Implementing Preservation Services over the Storage Resource Broker

Douglas Kosovic, Jane Hunter

School of Information Technology and Electrical Engineering
The University of Queensland
douglask@itee.uq.edu.au; jane@itee.uq.edu.au

Abstract

Many international institutions and organizations responsible for managing large distributed collections of scientific, cultural and educational resources are establishing data grids that employ the San Diego Supercomputer Centre's Storage Resource Broker (SRB) for managing their collections.

Over time, maintaining access to the resources stored within SRB data grids will become increasingly difficult as file formats become obsolete. Organizations are struggling with the challenge of monitoring the formats in their SRB collections and providing suitable migration or emulation services as required. Automated methods are required that notify collections managers of objects that are at risk and that provide solutions to ensure long term access. The problem is exacerbated by the often proprietary and highly eclectic range of formats employed by scientific disciplines – many of which are too uncommon to be considered by existing national digital preservation initiatives.

This paper describes our test bed implementation of a set of preservation services (obsolescence detection, notification and migration) over a heterogeneous distributed collection of objects, stored in SRB. It also provides an evaluation of the performance and usability of the *PresSRB* system, within the context of an environmental case study.

Introduction

Existing digital preservation projects have primarily been driven by the libraries and archives communities and as such, have tended to focus on library and cultural resources – stored in repositories such as DSpace and Fedora. More recently there have been projects focusing on preservation services for resources such as CAD drawings [1] and video games [2]. Our interest is in the preservation of large scale scientific data formats increasingly being stored within the San Diego Supercomputer Centre's Storage Resource Broker (SRB) by many scientific and research organizations.

SRB is a data grid middleware system that provides a uniform interface to heterogeneous data storage resources distributed over a network. It implements a logical namespace (that points to the physical files) and maintains metadata on data-objects (files), users, groups, resources, collections, and other items in an SRB Metadata Catalog (MCAT) which is stored in a relational database management system (e.g., PostgreSQL). Within the eScience domain, many communities are adopting SRB to implement data grids capable of managing the storage and movement of large scale collections of data and images. However, to date, no one has investigated the issues associated with maintaining long term access to digital files stored within SRB. Hence the objective of the work described in this paper is to investigate how preservation services (such as were developed within the previous PANIC [3] and AONS [8] projects) could be implemented over a collection of scientific data objects stored in SRB.

PANIC is a semi-automatic digital preservation system [3] developed at the University of Queensland that relies on semantic web services architecture. Preservation metadata associated with digital objects is generated at ingest and periodically compared with up-to-date software version, format version and recommended format registries. This enables potential object obsolescence to be detected and a notification message sent to the relevant agent. Preservation software modules (emulation and migration) were converted to web services and described semantically using an OWL-S ontology. Software agents enable the most appropriate preservation service(s) for each object to be automatically discovered, composed and invoked. The aim of PANIC was to leverage existing but disparate efforts by integrating a set of complementary tools and services including:

- Preservation metadata generation and extraction tools (e.g., JHOVE, the JSTOR/Harvard Object Validation Environment [4] and DROID [5], the National Archive's tool for performing batch identification of file formats)
- The Global Digital Format Registry (GDFR) [6]
- The UK National Archive's PRONOM project [7]

PANIC delivered a prototype system that successfully demonstrated the potential of a web services approach to automatic obsolescence detection, notification and migration. The aim of the AONS project [8] was to adapt the obsolescence detection and notification component of PANIC to generate a web service which could be applied

to multiple collection types (in particular, DSpace and Fedora) and which collections managers could easily subscribe to. AONS used preservation information about file formats and the software that these formats depend on, to determine if any of the formats within a collection are at risk. Up-to-date information about the current format and software versions was gleaned from authorized registries (PRONOM [7] and LCSDF [9]) and stored in a MySQL database. AONS then periodically checked the contents of the repository against the database to check for formats in danger of becoming obsolete. When any such formats were found, a notification report was sent to the repository manager. Because the interface between AONS and the repositories is simple, well defined and repository-independent, it was easy to deploy over different types of repositories (DSpace and Fedora).

