Cross-Domain Resource Discovery: Integrated Discovery and Use of Textual, Numeric, and Spatial Data – Annual Report 1 October 2000 – 30 September 2001

Summary Overview

This is the second annual report for the JISC/NSF funded project: "Cross-Domain Resource Discovery: Integrated Discovery and use of Textual, Numeric, and Spatial Data", also known as the Cheshire project.

One of the primary goals of the project has been to develop (and implement with real databases) a search and retrieval system based on international standards that is capable of searching cross-domain resources held in multiple locations. This goal has been achieved, although even more resources and services have yet to be added to the system, and truly seamless integration of these services will require further development of additional "middleware". We have developed and demonstrated technology for "harvesting" distributed databases and integrating the information in them into a hierarchical set of database selection servers that can be scaled to very large size without loss of performance.

The prioritized objective of the project has been to produce a next generation online information retrieval system based on international standards. We originally proposed to expand the current system ("Cheshire") and reimplement the basic components as CORBA distributed objects using both Java and C++. This proposition was modified to take account of new technological developments, as outlined in the first annual report: in particular it was decided not to implement CORBA (since server-to-server interoperability can be provided through other means) and it was decided to develop the client as a browser rather than as a Java application.

The result will be a distributed architecture in which the different components of the system can be recomposed on demand to different configurations. This will permit the dynamic implementation of different processing and retrieval methods as appropriate to given domains.

Current Strategic Relationships

Since the first annual report was issued (August 2000), the Cheshire system has been adopted by a number of large scale services in the United Kingdom as well as the rest of Europe. These include: the DIEPER Digitized European Periodicals project. See: http://gdz.sub.uni-goettingen.de/dieper/), NESSTAR (Networked Social Science Tools and Resources. See http://www.nesstar.org/, Now being continued as FASTER – Flexible Access to Statistics Tables and Electronic Resources. See http://www.faster-data.org/), MASTER (Manuscript Access through Standards for Electronic Records. See http://www.cta.dmu.ac.uk/projects/master/), ZETOC (Prototype for a new version of the Electronic Table of Contents from the British Library. See http://zetoc.mimas.ac.uk/), the Archives Hub (See http://www.archiveshub.ac.uk/), and the RSLP Palaeography project (See http://www.palaeography.ac.uk/). The system is also being used by the British Natural History Museum in London to develop new web-based services and as a prototype implementation for the metadata database of JISC data services hosted by MIMAS (Manchester Information and Associated Services).

Cheshire is currently being implemented as a service for the Resource Discovery Network (RDN), where it will be used to harvest RDN records from the various hubs using OAI and make them available for the DNER through the Cheshire Z39.50 target. (See below "*Integrating OAI and RSS metadata protocols into Z39.50 for the DNER"* for further discussion.)

The Arts and Humanities Data Service (AHDS) is now seriously considering the use of Cheshire for all of its related services. The research and development programme taken forward as part of the JISC/NSF initiative has provided both of these services with the basis for seamlessly searching across datasets and using advanced information retrieval techniques to get the users to the information they require.

We have also set up a number of Cheshire servers and clients to support the national infrastructure (in the U.K.) for various thesauri and name authority control access mechanisms. These include: support for the Library of Congress Subject Headings (LCSH), the UNESCO thesaurus, and the NRA authority files for Personal Names, Corporate Names, and Family Names. This is maintained for the general benefit of the library, museum, and archive domain, is free for all to use, and has been incorporated already into online data input forms (e.g. the Archives Hub).

During late 2001 and 2002 we expect to use the system to host all the data of the Online Archive of California, as that service migrates from proprietary software (DynaWeb) to standards-based software (Cheshire); and we also expect to use Cheshire to host a version of the MELVYL bibliographic database and part of the COPAC database. We have currently a testbed comprising OAC and COPAC datasets, and the MELVYL database has been copied to a server in Berkeley.

We are also developing new uses for Cheshire to demonstrate how the Z39.50 protocol can be extended to support applications in the non-bibliographic domain. During 2001 we released "Quick Step", a fully implemented Z39.50 email archiving tool. This is free for non-commercial or academic use.

