Cheshire 3 Framework White Paper: Implementing Support for Digital Repositories in a Data Grid Environment

Paul Watry
Univ. of Liverpool, NaCTeM
pwatry@liverpool.ac.uk

Ray Larson
Univ. of California, Berkeley
ray@sherlock.sims.berkeley.edu

Abstract

This paper outlines the research, development, and implementation plans for the Cheshire system as part of an overall digital library framework. Our plans are based on developing a high-level component framework extensible and flexible enough to accommodate radically different architectural models; one which supports large-scale preservation environments characteristic of content management systems such as DSpace and Fedora; and the support for semantic retrieval and natural language processes involving large-scale distributed datasets, served in a highly parallel environment. Rather than prototyping a single solution, we intend to follow a modular approach with a variety of text and data mining, ontological, document rendering, and text retrieval tools which can be used in various combinations according to need. In essence, the Cheshire digital library framework is designed to address the twin aspects of access and preservation in new and sustainable ways.

1. Introduction

The Cheshire digital library framework is one which seeks to integrate a number of recent advances arising from the data grid, digital library, and persistent archive communities in order to support highly scaled digital repositories across domains and data formats.

The primary thrust of our initiative is to take a number of existing data-grid, digital library, and persistent archive management systems, develop them as an integrative framework for grid-based digital repositories, and apply this framework in ways which will support content management systems and workflow environments.

Our work is aimed at providing better integration of publication and discovery mechanisms; preservation and technology evolution management; and interoperability across distributed resources.

Specifically, the framework will seek to integrate the data grid technologies of the Storage Resource broker; the digital library technologies of the Cheshire information retrieval system; a number of text and data mining capabilities; and incorporate the persistent archive technologies of the a document parser ("Multivalent").

Such an integration will result in a component-based framework which will facilitate access to containers of digital content in data grids. In doing so, the result will support a range of content management systems (for example, Dspace and Fedora) which, in future, will use the SRB.

The proposed framework will be defined in terms of object oriented components, all of which can fit internally into a number of workflow environments (for example, Kepler/Chimera).

We intend to deploy the Cheshire digital library framework on a number of SRB-testbeds across a range of domains, data formats, and models, all working in a data grid environment. The result should be of relevance to any SRB administered container of data. Over the summer of 2005, the San Diego Supercomputer Center intends to start benchmarking the results for a range of initiatives.

2. Overview

At Liverpool, we are currently working to integrate the information management technologies which have been developed across the data grid, digital library, and persistent archive communities, and apply these to an environment which will support highly scaled digital repositories.

Our aim is to develop and implement a system, based on existing and well-supported tools from these community, which will fulfill most or all of the objectives required for users of digital repository services.

Strategically, we are seeking to implement the abstract mechanisms needed to manage technology evolution, from characterizations of digital entity structures and semantics, to characterizations of standard operations on storage repositories and standard access mechanisms. The involvement of the data grid community is key, since it has already developed the

[*] This research was sponsored by a funding agency. Views and conclusions contained in this report are the authors' and should not be interpreted as representing the official opinion or policies, either expressed or implied, of the Government, or any person or agency connected with them.

client-server middleware (SRB) to provide a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets.

Our development plans, therefore, centre on integrating digital library and persistent archive systems into an existing framework. In doing so, we hope to leverage the substantial commitment the data grid community has made into providing comprehensive distributed data management solutions which support the management, collaborative sharing, publication, and preservation of distributed data collections.

Appendix I sets out an architectural diagram of the proposed architecture of the framework.

3. Research Agenda.

The research and development agenda may be summarized as follows:

1. The integration of the Chehsire system with a data mining tools, a document parser (Multivalent), storage abstract middleware (SRB); and a content management model (DSpace/Fedora) which will allow the system to be tailored to serve the specific retrieval and processing requirements of the higher education community. This integration will provide Cheshire support for distributed supercomputing; large-scaled preservation environments; advanced document processing and rendering capabilities; access to large distributed datasets; and the ability to query and update data in a distributed environment.
2. The integration of this system with a range of information extraction techniques (including high-level natural language processing, ontology-based searching of databases, semantic clustering, data analysis/data mining tools, and information retrieval procedures). We will draw upon our existing work in logic based formalisms to represent knowledge and map between ontologies, defining a semantic network, and use this to integrate semantic ontologies across different domains.
3. The use of computational linguistic algorithms (or statistical semantic grammars) to label entities and their relationships to texts.
4. The support of users' ability to synthesize queries and results across textual descriptions, databases, entities which may appear in text, and multimedia formats.
5. The development and implementation of user interfaces which support cross-domain resource discovery, data mining, and information synthesis, including the visualization of semantic structures for knowledge discovery. This includes Cheshire's existing support for narrowing searches to types of publications and time-spans, etc., as well as for schema integration, query optimization, and transaction processing.
6. The deployment of this across a distributed corpus using the large scale and production grid infrastructures at the San Diego Supercomputer Center and across JISC services. Collectively, the shared resources will encompass terabytes of data including full-text, multimedia, hypermedia, statistical, and semantic metadata.
7. The extension of the software and techniques across domains, data types (e.g. Geospatial and statistical data), and the model based integration of embedded software.