The aim of the work described in this paper is to investigate the deployment of:

1. obsolescence detection and notification services, followed by
2. migration services

over scientific data/objects stored within SRB. This work will be carried out through the development of the *PresSRB* prototype and its evaluation through an environmental case study.

The use of SRB also raises a number of new and challenging issues that need to be considered, including:

- The storage of preservation metadata within MCAT;
- Recommended format registries for data associated with specific scientific disciplines;
- Obsolescence detection and migration services for multiple versions of the same object, stored at distributed locations within a SRB data grid.

An Environmental Case Study

Remote sensing satellite images are typical of many scientific data sets. They are represented in a wide range of formats – both open and proprietary, depending on the organization or satellite operator producing the images. Formats include: CCRS, EOSAT, HDF (Hierarchical Data Format), Fast-L7A, CEOS, ERDAS Imagine and GeoTIFF (Geographic Tagged Image File Format). GeoTIFF is the most popular standardized file format for GIS applications – suitable for storage and transfer across operating system environments and applications. It is open, public domain and non-proprietary. GeoTIFF embeds georeferencing information (e.g. projection, datums and ellipsoids, coordinate values) as metadata within the TIFF (Tagged Image File Format) file [10]. As the GeoTIFF format is fully compliant with the TIFF 6.0 specification, applications that don't know about the GeoTIFF tags will be able to open them like any other TIFF file.

For evaluating the PresSRB system, we decided to work with existing users of SRB – the Centre for Remote Sensing and Spatial Information Science (CRSSIS) at the University of Queensland. CRSSIS are using SRB for the storage and analysis of Landsat 5 satellite images provided by the Queensland Department of Natural Resources and Water in the ERDAS Imagine (.img) file format.

The Landsat 5 satellite has an onboard sensor called the Thematic Mapper (TM). The TM sensor records the surface reflectance of electromagnetic (EM) radiation from the sun in seven discreet bands. Reflectance is the ratio of outgoing light reflected from the land surface to the incoming light from the sun. Mosaics of Landsat 5 satellite image data are provided as 6 layer ERDAS Imagine files. The various layers and corresponding wavelengths are shown in table 1.

Image layer	Landsat Band	Wavelength (μm)
1	1	0.45 - 0.52 Blue
2	2	0.52 - 0.60 Green
3	3	0.63 - 0.69 Red
4	4	0.76 - 0.90 near infrared
5	5	1.55 - 1.75 shortwave infrared
6	7	2.08 - 2.35 shortwave infrared

Table 1: *Landsat TM ERDAS Imagine Layers*

ERDAS Imagine is a commercial raster image processing and remote sensing geographic information system (GIS) application. The current version is 9.2 and is available for Microsoft Windows. In the past, both Unix/X-Windows versions were also available. For images that require more than 4GB of disk space, Imagine creates two files: the .img file contains the traditional superstructure, but the actual raster data is kept in a separate file which has an extension .ige [11]. The ERDAS Imagine file samples that were used in this case study ranged from 6.6 GB (28189 x 38828 pixels) to 20 GB (45144 x 72111 pixels).

The problem with ERDAS Imagine files are that they are very large, depend on proprietary software and are difficult to manage. To maximize long term access and availability, they should ideally be converted to GeoTIFF (both full-size and thumbnail for previewing).

Although the TIFF specification allows for multi-spectral imagery (more than 3 bands), many software applications are unable to handle multi-spectral TIFF files. To overcome this limitation, many satellite image providers deliver two GeoTIFF files, one with red, green and blue bands and another with near infrared, red and green bands [12]. For the PresSRB prototype, we present the collections manager with the option of either selecting 3 bands (RGB) or 4 bands (RGBA) for converted GeoTIFF files generated by the migration service. In addition, because TIFF 6.0 uses 32bit unsigned offsets, it is limited to a maximum files size of 4GB. Our SRB data grid contained a number of ingested GeoTIFF files that were bigger than 4 GB. For these files we used BigTIFF (aka TIFF-64) – a proposed standard for TIFF data bigger than 4GB in file size. There

are now a growing number of geospatial tools and libraries which are able to support it.

