfy their information needs for specialized topics ("subdomains") which may be non-textual or include cross-thesauri and trans-lingual support. The development and implementation of these techniques have enabled the system to develop automatically a "likelihood ratio weighting" associated with each searching term and each metadata value which will may lead the searcher more quickly to required information.
3. We have used these developments to research, develop, and implement support for a number of related neglected digital library problems involving metadata reuse. These include: Graphic display by time and place; Support for searching unfamiliar metadata; automated cross-language searching; improved methods of retrieving information with inadequate metadata.
4. The project has extended support for geo-temporal analysis, providing a means of making text, images, hyperlinks, etc. available on a map interface. These capabilities may be built on to give students and teachers in a variety of disciplines the ability to draw arbitrary selections of both local and internet-accessible resources and enable them to mix, match, and plot relationships of data using dynamic map displays. This accompanies the more general project objectives of metadata reuse which may be deployed in pedagogically inventive ways. The architecture is extensible and has been extended to cater to new web service protocols as and when they may be adopted.
5. The project has resulted in the development of a high performance, scalable, and extensible platform for information retrieval. Scalable performance has allowed us to explore the more resource-intensive information retrieval techniques described above.

More recent developments to the software are as follows:

1. *The automatic transformation of metadata.* We have extended the system so it is automatically able to transform existing data formats. While this presently is being used to transform data types such as EAD (Encoded Archival Description) into MARC, in the future we could use this capability to transform existing data into the IMS standards and to build descriptions of full records which will enable access to existing intellectual resources through portals or Virtual Learning Environments.

2. *The extension of the system to harvesters and non-Z39.50 services.* This objective is a prerequisite for relating information resources in a "protocol neutral" fashion so that web service protocols can be integrated seamlessly with Z-services. This has included the extending of the system's support for non-Z protocols (e.g. SOAP, OAI) as well as providing support for XMLSchemas, which may be provided for the integration of Cheshire-served datasets with commercially available portals. As an example of its cross-protocol functionality, Cheshire's support for SRW ("Search and Retrieve Web service" http://www.loc.gov/z3950/agency/zing/srw) is the most complete package available and Cheshire has been active in promoting and developing the protocol for wider use in the digital library environment. Cheshire already supports OAI (as part of the RDN) and SOAP.
3. *The creation of end-user tools to visualize the information retrieved.* We have developed capabilities to allow users to search across databases of unknown scope, to navigate, visualize, and interpret their results. This has included a number of visualization tools for spatial and temporal data.
4. *Design and prototype implementation of an extensible architecture for distributed collection search and retrieval.* We have developed a new design that accommodates distributed components of the system and paves the way for providing search and retrieval capabilities across a large range of systems ranging from very large central servers and distributed databases using a multitude of federated servers to personal systems managing collections for individuals. The design also provides expanded support for a variety of search protocols and capabilities as discussed in section 2 above.

As outlined in Section 2, the research and development of the system has achieved all the objectives outlined in the workpackages and stated in the deliverables. In particular:

1. We enable distributed queries through an inter-server discovery and search interface.
2. We style towards multiple interfaces and extensibility from the ground up, with clean plug-ins for interoperability.
3. We have executed maintainable, concurrent, high performance code.

5. Impacts

The project's primary impact has been the development of a system and framework ("Cheshire") which will support distributed information retrieval and which can be applied to the entire range of information retrieval needs for digital libraries across the entire spectrum of sizes and architectures currently in use or proposed (from very large distributed systems to handheld devices). Within this context, we have developed and demonstrated a solution to distributed search that is machine architecture and protocol neutral, that is a system that allows not only support of a multitude of distributed databases, but also supports a framework which enables support for virtually any of the current search and discovery protocols, operating on machines ranging from large multiprocessor servers to PDAs. A central objective of our research, therefore, has been to make such search capabilities and features a ubiquitous resource for digital libraries which can be exploited by systems and users.

Intellectually, the project has aimed to develop fundamental new approaches to information storage and retrieval, as well as to pioneer new approaches to information sharing and metadata reuse across systems and types of information. The result has the potential to change fundamentally the way that information systems and services are used by individuals and how these systems interact with each other.

