Cross-domain Resource Discovery: Integrated Discovery and use of Textual, Numeric, and Spatial Data

Ray R. Larson
(University of California, Berkeley)

Paul B. Watry
(University of Liverpool)

January 12, 1999

1 Introduction

The pursuit of knowledge by scholars, scientists, government agencies, and ordinary citizens requires that the seeker be familiar with the diverse information resources available. They must be able to identify those information resources that relate to the goals of their inquiry, and must have the knowledge and skills required to navigate those resources, once identified, and extract the salient data that are relevant to their inquiry. The widespread distribution of recorded knowledge across the emerging networked landscape is only the beginning of the problem. The reality is that the repositories of recorded knowledge are only a small part of an environment with a bewildering variety of search engines, metadata, and protocols of very different kinds and of varying degrees of completeness and incompatibility. The challenge is to not only to decide how to mix, match, and combine one or more search engines with one or more knowledge repositories for any given inquiry, but also to have detailed understanding of the endless complexities of largely incompatible metadata, transfer protocols, and so on.

This proposal describes an information access system that provides a new paradigm for information discovery and retrieval by exploiting the fundamental interconnections between diverse information resources including textual and bibliographic information, numerical databases, and geo-spatial information systems. This system will provide an object-oriented architecture and framework for integrating knowledge and software capabilities for enhanced access to diverse distributed information resources.

1.1 Overview

The purpose of this proposal is twofold, involving both practical application of existing technology, and theoretical examination and evaluation of next-generation designs for systems architecture and for intelligent assistance in the information retrieval task. For the first purpose we propose to develop and make ready for production a next-generation information retrieval system based on international standards (Z39.50 and SGML) which will be used for cross-domain searching, using the Arts and Humanities Data Service (AHDS), the CURL (Consortium of University Research Libraries), the Online Archive of California (OAC) and the Making of America II (MOA2) database as principal repositories. This system will serve as a model for developing efficient paradigms for information retrieval in a cross-domain, distributed environment. The second purpose will be addressed in the design, development, and evaluation of the distributed information retrieval system architecture, its client-side systems that aid the user in exploiting distributed resources and in the design and evaluation of protocols for efficient and effective retrieval in a internationally distributed multi-database environment.

The aim is to produce a robust, fully operational system ("Cheshire") within a three year period which would facilitate searching on the internet across collections of "original materials" (i.e., early printed books, records, archives, medieval and literary manuscripts, and museum objects), statistical databases, full-text, geo-spatial and multi-media data resources. This system will be based on the work done with the Cheshire II system in UC Berkeley Digital Library Initiative project, extended with additional capabilities and redesigned with a new system architecture. This standards-based client/server system will have important economies for libraries, museums, universities, and other information providers and the system produced will be made available without charge to non-profit, government and educational institutions.

The new extensions to this system will provide a platform and protocols to integrate databases with fundamentally different content and structure into a common retrieval, display, and analysis environment. These different database types, and some examples to be used in this project, include:

- Document databases which describe information about various topics ranging from news reports and library catalogue entries to full-text articles from academic journals including text, images and multimedia elements. (Oxford Text Archive, Performing Arts Data Service, California Sheet Music Project, CURL database, the Digital Archive of California and the Making of America II (MOA) database).
- Numeric statistical databases which assemble facts about a wide variety of social, economic, and natural phenomena (History Data Service, NESSTAR and UC Data).
- Geographic databases derived from geographic information systems, digitized maps, and other resource types which have assembled georeferenced view of the geographic features and boundaries including georeferenced information derived from place names (Archaeology Data Service, History Data Service, the UC Berkeley Digital Library database and the MOA database).

This proposal draws upon the continuing 3-year working relationship between the Division of Special Collections and Archives at the University of Liverpool Library and the Digital Libraries Program at the University of California, Berkeley, with the aim of extending it to other research-based repositories in the EU and USA. In addition to the development platforms of the AHDS, CURL, OAC and MOA2 databases, other repositories expressing an interest in the Cheshire project and in using the Cheshire software include Glasgow University (contact: Lesley Richmond), Oxford University (contact: David Price), Durham University (contact: John Hall), the Public Record Office (contact: Meg Sweet), the British Library (contact: Richard Masters) in the UK, the Archive Working Group at Yale University (Contact: Richard Szary) and the Institute for Advanced Technology in the Humanities at the University of Virginia (Contact: Daniel Pitti) in the US.

2 System Description

The system that we are proposing will build upon international standards and existing work in probabilistic information retrieval, and on the experience of the researchers in applying advanced retrieval methods to full-scale realistic databases.

2.1 A Functional Model

The continuing development of the Cheshire client/server system is based on a particular vision of how information access tools will develop, in particular, how they must respond to the requirements of a large population of users scattered around the globe who wish simultaneously to access the complete contents of thousands of archives, museums, and libraries, containing a mixture of text, images, digital maps, and sound recordings. Such a virtual library must be a networked-based distribution system with local servers responsible for maintaining individual collections of digital documents, which will conform to a specific set of standards for documentation description, representation, and communications protocols. We believe, based on the current directions of research and adoption of standards by libraries, museums and other institutions, that a major portion of this emerging global virtual library will be based on SGML (Standard Generalized

Markup Language), and especially its XML subset, and the Z39.50 information retrieval protocol for resource discovery and cross-database searching. (We also assume that the forthcoming versions of the HTTP protocol will continue to provide document delivery and hypertext linking services, and that SQL3, when finalized, will provide the low-level retrieval and data manipulation semantics for relational and object-relational databases)

The Cheshire retrieval system, in supporting Z39.50 "Explain" semantics for navigating digital collections, will allow users to locate and retrieve information about collections that are organized hierarchically and distributed across servers. It will enable coherent expressions of relationships among objects and collections, showing for any given collection superior, subordinate, related, and context collections. These are essential prerequisites for the development of cross-domain resources discovery tools, which will enable users to access diverse collections through a single interface.

