



**Storage and Infrastructure Work Package 1**  
**Facilitate distributed data management using Fedora.**

**Final Report**

**Version 0.1, April 2007**

**Lead Investigator:** Rafi M Feroze

**Prepared by:**

Rafi M Feroze      [rafi.feroze@jcu.edu.au](mailto:rafi.feroze@jcu.edu.au)

## **Executive Summary**

This document describes the work completed towards the fulfilment of requirements of the DART work package SI1 (*Facilitate distributed data management using Fedora*).

The objective of this report is to

- document the aims and objectives of the SI1 work package in terms of its contribution to the DART project,
- provide a detailed description of the work undertaken over the course of the project in order to achieve those aims,
- provide an archival record of any software, configuration instructions, hardware platforms that may have been built.

## Table of Contents

1	Introduction.....	4
2	Project Milestones .....	5
3	Technical Requirements.....	6
3.1	Fedora .....	6
3.2	SRB .....	6
3.3	Fedora-SRB Integration : SI1 Architecture.....	7
4	Project Outcomes.....	8
4.1	SRB managed Fedora Repository .....	8
4.2	Access to SRB managed External Content .....	8
4.3	Relational Digital Object - Access to data stored in Relational Database.....	9
4.3.1	Collaborations .....	10
5	Archival Storage of Project Deliverables .....	11
6	Publications .....	13
7	Terms of Reference.....	14
7.1	Glossary .....	14
8	Report Sign-off .....	15
9	Appendix.....	16
9.1	Appendix A - Screen shot describing Relational Digital Object .....	16
9.2	Appendix B – Meta Data returned to Fedora Server.....	18
9.3	Appendix C – Illustration of SI1 use case in DMQ4 work package.....	19

## 1 Introduction

SI1 work package was tasked with *investigating the appropriateness of Storage Resource Broker (SRB) as a candidate for integration (as the underlying data grid) into Fedora*. The focus of the investigation was to enable seamless *Data access & Data model integration for archival context management*.

A detailed analysis of functions provided by Fedora and SRB was carried out and the underlying architectures were studied in the first quarter of 2006. The results of this study together with proposals for integration and USE cases were documented and sent out for comment to stakeholders including the DART project Architect and external parties like the fedora development team at Cornell University. A copy of this document, can be found in the */doc* folder of the SI1 work package DVD<sup>1</sup>.

The analysis & review phase was completed by close of second quarter 2006, and the following observations were made.

1. Integration of SRB with Fedora at the storage level is best accomplished by extending the SRB lowlevel module under development by Edwin Shin at Cornell University. This module was in test phase and working on extensions to it required co-ordination with the fedora development team at Cornell.
2. DART work on extensions to the 'SRB lowlevel module' should focus on enabling fedora to disseminate SRB managed content outside of fedora repository. The proposed extensions complement functionality provided by SRB lowlevel module which uses SRB for the sole purpose of managing fedora's internal repository.

The above recommendations was in line with the objectives set out in the DART bid document for the Storage and Interoperability modules, which states :

*'DART project will need to integrate SRB with Fedora, both as a replacement storage layer for Fedora itself, and as a location for content outside a Fedora repository but managed by it'.*

DART extensions to the 'SRB lowlevel' module was completed by close of third quarter of 2006. These extensions permitted dissemination of data from SRB managed files and relational databases that were outside of Fedora repository.

Demonstrating the use of SI1 deliverable by using them in an application turned out to be the most difficult part. The analysis document proposed generic examples of use cases for the proposed extensions. Mid-way through the project, DART work package DMQ4 was identified as a suitable use case for SI1 work package. It was envisaged that once CIMA deployment was successful at IMB, data collected using CIMA application can be adapted for dissemination by DMQ4 demonstrator using fedora.

The usability of SI1 extensions to DART is demonstrated by its use in DMQ4 work package to disseminate internal SRB repository data and external relational datasets.

---

<sup>1</sup> See '*SI1 Requirements Analysis & Proposal.doc*' in the */doc* folder found in the SI1 work package DVD.

## 2 Project Milestones

The following activities were successfully carried out during the course of the project.

