SINBAD DIGITAL LIBRARY PRESERVATION USING IRODS DATA GRID

Marco Pereira, Marco Fernandes, Joaquim Arnaldo Martinsa and Joaquim Sousa Pinto
IEETA - Instituto de Engenharia Electrónica e Telemática de Aveiro, Universidade de Aveiro, Portugal

Keywords: Digital preservation, Digital libraries, Data grids.

Abstract: Digital libraries are important instruments to make knowledge accessible to everyone. Maintaining the information stored within a digital library accessible to the public while ensuring that it remains protected from outside threats is one of the key concerns in digital libraries research. This paper describes possible ways to integrate SInBAD, the University of Aveiro digital library system with a grid based preservation system.

1 INTRODUCTION

Digital preservation is a challenge faced by all digital libraries. The content that they store must be kept available to users in a format suitable to their needs but it must also be safe from catastrophic events that would lead to loss of information. A single copy of a file available on a centralised server is always at risk of disappearing due to data corruption, natural disasters, malicious attacks or human error, yet in spite of this risk some digital libraries currently in place were not designed to avoid it. A solution to this problem can be the use of data grids as storage space (Moore et al., 2005b). This approach is adopted by GRITO[1] project. Data grids themselves lack any kind of specialised management policies, that are in fact the key to preserve content, since the decision of when to recover content that has been potentially damaged, or when to search a digital library for newly added content can be as vital as keeping backups. GRITO aims to solve this problem, providing a stable grid platform for digital preservation. By integrating SInBAD[2] the current digital library system used by University of Aveiro[3] with GRITO's grid we wish to expand SInBAD's digital preservation capabilities.

In this paper we describe different strategies that can be used to achieve integration between SInBAD and GRITO's grid. We begin by providing an overview of what is digital preservation, proceed to a brief explanation of some related work and of the projects we are trying to integrate. We then describe possible ways to create a link between the projects, expose our conclusions and future work perspectives.

2 DIGITAL PRESERVATION

Digital information is generated at an high rate everyday. People casually take digital photos and create documents on their computers without ever considering that the information that are creating is at risk. Due to the high pace of technological evolution information created just a few years ago is at risk of being rendered unusable by obsolescence of the formats and equipment in which it is stored (Kenney et al., 2003). Digital obsolescence is not the only problem associated with digital information: traditional threats like natural disasters or malicious attacks can also cause information loss.

Digital preservation can be seen as the necessary steps to ensure access to digital information over extended periods of time (Webb, 2003), while maintaining digital information's integrity and authenticity (Ferreira, 2006). A digital preservation system is a system that allows the preservation of digital information respecting the above requirements. A reference model for the development of digital preservation systems is the International Organization for Standardization (ISO) standard 14721:2003 (OAI, 2002), a standard that provides activities, concepts and terminology to be used when developing an archive system.

[1]http://grito.intraneia.com
[2]http://sinbad.ua.pt
[3]http://www.ua.pt

Pereira M., Fernandes M., Arnaldo Martins J. and Sousa Pinto J. (2009).
SINBAD DIGITAL LIBRARY PRESERVATION USING IRODS DATA GRID.
In *Proceedings of the 4th International Conference on Software and Data Technologies*, pages 107-112
DOI: 10.5220/0002257601070112
Copyright © SciTePress

3 RELATED WORK

There are a number of projects that can be viewed as relevant in the area of digital preservation.

Chronopolis[4] is a project that proposes a preservation architecture "based on an integration of digital library, data grid, and persistent archive technology" (Moore et al., 2005a). Chronopolis proposed model is interesting, but since it relies on Storage Resource Broker as a grid middleware we can't fully adopt it.

PANIC (Hunter and Choudhury, 2006) is a project that aims to take independent developed digital preservation projects and integrate their tools into a grid framework. Tools collected this way are exposed via web services and with the help of orchestration tools can be used in preservation efforts. The tools that this project expose can potentially be used by our own integration efforts in order to, for example, provide format migration tools.

