# YourSRB: A cross platform interface for SRB and Digital Libraries

**Mathew J. Wyatt, Nigel G.D. Sim, Dianna L. Hardy and Ian M. Atkinson***

VeRG Lab, School of Mathematics, Physics and Information Technology, James Cook University, Townsville, Queensland, 4811, Australia.

Mathew.Wyatt@jcu.edu.au, Nigel.Sim@jcu.edu.au, Dianna.Hardy@jcu.edu.au, Ian.Atkinson@jcu.edu.au

## Abstract

*Many large scale research programs with massive data requirements (e.g. the Particle Physics Data Grid, the Biomedical Informatics Research Network) use SRB to manage research data collections. Despite this broad adoption, the potential of SRB and Data Grids is not fully utilized by the general research community. The reason for this under utilisation comes in part from the complexity of the Data Grid model and the lack of quality interfaces to SRB, making it difficult to use and configure.*

*YourSRB, is a system developed in the course of our research which provides an intuitive interface to SRB and aims to enable a broader community to benefit from SRB. The system also simplifies the use of SRB's awkward-to-initiate federation schemes, making SRB more suitable for use as the core of an organisational level research repository. In addition, our approach to federation makes the YourSRB system suitable for field workers who can off-line curate and annotate data, and subsequently federate their data with a primary store when network connectivity is available. YourSRB also implements data description and retrieval capabilities, enabling the effective management of massive data sets over multiple storage resources.*

## 1 Introduction

The development of new generation research facilities that will produce large volumes of data (e.g. the Australian Synchrotron Source) and the adoption of e-research methodologies by many research communities are anticipated to produce data management challenges. Hey and Trefethen (Hey and Trefethen, 2003) predict that the data generated by computer simulations, large instruments, sensors and satellites are likely to soon dwarf scientific data accumulated throughout the history of scientific exploration (Hey and Trefethen, 2003). Across the globe there are projects involving Astronomy, Bioinformatics, Environmental Science, Particle Physics, Medicine and the Social Sciences which are experiencing what has been described as the 'Data Deluge'. Examples of such projects include the LHC project at CERN (CERN, 2006) and the Global Ocean Observing System (GOOS) (GOOS, 2006).

This flood of data is not just limited to a few high profile, large-scale projects. The growth of video and other multimedia in disciplines ranging from the Arts to Education as well as in the Sciences has resulted in an approximate doubling of the world's stored data every 9 months (Kargupta et al., 2003). With this imminent flood of data apparent, there becomes a critical need to store, describe, maintain and understand the data collected. Regular storage repositories such as databases are not an adequate solution to the problem, as scientific data requires a high level of description, representation and collaboration for which databases are not ideally suited without considerable extension (Hey and Trefethen, 2003). In addition, conventional databases simply are not able to handle the Petabyte scales of data that are anticipated (Hey and Trefethen, 2003, Chervenak et al., October 2003) – no commercial database is yet of the scale of LHC or GOOS projects (Gray and Hey, 2001). In recent times Data Grids have been used to house Digital Libraries for e-Research information, and this model may be a possible solution to the management of the 'Data Deluge'.

The Storage Resource Broker (SRB) (SDSC, 2006f) is the archetypal Data Grid middleware and is commonly used to house Digital Libraries for e-Research. SRB is developed by the San Diego Supercomputer Centre (SDSC, 2006f). SRB provides a logical view of information stored across heterogeneous storage devices combined with a powerful metadata description and retrieval framework. SRB works in conjunction with the MCAT (Metadata Catalog), which is the core of the SRB system. As the name suggests, MCAT is a metadata catalog which sits on top of an SQL database and stores 4 types of metadata (Resource, Method, Data Object, User & Group). An MCAT server contains (or is contained within) a Zone where there can be only one MCAT server, but many SRB servers. A Zone can be thought of as a 'virtual' space where entities within that space are free to interact. Using this

system, it is also possible store and retrieve information across Zones by federating two MCAT servers together.

Federation of distributed storage resources is one of the important capabilities of SRB and from a technical standpoint is easy to achieve. However, there are many practical problems involved in SRB federation, and there is no straight forward way for users to access this powerful feature. The publicly available interfaces to SRB are not always cross-platform and can be difficult to use. This general absence of usability results in SRB generally only being used for large-scale, long-term projects that have access to the support resources required to develop and design specialised non-public interfaces. The overall effect is that the vast majority of researchers are not exposed to the value of Data Grids in their daily research activities.