For the conversion from Imagine to GeoTIFF format, we used GDAL (Geospatial Data Abstraction Library). GDAL is a translator library for raster geospatial data formats that is released under an X/MIT style Open Source license by the Open Source Geospatial Foundation [13].

GeoTIFF format is registered in PRONOM as a distinct format, but LCSDF has no format description for GeoTIFF. The ERDAS Imagine formats are not registered in the PRONOM, LCSDF or GDFR registries. Ideally there should be a recommended format registry for geospatial data, that provides best practice guidelines for the archival and curation of geo-spatial data.

System Architecture

Figure 1 illustrates the overall architecture of the PresSRB prototype and the various software layers that interact with the underlying SRB Data Grid.

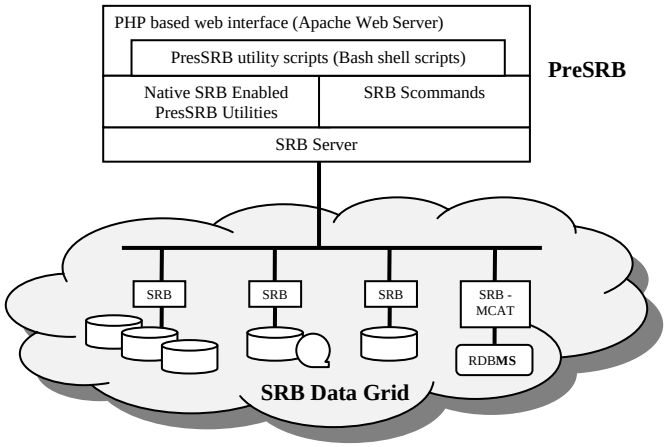

Figure 1: *PresSRB System architecture*

The PresSRB prototype was implemented on a PC running RedHat Enterprise Linux (RHEL) 4. Significant effort is required to setup and deploy SRB, so to simplify the deployment of SRB on other RHEL systems, SRB-3.4.2 was packaged into a number of RedHat Package Manager (RPM) files.

SRB Data Grid

The SRB is a data grid middleware system that provides a uniform interface to heterogeneous data storage resources distributed over a network. It implements a logical namespace (that points to the physical files) and maintains metadata on data-objects (files), users, groups, resources, collections, and other items in an SRB Metadata Catalog (MCAT) stored in a relational database management system.

SRB Scommands

Scommands are command-line SRB client utilities for accessing SRB data and metadata. Most *Scommand* names have a "S" prefix. *Scommands* are the most powerful and flexible of the SRB clients that come with the SRB source code. They are ideal for batch jobs, scripting and PHP wrappers.

Shell scripts are extensively used in the PresSRB prototype to wrap Scommands and the previously mentioned SRB-enable utilities, in order to provide much of PresSRB's functionality and to perform batch operations.

SRB-enabled Utilities

The simplest approach to providing SRB support to a non-SRB application is to copy or replicate data from SRB space to a local file space, and then provide the application with the local filename. This approach works very well when an entire file is to be processed, but if, for example, only the first 26 bytes of a 20GB ERDAS Imagine .ige file is required for format identification, it's very wasteful both with respect to bandwidth and performance.

For the PresSRB prototype, the Linux **file** command was used to identify file formats [14]. It was modified to make SRB client library calls instead of Unix file I/O calls. The benefit is that the stock **file** command reads at most 256kB of data to identify a file - so too does the modified **file** SRB-enable command hereafter referred to as **Sfile**.

Similarly the GDAL **gdalinfo** georeferencing meta-data extractor used on ERDAS Imagine and GeoTIFF files was modified to add native SRB support, as it also only needs to read a small portion of the files.

Each ERDAS Imagine file was migrated to two new files: a low resolution, 1% sized preview GeoTIFF file and a equivalent resolution GeoTIFF file.

For file format conversion in PresSRB, an Scommand-based shell script wrapper was initially used which retrieved a file from SRB space using the **Sget** command, then invoked the **gdal_translate** utility to perform the conversion on a local file and then upload a converted file to SRB space using **Sput**. For transferring large files using **Sput** and **Sget**, it was possible to take advantage of SRB's parallel I/O capabilities (multiple threads each sending a data stream over the network) which made SRB significantly faster than HTTP, FTP, SCP or even NFS.