To provide further information about Cheshire and its use in the United Kingdom (and elsewhere) we have set up a "Cheshire Resources" page which provides links to the services and tools enumerated above (this is in addition to the site home page). This page is available on: http://gondolin.hist.liv.ac.uk/~cheshire and has improved documentation for use of the Cheshire software. In addition, we have set up and are maintaining a Cheshire list-serve for developers and users of the system. This has been archived using the "QuickStep" tool; the archive and tool is publicly available via the "Cheshire Resources" page.

It may be appreciated that Cheshire is being deployed by many more sites than originally envisaged when the JISC/NSF application was submitted; but we believe strongly that such wide use is a vindication of the entirely standards-based approach that we have taken. We have tried to balance the development tasks with select support for use within the HE sector, as we believe that only through real applications of the Cheshire system can we come across problems needing to be solved or development tasks that need to be refined.

Recent Publicity Activities

In March 2001 we gave a seminar about the Cheshire project at UKOLN, attended also by staff of the Institute for Learning and Research Technologies ILRT (Bristol). This proved to be a very interesting event, since it gave us a chance to find out more about the activities of UKOLN and to determine more fully any role that the Cheshire project could play in the national infrastructure of digital library services (which is the ultimate aim of the project). It was interesting to note developments of the Open Archives Initiative (OAI) and, more broadly, infrastructural developments with various RSLP projects, particularly BACKSTAGE.

We will be closely involved with UKOLN related activities and have recently lent our support to the UKOLN application for "A National Focus for Collection Level Description Work in the UK". Similarly we expect to have increased involvement with ILRT who will be trailing the Cheshire system to see how it compares with other Z39.50 compliant systems and other forms of data distribution. We have suggested the development of a Z39.50 attribute set for the RSLP Collection Schema which could be used to support RSLP services and resources. This may be used for the RSLP project BACKSTAGE. As a result of the UKOLN seminar, we are corresponding with Peter Cliff from UKOLN, who is the Systems Developer for the Resource Discovery Network, about possible uses of Cheshire for the RDN.

Information on the database extraction and resource discovery method developed for this project was presented at the IEEE/ACM Joint Conference on Digital Libraries in Roanoke, Virginia in late June 2001 (Ray R. Larson, “Distributed Resource Discovery: Using Z39.50 to Build Cross-Domain Information Servers” See http://cheshire.berkeley.edu/papers.html).

Strategic Developments

The changes to the planned development of the Cheshire client and server were outlined in the first annual report, which contains also a fuller discussion of the development issues. We expect to have a beta-version of the new Cheshire client ready for use by October 2001, as per our deliverable list. This is based on an open source Mozilla browser and, when complete, the client should be layered seamlessly over versions of Netscape. EXPLAIN handling is now implemented; and a basic search interface is generated either from strobe or EXPLAIN. At present, the Cheshire browser displays "raw" unformatted records, but we will be introducing capabilities for style sheets to be downloaded from the XML/SGML data. The Cheshire browser is cross-platform, and worked under Solaris, Linux, Windows OS, and Macintosh OS X.

Due to the increased use of the Cheshire system (see above) we have developed (and are developing) a number of client-side records management utilities which were not originally specified in the JISC/NSF proposal. These include, for example, the capabilities of re-editing submitted records, etc., which users have requested.

We have also implemented retrieval and display of "components" of large XML or SGML documents. This will enable us to deliver either entire documents, or only selected components of those documents based in order of relevance. This was a very large development task, but one which is necessary for optimal retrieval of SGML/XML documents. To our knowledge, this is the first time this capability has been implemented within a Z39.50 based system.

In testing the new client, we have found a number of the commercially based Z39.50 systems lack many elements of the protocol, as least to a minimum version 3 standard. For example, the EXPLAIN database capability, which would normally be expected in any implementation, is often not supported. Similarly, there is often no support for SCAN, SORT, etc., even though we view these as a fundamental prerequisite for the establishment of a national distributed library infrastructure.