Generally, large research programs using SRB typically have the resources to devote to SRB administration and developing custom interfaces. Accordingly, the developers of SRB at SDSC are not focused on small-scale users of SRB and understandably tend to devote their efforts to supporting a small number of high profile projects.

YourSRB is an application that is designed to bring the valuable capabilities of SRB to more modest scale researchers who desire easy access and manipulation of data within a Data Grid environment. This is achieved by ensuring all data entered into the Data Grid is appropriately tagged and annotated, and by providing familiar interfaces for querying and retrieving data. These goals are essential if the Data Grid is going to be a long term solution to data storage and retrieval.

Our target audience includes research groups, small-medium scale project teams and distributed groups, as well as individual field based researchers. We anticipate that bringing the Data Grid to these groups will allow better information collaboration between colleagues. Moreover, as the notion of long-term, sustainable and generally accessible research repositories becomes more highly developed and embedded in the research culture, the requirement for Data Grids and an associated user framework will become essential.

## 2 What is YourSRB?

YourSRB is a cross platform interface to SRB which allows researchers to create, maintain and share information over the Data Grid. Its goal is to bring the broader research community closer together by forming collaborations *via* the Data Grid. YourSRB is designed not to work solely as an interface to SRB, but rather to be used as part of the entire Data Grid fabric. Using YourSRB, SRB can be made to act as a peer-to-peer file-sharing system. This is very advantageous, because data can be defined in the Data Grid where searching facilities are more powerful (owing to metadata descriptions) and collaboration between users can be achieved seamlessly via federations.

The proposed configuration places an SRB Metadata Catalogue (MCAT) and SRB server on a users local machine (notebook or PC) accompanied by the YourSRB interface, i.e. one SRB Zone per computer bundled with YourSRB. With this design, users have access to the Data Grid on their local machine regardless of their network connectivity status. This is particularly useful for researchers who collect data in remote locations and allows users to work within the scope of their own personal machine (not attached via a network to a server). When a user wishes to share information with other SRB Zones, they form a federation (trust relationship) between their Zone and another. This overall architecture is shown in Figure 1.

YourSRB has been developed using Java and uses the Jargon toolkit (SDSC, 2006b) from SDSC to interface to SRB services. The system also uses Chitter Chatter (VergLab, 2006), which is an extended API for Jargon developed at JCU. Chitter Chatter's extension allows for a seamless approach to the application of extended attributes such as annotations, and handles the creation and manipulation of metadata schemas. Chitter Chatter was developed to provide a common back-end between PGL (Personal Grid Library) (VergLab, 2006) and YourSRB. YourSRB has the ability to be run on Windows, Linux and MacOSX operating systems.

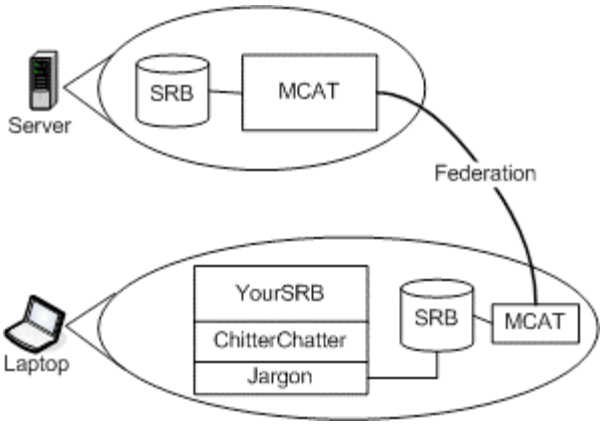


**Figure 1. YourSRB usage architecture**

## 3 Features of YourSRB

Three specific features of YourSRB are of particular importance: federation, user defined metadata schemas and searching.