Dec 05 - Jan 06

1. Review of Gfarm, SRB and OpenDAP was carried out, SRB found most suitable for integration into fedora at Repository level

Feb – Jun 06

1. A detailed review of Fedora and SRB was carried out to identify integration requirements. The outcome of this review was a Requirements and Analysis document.
2. Analysis document was sent out for comment to Project Architect and Fedora development team at Cornell. Comments received were discussed with project team and responses recorded.
3. Communication links with Cornell fedora development team was established. Review and Testing of development version of SRB lowlevel module was carried out.
4. SRB , Fedora and related software environments were set-up on local and server machines for development and support purposes.
5. Attended SRB workshop at Townsville as part of Pragma 06 conference. Met with and established communication links with SDSC team (Arcot Rajasekar and team) and JCU CIMA team (Ian Atkinson and team)

Jul – Oct 06 :

1. Phase 1 completed. SRB lowlevel module developed by fedora team was reviewed and tested to be suitable for DART project
2. Phase 2 completed: SRB managed datasets were integrated into Fedora as external datasets.
3. Phase 3 completed: SRB managed databases & shadow collections were integrated into Fedora for dissemination
4. The code was deployed on internal DART server 'datura' and external DART server 'maenad' in testing and implementation phase.
5. Installation instructions and other relevant documents were prepared.
6. Source code and handover documents were moved to SVN managed on internal DART server datura.

### 3 Technical Requirements

A brief discussion on the architectural and technical issues related to the work package is described below. For a detailed discussion on functional and architectural aspects of SII work package (& Fedora, SRB etc.) please refer to the Requirements Analysis document mentioned earlier. This document titled '*SII Requirements Analysis & Proposal.doc*' is available in the '*ldoc*' folder of the SII work package DVD.

#### 3.1 Fedora

Fedora provides a pluggable architecture. Its functions are implemented as modules that are configurable. Of particular interest to SII work package are the Storage and Access modules of fedora. The storage module is used to manage Fedora repository and is implemented on top of a simple file system. The Access module permits dissemination of data stored in the internal fedora repository and external *http* based web pages. The low level storage module provides the base functions for reading data (internal and external) that is disseminated by fedora. SII work package requirements are met by re-implementing this low level storage module to use SRB instead of file system for both storage and dissemination of internal and external data.

#### 3.2 SRB

SRB is best described as *a client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets*. SRB functions are exposed to potential clients by various mechanisms ranging from command line utilities to Application program interfaces (Java, C etc) and web based clients. Of interest to the SII work package is the Jargon Java API for SRB which can be used to implement and extend a new fedora low level storage module that uses data from SRB managed data sources. This permits fedora to disseminate data from diverse data sources ranging from networked file systems to relational databases.

The integrated Fedora-SRB system can leverage the best functions available in both Fedora and SRB.

Key functions of Fedora that is of interest to SII package include

- Digital Object creation and management
- Ontology based meta-data, in particular the Kowari-triple based resource index that ties together both internal Fedora meta-data and external Ontologies applied to the Digital objects.
- Version Control features that keeps an archive all changes made to the digital object
- Access control mechanisms allowing fine-tuned access to digital object elements.
- The 'External Content' methods. In particular , extending the 'external content' mechanism to access SRB data.

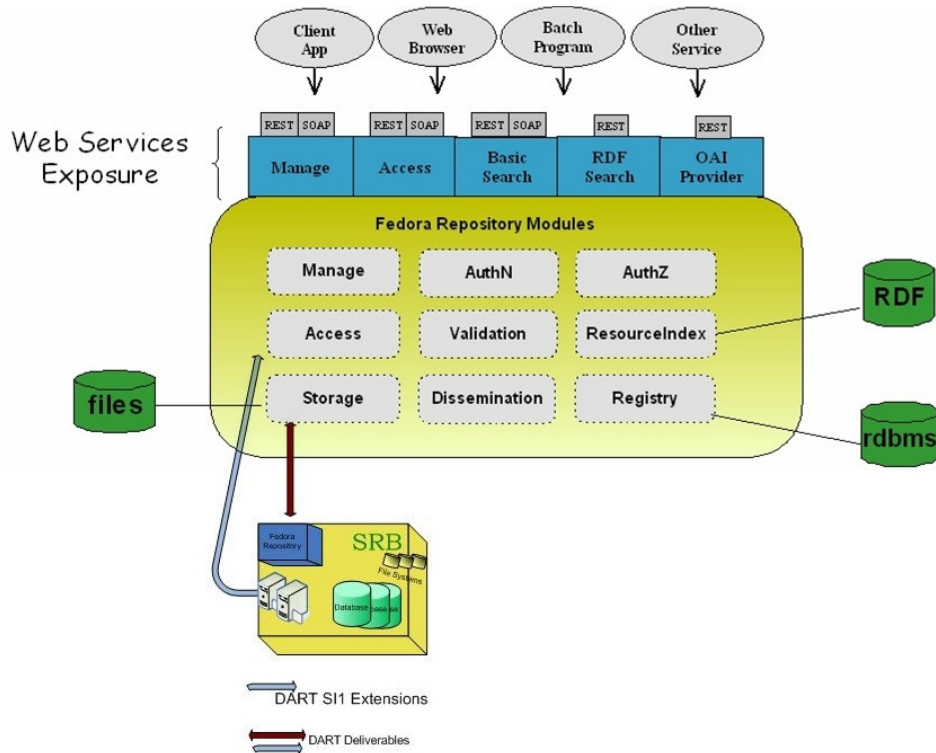
Key functions of SRB that the integration should aim to bring to Fedora include :

