Semantic Augmentation of the Storage Resource Broker

Stephen J. Jeffrey and Jane Hunter

Abstract— **Information discovery is looming as a major challenge with the growth of tera-byte size datagrids. In order to manage their distributed data collections, many scientific organizations are adopting the San Diego SuperComputer's Storage Resource Broker (SRB). The SRB's Metadata Catalogue (MCAT) focuses primarily on system or administrative metadata and supports domain-specific metadata through user-defined extensions. Although providing maximum flexibility, this approach will lead to interoperability problems when searching across distributed collections described using a variety of different user-defined metadata schemes. The aim of the work described in this poster is to semantically augment the SRB through an ontology and Resource Description Framework (RDF) descriptions in order to support arbitrary metadata schemata and to enhance the system's search capabilities. In particular we describe a semantic search engine and interface built on top of an OWL ontology, RDF instance data and a Jena reasoning engine that enables easier, more sophisticated searching, browsing, inferencing and retrieval of heterogeneous data stored using the SRB.**

***Index Terms*— Storage Resource Broker, Datagrids, Ontologies, OWL, Semantic searches**

I. Introduction

COMPUTATIONAL grids and networked instruments have lead to an explosive growth in our ability to generate and collect huge amounts of data. To support the increasing demand for shared access to large repositories of data, many projects have focussed on developing enabling technologies for constructing distributed datasets. The motivation for developing distributed archives is two-fold: redundancy and efficiency. Valuable datasets, particularly those containing irreplaceable historical data, need to be protected against equipment failure and both accidental and malicious damage by users and administrators. It is therefore critical that multiple copies of datasets are maintained at geographically separate locations. System redundancy must be designed so that the entire archive can be restored even when there is a complete loss or failure at the primary site. The other main issue driving the development of distributed datasets is efficiency.

Manuscript received August 15, 2005. This work was a component of the Australian GlobalGrid Project - Integrating Australia into Global e-Science, supported by DEST grant CG050091 of the *International Science Linkages* program.

S. J. Jeffrey is with the Advanced Computational Modelling Centre, University of Queensland, St. Lucia 4072 (e-mail: sjj@maths.uq.edu.au).

J. Hunter is with DSTC Pty Ltd, University of Queensland, St. Lucia (phone: +61 7 3365 4310; fax: +61 7 3365 4311; e-mail: jane@dstc.edu.au).

Computational grids and tele-instrumentation are forcing us to rethink the traditional paradigm of local analysis (computing and/or measurement) of local datasets. If the task necessarily involves access to large datasets or extensive I/O, it may not be feasible to copy the dataset to the user's local machine. In such situations the user ought to be guided by network efficiency considerations and also the available instrumentation or computational resources. It may be more appropriate to carry out the task (or parts thereof) using remote resources that can more effectively access the desired dataset. Scheduling of remote tasks is now becoming a simple and transparent operation as Grid technologies (such as the Globus toolkit [1,2]) mature.

As data grid technologies stabilize, we are faced with another issue: how to catalogue data so that it can be effectively searched and retrieved. Users have traditionally relied upon some form of descriptive metadata to determine what data is available and how it is organized. The metadata may be simple free-form annotation, it may be organized in fixed fields eg., Dublin-Core, or it may take a more formal domain-specific hierarchical classification format. Irrespective of how the metadata is organized, most of the current search methods involve key-word searching of the available metadata. Simple key-word searching has three serious limitations. Firstly, searching becomes very inefficient as the size of the dataset increases and becomes completely intractable on terabyte-size archives. Secondly, and more importantly, key-word searching requires both the user and archive creator to use the same term(s) when describing the data. This requires the user to have a precise knowledge of the vocabularies and classification schemes used within a particular discipline or scientific community. This problem is exacerbated as technical languages evolve and annotations by necessity become increasingly specific as datasets become more comprehensive. Thirdly, simple searching does not allow the user to retrieve material that may be logically related to material captured in the original search. Consequently, key-word searching is a serious barrier to information discovery.

Search methods can be improved by taking advantage of the relationships which are known to exist within the data and between different but related metadata terms and values. For example, the *parent:child* relationship could be used to search for all gene data associated with the offspring of a given parent. The search domain would be immediately reduced to the subset of children with the given parent. In contrast, a key-word search would need to search the entire domain for data items having both parent and child key-words and the user would then need to manually determine which items were indeed children of the given parent. Furthermore, not all children may store parent information, so the key-word search

for items containing both parent and child will miss those items. By using application-specific semantics, in this case the relationship between parent and child, the search can be restricted without the risk of excluding valid items from the search domain. The term *semantic search* is used to refer to search methods that take advantage of relationship information, that has been formally encoded within *ontologies*.