### 3.1 Federations

Currently, MCAT Zones are able to share information by forming federations. A federation can be thought of as a trust relationship between separate entities or authorities, in this case separate Zones. Federations are commonly used in many database management systems. In SRB, there are two forms of federation authentication: GSI Certificates and ENCRYPT1. Forming a federation using the ENCRYPT1 authentication scheme requires the user to manually perform six separate tasks and then run a Perl script, using either the terminal based Scommand tools, or MCAT admin tool. These complex tools are suited to

administrators of SRB and assume that the user has in-depth familiarity with SRB, leaving the majority of researchers with little hope of forming Zone federations without assistance. The YourSRB system automates this process for users by providing them with an interface to input authentication information, then performing the necessary federation formation in the background using the ENCRYPT1 authentication scheme.

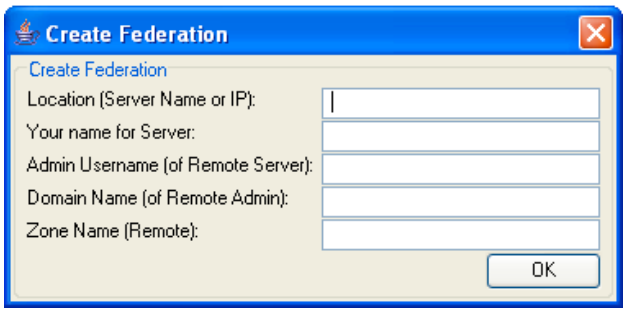


**Figure 2. Create Federation dialogue.**

In order to create a federation between the default Zone (which is named at the SRB installation stage) and another external Zone using YourSRB, the user selects the 'Create Federation' option from the file menu. Once selected the user is prompted with a Create Federation dialog box (Fig 2). The Create Federation dialog requires the user to input the following information:

- Address of remote zone – Used to locate the physical (or virtual) machine where the SRB Server resides
- Name for this zone – So the user can identify Zones in their federation list
- Admin user of remote zone – Needed to identify the trust relationship between the two zones
- Name of the domain in which the admin user of the remote zone resides – Needed to identify the trust relationship between the two Zones
- Name of the remote zone – Needed to identify the Zone to federate with

Once the user has entered the required information and clicked 'OK', the local MCAT server attempts to create a federation with the remote MCAT server. If the attempt to create a federation fails the user is re-prompted to change the details they supplied initially. For a federation to be complete, users from both Zones must create a federation to the other. Once the federation has been established by both parties, the user must run the Szonesync.pl script (distributed with the SRB and MCAT installation) which will retrieve user and resource information from the remote zone, and ingest it into the local MCAT. Automating synchronisation into YourSRB using Jargon is planned in future work.

## 3.2 Metadata and Descriptions

Metadata - data about data - is a core component of SRB and is used to describe digital objects stored in SRB. Before describing objects, it is necessary to decide what description format is required. Data within SRB is described using a metadata schema, or set of attributes which have unique values for different objects. The Dublin Core metadata schema, for example, contains 15 core elements that can be applied to any digital object (Initiative, 2006). The SRB MCAT stores the metadata attribute-value pairs for digital objects stored in SRB. The Windows interface for SRB (called InQ (SDSC, 2006a)) allows users to store these attribute-value pairs quite easily, but there is no facility for describing what attributes should be applied to these objects. Specifying the metadata schema is core to the YourSRB system. Users are given the option to create metadata schemas and apply them to all files within a directory. We are also using this general concept in another Gridsphere (web) based interface for SRB: Portable grid Library (PGL) (VergLab, 2006). When the user creates a metadata schema using YourSRB, the system stores the schema using an XML file. Users have the ability to apply many schemas to a Directory.

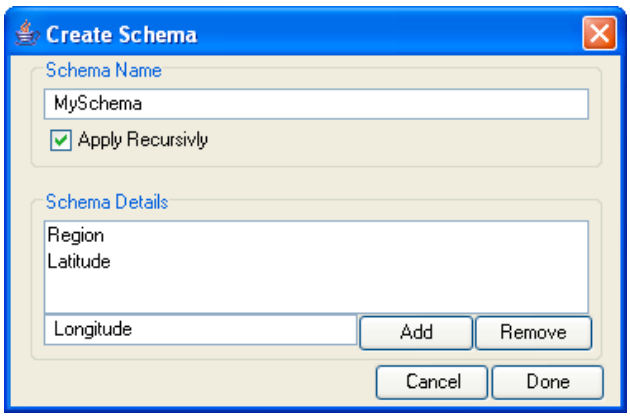


**Figure 3. Create metadata schema dialogue.**

In order to create a schema for a library/directory, the user selects the 'Create Schema' option from a menu and is prompted with a Create Schema dialog (Fig 3). Here the user is able to give the schema a name, and add/remove attributes.