- Access to distributed data sources managed by SRB. In particular the ability to define Digital object elements from diverse storage devices and locations.

- Access to Meta-data maintained and disseminated by the MCAT database. The analysis document discusses this matter in more detail.

### 3.3 Fedora-SRB Integration : SI1 Architecture

Figure 1, shows how SI1 work package extends the Fedora Architecture to include SRB as the underlying storage and dissemination source. By using SRB for disseminating external data (i.e., data managed outside of fedora), the work package extends the range of data available for dissemination to a variety of file systems and relational databases in addition to existing *http* based web content.



**Figure 1: DART SI1 Architecture**  
Integration of SRB at storage level in Fedora

## 4 Project Outcomes

A quick cross reference linking the SII extension of Fedora to DART objectives of this work package is given below.

- low level data store,
- *satisfies Data model integration, SRB archival and replication*
  - SRB managed external content
  - *satisfies Data access integration*
  - SRB managed database contents

-- *provides for additional Data access integration*

As a result of effort spent on SII work package, three new functions were made available to the Fedora application.

### 4.1 SRB managed Fedora Repository

SRB Lowlevel module, developed by Edwin Shin at Cornell University was reviewed and tested during the analysis phase of this work package. The final public version of this module was released in January 2007 as part of Fedora release 2.2.

The lowlevel module allows fedora to use SRB managed file systems for storing repository content. This module satisfies the data model integration requirements for the SII work package. In addition, by using SRB Archival and Replication functions provided at collection level, Fedora repositories can be freed from the process and volume intensive task of maintaining historical data.

Fedora 'lowlevel module' is easy to use. Two changes are required to an existing instance of the fedora server.

- include the relevant *srb\_lowlevel* JAR file in the fedora server's WEB\_INF library &
- modify Fedora Server Configuration file, *fedora.fcfg* to use the new SRB lowlevel module, instead of the file system module.

Instructions for above are available at the Fedora Website and in the *doc/install.html* file found in the SII work package DVD.

### 4.2 Access to SRB managed External Content

DART SII Extensions to SRB *lowlevel* module enables Fedora to access datasets and files managed by SRB servers external to Fedora repository. This new functionality is built on top of the existing Fedora access to HTTP based external files.

Networking multiple SRB servers and providing automated links and authentication mechanisms are complex issues that are best dealt by seasoned SRB managers and are out of scope of this work package. Having said that, the fact that SRB servers across the world can be linked to form a virtual network, provides users of DART SII extensions to Fedora a new set of tools to define and disseminate Digital Objects with content from distributed data sources across the world.

Fedora 'external content module' is part of the jar files delivered with SII work package. It is easy to use. Two changes are required to an existing instance of the fedora server.

- include the relevant *dartSII.jar* file in the fedora server's WEB\_INF library
- &
- modify Fedora Server Configuration file, *fedora.fcfg* to use the new SRBExternalContentManager module for the External Content module.

Additional instructions to set-up SRB parameters for the external content module is available in the *doc/install.html file* found in the SII work package DVD.

### 4.3 Relational Digital Object - Access to data stored in Relational Database

One of the unique features of SRB is that it allows users to treat relational databases as any other file system. This allows users to read and write to databases without need for database management clients. SRB introduces the concept of Shadow Database System to manipulate data stored in databases. By extending the functionality of SRB External Content module defined above, DART SII work package allows fedora users to access data stored in relational databases and disseminate the same.

