# Cross-Domain Resource Discovery:
# Integrated Discovery and Use of Textual, Numeric, and Spatial Data

Project Plan
Created: August 2000

Ray R Larson (University of California, Berkeley)
Paul Watry (University of Liverpool)

## 1. INTRODUCTION

This project plan outlines in detail the technical and administrative components of the JISC/NSF awarded project "Cross-Domain Resource Discovery: Integrated Discovery and Use of Textual, Numeric, and Spatial Data" jointly awarded to the University of California, Berkeley and the University of Liverpool in 1999.

The purpose of the grant is to enable the examination and evaluation of next-generation designs for systems architecture and for intelligent assistance in the information retrieval task.

Specifically, the grant is to develop and make ready for production a next-generation online 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 OPAC (Consortium of University Research Libraries), the HE Archives Hub, and the Making of America II (MOA2) and Online Archive of California (OAC) as principal repostories. Such a system is intended to serve as a model for developing efficient paradigms for information retrieval in a cross-domain, distributed environment.

In doing so, the following will be addressed:

❿ 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
- the design and evaluation of protocols of efficient and effective retrieval in an internationally distributed multi-database environment

The project aim is to produce a robust, fully operation system ("Cheshire") within a three year period which will facilitate searching on the internet across collections of "original materials" (e.g., 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 initially 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 educational institutions.

## 2. PROJECT AIMS AND OBJECTIVES

The project aim is to produce a fully operational system within a three year period aims to satisfy the following objectives:

Stage 1: Base client/server building
Stage 2: Metadata and format handling
Stage 3: Demonstration of production prototype for accessing each type of data and traversing between them.

Development of these in the course of the project will be addressed in the workplans, set out below.

The project goal 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. This will involve:

- Practical application of existing Digital Library technologies to some large-scale cross domain collections
- Theoretical examination and evaluation of next-generation designs for systems architecture and distributed cross-domain searching for Digital Libraries.

## 3. MANAGEMENT STRUCTURE

The management of the project will be divided between two sites: one at UC Berkeley and another at the University of Liverpool.

The relationship between University of Liverpool and the University of California has been formalized by a consortium agreement which outlines the roles and responsibilities of each partner.

The project consists of two principal investigators (PIs): Ray R. Larson at the University of California, Berkeley and Paul Watry at the University of Liverpool. The Principal Investigators have overal responsibilities for the project and are also responsible for the daily management of the project, coordination of activities across Liverpool and Berkeley, progress control and development of project plans, financial management, and liaison with external associations.

Each site will have a team of programmers dedicated to specified programming tasks, set out in the workpackages below.

The project will report to a convened Steering Committee which will be comprised of a broad cross-sectoral constituency of experts. The Terms of Reference for the Project Steering Committee outlines the roles and responsibilities of this committee. Members represent the central management of the project, the JISC/CEI, UKOLN, the Consortium of University Research Libraries, specialist partners, and other stakeholder communities.

The Steering Committee **Terms of Reference** are as follows:

1. To advise the project on behalf of the Higher Education Funding Councils for England and its territorial partners, and the Committee for Electornic Information, providing periodic reports to those agencies.
2. To represent the best interests of the broader higher education/further education community in the United Kingdom by advising the Project on how best to develop software commensurate with the information needs of scholarship and research, teaching and learning in educational institutions.
3. To receive periodic reports from the project on its progress, future project plans, and associated milestones and deliverables; to comment on such plans; and to verify that the project has the means to secure the realization of these plans.
4. To support the Project and to act as advocates for the project and its staff in furthering the project aims, with particular reference to the need to maintain for the project a sufficient level of visibility throughout the educational sector, information management, and research communities.
5. To review the annual report.

The project Steering Committee will meet annually, but will be provided with project documentation in more frequent intervals.

The current members of the Project Steering Committee are:
Reg Carr (CEI)
Andy Powell (UKOLN)
Paul Miller (UKOLN/Re:source)

Sheila Anderson (AHDS)
Derek Law (JISC/HE Archives Hub)
David Dawson (Re:source)
Frances Thomson (Liverpool University Library)
Richard Masters (British Library)
Julia Chruszcz (MIMAS)

There will be informal meetings of implementors, e.g. Peter Robinson's implementation at DeMontfort University and implementations at MIMAS, which will be supplemented in due course by a Cheshire User Group (CHUG).
The published reports and software outgoing from the project will be circulated to stakeholders in order to facilitate pick-up on a national and international scale.

## 4. PROJECT EVALUATION AND DISSEMINATION

### 4.1 *Initial Assessment*

This process focuses on assessing the project tasks and devising more detailed workplans which will enable the technical implementation of the project software. This process has been concluded and is set out in the detailed technical workplans set out below.

The main outcome of the initial assessment process has been to refocus the project's design objectives and reengineer Cheshire into a modern software system, with the hope of ensuring its future viability as a platform for information retrieval research. The work required to achieve these objectives are set out in the workpackages; the objectives themselves are discussed in more detail in Section 3 of the annual project report, available on the Cheshire web-site.