Our work has sought the broader impact of transforming the way that JISC services currently use and disseminates information. The developed systems and infrastructure comprising Cheshire will support not only the current generation of computing hardware, but will also support the "next generation" of computing devices embedded in everyday objects and so have the potential to impact all aspects of higher education services. The international nature of the collaboration has encouraged the greatest possible distribution and use of the systems and framework developed as part of JISC/NSF. This impact has been broadened by the open source nature of the software, which has encouraged independent development and use by Manchester Computing, UKOLN, etc.

The project's research innovations have had a number of specific impacts demonstrating the integration of access across domains. These include:

1. *Management of vocabulary control in a cross-domain context.* The Cheshire system is now able to map the searcher's notion of a topic to the terms or subject headings actually used to describe that topic in the database. The classification clustering technique developed as part of the project have been combined with probabilistic document retrieval algorithms to provide a cost-effective remedy which allows users to access unfamiliar metadata, including support for cross-thesauri and translingual retrieval. This is an alternative to the approach currently being investigated by HILT. The outcome has enabled a more direct connection between ordinary language queries ("query vocabularies") and indexing terms ("entry vocabularies") actually used to organize information in a variety of databases. These innovations are now implemented in a production environment as part of the Archives Hub, MerseyLibraries.org, etc., all of which support cross-thesauri retrieval without the expense associated with the development and maintenance of higher level thesauri. We are planning to implement this innovation as part of the JISC funded Information Environment Service Registry (IESR) which will be extended across all JISC datasets.

2. *Distributed access to existing metadata resources.* The Cheshire system has had a definite impact on the development and implementation of distributed information servers, most notably on the distributed version of the Archives Hub operating as a national service in the United Kingdom. This service uses the project's intellectual advances to enable individual archival repositories to host and maintain their own data, while at the same time becoming part of a distributed national service. The current JISC project to implement EAC (Encoded Archival Context) on a national basis relies entirely on the distributed infrastructure which resulted from the project outcomes. The MerseyLibrary.org service uses these same advances applied to bibliographic, archival, and museum object information held at distributed repositories. The Resource Discovery Network supports an architecture which relies on distributed OAI repositories which are harvested and served via Cheshire. The Los Alamos Biothreat Reduction Program also relies on the Cheshire distributed architecture to construct virtual information resources for forensics and information analysis.
3. *Navigating collections*. The data transformation capabilities (GRS-1) of the Cheshire system combined with the python and tcl scripting capabilities have made an impact on the ways that collections are visualized and navigated. For example, the Archives Hub service formats multilevel EAD documents in ways which will support a "drilling down" approach, which permits users to "drill down" from generic to specific description information. The impact of facilitating such access from collections to item level information was assessed in the RSLG report (paragraph 93d).
4. *Support for cross-domain clumps to facilitate resource discovery.* The project has made an impact on true and effective cross-domain resource discovery, most notably for the MerseyLibrary.org service which supports cross-searching of distributed datasets in different formats (EAD and MARC). This model is being investigated by a number of national services with a view to implementation on a production basis. The project has shown how to deliver entire SGML/XML resources, whether they be bibliographic, full-text, or multimedia, while at the same time supporting Z39.50 and SOAP protocols. This outcome has enabled a much broader range of integrated and complete information resources to be delivered to the user's desktop.

6. Future Developments

Future developments are, therefore, likely to focus on Cheshire's support for performing "ubiquitous" search, so that:

1. Information resources that are currently inaccessible (e.g., the "invisible" or "deep" web) can be made accessible and interoperable with other information resources.
2. Information resources that currently require metadata for effective search (e.g. multimedia information) can automatically acquire metadata through metadata capture, sharing, and reuse processes that leverage ubiquitous search.