Semantic search techniques have been developing for about a decade. The Resource Description Framework (RDF) [3] was an early attempt at constructing a formal language for describing data in a more semantically meaningful way. RDF Schema (RDFS) [4] was subsequently developed so that property information could be represented using RDF syntax. Ontologies were adopted as a "formal specification of a domain conceptualization" that can improve communication between humans and/or computers. Ontologies represented in RDFS were not sufficiently expressive, so the DAML and OIL ontology languages [5] were developed. The World Wide Web Consortium subsequently developed the Web Ontology Language (OWL) [6] as the semantic search standard, basing it upon RDFS and DAML+OIL. A number of groups have developed tools for creating, editing and processing OWL ontologies, most notably, FaCT [7], Pellet [8], Racer [9] and Jena [10]. Such tools are remarkable because they allow the user to perform semantic searches based on reasoning. Relationship information is stored in the ontology and the object or instance data is represented as RDF descriptions stored in a datafile. Reasoners, such as Fact, Pellet etc., infer information about data objects using the relationship information in the ontology. For example, consider an ontology that describes *cousin* as the relationship between the children of two siblings. If the datafile contains siblings `Person_A` and `Person_B`, and `Child_of_A` is the child of `Person_A` and `Child_of_B` is the child of `Person_B`, then the reasoner will deduce that `Child_of_A` and `Child_of_B` are cousins. The ability to dynamically infer information about the relationship between data objects can thus be used to create very powerful search tools.

Semantic search techniques will undoubtedly prove very useful for improved information discovery across distributed datagrids. As previously described, keyword searching has a number of limitations, particularly when applied to large heterogenous datasets that have been assembled, described and maintained by many different curators. The ability to reason using relationship information stored in ontologies enables semantic search engines to overcome many of the problems associated with existing search methods. Ontology-based searching uses the intuitive relationships between concepts to provide intelligent access to information.

In addition, as organizations adopt alternative grid data management approaches to SRB, such as the Grid Datafarm [11] or OPeNDAP [12], or integrate research output stored in digital repositories such as DSpace [13] or Fedora [14] with SRB data files, then the semantic interoperability layer we have built becomes even more important. It enables the relationships between disparate metadata schemes (such as the MCAT scheme, Dublin Core and METS [15]) to be formally represented within the ontology. The ontology can then be used as the mediator that facilitates federated searches across heterogeneous distributed multidisciplinary data stores and repositories.

In addition to improving data discovery and integration, the semantically-rich data descriptions that we are proposing will also be essential to the dynamic composition, orchestration and matching of optimum combinations of grid services or workflows to scientific data. This is a fundamental aim of the envisaged semantic web services architecture (WSRF) [16] of Grids of the future.

Hence in this poster we demonstrate how the datagrid middleware application Storage Resource Broker (SRB) [17, 18] can be semantically augmented by:

- Storing administrative and system metadata in MCAT [19] but storing the descriptive metadata associated with data files in a separate RDF data store;
- Defining the descriptive metadata terms and relationships between terms both within and across metadata schemas within an OWL ontology;
- Building a semantic search engine which uses the Jena reasoning engine to enable on-the-fly inferencing of implicit metadata and data;
- Providing an ontology-based search and browse interface that uses semantic descriptions and inferencing to identify the relevant files but then retrieves them using SRB and the core MCAT metadata.

Using simple example datasets, comprising experiments, miccroarray datasets and publications, we demonstrate how this approach enables easier, more intuitive and sophisticated searching, browsing, retrieval and integration of heterogeneous data stored using SRB.

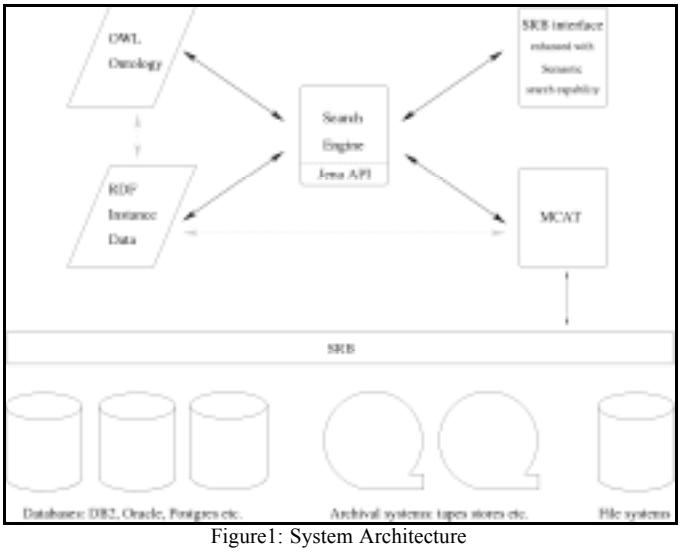

Figure1: System Architecture

REFERENCES

[1] Globus Toolkit, see http://www.globus.org/.
[2] I. Foster and C. Kesselman, "Globus: A Metacomputing Infrastructure Toolkit", Intl J. Supercomputer Applications, vol 11(2), pp. 115-128, 1997.
[3] RDF/XML Syntax Specification (Revised), W3C Recommendation, February 2004 http://www.w3.org/TR/rdf-syntax-grammar/
[4] RDF Vocabulary Description Language 1.0: RDF Schema, W3C Recommendation February 2004, http://www.w3.org/TR/rdf-schema/