We also tested adding native SRB support to the **gdal_translate** utility. This was beneficial for the conversion of large ERDAS Imagine files to BigTIFF/GeoTIFF file of the same resolution. This was because only 3 of the 6 bands of the ERDAS Imagine files are processed in the conversion.

Generation of small GeoTIFF preview files (scaled down to 1% in both the horizontal and vertical resolution from the original ERIDAS Imagine files) also had a significant benefit, since only a small "overview" needs to be read which is only a tiny fraction of the ERDAS Imagine file size. For some of the sample ERDAS Imagine files we used, they had up to nine overviews ranging from 34x84 to 8504x21269 pixels in size.

PHP Web interface

The SRB-3.4.2 source code comes with a sample ScmdWrapper PHP class which enables wrapping SRB Scommands. The SRB authentication, browsing, searching and ingestion components of the PresSRB Web interface employs the ScmdWrapper PHP class. A number of the PresSRB Scommand-based shell scripts which are invoked by a scheduler component are also able to be directly executed from the PresSRB Web GUI. This approach improves system performance (over the Tomcat-based Java approach used by AONS) because it requires less memory and overheads when dealing with large files.

System Implementation

Figure 2 below illustrates the main components of the PresSRB system. The six main components (described in the next 6 subsections) are:

1. Format identification and Preservation metadata
2. Format, Software and Recommendation Registries
3. Obsolescence Detection
4. Migration and Preview
5. Scheduler
6. Web GUI

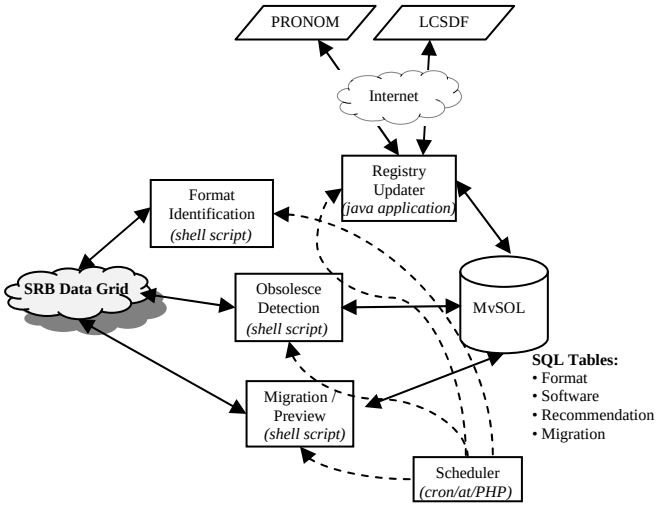

Figure 2: *PresSRB non-GUI components*

Format Identification and Preservation Metadata

Format identification in the PresSRB prototype involves identifying the file format of SRB data objects (i.e. files in SRB space), extracting georeferencing metadata if available and then populating the SRB data objects with PresSRB specific SRB user defined metadata. This is all performed by a shell script which invokes the following three commands:

- SRB enabled **file** command (**Sfile**) for file identification.
- SRB enabled **gdalinfo** for georeferencing metadata extraction.
- **Sufmeta** Scommand which provides the facility for inserting, deleting, updating SRB user-defined metadata (attribute name-value-unit triplets).

The **Sfile** utility is unaware of BigTIFF files and ERDAS Imagine .img and .ige files. In order to recognize these formats, the `magic/Magdir/images` file from the **file** source code needed to be modified to add signature recognition for these formats and then the file source code recompiled to generate a magic file with the new signatures.

The **Sfile** command currently recognizes GeoTIFF files as just generic TIFF files. Instead of modifying the **Sfile** command to specifically recognize GeoTIFF files, the shell script wrapper uses the **gdalinfo** command to confirm if they are GeoTIFF files.