Future research priorities will have the intention of making a number of extra search capabilities possible, such as:

1. Finding appropriate and timely information to aid in user tasks without explicit query formulation (and perhaps without the user even being aware that a search is being performed).
2. Automatic discovery and application of metadata from available sources to enable users to describe, more fully and accurately, their own work.
3. Exploiting and automatically combining multiple sources of information related to the task at hand.
4. Removing the need for information users to know about the structure and content of the databases and information services that they use.

The next stage of the project is to create a high performance, scalable, and extensible platform for information retrieval support. Scalable performance will support more resource-intensive information retrieval techniques. Over the next twelve months, we have proposed and will be implementing the following:

1. Distributed queries: transparent support for effective search across many diverse databases and resources
2. Interoperability, including support for existing protocols and simple extensions of communication protocols (such as using the existing, but typically unsupported "search" request in HTTP and the service search capabilities of GRID computing protocols).
3. Concurrency, scalability, and robustness: We want the search to be ubiquitous, so that search becomes a virtually invisible part of any digital library system.
4. Dynamic databases
5. Structured, maintainable code
6. Underprivileged deployment
7. Application of shared and sharable metadata to the description of databases and database contents, including currently "opaque" content such as multimedia information resources.

In future we hope to extend the system's capability for supporting NLP (natural language processing) techniques, which may be used to identify discussion of related or associated information and the use of information extraction techniques

for populating metadata databases with that discovered information. This is beyond the scope of the JISC/NSF grant allocation, but could extend from the existing Cheshire support for extracting geo-temporal references from texts and associating these with appropriate geographic coordinates and events.

Future developments could include adding system support for context sensitive reference linking via the OpenURL standard. This will enable users to create open links which are context sensitive and may be dynamically configured within portals or managed learning environments in a fully integrated manner. If possible, this would assist in the cost-effective integration of digital collections into institutional frameworks.

The project's capabilities for transforming data will enable existing metadata to be transformed to RDF. This could be used to combine information from multiple sources, e.g. Extending support for RSS feeds for the display of events; combine, store, and optimize feeds using different modules such as bookmarks and learning objects; enable users to personalize resources or interactions typically by profiling this information.

The planned technologies described above may enable students and teachers to deal more effectively with digital library and web based information resources, including new ways to visualize content, querying appropriate collections and organizing results, and exploratory analysis tools.

We plan to extend support for enabling users to find quickly the relevant metadata and information resources themselves to satisfy their queries, wherever these may be, e.g.:

1. How to select the appropriate databases or collections for search from a large number of distributed databases;
2. How to perform parallel or sequential distributed search over the selected databases, possibly using different query structures or search formulations, in a networked environment where not all resources are always available; and
3. How to merge results from the different search engines and collections, with differing record contents and structures (sometimes referred to as the "collection fusion" problem).

The goal of our future research is to make distributed retrieval an effective part of the entire network and computing infrastructure which is automatically invoked as needed, both explicitly by user interaction but even more commonly as part of the "invisible" computing infrastructure.

7. Research Innovations

The research component has been developed with a mind to advance our basic understanding of information.

1. By enabling users to cross-search different data formats, protocols, etc., we have been able to devise a system which can be used to bridge the current classification of information into categories of text, numeric data, and geospatial information. This outcome has strengthened plain-language access to otherwise hidden information.
2. The project has derived an advanced paradigm for information discovery and retrieval that exploits the fundamental interconnections between diverse information resources, including textual and bibliographic information, numeric databases, and geospatial information resources
3. The systems architecture permits advances in the design and evaluation of information retrieval systems due to its distributed component architecture and protocols which will allow researches easily to combine and test different components of information retrieval systems.
4. The software and the "bridges" developed in the project has provided a tool to foster the serendipitous discovery of new interdisciplinary knowledge by users of the system

The project set out to achieve as its primary innovation the merging of five areas of theory and practice which, until now, have remained separate. These are:

1. Text indexing and searching using an advanced probabilistic and Boolean search engine, and distributed search and retrieval from heterogeneous databases using the Z39.50 information retrieval protocol.
2. Effective management of different metadata vocabularies, including support for cross-thesauri
3. Production of tools for knowledge discovery and intelligent filtering
4. The automatic transformation of data, e.g. From MARC to RDF/IMS for existing large scale datasets
5. Interoperability and access to geospatial information Interoperability and access to numerical databases and their metadata encoded in the codebook DTD.