In its functionality, the proposed system will satisfy all of the recommendations of the National Resource Discovery Workshop (MODELS 4), and will address many of the research and development topics noted in the most recent JISC strategy review, as well as the development programmes of CEI and TASC.[1] It specifically addresses the critical issue of "vocabulary control" by supporting probabilistic "best match" ranked searching (as discussed below) and support for "Entry Vocabulary Modules" (EVMs) that provide a mapping between a searcher's natural language and controlled vocabularies used in the description of digital objects and collections. It also allows users to "navigate" collections (the "drilling down approach") through distributed Z39.50 "explain" databases and through the use of SGML as the primary database format, particularly for collection-level descriptions such as the EAD DTD. The system will follow the recommendations of the Third National Resource Discovery Workshop by providing fully distributed access to existing catalogues, and is designed to support cross-domain "clumps" to facilitate resource discovery. Finally, the proposed server anticipates the critical issue of displaying non-western character sets in its ability to handle UNICODE (in addition to the standard ASCII/ISO8859 character sets).

3 Background

3.1 Development History

The development of the Cheshire system began in the early 1990s at the University of California, Berkeley, as a means of testing the use of "probabilistic information retrieval methods" upon MARC bibliographic data. It was found that these advanced retrieval methods developed at Berkeley were far more effective than traditional Boolean methods (or vector space model methods) in accessing records from a bibliographic database. Needless to say, the deployment of these "probabilistic" retrieval algorithms has very important economies particularly in the searching of databases or documents such as EAD which normally do not use a controlled vocabulary.

The second version of Cheshire, currently deployed at both the University of Liverpool and the University of California, Berkeley, was designed to extend the format of the server to include SGML-encoded data. Because SGML is increasingly becoming the mark-up language of choice for research institutions, it was critical to extend Cheshire's capabilities to support the kinds of SGML metadata which is likely to be included in national bibliographies. These are: TEI (Text Encoding Initiative), EAD (Encoded Archival Description), DDI (for Social Science Data Services), CIMI (Consortium for the Interchange of Museum Information) records, as well as the SGML version of USMARC released by the Library of Congress (based on the USMARC DTD developed by Jerome McDonough for the Cheshire project).

The third version extends the use of SGML handling capabilities for these search indexes. This version was developed by Berkeley and Liverpool for the Arts and Humanities Data Service, enabling GRS-1 syntax

[1] MODELS 4: Integrating Access to Resources Across Multiple Domains (December 1996).
http://www.ukoln.ac.uk/models/models4.html. See also MODELS 3: National Resource Discovery Workshop: Organizing Access to Printed Scholarly Material; by Lorcan Dempsey and R. Russell.
http://www.ukoln.ac.uk/models/models3.html.

conversion for nested SGML data, component indexing and retrieval of SGML formatted documents, and automatic generation of Z39.50 Explain databases from system configuration files. The current version of the server is now able to include an element in an SGML record that is a reference to an external digital object (such as a file name, URL or URN) that contains full-text to be parsed and indexed, these can be local files or URL and URN referenced files anywhere on the internet. It also enhances the users' ability to perform somewhat less directed searching provided by Boolean and probabilistic search capabilities that can be combined at the user's direction. This version of Cheshire can display a number of data types ranging from full-text documents, structured bibliographic records, as well as complex hypertext and multi-media documents. At its current stage of development, Cheshire forms a bridge between the realms of purely bibliographic information and the rapidly expanding full-text and multimedia collections available on-line.

The system proposed here will support nested SGML DTDs, promote cross platform use, and support a far broader range of SGML document types, specified below. Specifically, this stage will involve the development of the Cheshire client (currently undertaken by researchers at Liverpool), the provision of a CORBA-based distributed object version of both client and server software (undertaken by researchers at UC Berkeley), the development of the Multi-Valent Document model by researchers at Berkeley and Liverpool to accommodate delivery of multi-media and GIS information without the need for plug-ins or helper applications, and application of Entry Vocabulary Modules (EVMs) enabling search support for unfamiliar metadata vocabularies including cross-language retrieval (http://sims.berkeley.edu/research/metadata/), to facilitate retrieval of information for large-scale, distributed information services.

The development and integration of these tools will enable extensive testing and comparative evaluation on the service providers that comprise the Arts and Humanities Data Service (AHDS), the CURL database, the Online Archive of California and the Making of America II database, with the aim of providing a production service for these entities and delivering production-quality versions of the Cheshire software for Higher Education Institutions in the United Kingdom and the United States.

4 Research Issues: Integrating Access to Resources Across Domains

The development of the Cheshire system outlined above has been driven in the belief that many of the recommendations of the National Resource Discovery Workshops (MODELS 3 and 4) can be catered for by a standards-based information retrieval system that will provide a bridge between existing on-line catalogue technology and databases and the explosively growing realm of network-based digital libraries with information resources including full-text, geo-spatial data, numerical data and multimedia. In the following sections, we will discuss each of these recommendations and the active research issues and design elements particular to the proposed Cheshire system that are associated with them.

These sections address the following research issues: 1) Distributed Object Retrieval Architecture 2) Management of Vocabulary Control in a Cross-Domain Context; 3) Distributed Access to Existing Metadata Resources; 4) Navigating Collections; and 5) Support for Cross-Domain Clumps to Facilitate Resource Discovery:

4.1 Distributed Object Retrieval Architecture

We see the architecture for the evolution of distributed information access systems as a highly flexible and dynamic system. In such a system both the *data* (digital objects instantiating information resources) and the programs that operate on that data (*methods*) to achieve the needs and desires of the users of the system for display and manipulation of the data (*behaviours*) will be implemented in a distributed object environment using CORBA for object management.