As a result, we strongly believe that the standard procurement process of digital libraries in the United Kingdom should be the subject of guidelines, for example based on the original procurement document for the AHDS. There also needs to be some way of determining whether these requested capabilities are actually being supported in a non-proprietary manner by the vendor. There seems to be no national validation service, even though considerable sums of money are being expended on Z39.50 systems. It may be that some form of "trip test" may be required to ensure that commercial vendors are actually supporting the Z39.50 protocol in a non-proprietary manner, as is required for interoperable digital library services.

Progress has been somewhat slower for the development of the java-based Cheshire III server due in part to a resignation of a staff member; but, more importantly, to a number of unanticipated technical problems with the JavaSpaces technology. The technical problems encountered were largely issues of performance: JavaSpaces transactions take considerably longer to complete than indicated by the literature. The solution for this is to retain SDLIP and Z39.50 as the primary interaction mechanisms for the distributed elements of the system. This will provide the performance needed for the system while ensuring continued compatibility with legacy systems supporting these protocols. However, we are making progress on development of the “next-generation” java-based server and expect to have a basic working version by the end of 2001. In addition to this work on the Cheshire III server, we have implemented a number of additions and changes to the existing Cheshire II server which will be rolled into the Cheshire III development. This work has focussed on adding additional features needed for the various services currently (and prospectively) using the system. The features added since the last report include:

- The ability to extract and separately index components of SGML or XML documents. This permits elements of large SGML/XML documents to be treated as if they were entire documents.
- More complete support for dynamic databases, including improvements to the facilities for updating and deleting records and index entries.
- More complete support for automatic generation of the IR-Explain-1 database providing server-level metadata on the Z39.50 accessible databases of a Cheshire II server.
- Numerous performance enhancements for the system, significantly increasing the speed of indexing, record conversion and display, and retrieval.
- Sorting of any retrieved set by any set of tags in the underlying XML records, or by the

elements used to create indexes. This facility also permits merging of multiple result sets from independent searches.

- Many other features to support special requirements of different databases and services.

These and future additions to the existing c-based Cheshire II server are now being used in services such as the Archives Hub and new requests for enhancements are being circulated to the Cheshire Steering Group for prioritization of the development process.

Currently we are experimenting with distributed search and retrieval, including the construction of collection "surrogates" derived from the Z39.50 SCAN services. This work was reported on at the Joint Conference on Digital Libraries in June, and will be demonstrated at the SIGIR meeting in New Orleans this September. A paper describing the efficiency (very good) and effectiveness (apparently also very good) of this method is being prepared for publication in the journal literature.

The project so far has focused on the design and performance objectives. We intend to test the system against datasets in the testbed during the final year of the project, and are currently testing against serveral large scale datasets, including COPAC (portion), Archives Hub, Online Archive of California, History Data Service, etc.. The goal will be to provide a working prototype allowing the user can search across different data, including EAD, MARC, CODEBOOK, etc. We intend to be able to provide cross-searching capabilities for the different metadata implied by these various data types.

Development of the Cheshire III Client:

The development of the Cheshire III client as a browser is progressing and is expected to be completed by the end of October 2001. We are currently using the open source third party project called "Protozilla", which has created a set of handlers to allow further protocols to be embedded within Netscape/Mozilla. We expect these handlers will be incorporated within the base Netscape/Mozilla distribution in the near future.

We are exploiting this package to allow Netscape/Mozilla and the existing Cheshire code to be glued together in a standardized way, rather than to implement a proprietary series of linking routines and duplicating the work of others.

The use of Cheshire and Protozilla within the Netscape/Mozilla framework permits us to maintain the state of multiple Z39.50 sessions at once, rather than the usual "fire and forget" method of http, which has resulted in the invention of proprietary, inefficient techniques such as Cookies and large server side activities databases.

Although the browser client will be the primary Cheshire III client, our strategy dictates that we maintain a number of different clients for different types of users. These include the existing Cheshire web client, a tcl client, and (in future) a Cheshire layer of MVD (Multi-Valent Document), an extensible document structure written as a Java application. However, the Cheshire III browser client is expected to be the most efficient and functional of all these and the one which we expect will be most used.