4. Integrating Data Grid, Persistent Archive, and Digital Library Technologies.

The Cheshire research agenda seeks to integrate the following technologies:

4.1. Data Grid Technologies.

Our framework will be primarily based on the data grid technologies provided by the SDSC Storage Resource Broker (SRB).

This is client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets. In conjunction with the Metadata Catalog (MCAT), it provides a way to access data sets and resources based on their attributes and/or logical names rather than their names or physical locations.

The SRB supplies transparent replication; archiving, caching, synchs, and backups; heterogeneous storage; container and aggregated data movement; bulk data ingestion; third party copy and movement; version control; and partitioned data management.

The SRB is emerging as the *de facto* standard for data-grid applications, and is used for the World University Network; the Biomedical Informations Research Network (BIRN); the UK eScience Centre (CCLRC); the National Partnership for Advanced Computational Infrastructure (NPACI); the BaBar collaboratory; NASA information power grid; among others.

4.2. Content Management Systems.

The digital storage model of the SRB is currently being used to extend the management capabilities of a number of content management models, in particular DSpace and Fedora.

This is intended to address the present limits of DSpace and Fedora, both of which assume local storage, and will ensure their collections can in future be of virtually unlimited size, and be stored, replicated, and accessible via federated grid technologies.

In supporting the SRB, we have designed the Cheshire framework specifically to use digital library technologies to exploit the integration of these systems.

This will facilitate digital content ingestion; search and discovery; content management; dissemination services; and preservation, within a data-grid environment.

4.3. Digital Library Technologies.

We seek to take advantage of the SRB-DSpace/Fedora integration cited above by providing for it a fully integrated document parsing, indexing, storage and information retrieval system ("Cheshire").

Originally created at UC Berkeley, the Cheshire system is now being co-developed at University of Liverpool and has recently gone through a major upgrade into its current stage, Cheshire 3 [1]. Its modular design and optimized coding offers a combination of highly advanced information management tools in a flexible and highly distributable environment capable of integration into any other system that supports Python.

The system itself is widely used in the United States and the United Kingdom for production digital library services including the distributed Archives Hub, the History Data Service, the Information Environment Service Registry, the Resource Discovery Network, the British Library ISTC service, and so forth.

We intend to extend the Cheshire system so as to use the data grid as a storage layer and permit the exchange of documents between the two systems in a federated environment.

This will result in Cheshire support for the full range of information retrieval and text mining capabilities, being of benefit to all users of the SRB.

The aim is to provide integrated support for searching highly scaled data, including life-cycle management capabilities for digital assets within preservation environments. To this end, we are hoping to evaluate the framework for use in the NARA preservation prototype.

4.4. Text and Data Mining Systems.

The Cheshire system is being used in the UK National Text Mining Centre (NaCTeM) as a primary means of integrating information retrieval systems with text mining and data analysis systems. The objective is to provide a platform which may be further developed in order to integrate text mining techniques and methodologies into workflows. This may be done as part of an internal Cheshire 3 workflow; or as external scientific workflow for systems such as Kepler-SRB (see below).

Our framework will seek to integrate the suite of text and data mining tools with the Cheshire environment, and implement these on highly parallel grid infrastructures to support a wide range of distributed digital library services. This will be used, in particular, as the basis for supporting the ontology-based searching of text datasets.

We intend to further supplement these capabilities by incorporating support for "high dimensional data" which is not particularly well handled by the current generation of text mining or machine learning tools. If implemented, such support would extend the capabilities of the Cheshire system to extract and relate semantic information, efficiently and effectively, beyond the current state of the art. This may be used as a means of discovering and filtering information which could be of relevance for further detailed analysis.

This added functionality will offer solutions to a range of problems likely to be of interest to the scientific and research communities. This includes support for the management and integration of data, combining traditional database technologies and knowledge representation techniques, including data modeling, knowledge representation, and query processing for model-based mediation, databases and workflows, and knowledge-based digital libraries and archives.

