A Semantic Search Engine for 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 San Diego SuperComputer's Storage Resource Broker (SRB). Indexing and retrieval of data stored in SRB is via SRB's Metadata Catalogue (MCAT). MCAT focuses primarily on system or administrative metadata but supports domain-specific metadata through user-defined extensions. Although this approach provides maximum flexibility, it will lead to interoperability problems when searching across distributed collections described using different user-defined metadata schemas. The aim of the work described in this paper is to semantically augment 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 and more sophisticated searching of heterogeneous data stored using SRB.**

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

I. Introduction

The Storage Resource Broker (SRB) is a datagrid application developed by San Diego Supercomputer Centre. It is middleware aimed at federating collections of distributed data and presenting them to the user as a coherent collection. A key component of SRB is the metadata catalogue (MCAT). MCAT is used to provide an abstraction mechanism so that users can access data via attributes or logical name, without the need to reference the physical name or location of a data object. MCAT stores both system and user-defined metadata. System metadata is used for SRB's internal accounting, whereas user-defined metadata specifies optional attributes that describe data objects using domain-specific terms. The user-defined metadata can be inserted in a number of different formats. Free-form text and attribute-operator-value triplets can be stored in a number of string and integer fields; fixed format metadata such as Dublin-Core is available; and additional tables can be added to MCAT.

SRB is being used by a broad array of scientific organizations [28] (including the International Virtual Observatory Alliance, BIRN, NASA, the Grid Physics Network and the US National Science Digital Library). Each of these organizations have implemented their own unique domain-specific user-defined metadata schemas that are extensions to the core MCAT metadata scheme. Current SRB search methods involve keyword searching of the available metadata. But this approach will have serious limitations as organizations want to share or integrate data sets. Firstly and most importantly, both the searcher and archive creator need to use the same metadata term and value when describing the data. Consequently the searcher is required to have a precise knowledge of the metadata schema and vocabularies used by a particular scientific community or organizaton. The problem is exacerbated as technical languages evolve and annotations become increasingly discipline-specific as datasets and protocols become more complex. Secondly, simple searching does not allow the user to retrieve material that may be logically related to material explicitly described by the search parameters. Consequently, keyword searching is a serious barrier to information discovery and integration.

Semantic search techniques have been developing for about a decade. The Resource Description Framework (RDF) [3] provides a formal language for describing data in a semantically meaningful way. RDF Schema (RDFS) [4] was subsequently developed so that structured information could be represented using RDF syntax. 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], based 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 OWL/RDF descriptions stored in a datafile. Reasoners, such as Fact, Pellet etc., infer information about data objects using the relationship information in the ontology. The ability to dynamically infer information about relationships between data objects can be used to create very powerful and sophisticated search tools.

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).

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.

In addition to providing intelligent access to information, ontologies can also be used for data integration. As organisations adopt a variety of grid data management applications, such as SRB, Grid Datafarm [11], OPeNDAP [12], DSpace [13] and Fedora [14], an interoperability layer will become essential. The individual applications will typically have disparate metadata schemes (such as MCAT, Dublin Core and METS [15]), which will severely limit one's ability to search and retrieve objects stored in a federated collection of data management applications. An ontology-based semantic interoperability layer could overcome this problem, as it would enable the relationships between the various metadata schemes to be formally represented within an ontology. The ontology could be used as a mediator that facilitates federated searches across heterogeneous and multidisciplinary data 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 future Grids.

In this paper we outline an architecture for semantically augmenting SRB. Using an example dataset and ontology, we demonstrate how this approach enables easier, more intuitive and sophisticated searching, browsing, inferencing and retrieval of heterogeneous data .

The paper is structured as follows. In Section II we outline previous related work. In Section III we describe the system implementation and architecture. In Section IV we describe the richer kinds of searches that our system enables and demonstrate its use. In Section V we discuss various issues relating to performance and implementation details. Concluding remarks follow in Section VI.