The basic architecture is a three-tiered division of data and functionality. The tiers are:

1. *The Client.* The basic client for the distributed Cheshire system can be any JAVA-enabled WWW Browser. The primary data delivery format will be as XML (or HTML for initial versions), and the methods for manipulating and navigating within the data will be implemented as CORBA-enabled JAVA applets, delivered on demand to the browser.

2. *The Application Tier* Applications for search and manipulation of data are distributed between the client and network servers (including the repositories) to provide distributed functionality (and to provide new behaviours to clients on demand from any compliant network server). The application tier or layer would both provide JAVA applets for execution on the client, as well as providing server-side methods invoked directly on objects in the repository either via direct CORBA invocations or indirectly via requests from other protocols (e.g. Z39.50 or Open Geospatial Datastore Interface (OGDI) for network access to heterogeneous geographic data held in multiple GIS formats and spatial reference systems(Gardels, 1996, 1997)). For example, a client browser might download an applet that can display MARC records, and invoke a server-side method to convert repository objects in XML to MARC format. We expect, for performance reasons, that many operations on stored objects will be server-side methods with primarily display functions on the client side.

3. *The Repository* Digital objects and metadata describing them will reside in the repositories tier or layer. Repositories can be implemented in a variety of ways, ranging from conventional Relational, Object-Relational, or Object-Oriented database systems and Text retrieval engines, to metadata repositories referencing physical collections in libraries.

4.2 Management of Vocabulary Control in a Cross-Domain Context

This is a key issue in the integration of resources across domains, as brought forward in MODELS 4. The underlying problem, recognized for over a decade, is that the current generation of on-line catalogues in most libraries do not do a very good job of providing topical or subject access to the collections (Matthews, et. al., 1983), The common result of many subject searches (up to 50%) is search failure, or "zero results". In a distributed environment, this is compounded by the lack of "vocabulary control" across domains, added to which is the tendency for users to use general wording or terms in subject queries, rather than specific ones.

In its initial form, the Cheshire system was initially designed to overcome these difficulties, and provide users with the tools for formulate effective queries. In its current configuration, the server is 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. This is provided for by a variety of search and browsing capabilities,[2] but the primary distinguishing feature is the support for probabilistic searching" on any indexed element of the database. This enables the use of a natural language queries to retrieve the most relevant entries in one or more databases, even though there may be no exact Boolean matches.

The results set of a probabilistic search is ranked in order of estimated relevance to the user's query. The search engine also supports relevance feedback, as well as automatically generated hypertext links that will allow the user to follow dynamically established linkages between associated records.

Support for probabilistic retrieval is critical to the success of a cross-domain server, insofar as it allows users to make effective queries even when there is no controlled vocabulary. In the case of the Social Science documents, for example, a user can make a successful "probabilistic search" on a given subject, where a traditional Boolean search would fail. The deployment of these algorithms is only a preliminary steps in managing vocabulary control. There are, naturally, many research issues to be addressed in finding the optimal mappings of user and document vocabulary to the controlled vocabularies used in descriptive metadata this is discussed further under "Search Support for Unfamiliar Metadata Vocabularies" below.

[2] Among the materials used for translating a searcher's query into the terms used in the databases are: elimination of unused words using field-specific stopword lists, particular field-specific query-to-key conversion or "normalization" functions, algorithms for reducing significant words to their roots or stems by converting suffix variations, such as plural forms of a word to a single form, as well as support for mapping database and query text words to single forms.

4.3 Distributed Access to Existing Metadata Resources

4.3.1 Data Mining: Dublin Core Metadata

One approach to semantic interoperability of distributed systems is to use a standardized set of metadata, such as the Dublin Core, for the description and retrieval of electronic resources from disparate data. For example, the Arts and Humanities Data Service (AHDS) currently operates on a model in which extended Dublin Core elements are used as the means of retrieving information from five different service providers. In practice, this has proved to be an inefficient way of effectively leveraging data from these services, since the data providers often interpret Dublin Core elements differently and also because Dublin Core elements only comprise a small part of the complex, rich data resources which could be available as a means of search and retrieval. To take one example, a full text search of the existing AHDS gateway (http://prospero.ahds.ac.uk:8080/ahds_live) for the term 'England' in the Dublin Core element set for the Oxford Text Archive produces only one 'hit'; whereas a Cheshire version of the TEI-header information of the same service (http://sherlock.berkeley.edu/OTA/) produces 258 'hits', ranked in order of relevance.

This is why many of the fundamental retrieval algorithms being developed as part of the Cheshire project, described below, are based on the premise that front-end prototyping will involve entire information resources, not simply restricted subsets based on Dublin Core metadata. These will provide a much richer platform for the development of retrieval strategies among large and complex data sets, and the inclusion of the Arts and Humanities Data Service (AHDS) in this project will bring particular expertise in this area, since this service currently is a practical implementation of the Dublin Core and the service providers are already experienced in a production environment in the use of Dublin Core for resource discovery (Miller & Greenstein, 1997).

4.3.2 Search Support for Unfamiliar Metadata Vocabularies

The next step beyond simple shared category lists like Dublin Core will be to provide support for enhanced retrieval of unfamiliar metadata (as distinct from Dublin Core Metadata), extending the findings of the DARPA-funded research project on *Search Support for Unfamiliar Metadata Vocabularies* (http://sims.berkeley.edu/research/metadata). This research, based on work from the Cheshire research projects, is attempting to go substantially beyond the state-of-the-art in developing systems that can construct linkages between natural language expressions of topical information and controlled vocabularies automatically. Today most such systems depend on the expensive human crafting of links within and between vocabularies.