An iRODS based preservation system (Hedges et al., 2009) was described in a paper published in the Future Generation Computer Systems journal. The paper describes the use of iRODS rules as support for digital preservation, an iRODS trait that we use in the GRITO project. This paper can be seen as a validation of the choice of iRODS as underlying grid technology.

4 SInBAD

SInBAD is the digital library and archiving system used by University of Aveiro. One of the key features of SInBAD's design is the emphasis placed on integration with existing subsystems. SInBAD uses a modular architecture (Almeida et al., 2006), (Fernandes et al., 2008) to provide access to heterogeneous content (theses, research papers, audiovisual material, posters,photos) to the members of the academic community. Different modules expose a coherent interface to the outside world via web services, and each module has its own web site. SInBAD also provides a portal that allows search in all modules. Content retrieval is also possible via web services, provided that one knows the correct identifier for the content.

The modular architecture allows content to be distributed throughout the university network. In spite of this apparent distribution, content is still located in the same geographical area, thus being in danger in case of, for example, natural disaster. Format obsolescence must also be taken into account: as digital container formats evolve, preserved content format should also evolve, so that it never becomes obsolete and thus rendered nearly useless (due to the lack of appropriate decoders). In order to cope with these perceived weaknesses the decision was made to integrate SInBAD with a long term digital preservation national project, GRITO.

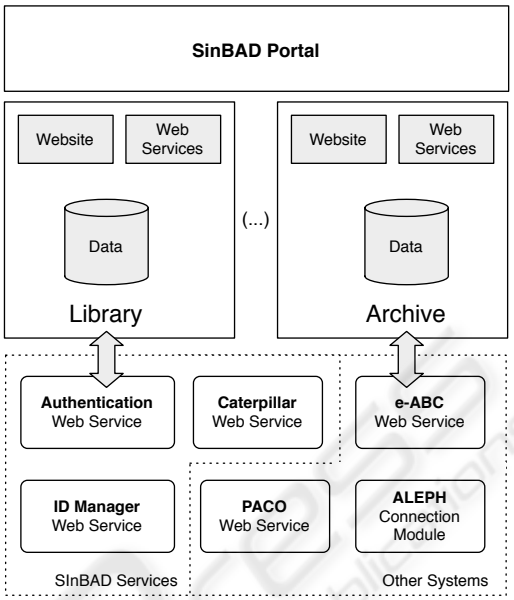

Figure 1: Generic SInBAD Architecture.

4.1 OAI-PMH

SInBAD implements the Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) (de Sompel and Lagoze, 2008). OAI-PMH is a protocol designed to provide an application-independent metadata harvest infrastructure.

OAI-PMH requests are represented by HTTP requests, and can take the form of POST or GET methods. In SInBAD's case we will be using the GET method. In this method a request is comprised of a base URL, followed by a question mark (?), followed by a list of arguments in the form of `key=value`. Multiple arguments are separated by an ampersand (&). Every request is required to have one verb as argument, in order to specify the action that should be executed. In SInBAD's case the base URL is http://sinbad.ua.pt/OAI/ (a test form is available in the same address).

When a request is issued to a OAI-PMH compliant repository a response will be generated in the form of a XML encoded UTF-8 byte stream. The resulting XML can then be parsed and used as an application deems necessary.

One possible use for the OAI-PMH is the retrieval

[4]http://chronopolis.sdsc.edu/

of identifiers, that can be used to access content using several methods exposed in SInBAD's web services.

5 GRITO

GRITO is a Portuguese digital preservation project, funded by FCT (GRID/GRI/81872/2006). It aims to use data grids as a support for digital preservation. In order to avoid investment in expensive resources GRITO plans to leverage existing resources in research grids (by extending them to allow their use for digital preservation) while also creating a grid cluster exclusively dedicated to digital preservation. The final goal of the project will be to integrate both types, creating a low cost digital preservation grid that can be used by public institutions.