II. Previous related research

As far as we are aware, no implementations have been developed that extend the SRB/MCAT search interface to enable more intelligent semantic searching through the use of an ontology(ies), a reasoning engine and RDF descriptions. Such an approach would facilitate semantically rich data descriptions and enhanced semantic interoperability that is complementary to the data virtualization services provided by SRB.

Related to our objectives is the DSpace/SRB Integration project funded by NARA. It is a collaborative project between MIT, SDSC and UCSD that aims to facilitate distributed data management by replacing DSpace's persistent datastore with SRB. It overcomes semantic interoperability problems by using METS as the common metadata schemas to enable authenticated exchange of data objects between systems. Cornell University's Fedora project is also investigating replacing its datastore with SRB. Fedora provides a simple search interface using Dublin Core metadata as well as a RDF-based Resource Index for querying digital object relationships. Neither of these projects provides an ontology-based search interface to digital objects in SRB that mediates across heterogeneous metadata schemas.

A growing number of initiatives, projects and workshop reports are elucidating the necessity for semantic data grids and the more general semantic grid concept [20-24]. These activities specifically acknowledge the importance of ontologies and RDF in enabling scientists to search, discover, manage and integrate data from large distributed scientific data archives. Semantic data grids have been developed and implemented for a number of specific disciplines e.g., earth sciences [25] and combinatorial chemistry [26]. These implementations have been built on community-specific databases. They don't provide a generic, discipline-independent solution through semantic augmentation of SRB.

III. Implementation and architecture

Currently SRB has a command line interface (known as S-commands) and a number of graphical interfaces. Users can search and retrieve objects via object name or via metadata content. In the current context we are mainly interested in identifying objects via their user-defined metadata fields. In this respect, the search facilities offered by SRB (and most other similar applications) are limited to direct keyword matching, keyword matching using wildcards ('*') and logical operations on attribute values.

The limitations of the existing search interfaces can be conveniently demonstrated using the command line interface. Consider for example, the S-command:

```
SgetD -A "DCOMMENTS like '*value*' "    'name*'
```

which will locate all objects with name matching the wildcard expression '*name** ' and the comment field containing the string '*value*'. The limitations of keyword matching were outlined earlier in Section I. Search mechanisms based upon attribute-operator-value triplets enable the user to focus the search on a restricted domain, but still require (i) the given attribute to be present in the metadata; and (ii) the user to have knowledge of the metadata format. Both requirements are a barrier to effective information discovery. The limitations of the existing search mechanism in SRB motivated us to develop a semantic search engine for the application.

The semantic search interface that we have developed is an extension of the MySRB web interface developed by SDSC [29], shown in Figure 1. Our extension has two components: (i) the graphical MySRB web interface, augmented with semantic search functionality; and (ii) an independent search

engine. A high-level view of the overall system architecture and its components is illustrated in Figure 2.

Figure 1. Existing MySRB search interface

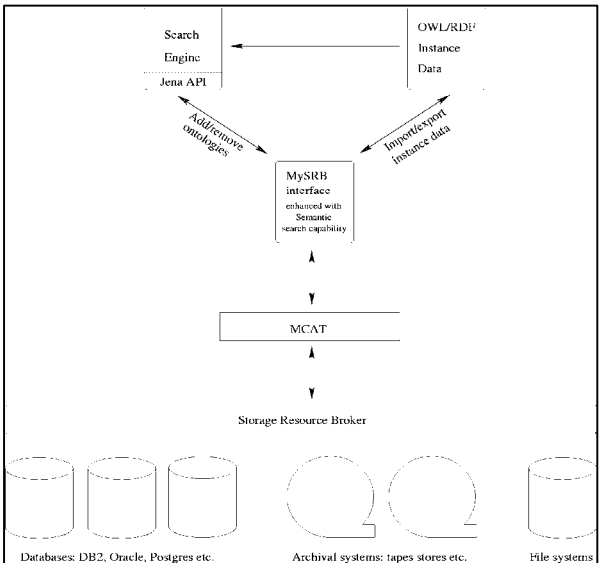

Figure 2. System architecture

Our search engine is implemented using a client-server architecture and is written in Java for portability. The server runs as a persistent thread on the machine that hosts the SRB metadata catalogue. Jena is used for all semantic operations. When the server is started Jena is called to load ontology(ies) and instance data. An inferencing model is created and various class properties are stored. These properties are subsequently used for configuring the search interface. Both the ontology and instance data are stored in XML format using OWL/RDF labels.