The outcome will be integrated support for:

1. Text and Data Mining capabilities, including Latent Semantic Analysis, Support Vector Machines, naïve Bayes Networks, Genetic Algorithms, and recursive feature algorithms.
2. Information Extraction capabilities, including text conversion, text zoning, text segmentation, term extraction, ontology lookup, parts of speech tagging, named entity recognition, template extraction (finding properties of named entities), fact extraction (finding relations between entities), temporal information extraction.
3. Information Retrieval capabilities, including logistic regression techniques, Boolean and proximity searching, relevance feedback techniques.

These technologies may be used as a means of *information retrieval* (searching for information already known) as well as for *knowledge discovery* (using data mining methods to discover new knowledge, previously unknown).

The support of these capabilities is intended to satisfy the aim of those working within a generalized SRB-based architecture which can be used by domain experts to characterize their knowledge about a collection.

4.5. Workflow Environments

Workflow environments such as Kepler-SRB are designed to allow researchers to design their own scientific workflows and execute them efficiently using emerging Grid-based approaches to distributed computing. Their objective is provide a software system which will give scientists in a variety of disciplines access to scientific data and a flexible means of executing complex analysis on those data. These will enable users to manage interactions with databases and the processing of query results; manage execution of applications within a computational grid; and describe and execute workflow templates.

The proposed digital library framework is designed to enable a comprehensive data environment for users of these tools within federated digital repositories.

In particular we intend to provide a platform which may integrate text mining techniques and methodologies into workflows. This may be done as part of an internal Cheshire workflow; or as external scientific workflow for systems such as Kepler-SRB.

4.6. Digital preservation technologies.

Another strategic emphasis will be to incorporate the multivalent document model and parser into the Cheshire system, and use this as the means of implementing a long-term preservation environment which does not rely on emulators or converters to retain the content and format of legacy documents.

Currently, many digital preservation systems rely on a form of emulation which consists of migrating the original application forward onto new architectures; this requires wrapping the application so that it can issue modern operating system calls.

The problem with this approach is that there is no guarantee that operating calls will be available on modern systems that correspond to the original operating system. Thus, the wrapping technology must be constantly migrated to new operating systems. Over a long period such techniques may progressively degrade data.

We are instead proposing to implement a more abstract model, based on the integration of a document parser ("Multivalent") [2], born at UC Berkeley but now being developed at Liverpool, which would allow us to parse original documents in modern languages most likely to be supported by operating systems over time.

The support for multivalent architecture would allow the system to keep documents alive, e.g. directly viewable in their original state, through the development of what are known as "media adapters" implemented as part of the extensible multivalent document infrastructure.

More generally a Cheshire interface to the Multivalent model will facilitate a more sophisticated, document interaction for users of eprints and digital library services: these may extend to Cheshire support for different media types and formats (including images and video) which may be annotated or searched with text based techniques; geographic information systems (GIS) visualizations that compose several types of data from multiple datasets; distributed user annotations that augment and may transform the content of the conceptual document; and support for true UNICODE rendering for non-western texts (e.g. Arabic, Cambodian) and pre-reformed versions of languages (e.g. Greek, Chinese), all of which are not well served by the current generation of digital library systems.

The outcome will be Cheshire support for:

1. Long-term and sustainable preservation of digital entities for SRB and Dspace/Fedora;
2. A platform to migrate collections using an abstract document model, ensuring authenticity of archived data;
3. The managed development of "media adapters" enabling documents to be directly viewable in their original state;
4. Sophisticated document interaction for eprints and digital library services, extending to different media types and formats.

5. Conclusion.

The Cheshire digital library framework extends the functionality of the current generation of digital library systems to form a comprehensive end-to-end knowledge management system, fulfilling a variety of functions,

The integration of Cheshire digital library services with SRB will provide users with a platform to service large scale archival repositories and content management models, such as DSpace and Fedora.

The framework is design to ensure the accessing of content within these repositories (the "containers" of the SRB).

The integration with Multivalent document model will provide Cheshire with an abstracting mechanism which will serve to preserve and render collections of digital documents in ways which are not vendor-specific, and therefore ensure the access and authenticity of archived data across software and hardware over time.