An good use case for this Relational Digital Object can be found in the demonstrator developed for DMQ4 work package. The data for this demonstrator is collected in an X-Ray crystallography lab at IMB. Among other things this data includes temperature and humidity values from sensors placed in the X-Ray lab. Processing requirements on this data include determination of minimum/maximum/average values over any given period of time. Storing this data in a simple relational table and using the built-in SQL functions to manipulate them is the best solution for this problem. By enabling fedora digital objects to send SQL queries to SRB managed relational databases, the same data can now be disseminated to a wider audience through web browsers without need for any database client. A typical Digital Object for the protein crystallography experiment will then include data streams for X-Ray diffraction pattern, Camera Images (fetched as files) and sensor data (fetched as relational digital object). A sample screen shot describing a fedora Relational Digital Object is given in appendix 1.

Fedora 'Shadow Database module' is part of the jar files delivered with SII work package. It is easy to use. Two changes are required to an existing instance of the fedora server.

- include the relevant dartSII.jar file in the fedora server's WEB\_INF library
- &
- modify Fedora Server Configuration file, fedora.fcfg to use the new SRBDBShadowExternalContentManager module for the External Content module.

Additional instructions to set-up SRB parameters for the SRB Shadow external content module is available in the *doc/install.html* file found in the SII work package DVD.

#### Architecture Issues

To use the deliverables provided by SII work package, the following applications must be installed either on local machine or available over the network.

#### Fedora Server

SII deliverables were tested and found working against Fedora server version 2.1.2. This version of Fedora is available at <http://www.fedora.info/download/2.1.1/>.

Review of beta code for Fedora 2.2, revealed no inconsistencies with the dart SII deliverable. Hence the Jan 2007 release of Fedora version 2.2 should work with minimal configuration changes, if any.

#### SRB Server

Fedora SRB integration was tested with SRB server version 3.4.2. Latest SRB servers can be downloaded from <http://www.sdsc.edu/srb/index.php/Downloads>.

Since SII integration work relies mostly on the Jargon Java API for SRB and not on the Server itself, latest version of SRB should work fine with SII deliverable. To obtain specific version of SRB tested for this work package, please contact the SDSC SRB team. Dart is unable to provide SRB code due to licensing restrictions imposed by SRB.

#### Jargon API

SI1 deliverables were tested against Jargon jar file provided with the source code for SI1. As with SRB, Jargon website provides link to the latest release version (<http://www.sdsc.edu/srb/jargon/>).

#### **4.3.1 Collaborations**

The deliverables of this work package depend on two evolving Applications, Fedora and SRB.

Fedora functions are modular and configurable. Provided the base architecture dealing with repository data management is not modified, the deliverables of SI1 work package can be used with future versions of Fedora. Users of this work package are encouraged to register with Fedora user group to keep abreast of changes to fedora.

SRB and its Jargon API are very stable. No major modification is envisaged to SRB's virtual file hierarchy based on collections and MCAT meta-data. Fedora integrates SRB at storage level by replacing the local file system with SRB managed collection hierarchy. MCAT meta-data fetched from SRB and returned to Fedora server is limited to those mentioned in Appendix B.

## 5 Archival Storage of Project Deliverables

The following is the record of Software, configuration instructions, Analysis and Review documents created as part of work done for SII work package. A copy of all the files mentioned below is available in the SII work package DVD.

1. The document that you are currently reading is titled, *SIIFinalReport.doc* and is available under the Root(i.e., / folder) of the DVD.
2. The following documents are available under the '/doc' folder.
  - *install.html* : provides installation instructions for using SII libraries
  - *InitialReview-Gfarm-Opendap-SRB.html* : Documents information gathered on Gfarm, Opendap and SRB during the early stages of the project.
  - *SII Requirements Analysis & Proposal.doc* : Requirements Analysis document for SII work package.
3. The following 'jar' files are available under the '/dist' folder. This forms the core binary deliverable for using SII workpackage.
  - *dartSII.jar* : this 'jar' contains all the java classes that were coded for the SII work package. For detailed description on individual classes, please refer to the javadoc files discussed below.
  - *The following two jar files were not created by DART SII work package, Classes in dartSII.jar above extend functionality provided in the two classes below. They are required for successful deployment of SII work package and hence are described below :*
    - *jargon.jar* : this 'jar' contains jargon classes that were used for testing java source packages in dartSII jar file. This jar file was created by external party and is open source and latest version can be obtained from SDSC's Jargon website.
    - *srblolevel.jar* : this 'jar' was created by fedora developers at Cornell and provides methods for integrating fedora repository in SRB.
4. Documentation describing Java API's resulting from the 'dartSII.jar' package is available in */src/javadoc* folder of SII work package DVD. A list of important java classes that were coded as part of SII work package is given next.
5. All source code developed for the SII work package is available under the */src/java* folder of the SII work package DVD. Refer to relevant javadoc (above) for usage information. The classes include
  - *SRBShadowDBWriteFile.java* : Package *edu.sdsc.grid.io.srb* . This class extends JARGON class *SRBShadowFile* to provide write and append functions on database shadow Files.
  - *SRBExternalContentManager.java* : Package *fedorax.server.module.storage. lowlevel.srb*. This class extends Fedora's *DefaultExternalContentManager* class to provide access to SRB managed external datasets.
  - *SRBShadowExternalContentManager.java* : Package *fedorax.server.module.storage. lowlevel.srb*. This class extends the *SRBExternalContentManager* class to provide access to SRB managed Shadow files .
  - *SRBShadowExternalDBContentManager.java* : Package *fedorax.server.module.storage. lowlevel.srb*. This class extends *SRBExternalContentManager* to provide access to relationaldata stored in SRB managed databases.