When the user clicks 'Save', a schema is created in the form of an XML schema document. The document generated is then saved in the Schema Repository directory created by YourSRB in the user's home directory. The user is then able to apply the newly created schema to any selected directory. The user is also given the option to apply the schema recursively, this means that the schema will apply to all files inside directories down the tree hierarchy. It is future work to allow the user to apply schemas to different files types based on the file extension. Currently the metadata schema creator only allows you to pre-specify the attribute names. It is future work to add value type restrictions to the schema.

## 3.3 The power of search

InQ allows users to create queries for searching metadata on objects stored in SRB, which can yield exact results if the user has the patience to create the query, and also knows the metadata formation and

terms. However, the majority of researchers are possibly more familiar with keyword search engines such as Google and Yahoo. This is why YourSRB offers two forms of searching: a simple search, and an advanced search. The simple search allows users to query the metadata based on a keyword and some other properties (file type, file size etc.). The Advanced search provides an interface similar to InQ, where the user can create advanced queries which search specific attributes.

### 3.4 Interface

Figure 4 is a screenshot of the YourSRB user interface. On the left hand column is a conventional windows-like hierarchical view of the files and directory structure within the SRB collection. The central column is an icon view of each of the files. Clicking on any displayed files reveals the metadata associated with these (presented in the right hand column). Depending on access rights users can edit this metadata, upload or download files between the local file system and local SRB instance. The interface also supports drag and drop file transfers. Tabs on the Left hand side reveal the simple and advanced search interfaces to the users. These searches use the MCAT metadata store so that metadata and annotations are searched as well as simple file names.

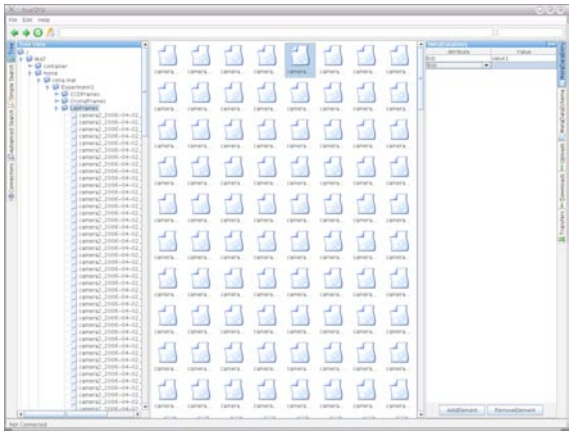

**Figure 4. YourSRB user interface.**

## 4 Collaboration Scenarios

The following are a set of scenarios for user collaboration and federation within YourSRB.

### 4.1 Ideal Local Scenario

An MCAT Zone named ZoneA exists at a University, and is used as the main storage archive and Digital Library repository for many research activities. An MCAT zone named ZoneB is located on a PC in the office of a researcher at the University. An MCAT zone named ZoneC is located on a notebook PC of a scientist who makes field trips. The scientist takes the notebook with her on field trips to store information about experiments and samples taken at various locations. In order to connect to ZoneA, owners of ZoneB and ZoneC would connect to their local network, then using YourSRB they would request to make a connection to ZoneA. Based on the authentication details provided, Zones B and ZoneC would attempt to federate with ZoneA. Once federated, users will be able to treat ZoneA as an extension of their current zone. Once federated with ZoneA, theoretically a 'silent' federation should exist between ZoneB and ZoneC, allowing ZoneB to treat ZoneC as an extension of its Zone - vice versa (Fig 5). This silent federation would be highly desirable in situation such as large scale collaborative research, where all the data is collectively owned, which in a more ad hoc federation it may not be desirable. A way of mediating these silent federations is proposed, which uses GSI certificates to specify who is allowed to silently federate. For instance, all silent federators must fall under the same certificate authority, and must match in the organisational unit portion of their distinguished name.