### 4.2 *Formative and Summative Evaluation Strategies*

As the software is developed, we propose to establish focus groups to provide feedback involving all stakeholders in the project. These focus group evaluations will focus on the use and usability of the system, and its effectiveness in the context of systems performance. The conclusions reached will inform the design criteria for the project.

In addition, we intend to use two methods of data collection, transaction monitoring and online questionnaires, to obtain information on how the system is used, and on the reactions and opions 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 online questionnaires, which may 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 online catalogue use and users.

### 4.3 *Dissemination:*

Dissemination of the current version of "Cheshire" is ongoing, with an increased take-up in the United Kingdom, including participants in the HE Archives Hub, MIMAS data sets, local implementations (e.g. Warwick University), EU Funded projects (e.g. MASTER), and various JISC services (e.g. History Data Service). It is expected that as the new code is written upgrades would be made to all of these stakeholders in an effort to increase take-up even further. The Annual Report of the project will outline in more detail implementation strategies, as well as the proposed migration paths for the redesigned Cheshire system.

## 5. PROJECT START DATE

The timetable is based on a project start date of November 1999.

## 6. PROJECT PHASES

Broadly speaking, the project is divided into three phases which will be used to delineate tasks and deliverables. These phases can be summarized as follows:

Phase 1: (Months 1-18) will be spent analysing the tasks to be done, familiarizing ourselves with the codebase, and scripting the initial prototype version of the code in a way which will ensure the integration of the Berkeley/Liverpool work. This will include the base client and server building.

Phase 2: (Months 18-30) will be spent refining and debugging the codebase, including extra support for GIS, and limited testing on databases supplied by AHDS and MIMAS. This will include work on the "metaprotocol" and the creation of a high performance, "network of workstations" style operation.

Phase 3: (Months 30-36) will be spent assimilating feedback, refining the system further, and seeding implementations throughout the University of California and select UK Higher Educational institutions.

These phases are broken down into discreet workpackages, as listed below:

## 7. PROJECT WORK PACKAGES

Each work package listed below is broken down into main elements and outputs. These work packages describes in more detail the technical work ongoing for the project and are intended to *supplement* the information outlined in the intial project proposal and in the annual reports.

## *WORKPACKAGE 1: Base server/client building(Phase 1: Months 1-18)*

Introduction

Before starting to work on the project we first commenced an initial assessment exercise to assess the design objectives in light of new software technologies which have evolved since the proposal was originally written. A number of these technologies can facilitate achieving the project objectives. These are set forth in Section 3 of the Annual Report, but are briefly referred to here in the context of revised workpackages.

This is a client/server system with the development of the client largely being written at the University of Liverpool and the server at the University of California, Berkeley. After some investigation, we decided not to adopt some of the original technologies incorporated in the original proposal, as follows:

*Changes in Client Design:*

Although the project proposal implied a Java based system for the client, we have subsequently decided to utilize the Mozilla framework being developed in an Open Source fashion by the Netscape Corporation. This has greatly increased the existing resources available for use in the creation of the client.

The client must be able to be used under any operating system, be that Windows, Linux, MacOS, as well as on as many hardware platforms as possible as well. Mozilla has been designed to fit this from the ground up, with the ability to be cross-platform of primary concern in the implementation strategy.

The client interface must be familiar to as many users as possible. As can be seen in the information technology marketplace every day, most new products have the same style of interface as those that have gone before in order to make the learning curve for consumers as brief as possible. By simply extending Mozilla to handle the Z39.50 protocol, this learning curve will be minimalised as most of the functionality of the client is already present in the original, and thus already familiar.

Mozilla is made up of small discrete components that fit together into a larger whole. As such, adding additional functions to it is just a matter of writing a component that works together with the other components already written. This reduces the development time as well as ensuring that when Mozilla is updated, the changes needed to keep the Z39.50 component synchronised will be minimised while still having access to all future advances in the Mozilla framework.

Mozilla implements XPI - the Cross Platform Installer. This is a means of having new components or User Interface modifications installed automatically by clicking a button on a web page or similar method. As in any system accessable via the Internet, clients will be run over network connections of different speeds. The startup time for a fully implemented java client run inside a conventional web browser over a typical modem connection from a home PC would be unbearably slow. By only requiring a once off install of the Z39.50 component and User Interface, this is minimised.

Finally Mozilla supports all of the current relevant standards. It supports the Document Object Model level 1 (DOM1), HTML version 4.0, ECMAscript (standardised javascript), Resource Discovery Framework (RDF), eXtensable Markup Language (XML), Cascading Style Sheets (CSS) and so forth. Adherence to standards is essential to ensure that the product remains usable with different servers and data types.

*Changes in the Server Design:*

Although the original project proposal outlined support for CORBA (a distributed objects architecture used for building distributed systems), we have subsequently decided that support for Z39.50, SDLIP (Simplified Digital Library Interoperability Protocol), and JavaSpaces will provide all the server-to-server interoperability required for the Cheshire project.