The Microsoft Corporation has recently dropped support for Java in its new products, including the new Windows XP operating system. We feel this validates our original decision to develop and implement the Cheshire III client as a browser in a way which conforms completely to international standards. Although we expect to develop support for Cheshire within a Java application (MVD), this will be a slower, less functional version.

Related JISC Funded Projects

Cheshire now fully supports all of the development aims of the Join-Up and Z-Gate Projects funded through the JISC 5/99 call. In many cases, the system can now extend the capabilities originally envisaged for Join-Up and Z-Gate projects and is incorporating these capabilities for national services throughout the United Kingdome. Although Cheshire's current funding from JISC/NSF is primarily purely research led, its widespread adoption by these and other services represents considerable return on investment.

In particular:

- Cheshire will be used as a Z39.50 system for the Resource Discovery Network and the Arts and Humanities Data Service portals, supporting cross-searching of datasets hosted at the JISC national datacenters.
- Cheshire now extends the concept of ZBLSA (a Z39.50 broker associating journal citation records to full text articles held by non-Z39.50 servers) by supporting delivery of full text and multimedia using Z39.50.
- Cheshire supports all of the request and deliver options of Docusend.
- Cheshire is being trialed as a prototype for ZETOC, a MIMAS service providing Z39.50 compliant access to the British Library's Electronic Table of Contents (ETOC).
- Cheshire already supports the interface of Bath Profile clients to non-Bath Profile targets.
- Cheshire is being used to develop the metadata database of JISC services hosted by MIMAS
- Cheshire now supports all of the proposed functionality of JAFER (Java Access for Electronic Resources) and can be extended allow the creation of dynamic internet-based learning aids and portals within the extensible MVD (Multi-Valent Document) framework (part of the California Digital Library project).
- Cheshire is already a participant of the National Focus for Collection Level Description Work in the UK interoperability focus.

In addition, the research and development aspects of Cheshire are expected to make an original contribution to the JISC funded HILT (High-Level Thesaurus) project, which aims to research, report, and make recommendations on the problems of cross-searching and browsing by subject across a range of services and data types. In particular, our JISC/NSF funding has enabled us to prototype enhanced support for metadata and vocabulary in a cross-domain context without the need for handcrafting expensive cross-walks between different thesauri. We are about to implement this as part of the Archives Hub service and plan to implement these techniques for the Resource Discovery Network, the Arts and Humanities Data Service, and various other JISC datasets. It is expected that the techniques evolved from our pure research project will, practically speaking, leverage existing invements in metadata by making existing metadata more accessible; by generating new metadata through automatic metadata assignments; and by transferring existing metadata to apply to other, additional datasets. (See section below on Entry Vocabulary Modules.)

Enabling the DNER: The Role of the "Cheshire" Information Retrieval System.

On 14 June 2001 we responded to the document "The DNER Technical Architecture" by issuing a statement defining a role of the "Cheshire" information retrieval system within the emerging architecture of the DNER. This has been published on our web site and outlines how the research and development objectives funded through JISC/NSF could make a contribution to the emerging information landscape within the United Kingdom.

Essentially:

- The outcomes of the JISC/NSF funded project have ensured that the Cheshire system will support all of the information retrieval and metadata protocols likely to form the DNER, including Z39.50, SDLIP, OAI, RSS, and HTTP, and can be extended to support many others. Cheshire can also support user-profiling and authentication as required by the DNER.
- Cheshire will support XML and SGML as the primary database format of the underlying search engine, which means that the system will provide a common storage format for the variety of data most likely to comprise the DNER, including images, full text, bibliographical, etc. In this way it extends a number of capabilities originally envisaged for the Join-Up and Z-Gate projects.
- Cheshire will support advanced information retrieval techniques to enable end-users to efficiently discover information (e.g. probabilistic searching, relevance feedback, automatically generated hypertext links, nearest neighbour searches, support for Entry Vocabulary Modules). These advanced retrieval techniques rely on extending the research

objectives outlined in the original JISC/NSF proposal.