The search client is a Java application that connects to the server, runs the search operation and returns the results to the user interface. A new client is started for each search operation and terminates once the results have been returned. It is a short lifetime process and like the server, it (usually) runs on the MCAT machine.

Figure 3. The semantically augmented MySRB interface

The MySRB web interface has been modified to incorporate additional functionality for communicating with the search engine and allowing the user to construct a semantic search request. The modified MySRB interface is shown in Figure 3. The user interface functions as follows:

1. The user loads the MySRB web interface in their browser and connects to the inferencing server. The interface is identical to the original MySRB interface with the exception of two text boxes in the lower left frame. The user must enter the domain name of the machine hosting the semantic search engine and the port number to establish the connection.
2. The user adds or removes ontologies. Ontologies that are currently loaded in the search engine are displayed in a selection box, and the user may elect to remove one or more. The user may also add new ontologies. By adding a series of ontologies the user may construct a hierarchical description of the desired domain to an arbitrary level of complexity.
3. The user builds instance data. SRB administers objects via system and user defined metadata stored in SRB's native metadata catalogue. Sufficient information must be exported to the semantic search engine so that it can build instance data and perform inferencing. User defined metadata can be added, modified and deleted via the original MySRB interface. Users can optionally flag metadata as being semantically relevant, in which case it will be exported to the search engine when the user requests instance data be built. It should be noted that while instance data will be generated for all objects in the specified SRB collection, only the semantically relevant subset of metadata is exported.
4. The user may optionally save the instance data that is currently loaded in the search engine. Building instance data from SRB metadata (Step 3) is a relatively slow operation, so it is advisable to save the instance data when possible. The data is saved in OWL/RDF format and can be loaded as an ontology (Step 2) when next required. The

user may then commence using the semantic search engine once the desired ontologies and instance data have been loaded. The user may specify any or all of the following: SRB datatype (if an appropriate ontology has been loaded), object class, properties of the selected object class and free-format comments/annotations (see Figure 4). Search parameters are passed to the inferencing client which connects to the inferencing server. The server performs the search and returns the results to the client, which passes the instance data back to the MySRB interface. Objects satisfying the search parameters are then displayed in the MySRB user interface, whereby users may browse the results or continue searching.

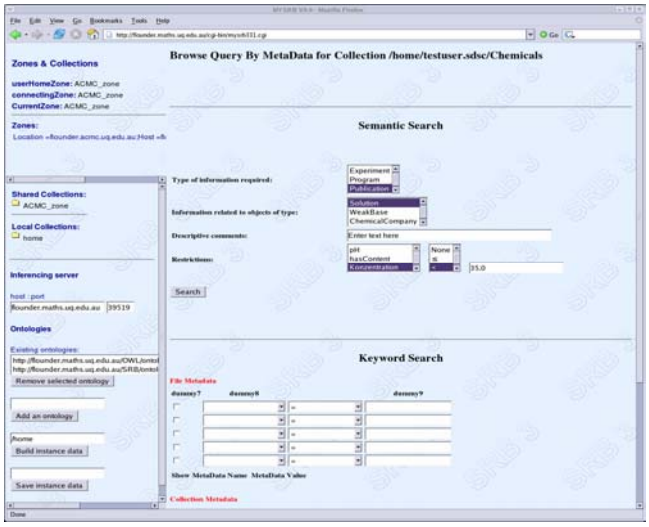

Figure 4. The Semantic Search User Interface

We have designed the semantic extension to be independent of the core SRB application, so existing installations do not require rebuilding. Relatively minor changes were made to the existing MySRB interface, and a new search engine was developed. The search engine is a stand alone application and does not interact directly with either SRB or MCAT. All interaction with SRB and MCAT is via the user interface. The modifications to MySRB were necessary to support the added semantic functionality.