For the purposes of this project we propose continued development of the Cheshire client and application layer middleware to provide sets of "entry vocabulary modules" based on the controlled vocabularies of our testbed databases. These "EVMs", will accepts natural language expressions of user's queries and will generate a ranked lists of controlled vocabulary headings most likely to be useful for that search. This will have three uses:

1. as a prompt when searching an unfamiliar vocabulary
2. as computer-aided or automatic indexing of data resources using existing controlled vocabularies
3. to extend searches, using derived information of found records as a basis for finding similar records in another database

When used in conjunction with the existing Cheshire algorithms for probabilistic indexing and retrieval, these EVMs provide descriptive surrogates can be used to match user or document terminology to corresponding controlled vocabulary terms.

4.4 Navigating Collections (the “Drilling Down Approach”)

One of the primary considerations brought forward in the discussion of search models during the MODELS workshop 4 is the use of Z39.50 to support a “drilling down” approach, which would permit users to “drill down” between generic and domain-specific descriptive information. The difficulties of this in the context of Z39.50 are cited in the MODELS 4 recommendation for further work [item 2.3].

In designing the second and subsequent versions of the Cheshire system, we faced the question of how to provide a search engine that could support a navigational record schema, that could be used on both simple text and complex structured records, and also support complex multimedia documents and databases. In answer to this, it was decided to adopt SGML as the fundamental data storage type for the Z39.50 client/server. Virtually all data manipulation for the database has been generalized as processes acting on SGML tags or sets of tags. Instead of having to develop new routines to manipulate each sub-element of a new datatype, the developer only needs to provide a DTD and a conversion routine to convert the new data type to SGML. The built-in file manipulation and indexing routines can then extract and index any tagged sub-elements of the data type for access.[3]

In the proposed distributed system architecture, this functionality will be part of the application layer and will be available to both client and repository manipulation of the SGML/XML data via CORBA distributed methods.

In using SGML tagging for all data in the database and by adopting the SGML DTD language to define the structure of each data file, it is possible to use a common format for data types ranging from full-text documents, structured bibliographic records, to complex hypertext and multimedia documents (using the HTML DTD that defines the elements of the WWW “pages”). This has important economies in the delivery of resources across domains.

We propose to support a far broader range of SGML document types, and to provide JAVA methods for display of SGML documents on the client. The obvious candidate for this is DSSSL (Document Style Semantics and Specification Language, international standard: ISO/IEC 10179), although the use of XML (Extensible Markup Language) with XSL (Extensible Style Sheets), a restricted subset of SGML with additional formatting capabilities will also be supported. At a practical level, by creating style sheets for the most commonly used SGML data types (EAD, CIMI, DDI, TEI), it will be possible to deliver visual representations of nested data using multiple DTDs. In order to achieve this, the participants will have to agree on some common visual representation of data, requiring consultation among institutions. The University of Liverpool has already begun by developing and distributing a prototype conversion programme which will format and index archival finding aids encoded in EAD. The functionality of this programme will become part of the client-side methods for the Cheshire client (http://gondolin.hist.liv.ac.uk/˜azaroth/ead2html.html).

4.5 Support of Cross-Domain Resource Discovery

Cross-Domain Resource Discovery is the area of primary concern for this proposal. Its importance has been emphasized by both MODELS workshops and addressed in phase 3 of the JISC eLib programme in the UK and has been an important area of research in the NSF/NASA/ARPA Digital Library Initiative projects in the US. Indeed, the idea of a National Scale Resource Discovery system for the UK is based on groupings or clumps of OPACs, whether they be physical or virtual clumps, predicated on the use of Z39.50.

Although the sheer variety of data types produced for indexes, catalogues and archival listings across domains remains a challenge, evidence demonstrates that most institutions are now using SGML as the mark-up language of choice. The recent report for BIBLINK, for example, documents the wide-spread use of SGML as the encoding format for the kinds of metadata which are likely to be included in national bibliographies. These include TEI, EAD, CIMI, and DDI, as well as the SGML version of USMARC released by the Library of Congress (and originally developed as part of the Cheshire project).

[3] A basic configuration file, itself an SGML document, defines the physical database elements, including the locations of data files, which SGML DTD describes the file, and information on which indexes to create and the elements they should contain.

Many of the principal international bibliographic data carriers such as Research Libraries Group (RLG) and OCLC are now developing strategic initiatives predicated on the use and development of SGML formatted applications. There is currently a movement to resolve interoperability issues around the CIMI, EAD, TEI, DDI, and MARC DTDs.

Despite this, relatively little work has been undertaken into the use of production level SGML and XML-based search engines using the Z39.50 information retrieval protocol. This makes the development, testing, and implementation of a Z39.50 client/server using SGML, as described in this proposal, a critical priority for development. It is, in our view, an absolute prerequisite if the recommendations coming out of the MODELS workshop 4 are to be implemented.

The proposed Cheshire system, with its potential to display hierarchically organized information about digital collections distributed across servers, is probably the only working model for cross-domain resource discovery that is entirely standards-based. In its use of SGML, it allows institutions to localize their own descriptive information, while permitting remote users distributed access via a structured information retrieval protocol (Z39.50). The use of DSSSL/XSL permits users to navigate consistently across collections, which may not have conventionally defined structures, provided they can be expressed in SGML.

In particular, it is proposed to extend the SGML/XML handling capabilities of Cheshire to exploit an extraordinary range of documents, providing CORBA-based methods for extracting and indexing their contents available to any client, application or repository that conforms to our overall architecture for distributed information retrieval systems.