The grid middleware that will be used is iRODS[5], developed by the San Diego Supercomputer Center (SDSC). iRODS is a grid middleware that provides management policies in the form of rules. In iRODS management policies can be changed in run time (Rajasekar et al., 2006), and its functionalities can be tailored to specific needs. New rules can be created by combining and creating microservices (small procedures written in C programing language). It can also function in a federation of grid clusters and present collections contained in them as one, a trait that will help GRITO integrate grids under the control of different institutions on a single preservation grid.

iRODS offers several means to interact with the grid. Out of the box provides a set of command line tools (itools) inspired by known UNIX commands (like icp to copy files, or irm to remove files). Interaction with the grid can also be achieved by using a PHP client API (Prods), a Java based API (Jargon[6]) and on Linux systems collections can be mounted as a regular folder with the use of a FUSE module.

This middleware was not specially designed to handle some of the specific requirements that digital preservation scenarios impose, but since it is open source it could be extended (Barateiro et al., 2008) to accommodate such requirements.

6 INTEGRATING SINBAD AND IRODS

Since SInBAD is a deployed production system it is a requirement of the integration that as little changes as possible should be made to it. This excludes any possible integration scheme were SInBAD would be changed, requiring the creation of a transparent digital preservation system.

In order to place content from SInBAD into the preservation grid it is possible to adopt a strategy where no changes are made to SInBAD itself, since SInBAD already exposes all the needed interfaces as web services. In the future SInBAD must be able to recover content from the grid. To achieve that goal some modifications will have to be made, yet these changes should be kept only to the essential for content recovery.

The use of the OAI-PMH allows us to obtain information about the content that was added to each collection since a specific date (along with information on how to retrieve it). We can then use the information retrieved with OAI-PMH on how to access content to use SInBAD's existing web services. These web services allow us to retrieve the content stored in the digital library (we must also retrieve any remaining associated metadata) and place it into the grid, thus achieving integration without disturbing the existing SInBAD infrastructure.

In this integration scheme iRODS with an extended service set (Barateiro et al., 2008) will be responsible for the preservation of any content harvested from SInBAD. To harvest content from SInBAD two approaches can be taken:

- Creation of an iRODS microservice.
- Creation of an Intermediary system.

The merits and drawbacks of each approach will be discussed in the following sections.

6.1 iRODS Microservice

Creating an iRODS microservice (illustrated by Figure 2) is an option to achieve integration between iRODS and SINBAD. iRODS already provides the necessary tools to enable a rule to be executed periodically, so periodic harvest of content can be considered a trivial task.

The integration process would become a two microservices rule:

1. A microservice would contact SInBAD using the OAI-PMH protocol, generating a list of targets to be retrieved.
2. A microservice would take as input the list of targets and use SInBAD's web services to retrieve and store content into the grid.

Since iRODS microservices must be written in C we consider that ideal libraries to deal with these steps

[5]https://www.irods.org/index.php/

[6]http://www.sdsc.edu/srb/jargon/

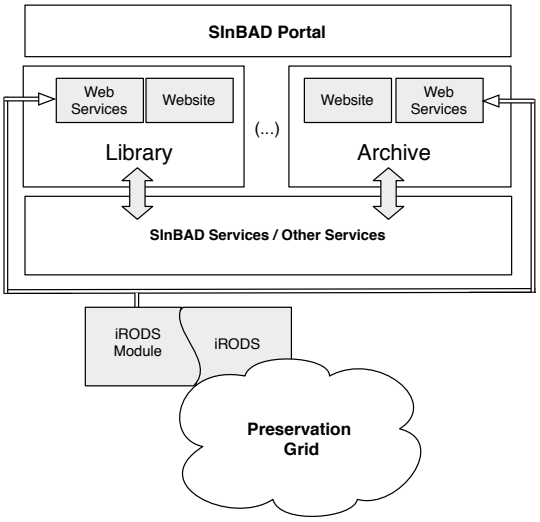

Figure 2: SInBAD/iRODS integration using a microservice.

are libcurl[7], expat[8] and gsoap[9]. All of these libraries allow us to develop the needed microservices with minimal effort, by providing multiple file transfer protocols, XML parsing and webservices support respectively.

The advantage of this approach would be the rational use of resources. The iRODS server (that is assumed to be always accessible) is an underused resource on a digital preservation scenario. By using a rule to periodically harvest SInBAD's content we could use some otherwise unused processing cycles, releasing SInBAD (a system that is accessible to the public) of that task, and since content would be harvested directly from SInBAD to the preservation grid, network traffic would be reduced.