The server design will now incorporate a number of new software technologies, including Java Programming Language; Java RMI Remote Method Invocation; Java HotSpot Server; Berkeley DB 3.1; JavaServlet Pages; DOM Compliant XML parsers; Forte Integrated Development Environment; Java Naming and Directory Interface; SDLIP; and JavaSpaces. These software technologies and their use in the Cheshire project are discussed in detail in Section 3.3 of the Annual Report.

In essence, this workpackage will focus on building the client/server package which incorporates these new technologies. The new version of the Cheshire system will be designed for parallelism and scalability. Finally, the workpackage addresses the migration path. The strategy will be to adopt an incremental approach where a minimally functioning system will be constructed and will evolve into the full system as development continues. This is outlined in Section 3.4 of the Annual Report.

*Elements:* Elements of this 18 month workpackage include:

➓ Planning and familiarization phase
➓ Z39.50 URL phase
➓ Z39.50 scripting phase
➓ Interface building phase
➓ Integration and verification phase

These are discussed in further detail in the project's annual report.

*Outcomes:* This workpackage will enable design objectives of the Cheshire client/server to be met. These include support for distributed queries, interoperability, concurrency, web-based system administration, dynamic databases, maintainable code, underpriviledged deployment, and a focus on high performance "network of workstations" style operation.

*WORKPACKAGE 2: Metadata and Format Handling(Phase 2, Months 18-30)*

Introduction

Our initial work in Cross-Domain Resource discovery has concentrated on using the facilities of the Z39.50 information retrieval protocol to implement what we are calling a "Meta-Search" capability using the existing Z39.50 severs and resources.

A number of existing attempts at distributed search and resources discovery (including the AHDS implementation) have relied on broadcasting of search requests to all servers making up the distributed resources to be searched. There are a number of practical problems with this approach, outlined in Section 4.5 of the Annual Report, which make it impractical for large-scale implementations.

For this workpacakge, we intend to use the SCAN service of Z39.50 servers to build combined indexes containing information "harvested" from the individual servers. This will enable us to interrogate many more targets than is at present the case. Once this capability is introduced, it will be easy for any Cheshire serer to function as a meta-search server for some group of other servers.

This workpackage will explore the functionality of the Z39.50 SCAN operand within the Cheshire system and investigate how to merge search results from multiple sources to enable the system to work efficiently with potentially thousands of nodes.

*Outcomes:* Automating the Meta-indexing process
Enabling efficient cross-domain resource discovery across hundreds or thousands of networked nodes.
Better support for cross-domain information needs in a networked environment.

*WORKPACKAGE 3: Entry Vocabulary Modules (Phase 2, Months 18-30)*

This workpackage is based on work in "Search support for unfamiliar Metadata Vocabularies" sponsored by DARPA, and will enable intuitive mapping from natural languages to controlled vocabularies to enable users to data-mine.

Months 12-24
*Elements:* Implement prototype entry vocabulary modules, including work on translingual mappings.

Months 24-36
*Elements:* Test and evaluate improved Entry Vocabulary Modules. Integrate with the

Cheshire system.

*Outcomes:* Support for Entry Vocabulary Modules to enabled enhanced subject access across different databases.

*WORKPACKAGE 4: Support for the GEO profile for geographic information retrieval (GIR) applications (Phase 2, Months 18-30)*

This workpackage is concerned with providing access to georeferenced information sources. It includes all of the areas of traditional IR research with the addition of spatially and geographically oriented indexing and retrieval.

Months 12-24:
*Elements:* Design and implement preliminary support for geographic coordinate search and retrieval.

Months 24-36:
*Elements:* 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.

*Outcomes:* Support for GIR applications

*WORKPACKAGE 5: Use of current (Z39.50) and new (SDLIP) Protocols For Access to Other Metadata Systems (Phase 2, Months 18-30)*

Months 1-36
*Elements:* Support for common semantics (e.g. Dublin Core, Bath Profile). Work with the UKOLN RSLP Collection Description Project. Support for multiple simultaneous registered Z39.50 profiles. Research into a metaprotocol for communicating information about databases, search elements and collections (initially based on Z39.50 Explain).

*Outcomes:* Cross domain use of Cheshire system.

*WORKPACKAGE 6: Testbed Applications (Phase 3, Months 30-36)*

The deployment of the Cheshire software in an actual working environment as a "proof of concept" is critical to the success of this project. Following on from initial development of the server and client, we propose to test the system against a number of JISC datasets held at MIMAS and the Arts and Humanities Data Service as well as datasets at the University of California. These include: COPAC, HE Archives Hub, History Data Service, OAC, MOAC, among others. These testbed applications will serve as the foundation for formative and evaluation activities, which will indicate usefulness of the system, and will result in a final project report for publication. The testbed activitiy is likely to begin earlier than scheduled (Month 30).

*Elements:* Use of system in a working environment
Evaluation activities
Project report

*Outcomes:* Refinement of system
Robust testing
Final report

## 8. KEY DELIVERABLES

| | |
|---|---|
| **Month 18:** | Production of base client and server code (Phase 1) |
| **Month 30:** | Support for metaprotocol (Phase 2) |
| **Month 36:** | Testbed Implementations (Phase 3) |
| **Month 36:** | Publication of project report |