IV. Demonstration of Enhanced Search Capabilities

The semantic augmentation of SRB is demonstrated using a simple ontology that is a subset of the Gene Ontology (GO) based around the term "apoptosis" together with data that we have extracted from the ArrayExpress database. Figure 5 below illustrates the ontology we have used.

We have extracted a series of experiments, array data, protocols and publications related to *apoptosis* from the ArrayExpress database and stored them in SRB. In addition we have extracted, modified and manually generated user-defined metadata describing the objects – this is stored as instance data accessible by the semantic search engine.

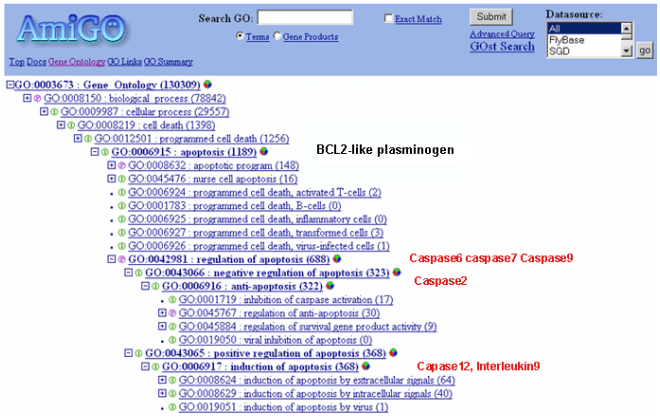

Figure 5. A Sample ontology – a subset of the Gene Ontology

An example of the enhanced capabilities of the semantic search is the ability to locate objects of any type (publications, experiments, arrays or protocols) related to a particular term or related terms. For example, the semantic search interface allows you to locate all digital objects (or specific type of data) related to "apoptosis". The user interface and the list of instances satisfying the search criteria are shown in Figure 6.
The search not only returns those objects that explicitly contain the value "apoptosis" in the subject field of their metadata. The semantic reasoning engine uses the subclass, equivalent class and equivalent property definitions in the ontology to retrieve additional objects about the topic – even when their metadata does not explicitly contain this value.

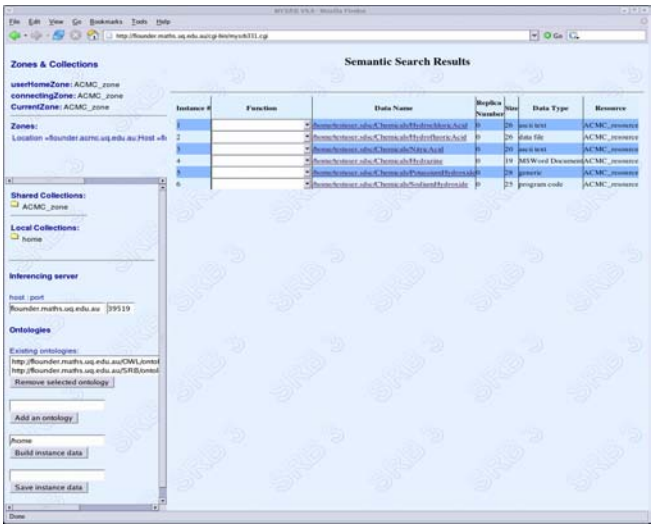

Figure 6. Browsing the results returned by the search engine

For example, with reference to the abridged Gene ontology in Figure 5, "apoptosis" is defined as a type or sub-class of "programmed cell death". "Programmed cell death of activated T-cells" is a subclass of apoptosis. "Inhibition of caspase activation" is also a sub-sub-class of apoptosis. The latter are specialisations of the parent superclass. The inferencing agent (Jena) is able to correctly identify and retrieve both those publications, arrays, experiments, protocols explicitly assigned to "apoptosis" as well as objects about "inhibition of caspase activation" for example. The ability to infer information about objects is extremely

powerful as it allows the search engine to locate objects which match the search criteria, even if the object's associated metadata does not explicitly match the search parameters.