5 Testbed Development

To help develop the appropriate technologies we propose to use two large-scale information services sponsored by JISC in the UK which offer complementary data formats: The Arts and Humanities Data Service (AHDS) and the CURL (Consortium of University Research Libraries) databases and two large-scale distributed object databases in the US (The Online Archive of California, and the Making of America II databases). Although these will be the focus of the present proposal, we will also be bringing together a consortium of related data providers who may wish to test data using the proposed Cheshire system: these include government agencies (the Public Record Office); Universities (Glasgow, Durham, Liverpool, and Oxford); as well as hybrid library projects sponsored as part of the eLib 3 programme in the UK and the Archive Working Group at Yale University (including other participants in the US, UK and Australia) the Institute for Advanced Technology in the Humanities at the University of Virginia (Contact: Daniel Pitti) in the US. We will also be providing the Cheshire technology to and participating in the development of the NESSTAR (Networked Social Science Tools and Resources) project.

The NESSTAR project (http://dawww.essex.ac.uk/projects/nesstar.html) is combining the skills and knowledge of the three main partners, The Data Archive in the UK, Danish Data Archives and Norwegian Social Science Data Services with assistance in significant areas of user analysis, usability, user validation, evaluation and quality assurance from the Institute of Journalism in Norway, ASEP and JD Systems in Spain, Central Statistics Office in Ireland and Aarhus University in Denmark. The Council of European Social Science Data Archives is a sponsoring partner for the project. The project is funded by the European Commission under the Information Engineering sector of the Telematics Applications programme.

It is our intention that the technology developed as part of this research proposal will serve as the basis for full-scale information systems of international prominence.

We chose the AHDS, CURL, OAC, and MOA2 databases as the focus of our work for several reasons: The data sets are large and of a diverse nature; users of these services represent a broad range of technical expertise; both have a well-founded administrative structure with existing user-evaluation mechanisms (thus reducing research overhead costs); finally, the proposed system would give considerable added value to the repositories themselves, which already comprise valuable national and international resources.

Both PI's have had a long-standing connection with the AHDS (An earlier version of Cheshire was installed by the PI's for the History Data Service – one of the AHDS providers). This forms part of the AHDS gateway. In addition one of the principal investigators, Paul Watry, is a member of the CURL RDD Committee, which includes development of the COPAC service in its remit.

A brief description of the data services making up the core testbed for this project follows:

The CURL Database The Consortium of University Research Libraries currently gives access to its bibliographic database via COPAC, a Z39.50 service funded by JISC and supported by Manchester Information Datasets and Associated Services (MIDAS). The COPAC service currently consists of some 3.5 million MARC records held in a central server at MIDAS, but there are plans to extend this database to non-bibliographic data resources, such as full-text and EAD-encoded documents.

Arts and Humanities Data Service The Arts and Humanities Data Service (AHDS) is a national service funded by JISC to collect, describe, and preserve the electronic resources which result from research and teaching in the humanities. This research project will focus on the current production service of four data services. The targets for AHDS include:

1. Archaeology Data Service (ADS): The ADS uses a proprietary DBMS system (Fretwell-Downing's VDX system) to store extensive data incorporated as part of the resource, such as geospatial images, aerial photography, and CAD images.
2. Performing Arts Data Service (PADS): PADS is currently using Hyperwave, an object oriented information retrieval system, to locally store and retrieve information retained in a variety of formats.
3. Oxford Text Archive (OTA): The OTA holds its entire corpus as SGML documents.
4. History Data Service (HDS): The History Data Service holds its data as SGML documents which point to a number of numeric and alpha-numeric data, text, digitized boundry data, and images converted from historic source documents into computer-readable form.

These four services are available via the AHDS gateway with access points determined by an extended version of Dublin Core. In addition, there are full-text versions of the History Data Service and the Oxford Text Archive (TEI-header information only) available via Cheshire clients and servers.

1. We intend, first, to convert the metadata from the ADS and PADS databases to SGML. (Full-text from OTA and metadata from HDS are already encoded in SGML and require no conversion.)
2. Then, in order to further extend the capabilities of these databases, we intend to develop a Cheshire client to access methods for all primary data types (text, image, and map-oriented data), indexing each document by as many methods as are applicable. For example, photographs will be indexed not only by the content of their images, but also by their text, their pre-assigned categories, their location, and so forth. This will necessitate development of a Cheshire client that will support formatting of SGML documents, using DSSSL or XSL. These indexed documents will provide a basis for testing retrieval algorithms as well as for pre-processing and post-processing retrieval results sets.
3. It should be noted that a Cheshire query for these data resources may necessitate interaction with one or more interoperable clients displaying a different document data type: for example, geographic information system (GIS) data sets. A GIS-oriented browser delivered in response to a Cheshire query will make it easy to ask for information pertaining to a geographic region. In devising such a system, we will be integrating the access tools developed for the DARPA-funded Berkeley Digital Libraries project into the Cheshire environment. For example, a related client is being developed by researchers at UC Berkeley to display a union of aerial photographs from UC Santa Barbara (Project Alexandria) and UC Berkeley databases.

4. Finally, we intend to use the HDS data as a testbed for integrating numerical statistical databases and geographic databases within the Cheshire system.

Online Archive of California : The Online Archive of California Project, is a two-year pilot project to develop a UC-wide prototype union database of 30,000 pages of archival finding aid data encoded using the Encoded Archival Description (EAD) SGML document type definition. This database will serve as the foundation for the development of a full-scale digital archive for the University of California System (UC) available via the Internet to diverse user communities.

Making of America II : The Making of America II is a Digital Library Federation project to continue and extend research and demonstration projects that have begun to develop best practices for the encoding of intellectual, structural, and administrative data about primary resources housed in research libraries. The Making of America II Testbed Project collection will be ”Transportation, 1869-1900”, particularly the development of the railroads and their relationship to the cultural, economic, and political development of the United States. It will comprise multimedia information coordinated by SGML metadata and “hub” documents.