The shell script wrapper maps the output of the **Sfile** command on a SRB data object to a PRONOM Persistent Unique Identifier (PUID) for a set of known file formats. If the shell script is able to determine the PUID of a SRB data object, it then inserts a SRB user defined attribute called "pressrb_PUID" on that data object with the value set to the PUID. If the **Sfile** command is able to determine the mime type of a data object, then a "pressrb_mime_type" is also inserted.

To avoid name space collisions of SRB metadata attribute names which might already be used in an existing SRB data grid, the attribute names are prefixed with "presrb_".

As PUIDs for BigTIFF and ERDAS Imagine file formats haven't been assigned yet in the PRONOM repository, we are currently using a local format repository MySQL table to define the interim PUIDs listed in Table 2. We intend submitting a request to the managers of PRONOM to have PUIDs assigned for these formats.

PUID	Format Name	Extension
x-fmt/10000	BigTIFF	tif, tiff
x-fmt/10001	BigTIFF/GeoTIFF	tif, tiff
x-fmt/10002	Erdas Imagine	img
x-fmt/10003	Erdas Imagine - Large Raster Spill File	ige

Table 2: *Local File Format Repository*

For example, given the following ERDAS Imagine SRB data object file:

```
srb:/home/srb.demo/sample.img
```

The format identification shell script wrapper will invoke the **Sufmeta** command similar to command-line that follows, which sets the "pressrb_PUID" attribute to the appropriate PUID value:

```
Sufmeta –d pressrb_PUID "x-fmt/10002" sample.img
```

As we are dealing with very large raster images, the output target file can easily exceed the maximum file size permitted for that format (e.g., a TIFF file cannot exceed 4GB – if it exceeds this size, it should be represented as a BigTIFF file). Metadata that can predict the file size for converted files is very important for the migration service.

Table 3 shows sample raster image metadata for an ERDAS Imagine file (name, value, units triplets). This metadata was extracted using **gdalinfo**. This file is to be migrated to a GeoTIFF image (34014 pixels x 85075 lines x 3 bands/bytes). The estimated size of the output file will be larger than 4GB. In this situation, a warning message is displayed which includes the estimated number of files of excess size.

Name	Value	Units
x_resolution	34014	pixel
y_resolution	85075	pixel
num_bands	6	

Table 3: *General Raster Image Metadata*

Georeferencing metadata is also extracted from ERDAS Imagine files using the **gdalinfo** command. Table 4 lists the georeferencing metadata for a sample SRB data object that is stored in MCAT. This georeferencing metadata is not currently being used in PresSRB as preservation metadata. But it may be required in the future to prevent lossiness - if we convert to file formats which only supply limited georeferencing metadata.

Name	Value	Units
proj_coords	Transverse Mercator	
latitude_of_origin	0	Degrees
central_meridian	141	Degrees
scale_factor	0.99959999	
false_easting	500000	M
false_northing	10000000	M
pixel_size_x	25.0	m/pixel
pixel_size_y	-25.0	m/pixel
x_axis_rotation	0	Degrees
y_axis_rotation	0	Degrees
easting	178412.5	M
northing	8903562.5	M

Table 4: *Georeferencing Metadata*

Format, Software and Recommendation Registries

PresSRB re-uses the AONS I registry Java code with some minor modifications. This registry interface performs the role of retrieving:

- the format and software information from PRONOM registry; and
- format, software and recommended format information from the LCSDF registry.

The AONS I registry code is a web crawler which retrieves data from the PRONOM and LCSDF web sites via HTTP and transforms it into XML which is then inserted into the format registry, software registry and recommendation registry (MySQL tables) [8].

Because the BigTIFF and ERDAS Imagine formats are not represented in the PRONOM and LCSDF registries, and GeoTIFF is not represented in the LCSDF registry, MySQL tables were created to supplement the external registries for these formats. For example, the local format registry table is populated with data based on the values shown in Table 2.

Obsolescence Detection

Because PresSRB re-uses the AONS I format registry, software registry and recommendation registry, the four types of obsolescence warnings generated by AONS I are also generated within PresSRB. These are:

- Format has a new version
- Format not supported by any software
- Format is proprietary
- Format supported by obsolete software

Figure 3 illustrates an example report warning of proprietary format obsolescence.