The drawback of the approach is the tight integration with iRODS. By creating an iRODS module we are creating specific code that would be hard to reuse and maintain in case of major platform change (extensive modifications in SInBAD or grid middleware change). This tight integration would mean that in the future content recovery would be made by executing a new set of rules in the iRODS engine, that would themselves invoke the necessary services in SInBAD.

6.2 Intermediary System

Creating an intermediary system to serve as proxy between SInBAD and iRODS is another option. This approach (illustrated by Figure 3) follows the current subsystems driven SInBAD philosophy, and would interact with SInBAD as an back-office administration tool. It would be a modular system, allowing the use of other preservation technologies like, for example the use of a peer-to-peer based preservation system in addition to the iRODS based system.

This intermediary system would follow a three step approach:

1. Contact SInBAD using OAI-PMH and discover all the content added since the last contact.
2. Fetching all the new content and corresponding metadata.
3. Submit content/metadata pairs to the iRODS grid.

Each of these steps can be seen as a module, the first step as a module that provides access to the OAI-PMH protocol, the second step a module that provides access to SInBAD's digital contents and the third step as a module that allows the use of the iRODS infrastructure. Although the majority of SInBAD's subsystems are written in .net framework, all the resources needed to integrate SInBAD with GRITO can be accessed in a platform neutral way due to the use of web services. This allows us to write this intermediary system in Java in order to use the provided Jargon API, therefore avoiding the use of the command line tools. The Jargon API provides the necessary methods to place content in the grid, to associate metadata to content and allows the creation of queries (based on the available metadata) that can be used to selectively retrieve content from the grid.

To avoid duplication of management policies a rule could be created in iRODS that when executed would trigger the preservation process, by invoking a web service of the intermediary system. This approach assumes that all management policies (content harvest, format migration) will be placed on the iRODS middleware.

The advantage of this approach would be flexibility. In the event of a grid platform migration only the submission mechanism would need to be changed, and in the event of a major SInBAD change no microservices would have to be changed on the iRODS side. This flexibility extends into the future, since this system can be seen as an administration tool, it can be extended to manage content recovery (detailed in section 8).

The drawback of this approach would be resource usage. The creation of a proxy system would generate more network traffic than the iRODS module, since content would be retrieved to this intermediary system and only after submitted to the grid (assuming that this application is running on a different machine than the iRODS rule server). In a digital library scenario,

[7]http://curl.haxx.se/libcurl/
[8]http://expat.sourceforge.net/
[9]http://www.cs.fsu.edu/ engelen/soap.html

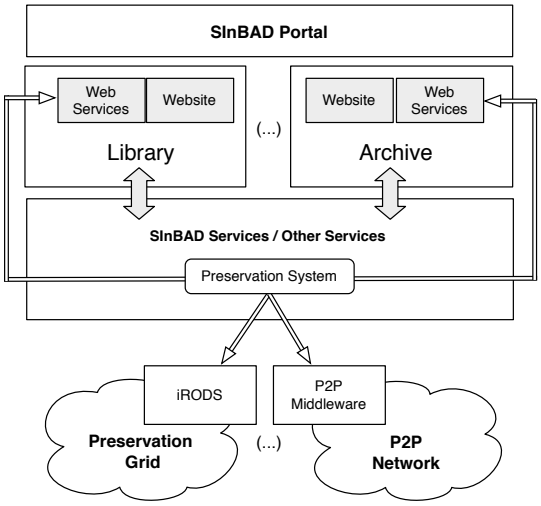

Figure 3: SInBAD/iRODS integration using an intermediary system.

this drawback can be made less relevant by scheduling content harvest to periods of low network activity.

7 CONCLUSIONS

Integration between SInBAD and iRODS will be an interesting addition to the SInBAD architecture Table 1 presents a comparison between integration options. Since we wish to promote flexibility in the adopted solution, the option of creating an intermediary system is the best way to integrate SInBAD with IRODS. In order to retrieve content from the grid both alternatives require the same changes in SInBAD, but the intermediary system follows more closely the current SInBAD's philosophy. We consider that the possibility of integration with different preservation mechanisms (such as peer-to-peer networks, or content conversion services) that the intermediary system provides is the key factor in this decision.