In addition to using data from the data services cited above, we will also focus on identifying additional text and bibliographic resources to interact with the numerical and geospatial data sources. We plan to include, for example NESSTAR (discussed above) and the UC Data archive, and the existing Digital Library Initiative database at Berkeley as part of these extended resources.

6 Project Development Details

6.1 Systems Development and Integration Design and Development Methods

The specific aims of the development work over the three-year duration of the project are as follows:

1. To encode the client/server library in Java to promote cross-platform use. Specifically, the system will require a Z39.50 communications class library, and an SGML parser class library all written in Java or, minimally, a graphic user interface (gui) encoded in Java making calls to the C code. We intend to begin incrementally, encoding the GUI first before progressing to the other class libraries. This implies the development of entity management work of opening, closing, and editing files. We intend to use Native Methods to tie in other functionality, including the Z39.50 libraries, the SP SGML parser, and the Jade DSSSL engine.

2. The client software will have to support a far broader range of SGML document types than it currently does. For this to prove practical, some convention for display of SGML documents needs to be implemented for the client. The obvious candidate is DSSSL, which would imply also implementing a DSSSL engine in Java. The JADE system from James Clarke is implemented in C++ and would have to be ported to Java. This would be available as a CORBA service for any SGML/DSSSL data.

3. The collections proposed will produce some very complex SGML documents, with some containing binary data. This will require the delivery of multiple files to the client for display of a single document, possibly including separate files for the SGML declaration, the DTD, a DSSSL style sheet, the basic document, and multiple referenced files containing binary data. At a practical level, this means that for a client to display a document retrieved, in response to a query, it will have to obtain the SGML declaration, the DTD, any DSSSL/XSL specifications, and the base document, and invoke appropriate services (methods) to parse them. It will also resolve SGML entity references from the parsed base document to other files comprising the document and retrieve those additional files (we plan to use extensions of Z39.50 to accomplish this initially), and finally utilize the additional retrieved document

components (including any DSSSL style specifications) to generate a display of documents for the users. The mechanisms used both to refer to different files composing a single SGML Document (URN's, etc.) and locate and retrieve the component parts, will need further specification. It is acknowledged that specifying these mechanisms will be both a technical and political issue, as control of and access to files from various institutions will have to be arranged within a single, technical framework.

4. We are assuming that the clients will be Java-based and will employ Java's support for UNICODE to enable the display of character sets outside the limited repertoire of ASCII and ISO8859 Latin-1. Second, the indexing and retrieval methods of the applications layer (servers) will need to be recoded to handle UNICODE character sets, in addition to standard ASCII/ISO8859 sets. Font sets for all languages present with the various collections will need to be either located or produced for inclusion in the client software

5. Given the wide variety of SGML input mechanisms over distributed sites (including SGML editors, word processors, and DBMS, involving varying degrees of detail in a variety of media), we intend to examine encoding practices under SGML with a view to cross-domain standardization. For example, we plan to consult with appropriate authorities and participants of Yale Archives Group on the on research issues relevant to the distributed access of archives on-line for EAD and MARC-AMC encoded materials. These include: ISAAR (CPF); management of vocabulary control; metadata; and content mapping between ISAD(G), EAD, MARC-AMC, and MARC-AMC DTD.

6. Throughout this project, we will use the SGML handling capabilities of the Cheshire system to exploit different document types and to extract and index their contents and to create (Entry Vocabulary Modules) EVMs for intelligent mapping from natural language to the controlled vocabulary used in databases. This will include the implementation of Entry Vocabulary Modules for access to standardized classification systems (such as the Library of Congress Classification) and other controlled vocabularies used in participating databases.

7. One particularly relevant extension to this work will be the DDI DTD, used by the History Data Service (HDS) and by NESSTAR, to provide electronic "codebooks" for numerical databases. In this respect, we will seek to integrate numeric statistical databases which assemble facts about a wide variety of social, economic, and natural phenomena and document databases which describe information about various topics, ranging from new report and library catalogue entries to full-text articles from academic journals. Both Berkeley and Liverpool have been collaborating with the History Data Service and NESSTAR (Networked Social Science Tools and Resources) project in Europe to use the Cheshire system for storage and retrieval of DDI encoded data about numerical databases.

8. Given the Archaeology Data Service's (ADS) focus on environmental information (archaeological digs), an area of particular interest in terms of this project is geographic information systems (GIS). A primary development goal will be to integrate geographic databases (geo-data) with text and numeric data. These constructs are not only of a conceptual nature, but will form a working prototype within the AHDS itself.

9. We plan to convert the metadata associated with the multi-media resources of PADS and ADS to SGML. The converted version of this metadata will be used for testing planned extensions for CORBA. We intend to draw on the work of researchers at UC Berkeley on "ZQL" for SQL access to object-relational DBMS via Z39.50.

10. The integration of diverse information resources in the production environment that extends beyond the participating databases described above will require research into the design a "Metaprotocol" for describing the protocols needed to interact with the distributed system components and databases which might be used by databases and systems outside the project. The metaprotocol could be used to describe the data objects transferred by a given protocol, and the expected interactions between distributed components. Information in the metaprotocol might be used to allow a new protocol to be dynamically incorporated into the system. We plan for an initial version of this protocol to be based on the Z39.50 Explain service which can provided detailed information about the databases, search capabilities and data attributes of Z39.50 servers. This will need to be extended with support for CORBA method descriptions (e.g., IDL and the CORBA Interface Repository).

7 Performance Goals and Research Plan

Over the three years of this project there are several major objectives, each with a number of research tasks and milestones that will be critical to the success of the project as a whole. The systems development work will be divided between the University of California, Berkeley and the University of Liverpool. In general, client design and development will be done at Liverpool; server design and development will be done at Berkeley.

The project is designed in three phases:

1. Year One: Design of prototypes for access to each type of data;
2. Year Two: Design of preliminary prototype for traversing between the types;
3. Year Three: Demonstration of production prototype for accessing each and traversing between them.