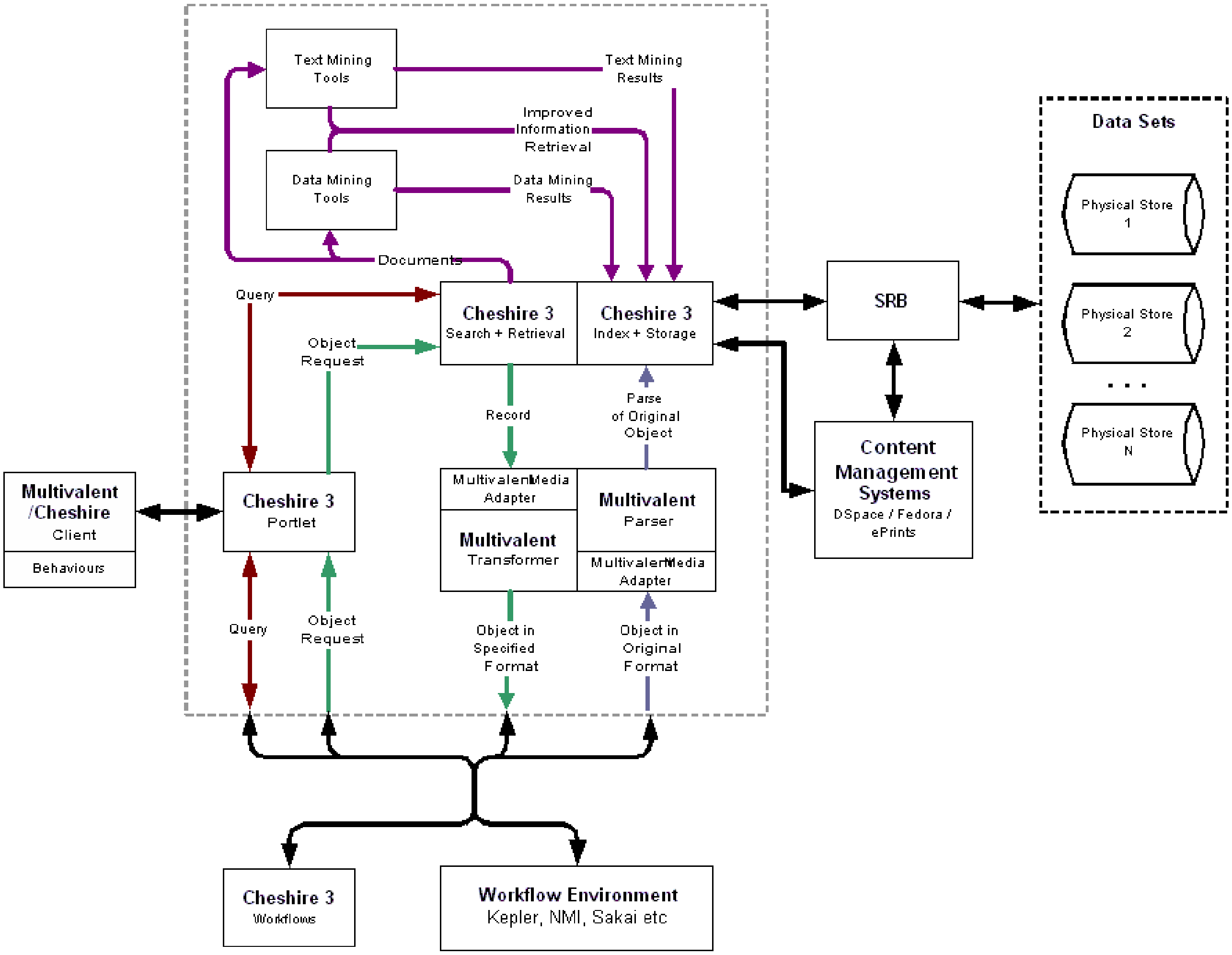
The integration with text and data mining tools will provide users with new data mining capabilities, based on dimensionality reduction techniques, which may be used to construct semantic and structured ontologies based on latent semantic analysis. Although such techniques have long been recognized as a powerful way to mine text, to date they have never been implemented as part of a scaled production system.

A primary outcome of the Cheshire digital library framework will be the application of these advanced text mining and rendering capabilities within large-scale preservation environments and current digital library services.

6. Acknowledgements.

Development of the Cheshire system was supported in part by the National Science Foundation and the Joint Information Systems Committee (U.K.), under the International Digital Libraries Program award IIS-9975164.

Appendix 1 (Architectural diagram).



References

[1] R. R. Larson and R. Sanderson, "Grid-Based Digital Libraries: Cheshire3 and Distributed Retrieval", *JCDL* 05, (June 7-11 2005.).

[2] T. Phelps, "Multivalent Documents: Anytime, Anywhere, Any Type, Every Way User-Improvable Digital Documents and Systems", Ph.D. Dissertation: University of California, Berkeley (1998).