Combined, these capabilities will greatly simplify the process of access and discovery for end users.

Effective searching of multiple Z39.50 targets within the DNER

One of the primary outcomes of the research funded through JISC/NSF has been the development of techniques enabling effective searching of multiple Z39.50 targets. To date, service providers have relied upon simultaneous "broadcasting" of search requests to all servers (or "targets") making up the distributed sources to be searched.

One of the chief drawbacks of such "broadcast searches is that all systems must be searched before the user or search controller can determine which systems are most likely to provide the results that the user is seeking. This is highly ineffecient and particularly impractical for large-scale distributed datasets.

The extent of the DNER requires that a technical solution needs to be found and implemented as a matter of priority.

We have examined this as a research objective within the context of the JISC/NSF project and have come up with what we believe to be an original solution. Instead of using "broadcast" searches, we have instead used the SCAN service of Z39.50 servers to build combined indexes containing information "harvested" from individual servers. This is a highly efficient, standards-based way of searching across any number of distributed datasets and could easily be implemented for the whole of the DNER.

Integrating OAI and RSS metadata protocols into Z39.50 for the DNER

In future some of the JISC-funded collections may be non-Z39.50 available datasets, but support the OAI (Open Archives Initiative) or RSS (RDF Site Summary) metadata protocols (assuming that RSS is expressed in the XML format). Neither OAI nor RSS are information retrieval protocols. This means that, in order to be included within the DNER, there must be some method of "harvesting" these OAI and RSS datasets and providing search and retrieval of the metadata they contain.

We have developed the research objective above to use the Cheshire Z39.50 system to "harvest" high level metadata within the DNER, using the SCAN service of Z39.50. This metadata can then be used to quickly locate appropriate repositories which could form the basis of a selective Z39.50 cross-search which could retrieve the full text, etc., of the relevant individual objects.

We have implemented this already for the datasets comprising TREC (Text Retrieval Evaluation Conference) and various JISC datasets with results which were surprisingly accurate and fast.

The above technique can be extended further to enable the whole of the DNER to conform to the Z39.50 information retrieval protocol, with Z39.50 targets seamlessly interoperating with each other. Therefore, far from having an unsustainable number of different protocols to support, the end user would only have to interact with a single Z39.50 based interface which would efficiently knit together the entire range of DNER data resources.

This does not mean that the Cheshire system has to be used by all of the information resource providers, since the Z39.50 services already have the capability of interoperating with each other. The new client will be able to search an independently maintained database of proxy Z39.50 Explain records to compensate for those services which do not fully support the Explain standard.

The above scenario would make it simple for the DNER office to maintain a list of Z39.50 servers (including OAI and RSS harvesters). It would reduce substantially the managing and maintenance cost of the DNER, while increasing functionality for the end-user.

Such an approach is interesting because it is based on advances in pure research into information retrieval issues, rather than the use of conventional Z39.50 toolkits.

Metadata Reuse: Entry Vocabulary Modules (EVMs)

One primary research objective of the JISC/NSF project is to enable the enhanced retreival of unfamiliary metadata using what we call "Entry Vocabulary Modules", or EVMs. This capability, growing out of the Cheshire project, is really a method of constructing linkages between natural language expressions of topical information and controlled vocabularies automatically.

One of the more common challenges facing any end user is in navigating various data sources which might use different thesauri. The Archives Hub is a case in point: data contributors follow either the LCSH (Library of Congress Subject Headings) or UNESCO thesauri. How do users unaccustomed to using either thesauri find out the information of interest to them? One solution, currently examined by HILT, is to create higher level cross-walks between the various thesauri. This is a very expensive solution which, in many cases, proves to be unsatisfactory as well as being expensive to maintain.

A key objective of the project is to discard this methodology and instead develop more research-oriented methods of providing access to these subject headings, no matter how unfamiliar and bewildering they may be to the end user, by automating the process of association between natural languages and their subject headings.