Year 1 Tasks

- Design and implement a preliminary CORBA-based version of the Cheshire system on the existing C code base, including preliminary design for the distributed component protocols (Berkeley).
- Preliminary design of user interface; encoding Z39.50 client library in Java, including support for UNICODE. (Liverpool). 18 months for the first pass, and 36 months for production level code.
- Implementation of DSSSL engine in Java. (Liverpool). 12–18 months, using the JADE system ported from C++.
- Implement prototype Entry Vocabulary Modules.
- Design and implement preliminary support for geographic coordinate search and retrieval in the Cheshire system.
- Develop a matrix of data types, and of hardware/software environments to ensure a full range of problems; formalize existing in-house data and net-based resources to create a defined testbed for this research.

Year 2 Tasks

- Design and implement a Java, C, and C++ CORBA-based distributed component version of the Cheshire III server (Berkeley).
- Testing and performance evaluation of the distributed components protocols for the distributed Cheshire system. (Berkeley).
- DSSSL specifications for EAD, TEI, DDI; Z39.50 attribute profile for EAD. (Liverpool).
- Cross institutional standardization of SGML encoding practices. (Liverpool).
- Test and evaluate improved Entry Vocabulary Modules. Integrate with the Cheshire system. (Liverpool and Berkeley).
- Modify current OGDI/gltp software to accept input from the indexing/searching engine and to acquire metadata, geospatial objects, and attribute information from remote geographic datastores; work to inform the OGC abstract model definition process for catalogue services.
- Design for the MetaProtocol for description and dynamic adaptation to new telecommunications protocols.
- Test Cheshire interoperability with other OGDI/gltp programmes.

Year 3 Tasks

- Design and implement an improved version of CORBA-based version of the Cheshire system, incorporating support for the MetaProtocol as well as full support of the component architecture.
- Implement and test improved user interface with users.
- Evaluation of use and effects on user work of the system.
- Document and report on the system, its use and value.
- Demonstrate interoperability across multiple systems and protocols.

7.1 Deliverables

1. The production of a robust, scalable on-line information retrieval system based on international standards (Z39.50 and SGML) which can be used as a delivery mechanism for cross-domain resource discovery projects. This system will be portable to a range of system platforms, and will support the navigation of digital collections, as specified in *Z39.50 Profile for Access to Digital Collections*. Navigational support will enable users to retrieve lists of related collections, including parent, superior, and context collections, brief descriptions of these collections, and descriptions of their relationship to the subject collection.
2. The Arts and Humanities Data Service has agreed to serve as hosts for the code, to be made available via ftp. In addition, UC Berkeley will continue to be a host in the United States and Liverpool will continue to be a host in the United Kingdom. Installation will be possible for any institution running a Unix or Wintel NT-based system, and the source code will be self-documenting.
3. Consortium participants (AHDS, CURL, OAC and MOA2) will participate in a trial service involving a wide range of databases. The demonstration of this system extended over a number of sites in the United Kingdom will show how a standards-based information retrieval system can support distributed access to existing catalogues, statistical databases, multi-media, GIS, and full-text data.
4. Z39.50 formatting for commonly used SGML formats will be produced for server/client use based on DSSSL or XSL.
5. Consortium members will document progressive stages of the project, demonstrating the development of each of the consistent components, taking a model, standard, or prototype element and extending and adapting that element to production-level, and showing how the overall structure conforms to a particular vision of the next generation of on-line catalogues and similar on-line information systems. It is intended to publish further findings on retrieval algorithms, the use of SGML structured documents as database objects, management of distributed objects for information access and retrieval, user interfaces, and user reactions to the advanced on-line catalogue.

8 Project Management Arrangements

The Principal Investigators will be Professor Ray R. Larson, School of Information Management and Systems, University of California, Berkeley for the US/NSF sponsored half of the project and Dr. Paul Watry, Special Collections and Archives, University of Liverpool for the UK/JISC sponsored half. The host US institution will be the University of California, Berkeley and the host UK institution will be Liverpool University. The additional consortium partners will include the Arts and Humanities Data Service Executive (Director: Daniel Greenstein), the History Data Service (Sheila Anderson), the Oxford Text Archive (Michael Popham), the Performing Arts Data Service (Celia Duffy and Carona Boehm), the Archaeological Data Service (Julian Richards), and the Consortium of University Research Libraries (contacts: Ric Collins and Julia Chruszcz).

It is proposed to create a management committee, chaired by a University Librarian, that will meet at regular intervals with project site representatives.

Prior to making a formal proposal, all participants will attend a project meeting chaired by a University Librarian to draw up a memorandum of agreement avoiding any potential conflicts. All participants will be formally committed to implementing and testing the system, and participating in design of the DSSSL/XSL formats.

9 Project Evaluation and Dissemination

As outlined in the Tavistock report(Stern & Kelleher, 1995), the project will deploy formative and summative evaluation strategies as outlined below:

9.0.1 Formative Evaluation Activities

We intend to use two methods of data collection, transaction monitoring and on-line questionnaires, to obtain information on how the system is used, and on the reactions and opinions of users about system features and capabilities. The data collected by transaction monitoring can be used to reconstruct the users' interactions with the system for detailed analysis of the types of searches conducted and their results.

Transaction monitoring will be supplemented by an on-line questionnaire, which can provide insight into what users feel about certain features and the user interface. The questionnaire will be based on the user questionnaire developed for the Council on Library Resources national survey of on-line catalogue use and users.[4]

We intend to provide a continuous assessment of the project and dissemination of information about it in publications and in workshops. Already disseminated are articles which appeared in the *Journal of the Society of Archivists* and to *JASIS*. The role of the AHDS in ongoing information dissemination will be critical to our strategy of broadening the relevance and use of the access tools described in this proposal.