PresSRB Obsolescence Report
SRB Collection: /A/home/srb.demo/2002/
Format is proprietary

PUID (PRONOM Unique Identifier)	Name	Total affected items	Number of unmigrated files	Recommendation
x-fmt/10002	Eradas Imagine	6	6	Migrate file(s) to non-proprietary GeoTIFF file(s)

Figure 3: Sample *PresSRB Obsolescence Report*

In the future, PresSRB will also generate the following warning:

• Format has no encapsulated georeferencing metadata

This situation would occur when a raster image file which has no embedded georeferencing metadata is co-located in an SRB collection with a separate file which contains the metadata.

Input Format	Output Format	Quality Ranking	GDAL Input Driver	GDAL Output Driver	Description
ERDAS .img	GeoTIFF	10	ERDAS	TIFF	If RGBA photometric interpretation option is selected for the converted files, the files may not be able to be opened or handled correctly by many software applications
ERDAS .img	JPEG2000	9	ERDAS	JP2KAK	This converter requires GDAL to be built with the Kakadu SDK which can be purchased from http://www.kakadusoftware.com/
ERDAS .img	JPEG2000	5	ERDAS	JPEG2000	This converter does not support generating output files that are > 2GB. Consider using the commercial Kakadu (JP2KAK) GDAL driver instead if JPEG2000 is required.

Table 5: *Migration table (PUIDs replaced with actual format name for clarity)*

Migration and Preview

The migration service converts raster image files from one format (ERDAS Imagine files) to another (GeoTIFF), preserving the resolution of the original image file, although not necessarily preserving the number of bands of the original.

The preview service provides a scaled down GeoTIFF version of the original geospatial raster files - the intention being to provide small preview files which are only a few MB in size as opposed to GBs in size

Unlike file systems which do not allow a folder and a file to have the same path name, SRB does allow data objects to have the same logical path name as SRB collections. So for example, given the following ERDAS Imagine SRB data object:

```
/home/srb.demo/sample.img
```

SRB allows us to generate the following migrated and preview data objects located in a sample.img data collection, while leaving the original data object intact:

```
/home/srb.demo/sample.img/geotiff_rgb.tif
/home/srb.demo/sample.img/preview_geotiff.tif
```

Currently the migration and preview services are limited to the input and output formats supported by GDAL [15]. Furthermore, they are limited by the drivers GDAL was built against and the features that this driver supports. For example, the JPEG2000 driver based on the free JasPer JPEG2000 library has a maximum file size support of 2GB, while the JP2KAK JPEG2000 driver which requires the commercial Kakadu library has an unlimited file size.

The migration service consists of two main components: a discovery component and a conversion provider component.

The intention of the migration discovery component is to present the collection manager with migration options for a specified file format. This is achieved by querying the migration MySQL table (similar to what is shown in Table 5.) which lists the various GDAL conversions and provides a description of the quality of the conversion. It contains a quality ranking field with a possible value between 0 and 10 which used to rank the available conversion providers presented to the collection manager.

The output formats supported by GDAL utility is listed using the `--formats` switch. For example the JP2KAK JPEG2000 conversion provider will be listed, but it won't be a selectable option. However information on where to purchase the Kakadu JPEG2000 library is presented.

If there is a reduction in the number of bands from the original to the migrated file (e.g., 6 bands in ERDAS Imagine and 3 bands in GeoTIFF), then this information is also presented to the collection manager.

If any of the generated GeoTIFF files are > 4GB, warnings are displayed that warn that some of the files will need to be converted to BigTIFF.

Scheduler

The PresSRB services were designed to be either executed directly from the command-line or via a thin PHP-based web wrapper. Because they can be invoked via the command line, the standard scheduling services available on unix-like operating systems can also be used.

For scheduled operations that are required to be executed periodically in some sort of recurring pattern, the **cron** scheduler is used. For operations which are required to be performed once at some time in the future, the **at** command is used. Once a scheduled job is completed, a email summary is sent to the collection manager.

Web GUI

The PresSRB GUI consists of a number of components which are described below.