To facilitate this, we have exploited the capabilities of the Cheshire system's support for probabilistic information retrieval on any indexed element of the datase(s). This means that we can use a natural language query (for example, plain English) to extract the most relevant entries in one or more databases. From this information the server can *automatically* present to the user a cluster of subject headings which might be relevant to their inquiry. The user then can select the subject heading or combination which is most appropriate and then use this as a basis for a more effective subject search across the different databases.

This capability has been remarkably effective in enabling users to map their query to the controlled vocabularies (subject headings) used in descriptive metadata; much more so than traditional boolean methods. But a greater (and unanticipated) benefit may be that we are now able to cross-search different thesauri and automate associations between them and the user's inquiry.

We are now experimenting use of this to facilitate automatic subject retrieval across any number of thesauri supported by a number of distributed datasets. Although there is some more development to be done in this area, the initial results appear to bear out the initial promise of cross-searching different thesuari without the usual overheads of handcrafting links of and between different vocabularies. This could form a substantial contribution to the DNER infrastructure.

As a result, we have now incorporated this as part of the Archives Hub service and plan to extend it to other JISC datasets comprising the DNER.

Future Research Issues: Metadata Reuse and Geospatial Information

The initial findings suggest that further research in this area will greatly facilitate access to metadata, particularly those describing geospatial datasets. Although the project is currently seeking to integrate access to numerical, textual, and geodata we are now in a position to take forward a number of research issues which build upon our present findings:

Specifically, we could now examine methods of mapping geographic place names in text (natural languages) to probable geographic coordinates; and 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 the 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, all of which can be used to facilitate 1) Time adjacency searching ("within 10 years of 1537"); 2) Place-name disambiguation (relating place names to gazetteer entries); 3) Geographical adjacency searching ("within 50km of Paris"); and 4) Graphic display by time and place.

Arguably, such metadata reuse in the context of geospatial searches could, in future, result in dramatically enhanced retrieval of unfamiliar metadata, thus producing an increased return on the existing large scale investments in digital libraries.

Advisory Committee

As part of the mandated JISC activities for this project we formed an steering committee consisting primarily of people concerned and involved in the development of distributed information services in the U.K. The members of the steering committee are:

Sheila Anderson (Director of the Arts and Humanities Data Service)
Lynne Brindley (Executive Directory, The British Library)
Reg Carr (CEI Chair and Head Bodliean Library, Oxford University)
Julia Chruszcz (Director of National Data Services, Manchester Computing)
David Dawson (ICT Advisor, Re:Source)
Catherine Grout (Assistant Director, Development, JISC DNER Office)
Derek Law (Chair) (Librarian of Strathclyde University Library)
Ray Lester (Department of Library and Information Services, Natural History Museum)
Paul Miller (Interoperability focus, UKOLN)
John Perkins (Executive Director of the Consortium for Museum Intelligence -- CIMI)
Andy Powell (Assistant Director and Team Leader – Distributed Systems and Services, UKOLN)
Frances Thomson (University Librarian, University of Liverpool)
Ray R. Larson (PI)
Paul B. Watry (Co-PI)

The first meeting of the steering group was held at the British Library in London on 18 April 2001. The "terms of reference" for the steering committee are:

Terms of Reference for the Cheshire Steering Committee

1. To advise on the project on behalf of the Higher Education and Further Education Sectors, the various non-HE Sectors which hold a stakeholding interest in the project, and to the Joint Information Systems Committee.
2. To provide a reporting framework for the JISC, the DNER office, the National Science Foundation and related organizations.
3. To receive periodic reports from the Principal Investigators of the "Cheshire" project.
4. To represent the best interests of the Higher and Further Education sectors.
5. To represent the broader interests of museums, libraries, archival repositories, all of which have a stakeholding interest in the project
6. To act as an advocate for the project and its aims
7. To increase visibility of the project within the UK HE and FE sectors as well as the wider communities of museums, archival repositories, libraries, etc.
8. To read and comment on the annual report and any interim reports.

The meeting saw a need for additional support for adopters of the Cheshire technology, either through commercialization of the software and services or through further support from JISC.