9.1 Summative Evaluation

The summative evaluation will focus on the use and usability of the system, and its effectiveness in the context of system performance.p This project will support valuable analysis of information systems, retrieval algorithms, and data structuring in a production environment. The conclusions reached, by both internal and external evaluators, may impact on future investment and resource allocation decisions, as well as the design criteria for future information access systems.

10 Intellectual Advances

A primary focus of the research in our project is the development of better content-based access methods, along with better paradigms of user interaction with content once located. The research will, consequently, advance our basic understanding of information science in the following ways:

- The current classification of information into categories of text, numeric data, multi-media and geospatial information will be greatly dissolved. This will in turn greatly strengthen plain-language access to otherwise hidden information

[4] The on-line questionnaire and transaction logs will allow for analysis of the different types of searches conducted on the system. one goal is to determine if trends in searching shift from type to type (e.g. title to subject or topic) with the use of a probabilistic catalogue. For the user questionnaire developed for the Council on Library Resources *see* (Matthews, et. al.,1983).

- The system which will provide a new paradigm for information discovery and retrieval, exploiting the fundamental interconnections between diverse information resources
- The system and the "bridges" built as part of the project will provide a tool that will foster the serendipitous discovery of new interdisciplinary knowledge the by users of the system, including teachers, researchers, students, and ordinary citizens

The originality of the proposed work is two-fold:

- In designing ways to traverse between very heterogeneous data environments; and
- Adopting a component-level approach to information retrieval systems

We believe that the architecture briefly outlined above represents and fundamental and inevitable step in the design of information access and retrieval systems. Each aspect of the indexing and retrieval process carried out by the system can be implemented in a shared and distributed fashion facilitating the application of new combinations and sequences of retrieval and indexing algorithms for the differing needs of particular data sources or user requirement. This will provide a platform for systematic analysis and exploration of new information retrieval methods and permit examination of retrieval algorithms at the component instead of the system level.

11 Conclusions

The provision of a production-level version of this retrieval system will certainly have important economies in the development and implementation of digital libraries in receipt of funding. More than this, the issues of navigation, indexing, and searching raised as a result of this funding allocation would be an important step in determining how standards-based, large-scale network resources can become comprehensible to a diverse user community with relatively little in the way of resources, guidance, or assistance.

In today's ever-broadening landscape of on-line databases, web sites, search engines, and protocols, it has become increasingly difficult for users to know where, how, and what to search to satisfy their needs for information and knowledge. Today's information seekers must have extensive knowledge and skills to navigate these on-line resources and to extract the knowledge relevant to their needs.

We have proposed a standards-based, next-generation on-line information system to aid the user in exploiting the fundamental interconnections between diverse information resources, including textual and bibliographic information, multi-media, numerical databases, and geospatial information resources. It can provide a platform and protocols to integrate databases with fundamentally different content and structure into a coherent common retrieval, display, and analysis environment to promote the discovery and use of new knowledge.

A Prior NSF Support

The PI (Larson) was a faculty investigator on the NSF/NASA/ARPA Digital Library Initiative Project (IRI-9411334) entitled "The Environmental Electronic Library: A Prototype of a Scalable Intelligent Distributed Electronic Library". P.I.s: Robert Wilensky and Michael Stonebraker. $4 million, 9/94-8/98.

This project (involving many faculty and graduate student researchers) developed a prototype digital library focused on the California environment. The Library and its contents are accessible via http://elib.cs.berkeley.edu. The project explored many areas ranging from document decoding and scanning to information categorization and retrieval. The project showed that it was possible to build an effective and useful large-scale digital library of scanned document images, and to enliven those images through developments in Optical Character Recognition, and the development of a new model for online documents called "Multivalent Documents". The project also showed that effective retrieval from OCR data (with relatively high misinterpretation of characters and words) is possible to implement (The Cheshire II system is the primary text search and retrieval engine for the project). The project also made several breakthroughs in computer vision research and in the design of scalable information retrieval methods for image retrieval (also using Cheshire II in conjunction with selection and matching methods developed by computer vision researchers). The project resulted in many publications, many of which are accessible from the project web site. Specifically relevant publications for this project are:

M.K. Buckland & C. Plaunt. "On the construction of selection systems." *Library Hi Tech*, 48 (1994):15-28.

M.K. Buckland & Plaunt, C. "Selecting Libraries, Selecting Documents, Selecting Data." In *Proceedings of the International Symposium on Research, Development & Practice in Digital Libraries 1997, ISDL 97, Nov. 18-21, 1997, Tsukuba, Japan.* Tsukuba, Japan: University of Library and Information Science, 1997, Japan. Pp. 85-91. (http://bliss.berkeley.edu/papers/isdl97/isdl97.html)

Ray R. Larson, "Geographic Information Retrieval and Spatial Browsing" In: L. Smith and M. Gluck, Eds. *GIS and Libraries: Patrons, Maps and Spatial Information*, Urbana-Champaign : University of Illinois, 1996. (p. 81-124)

Ray R. Larson, Jerome McDonough, Lucy Kuntz, Paul O'Leary, and Ralph Moon, "Cheshire II: Designing a Next-Generation Online Catalog." *Journal of the American Society for Information Science*, 47(7) (July 1996), p. 555-567.

Ray R. Larson and Jerome McDonough. "Cheshire II at TREC 6: Interactive Probabilistic Retrieval." In E.M. Voorhees and D. K. Harman, Eds. *Information Technology: The Sixth Text REtrieval Conference.* Gaithersburg, MD : NIST, 1998. (NIST Special Publication 500-240)

Ray R. Larson. "Interactive Probabilistic Retrieval: Cheshire II at TREC 7." in TREC 7 notebook (to appear in TREC 7 proceedings).