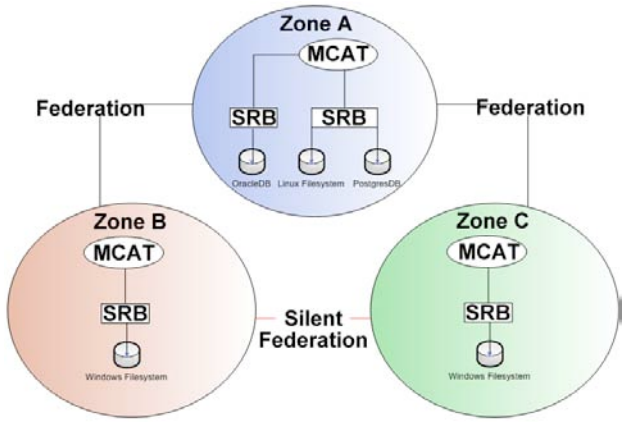


**Figure 5. Ideal local federation scenario.**

### 4.2 Ideal World Scenario

An MCAT Zone named SDSC is located on a server at the San Diego Supercomputing Centre. An MCAT Zone named JCU is located on a server at the James Cook University in Australia. An MCAT Zone named AIMS is located on a server at the Australian Institute of Marine Science in Townsville. Zones JCU and AIMS have a federation directly with the world server but not

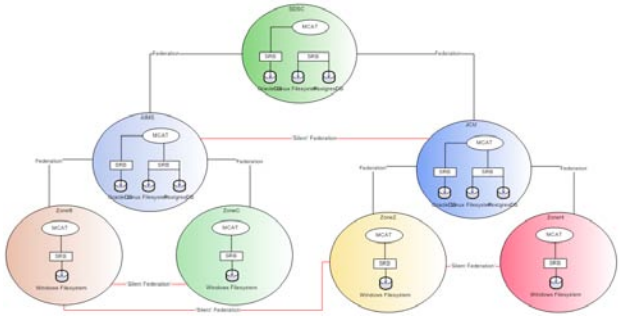

**Figure 6. Ideal world federation scenario.**

to each other. The JCU and AIMS Zones then have other Zones federated to them (as in the case of the local architecture). Ideally, a 'Silent' federation should exist between JCU and AIMS. If this is the case, then a 'Silent' federation should then exist between the Zones federated with JCU and the Zones federated with AIMS (Fig 6). Ideally, this is how YourSRB would provide the peer-to-peer sharing of information.

### 4.3 The real scenario

The architectures discussed in sections 4.1 and 4.2 are the ideal architectures YourSRB would use in order to act as a peer-to-peer data sharing system. Unfortunately these architectures can not yet be achieved due to limitations within the SRB federation implementation. Presently when a federation is formed between two Zones, the federation exists between those two Zones only. For example, if ZoneA is federated with ZoneB and ZoneB is federated with ZoneC, ZoneC does not have a 'Silent' federation with ZoneA. To achieve the scenarios in the previous sections, it is required that there be a 'Silent' federation between every single Zone. For a group of 4 Zones in the Local Architecture for example, the minimum number of federations which need to exist would be 3. In reality, with each Zone being federated to the next, there is a minimum of 6 federations required. It also requires each Zone to be aware of all other Zones. In the local architecture discussed previously, a Zone would only need to know that one other Zone exists. We have explored the prospect of modifying the SRB federation system to incorporate these improvements, and while it is technically possible we have not developed this code. We are presently in a planning phase in order to implement these changes to SRB. Despite the drawback of not yet having a try peer-to-peer federation model, there are still many use cases where the model we have implemented is viable.

## 5 Related work

There are currently very few interfaces publicly available for SRB: InQ (SDSC, 2006a), MySRB (SDSC, 2006d) & MCAT Admin (SDSC, 2006c), and the SCommands (SDSC, 2006e). While four interfaces may appear to be sufficient for any given software tool, all of the current interfaces are variously flawed in usability, performance and functionality. InQ (SDSC, 2006a) is an application developed by the SDSC (San Diego Supercomputing Centre) specifically for Windows. InQ (or Inquisitor) allows users to connect with an SRB server of choice. The InQ application has the following features:

- Authenticate
- Create Collections/Containers
- Switch between storage resources
- Associate metadata with objects
- Upload and Download Files
- Modify Access Permissions
- Run search queries

InQ closely resembles the Windows Explorer application on Windows. Although InQ is a well designed application which is easy to use, it has limitations. InQ is developed specifically for the Windows operating system, which leaves Linux and Mac users unable to use the product. It also has rather limited functionally in terms of SRB federation, and has only a rudimentary awareness of metadata.