The preceding example is a simple demonstration of the power of semantic searches. The full advantage of semantic searching will become more apparent as domain specific ontologies are developed and used in conjunction with ontologies from other disciplines. Relating ontologies will increase the reasoner's capacity to infer relationships between multidisciplinary datasets, enabling users to discover and integrate related information that would be very difficult to find using simple keyword searches. In addition, semantic descriptions of data will enable the matching, choreography and scheduling of the optimum data processing methods and/or analytical services on the Grid.

V. Discussion

In the previous section we described and demonstrated our implementation of a semantic search engine for information discovery and integration in SRB. The example illustrated two key concepts (*ie.* terminology mapping and reasoning) which demonstrate the power and utility of augmenting heterogeneous data repositories with semantic search capabilities. The added functionality does however impose a performance penalty. Inferencing is a relatively slow operation and in our current implementation we have not attempted to optimise inferencing speed. Rather, the system architecture was designed to maintain a logical separation between the inferencing system, SRB and any existing user defined metadata. While system performance could be improved (fast RDF stores are currently under development [30]), users will not incur significant delays when searching for objects. Most delays occur during the initialisation phase when ontologies are being loaded. Adding and removing ontologies is a slow operation as the inferencing model must be built/rebuilt, and cached information such as classes, properties *etc*. must be updated.

Our current implementation requires us to duplicate the semantically-relevant components of the user-defined metadata by exporting it to an OWL/RDF datafile that can be loaded by the inferencing agent. Duplication of metadata is inefficient and requires ongoing maintenance - whenever the user adds or removes an object, the exported metadata must be updated. Currently this is a manual operation, but it could easily be automated as an internal SRB process.

The inferencing system is loosely coupled to the core SRB application and MCAT. The system architecture was designed in this manner so existing installations could be semantically augmented with minimal disruption. However the cost of this convenience is the need to export (and therefore duplicate) the semantically relevant components of the user defined metadata. Data duplication could be overcome by tightly coupling the inferencing system with SRB and MCAT. The semantic search facility could be built into SRB by storing the ontology(ies), instance data and user defined metadata directly in MCAT, in a format that could be efficiently accessed and processed by the inferencing agent. Explicitly coupling SRB with the search system would require substantial modification of both SRB and the Jena reasoning code. While this would involve significant effort, the benefits would include improved search performance, elimination of duplicated metadata, and reduced administrative overhead.

VI. Conclusions

While the major strength of SDSC's SRB is its seamless virtualization of data file location, we believe that the potential heterogeneity of its user-defined metadata is a weakness that will adversely affect the discovery and interoperability of data stored in SRB. To overcome this limitation we have developed a semantic search engine for SRB. The search engine uses the Jena reasoning agent to perform inferencing on an OWL ontology and associated instance data. The search system has been constructed as an extension to the MySRB interface, without requiring modification to either SRB or its metadata catalogue. The merits and limitations of this approach have been discussed.

Our implementation of a semantic search system for SRB has shown that it is possible, and indeed viable, to extend the existing SRB application with a layer of semantic web technologies to enable ontology-based semantic searches. Semantic searches will become an invaluable tool for information discovery as multidisciplinary datagrids continue to expand. While our current implementation has a number of limitations, it serves to demonstrate how and why ontologies and semantic inferencing will become the tools of choice for discovering new and complex relationships in datagrids that span geographic and disciplinary boundaries.

Acknowledgements

The work reported in this paper has been funded in part by the Co-operative Research Centre for Enterprise Distributed Systems Technology (DSTC) through the Australian Federal Government's CRC Programme (Department of Education, Science, and Training).

References

[1] Globus Toolkit, see http://www.globus.org/.
[2] I. Foster and C. Kesselman, "Globus: A Metacomputing Infrastructure Toolkit", Intl J. Supercomputer Applications, **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] N. Yamamoto, O. Tatebe and S. 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 and 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, **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 and 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 and 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] Kowari Metastore, see http://www.kowari.org.