Authentication. An authentication Web page is used to authenticate content managers Only the authenticated content managers have the right to generate the

obsolescence reports and execute migration services. Other users are only permitted to browse, query and view the collections and reports.

Browsing. A simple Web interface for browsing through a SRB collection is provided. When a content manager selects a collection or data object, they also have access to menu items that allow them to invoke and schedule the obsolescence and migration/preview services.

Querying. The standard SRB user-defined metadata queries can be performed via a simple web interface which returns a list of SRB Objects that match the query.

Ingestion. The PresSRB prototype provides a web form to enable ingestions of individual files into SRB space. File upload uses the multipart POST method. A PHP script receives the upload file and then ingests it into SRB space using the **Sput** Scommand. For PHP, the default maximum file size value is only 2MB. Thi would need to be increased for the handling of larger files. Although this can be set as high as 2GB, it's much more efficient to use a dedicated SRB client or the *Scommands* to ingest very large files. Most web-administrators would also discourage setting this value too high.

Obsolescence The obsolescence web page displays the last generated obsolescence report. Generating new reports is handled by the scheduler (described below). Figure 5 shows the scheduler page for the obsolescence detection scheduler. Users are able to specify the frequency of job execution via the GUI.

Migration and Preview. The dynamically generated Migration Web page is shown in Figure 4. For a given collection, this page displays the current file formats requiring migration (and the number of files) and provides a pull-down menu of migration services that can be applied. When converting to GeoTIFF format, users can specify the number of bands (RGB or RGBA). It also lists the recommendations from the registries (if available) and any issues associated with the migration service. Selecting the Schedule button, will schedule the underlying migration script.

Scheduler. The Scheduler enables users to schedule either obsolescence detection or migration scripts. A crontab configuration file specifies how to execute commands on a particular schedule and the addition, modification and removal of jobs. A PHP wrapper for the crontab command is used within the PresSRB Web GUI; so that users won't need to know the intricacies of the crontab file syntax. Figure 5 below illustrates the scheduler Web user interface.

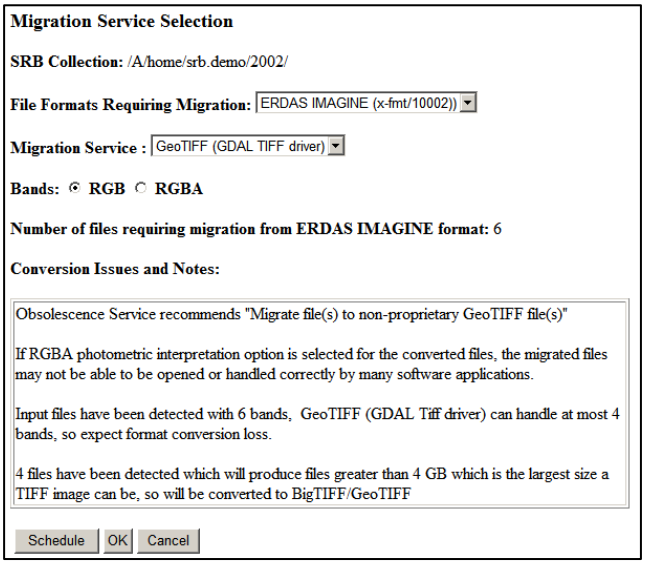

Figure 4: *Migration Service Selection GUI*

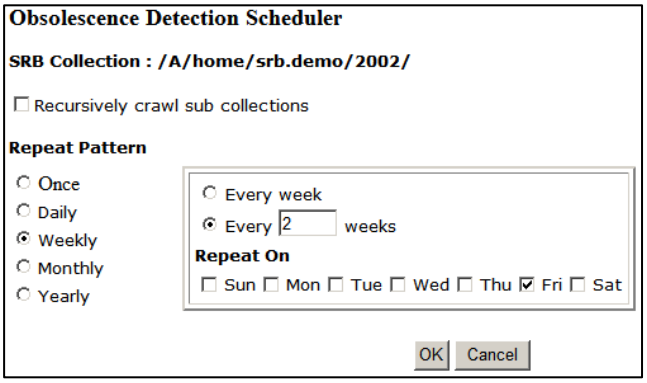

Figure 5: *PresSRB Scheduler GUI Interface*

Discussion and Conclusions

One of the greatest challenges with regard to scientific data is the lack of preservation services or support for many scientific data formats. There is a need for an initiative to establish a registry of recommended formats for scientific data within different disciplines. More specifically, there are currently no registries which provide recommendations for GIS based file formats. For example, neither PRONOM nor LCSDF support ERDAS Imagine files. Within PRONOM, GeoTIFF is just a placeholder and no real information is provided. In LCSDF, GeoTIFF isn't registered, even as a sub-format of TIFF.