MySRB (SDSC, 2006d) is a web based interface to SRB which also developed by the San Diego Supercomputing Centre.

MySRB allows the viewing / creation / deletion / ingestion of files, directories and metadata. The MySRB system is important because there are times when a web based interface to SRB is essential. MySRB is however complex and difficult to use for casual users, the result of it being principally designed as a demonstrator tool. Again this limits the use of MySRB as a metadata aware digital library. The application offers functionality, but lacks user focused aesthetics and ease of use. MySRB has the potential to be a valuable tool, but much work remains to be done to make generally usable.

The MCAT Admin utility is Java based GUI which is used to administer the MCAT. MCAT Admin offers the following functionality:

- Create/Display/Modify Zones
- Display/Add/Modify/Delete Users
- Display/Add/Create/Delete Resources
- Display/Add/Delete Locations
- Display/Add Tokens and Domains

The MCAT Admin tool is useful in administering the MCAT server. However, as in the case of the other interfaces, the MCAT Admin tool is poorly designed and lacks in aesthetics. The tool is aimed at use by administrators and is very complex to use. An alternative to using the MCAT Admin tool are the SCommands.

The SCommands are a command line interface to SRB and MCAT. The SCommands are provided for those users who require command line access. As with most CLI's, the SCommands provide a fast way of accessing the SRB and MCAT. In total, there are 73 SCommands which provide a full set of functionality for accessing the SRB and MCAT.

## 6 Implementation Issues

Most of the issues that arose during development of YourSRB were due to lack of documentation and examples associated with the SRB developer tools. Much of the existing SRB documentation is stored as plain text files or is briefly mentioned in research papers. The ENCRYPT1 authentication was used for federations because documented instructions exist to implement this feature. While federation via GSI certificates is possible it has not yet been introduced to YourSRB.

The Jargon API provides near complete functionality, but in some areas it is poorly designed and documented. The srbModifyZone(//..) function is one such example. This function (SDSC) function takes 8

input parameters, the final parameter is an option (or 'key'), and depending on the key, 4 of the 8 parameters have different roles when the function is called. Such practices make the Jargon API less scalable, as changing the function will cause problems related to backwards compatibility with the existing code base.

As a result of the poor documentation, the MCAT Admin tool was decompiled and examples on how to call functions taken from the code, in order to implement much of the functionality of YourSRB.

## 7 Application use-cases

The following usage scenarios have been summarised from scientific groups anticipating the availability of YourSRB. These cases have been used to guide its development. At present we are working with these and other groups developing and refining the capability of YourSRB.

### 7.1 Materials Science

Ceramic state prediction. This materials science application (Prof. Chris Berndt, JCU) is a blend of chemistry, physics and engineering, where the main focus is to probe and analyse the structure of materials based on external factors. A particular application is the analysis of bio-ceramic coatings that can be applied to prosthetic bone joints.

A typical experiment would consist of choosing a ceramic, spraying the material onto a surface at a specific temperature, angle, velocity etc. Once the ceramic has been applied to a surface, micrograph images are recorded and the ceramic surface is pressure tested for strength and durability. The nature of the coating morphology and ceramic composition vary with respect to the distance of application, the particle size, and the velocity in which the ceramic is applied (Callus and Berndt, 1999). It has also been observed that the coating morphology is directly related to the strength and bio-acceptability of the coating.

Based on experiments already completed, the ceramic state predictions project attempts to predict coating morphologies, strength and bio-acceptability through specialised image analysis and data mining techniques. Currently, a digital library of ceramic microstructures useful for analysis exists (http://www.udri.udayton.edu/) but for future development, problems lie in its lack of maintenance, centralised location, difficulty of use, primitive search tools, no API and a lack of automation.

The ceramic state prediction project aims to allow increased of collaboration between scientists and more advanced experimental selection techniques. Without the use of grid based and semantic technologies this goal is unachievable.