- *SRBExternalMetaData.java*  
*SRBFilePipedOutputStream.java* & *utils.java*  
Package *fedorax.server.module.storage.lowlevel.srb*. These classes are used by the above classes for meta-data fetch and utilities.
  - *ExternalNetworkContent.java* :  
Package *fedorax.server.module.storage.lowlevel*. This is another implementation of the fedora low level file system to allow network URLs beginning with '//' to be used when defining digital objects. This allows fedora server to disseminate content available over network, instead of being limited to local file system. This code is not directly relevant to SI1 work package, but was coded by the lead investigator to help resolve a problem raised in the fedora user newsgroup.
  - *build.xml* : Build file, used to compile code and generate 'jar' files.
6. Architecture Diagrams are available in */diag* folder of the SI1 work package DVD.
- *SI1-Arch.jpg* : Illustrates how SRB and Fedora integration at storage level work.
  - *DMQ4-DC.png* : Illustrates how the data capture methods for DMQ4 work package use SI1 work package deliverables.

## **6 Publications**

No journal publications or conference proceedings were created from this work.

In addition to documents and diagrams detailed in the previous section , a power point presentation was provided during DART Workshop held at Brisbane in April 2006.

This presentation described the problem and listed options that were under review at that time.

The presentation document can be found in the */doc* folder found in the SI1 work package DVD under the name *SI1-WorkshopPresentation.pdf*

## 7 Terms of Reference

### 7.1 Glossary

<b>Acronym</b>	<b>Definition</b>
Fedora	Flexible Extensible Digital Object Repository Architecture
SRB	Storage Resource Broker
SDSC	San Diego Supercomputer Centre
MCAT	Meta-data Catalogue
IMB	Institute of Molecular Biology, St Lucia, Australia

## 8 Report Sign-off

It is agreed between

Lead Investigator : Rafi Mohamed Feroze

and

Chief Investigator : Xiaofang Zhou

and

DART Project Director : Andrew Treloar

That the **Final Report Document** for the DART SII (Facilitate distributed data management using Fedora) gives a full account of the work undertaken for the DART Project.

Rafi M Feroze	<a href="mailto:rafi.feroze@jcu.edu.au">rafi.feroze@jcu.edu.au</a>
Xiaofang Zhou	<a href="mailto:zxf@itee.uq.edu.au">zxf@itee.uq.edu.au</a>
Andrew Treloar	<a href="mailto:andrew.treloar@monash.its.edu.au">andrew.treloar@monash.its.edu.au</a>

- has been read and reviewed by all parties,
- shows that the work package SII has been completed satisfactorily,
- clearly outlines the functionality that was delivered.