[5] DAML+OIL, March 2001 http://www.daml.org/2001/03/daml+oil-index.html
[6] M.K. Smith, C. Welty and D.L. McGuinness, "OWL Web Ontology Language Reference", W3C Recommendation 10 Feb 2004 http://www.w3.org/2004/OWL/.
[7] FaCT: Fast Classification of Terminologies Description Logic classifier, see http://www.cs.man.ac.uk/~horrocks/FaCT/
[8] Pellet OWL Reasoner, see http://www.mindswap.org/2003/pellet/index.shtml.
[9] RacerPro Reasoner, see http://www.racer-systems.com.
[10] Jena – A Semantic Web Framework for Java, see http://www.hpl.hp.com/semweb/jena2.htm.
[11] Naotaka Yamamoto, Osamu Tatebe, Satoshi Sekiguchi, "Parallel and Distributed Astronomical Data Analysis on Grid Datafarm", Proceedings of 5th IEEE/ACM International Workshop on Grid Computing (Grid 2004), pp.461-466, 2004
[12] OPeNDAP: Open-source Project for a Network Data Access Protocol http://opendap.org/
[13] M.Smith, "Eternal Bits: How can we preserve digital files and save our collective memory?" , IEEE Spectrum, August 2005
[14] C. Lagoze, S. Payette, E. Shin, C. Wilper, "Fedora: An Architecture for Complex Objects and their Relationships," forthcoming in Journal of Digital Libraries, Special Issue on Complex Objects, Springer 2005. http://www.arxiv.org/abs/cs.DL/0501012
[15] METS (Metadata Encoding and Transmission Standard) : An Overview and Tutorial http://www.loc.gov/standards/mets/METSOverview.v2.html
[16] The WS-Resource Framework http://www.globus.org/wsrf/
[17] R. Moore, A. Rajasekar and M. Wan, "Data Grids, Digital Libraries and Persistent Archives: An Integrated Approach to Publishing, Sharing and Archiving Data", Proceedings of the IEEE, vol 93, pp. 578-588, March, 2005.
[18] San Diego Supercomputer Center's Storage Resource Broker, see http://www.sdsc.edu/srb/.
[19] MCAT – A Meta Information Catalog (Version 1.1) http://www.npaci.edu/DICE/SRB/mcat.html
[20] Semantic Grid Community Portal http://www.semanticgrid.org/GGF/
[21] Towards a Semantic Data Grid for Systems Science http://www.emsl.pnl.gov/sdg/index.htm
[22] A. Woolf, R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, B. Lawrence, R. Lowry, K. O'Neill, "Semantic Integration of File-based Data for Grid Services", Workshop on "Semantic Infrastructure for Grid Computing Applications", CCGrid 2005, Cardiff (May 2005). http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS//Vol-138/paper5.pdf
[23] Z. Xu, M. Karlsson, C. Tang and C. Karamanolis. *Towards a Semantic-Aware File Store.* In the Proceedings of HotOS-IX, May 2003, Kuai, HI. http://www.hpl.hp.com/personal/Magnus_Karlsson/papers/hotos.pdf
[24] V. Christophides, C. Houstis, S. Lalis, and H. Tsalapata, "Ontology-driven Integration of Scientific Repositories", In Proc of the Fourth Workshop on Next Generation Information Technologies and Systems (NGITS'99), Zikhron-Yaakov, Israel, July 1999
[25] K. Houstis, S. Lalis, V. Christophides, D. Plexousakis, E. Vavalis, M. Pitikakis, K. Kritikos, A. Smardas, "A Data, Computation and Knowledge Grid: the case of the ARION system", In Proc. of the 5th International Conference on Enterprise Information Systems (ICEIS2003), pp.359-365, Angers, France, April 23-26 , 2003
[26] K. Taylor, D. De Roure, J. W Essex, J. G Frey, R. Gledhill, S. W Harris A Semantic Datagrid for Combinatorial Chemistry, Grid 2005 - 6th IEEE/ACM International Workshop on Grid Computing, Seattle, Washington, Nov 2005
[27] SIMILE project – Semantic Interoperability of Metadata and Information in Unlike Environments http://simile.mit.edu/
[28] Current Projects Using SRB http://www.sdsc.edu/srb/Projects/main.html
[29] A. Rajasekar, M.Wan and R. Moore, "MySRB & SRB - Components of a Data Grid", The 11th International Symposium on High Performance Distributed Computing (HPDC-11) Edinburgh, Scotland, July, 2002
[30] Wine Ontology, see http://www.w3.org/TR/owl-guide/wine.rdf