### 7.2 Maritime Archaeology

The Maritime Archaeology data sharing project (Dr. David Rowe, JCU) is intended to facilitate maritime archaeologists to gain access to data held in widely distributed databases with dissimilar database structures and allow the discovery and download of the data in a usable format. This is a cross-disciplinary study between archaeology and information technology and utilizes e-Research methods and tools including the semantic web (Hardy et al., 2006). The end product of this research is a prototype system which allows maritime researchers to share archaeological information across geographically diverse locations, while implementing strict rights management rules. Issues being addressed include data quality concerns, security issues, and perceived conflicts within the maritime research community regarding allowing access to sensitive data.

The use of the semantic web is of particular importance in this project concerning the discovery of resources. Automatic and manual meta-data creation and harvesting are being implemented in order to make search queries more productive and targeted. SRB is used to manage access to system resources, enforce rights management rules and maintain the semantic links and descriptions of the data sources. Although SRB may not be the best and only choice for federating databases, it does allow us to also integrate other digital objects, such as images and written documents into the repository. This is important, as the possible next states of this project include integration of an artefacts database, including photos. This ability to neatly integrate these two types of digital stores is unique.

This is one of the first uses of this technology in regards to the archaeological field in Australia. Large amounts of data currently reside in databases that are unavailable for research due to artificial constraints imposed by the lack of information technology solutions. This research allows this data to be accessed and therefore be used in further archaeological inquiry and exploration.

## 8 Conclusion and future work

While storage mechanisms for maintaining extremely large datasets have become readily available, systems which provide access and structure to the heterogeneous data have not kept pace. SRB provides a scalable, robust application for federating datasets, but is difficult for the 'average' researcher in the field to utilize. Interfaces to SRB such as inQ and MySRB have not been designed with this sort of audience in mind. In addition, cross-platform tools are not available which combine all of the functionality of these applications. YourSRB, although still in the development stages, begins to bridge this gap. Future work will include: incorporation of an integrated security model involving Shibboleth, construction of a full peer-to-peer federation model, additional development of the metadata schema model and performance/load testing on the first release candidate. YourSRB is currently in beta testing and release version 1.0 is scheduled for release in January 2007.

## References


CALLUS, P. J. & BERNDT, C. C. (1999) Relationships between the mode i1 fracture toughness and microstructure of thermal spray coatings. *Surface and Coatings Technology*.

LHC Project, http://lhc.web.cern.ch/lhc/, accessed June, 2006.

CHERVENAK, A., DEELMAN, E., KESSELMAN, C., ALLCOCK, B., FOSTER, I., NEFEDOVA, V., LEE, J., SIM, A., SHOSHANI, A. & DRACH, B. (October 2003) High-performance remote access to climate simulation data: a challenge problem for data grid technologies. *Parallel Computing*, 29**,** 1335-1356.

Global Ocean Observing System, http://ioc.unesco.org/goos/, accessed May, 2006.

GRAY, J. & HEY, T. (2001): In Search of Petabyte Databases. *HPTS Workshop, Asilomar*

HARDY, D., SIM, N. G. D. & ATKINSON, I. M. (2006): Searching Heterogeneous and Distributed Maritime Databases: A Technology Prototype. *Bulletin of Australian Institute of Maritime Archaeology*,

HEY, T. & TREFETHEN, A. (2003) The data deluge: An e-science perspective. *Grid Computing***,** 809–824.

Dublin Core, http://dublincore.org, accessed May, 2006.

KARGUPTA, H., JOSHI, A., SIVAKUMAR, K. & YESHA, Y. (2003) Data Mining: Next Generation Challenges and Future Directions. *AAAI/MIT*.

InQuisitor, http://www.sdsc.edu/srb/index.php/InQ, accessed May, 2006.

Jargon, http://www.sdsc.edu/srb/index.php/Jargon, accessed May, 2006.

MCAT Admin, http://www.sdsc.edu/srb/index.php/Admin_Tool, accessed May, 2006.

MySRB, http://www.sdsc.edu/srb/index.php/MySRB, accessed May, 2006.

Scommands, http://www.sdsc.edu/srb/index.php/Scommands, accessed May, 2006.

Storage Resource Broker (SRB), http://www.sdsc.edu/srb, accessed May, 2006.

Personal Grid Library Project, https://plone.jcu.edu.au/hpc/staff/projects/hpc-software/personal-grid-library, accessed May, 2006.