Also particularly within the scientific domain, we are entering an era in which many files formats are reaching their file size limits (either 2 or 4 GB). In these cases, the current file format can no longer be used as a recommended file format for migration. Additionally, for many scientific datasets, the generic preservation metadata is inadequate and needs to include specialised metadata, e.g., for raster images, the resolution of the original file is

important. Georeferencing metadata is also significant but not currently supported in standardized preservation metadata schemas.

To conclude, SRB is an ideal infrastructure for dealing with large files, especially because it expedites the copying and replicating of large scale files using parallel data transfer. It provides an ideal infrastructure for preservation based on the LOCKSS [16] approach. However it is lacking in support for the preservation services required to ensure long-term access to many of the file formats being stored within SRB. Within the PresSRB prototype described in this paper, we have implemented and evaluated obsolescence detection and notification services and migration services for a particular environmental case study. We have demonstrated how this can be achieved by integrating external services using wrappers around the SRB Scommands and by adding native SRB support to existing utilities. Significant further work is required to evaluate and implement similar approaches for data archived within SRB by other scientific disciplines such as astronomy, biology and earth sciences.

References

1. Ball, A. and M. Patel, *Approaches to Information Curation in Engineering*, in *Knowledge and Information Management Through-Life*. June 2008: Institution of Mechanical Engineers, London.
2. Guttenbrunner, M., C. Becker, and A. Rauber, *Evaluating Strategies for the Preservation of Console Video Games*, in *iPRES2008*. 2008: London, UK.
3. Hunter, J. and S. Choudhury, *A Semi-Automated Digital Preservation System based on Semantic Web Services.* Joint Conference on Digital Libraries, JCDL, 2004: p. 269-278.
4. *JHOVE - JSTOR/Harvard Object Validation Environment*. [cited; Available from: http://hul.harvard.edu/jhove/.
5. The National Archives of the UK. *DROID Digital Record Object Identification*. [cited; Available from: http://droid.sourceforge.net/wiki/.
6. *Global Digital Format Registry*. [cited; Available from: http://www.gdfr.info/.
7. *PRONOM*. [cited; Available from: http://www.nationalarchives.gov.uk/pronom/.
8. Curtis, J., et al., *AONS - An Obsolescence Detection and Notification Service for Web Archives and Digital Repositories.* Special issue on Web Archiving for the New Review on Hypermedia and Multimedia (JNRHM), January 2007. **13**(1): p. 39-53.
9. *LCSDF (Library of Congress Sustainability of Digital Formats)*. [cited; Available from: http://www.digitalpreservation.gov/formats/.
10. Ruth, M. *GeoTIFF FAQ Version 2.3*. 2005 [cited; Available from: http://www.remotesensing.org/geotiff/faq.html.
11. *Erdas Imagine .ige (Large Raster Spill File) Format*. [cited; Available from: http://home.gdal.org/projects/imagine/ige_format.html.
12. Beaty, P. *What is wrong with my GeoTIFF?* 2008 [cited; Available from: http://field-guide.blogspot.com/2008/05/what-is-wrong-with-my-geotiff.html.
13. *GDAL - Geospatial Data Abstraction Library*. [cited; Available from: http://www.gdal.org/.
14. *Fine Free File Command*. [cited; Available from: http://www.darwinsys.com/file/.
15. *GDAL Raster Formats*. [cited; Available from: http://www.gdal.org/formats_list.html.
16. *LOCKSS (Lots of Copies Keep Stuff Safe)*. [cited; Available from: http://www.lockss.org/.