The following are innovation outcomes in research issues and design which support the text discovery and retrieval elements of the project:

1. *Tools for knowledge discovery.* To enable intelligent metadata reuse within the digital library (and for managed learning) environments, and to allow natural language processing to increase semantic usefulness of this data. The project has resulted in innovative ways of optimally mapping user and document vocabularies to the controlled

vocabularies used in descriptive metadata, including advances in natural language processing and devising statistical associations between human and metadata vocabularies. This has improved the effectiveness of existing metadata resources and will reduce the reliance on expensive "handcrafted" links between metadata vocabularies.

2. *Access, Retrieval, and Filtering Information:* The Z39.50 technology utilized advanced information retrieval techniques to permit users to achieve the depth and diversity of information formerly limited to highly skilled experts. The research component of the project has focused on retrieval techniques which will permit users to achieve the depth and diversity of information formerly limited to highly skilled experts.
3. *Advanced methods for Services Related to Access:* The outcomes of the project have resulted in extending the software capabilities for protocols other than Z39.50 to facilitate true interoperability among diverse distributed databases. We have developed the Z39.50 system to act as a transformation engine, turning existing Z39.50 and other resource descriptions into RDF and IMS specifications. This technique may in future contribute to the development of teaching and learning resources, integrating geospatial, bibliographic, and other domains within that framework. The Cheshire system may now in effect locate, harvest, and index descriptions using a Z39.50 server, providing Z39.50 SOAP, HTTP, SRW, etc., interfaces which can enhanced the information interchange required by portals, MLEs, etc. We have implemented partial XMLSchema support within the Z39.50 framework for this to fulfill its true potential to assimilate a vast network of resources into portals, managed learning environments, etc.
4. *The development and integration of cross-media standards and metadata.* We have developed a new method of metadata reuse which is based on the technological advances resulting from the systems architecture: in particular, the probabilistic retrieval capabilities of the system to "cluster together" metadata elements, making it easier for the user to retrieve the most relevant entries in one or more datasets using pseudo-relevance feedback techniques. These technique will facilitate access to metadata describing numerical datasets, geospatial datasets, text and bibliographic datasets. They may be used to extend the existing capabilities for mapping multiple languages to terms used in topical metadata. We have provided preliminary client support for making visual representations of temporal and spatial information using a GIS viewer (TimeMap). The architecture itself is standards-based with support for SGML/XML (including XMLSchemas) and the Z39.50 information retrieval protocol among others (e.g. OAI, SOAP, SRW, etc.) which may further encourage the archiving and preservation of this metadata in a standards-based framework.
5. *Generic research in content technologies.* The project has also extended natural language processing to support the automatic generation of search terms, which may be extended to foreign language search terms. This may be in future be extended to develop and implement a multilingual query translator which, when combined with the Z39.50 search and retrieve protocol, could be programmed to broadcast the translated query in many languages to search appropriate foreign language catalogues. This could include the language of special communities and dialects.
6. *The integration of geographic information into the web-based service environment (e.g. Information Visualization).* The project has integrated techniques of natural language processing with a GIS information visualization system (TimeMap).
7. *Access to Digital Resoruces.* An outcome of the Cheshire project is to exploit the content of distributed repositories in an interoperable, standards-based framework. This will be the basis for extending the technology to support generic information brokers or gateways already funded, accommodating a variety of cultural and scientific datasets. The integration of scientific and cultural resources into VLEs/MLEs may result in improved access for teachers and learners while at the same time resulting in cost-effective operations of services which can remain totally distributed, rather than federated or hybrid.

A primary research objective of this project has been to bring complex discovery techniques to bear on narrowing, organizing, and rectifying the resources that follow a first-level search. The development and implementation of this as part of the Cheshire system through the techniques cited above has been undertaken to ensure that users can create the most accurate and precise views of information possible based on their interest, and the project itself is intended to achieve the highest degree of interoperability.