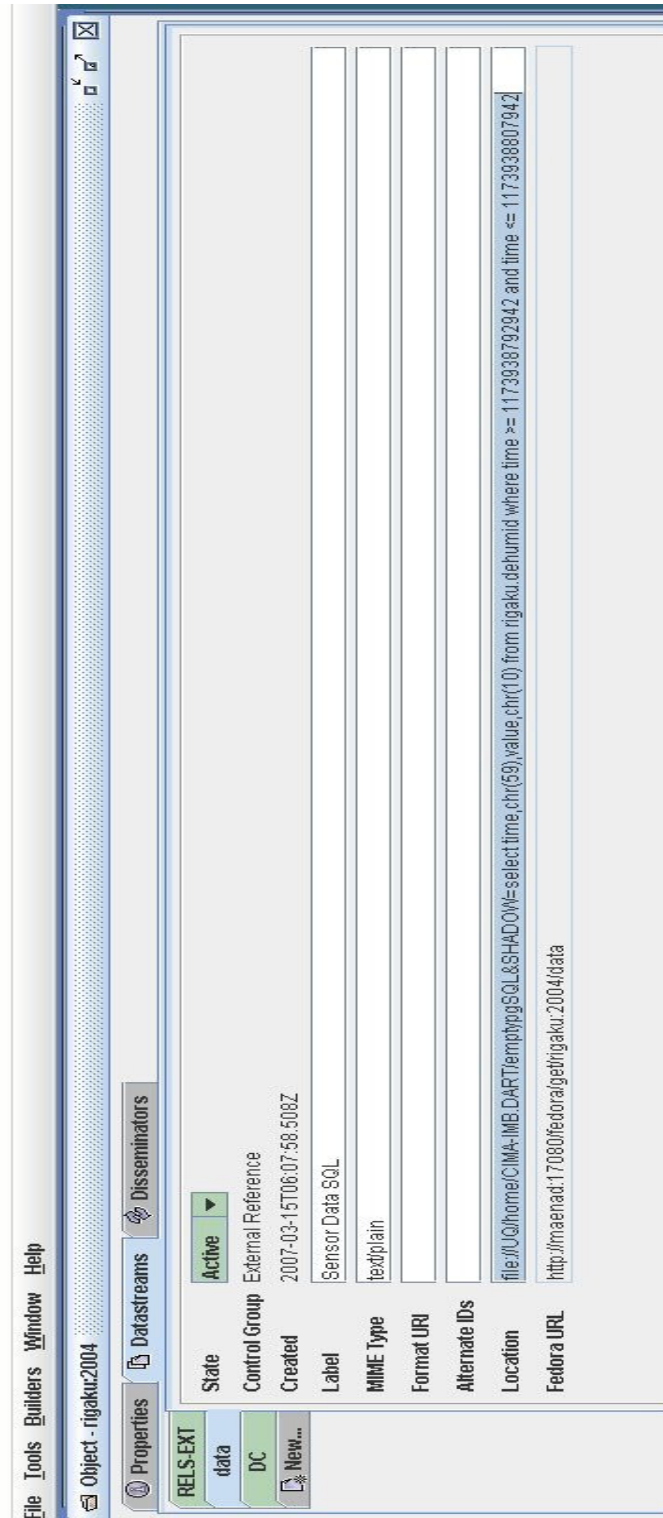
Dated this dd<sup>th</sup> day of mmmm 20yy

\_\_\_\_\_  
Signed by name of CI for and on  
behalf of the Chief Investigator

\_\_\_\_\_  
Signed for and on behalf of DART by the  
Project Director Andrew Treloar

# Appendix

## Appendix A - Screen shot describing Relational Digital Object



**The foxml for datastream Relational Digital object is given below :**

```
<foxml:datastream CONTROL_GROUP="E" ID="data" STATE="A" VERSIONABLE="false">
  <foxml:datastreamVersion CREATED="2007-03-15T06:07:58.508Z" ID="data.0"
  LABEL="Sensor Data SQL" MIMETYPE="text/plain" SIZE="0">
    <foxml:contentLocation REF="file://UQ/home/CIMA-
  IMB.DART/emptygSQL&amp;SHADOW=select time,chr(59),value,chr(10) from
  rigaku.dehumid where time &gt;= 1173938792942 and time &lt;= 1173938807942"
  TYPE="URL"/>
  </foxml:datastreamVersion>
</foxml:datastream>
```

## Appendix B – Meta Data returned to Fedora Server

SRB provides a variety of meta-data for collections and datasets managed by it. This data is stored in a Meta data Catalogue (MCAT). Since Fedora provides means for meta-data management internally using Foxml and RDF, MCAT meta-data is useful only for the externally managed SRB datasets (i.e., outside of fedora repository).

The following MCAT meta-data are returned to the SRB server when external datasets are fetched using the SI1 extensions to srbLowlevel module. The meta-data is returned as part of *header properties* of the MIMETypedStream object returned by the External Content Manager. However, since Fedora does not use this data, at present the data is not made available to the end user. It is hoped that future releases of fedora will make use of this data.

SRBMetaDataSet.FILE\_TYPE\_NAME,  
SRBMetaDataSet.FILE\_NAME,  
SRBMetaDataSet.FILE\_COMMENTS,  
SRBMetaDataSet.SIZE,  
SRBMetaDataSet.DEFINABLE\_METADATA\_FOR\_FILES  
which is a set of user defined metadata (name = value)

## Appendix C – Illustration of SI1 use case in DMQ4 work package