8 FUTURE WORK

After this analysis, active development of a working prototype of the intermediary system has begun. Placing content into the preservation grid is only a step for a complete preservation scenario. For the complete scenario we must be able to recover content that was placed in the grid, and this will lead to the creation of a new recovery web service within SInBAD, because the existing web services are not recovery friendly.

The current behaviour of SInBAD is to create a new identifier each time content is added to a repository. This behaviour does not allow the use of the existing methods to recover content using the preservation system. New methods that allow authorised users to repair compromised content with the help of the content stored in the preservation system must be created. These methods can be added to existing web services, or grouped in a new web service. As was mentioned before SInBAD is a production system, so the preferred approach will be to group all the needed methods in a new web service in order not to disturb the rest of the system.

Table 1: Comparison between iRODS microservice and an Intermediary system.

	iRODS Microservice	Intermediary system
Programming language	C	multiple (C/Java)
Resource usage	low	moderate
Integration with SInBAD	tight	loose
Integration with digital preservation system	tight	loose
Ability to deal with changes in SInBAD	moderate	moderate
Ability to deal with changes in management policies	high	high
Ability to deal with changes of digital preservation system	none	high
Perceived maintenance effort	moderate	low

ACKNOWLEDGEMENTS

This work was funded in part by the Portuguese Foundation for Science and Technology grants SFRH/BD/23976/2005 and GRID/GRI/81872/2006.

REFERENCES

(2002). *Reference Model for an Open Archival Information System (OAIS), Blue Book, Issue 1*. CCSDS - Consultative Committee for Space Data Systems. http://public.ccsds.org/publications/archive/650x0b1.pdf.

Almeida, P., Fernandes, M., Alho, M., Martins, J. A., and Pinto, J. S. (2006). Sinbad - a digital library to aggregate multimedia documents. In *AICT-ICIW '06: Proceedings of the Advanced Int'l Conference on Telecommunications and Int'l Conference on Internet and Web Applications and Services*, page 173, Washington, DC, USA. IEEE Computer Society.

Barateiro, J., Antunes, G., Cabral, M., Borbinha, J. L., and Rodrigues, R. (2008). Using a grid for digital preservation. In *ICADL*, pages 225–235.

de Sompel, H. V. and Lagoze, C. (2008). The open archives initiative protocol for metadata harvesting. http://www.openarchives.org/OAI/ openarchivesprotocol.html.

Fernandes, M., Almeida, P., Martins, J., and Pinto, J. (2008). A digital library framework for university of aveiro. In Borwein, J., Rocha, E., and Rodrigues, J., editors, *Communicating Mathematics in the Digital Era*, pages 111–123. A K Peters.

Ferreira, M. (2006). *Introdução à preservação digital – Conceitos, estratégias e actuais consensos.* Universidade do Minho, Escola de Engenharia.

Hedges, M., Blanke, T., and Hasan, A. (2009). Rule-based curation and preservation of data: A data grid approach using irods. *Future Generation Computer Systems*, 25(4):446 – 452.

Hunter, J. and Choudhury, S. (2006). Panic: an integrated approach to the preservation of composite digital objects using semantic web services. *Int. J. Digit. Libr.*, 6(2):174–183.

Kenney, A. R., McGovern, N. Y., Entlich, R., Kehoe, W. R., and Olsen, E. (2003). Digital preservation management. http://www.library.cornell.edu/iris/tutorial/dpm/.

Moore, R., Berman, F., Middleton, D., Schottlaender, B., JaJa, J., and Rajasekar, A. (2005a). Chronopolis - federated digital preservation across time and space. pages 171–176.

Moore, R., Rajasekar, A., and Wan, M. (2005b). Data grids, digital libraries, and persistent archives: An integrated approach to sharing, publishing, and archiving data. *Proceedings of the IEEE*, 93(3):578–588.

Rajasekar, A., Wan, M., Moore, R., and Schroeder, W. (2006). A prototype rule-based distributed data management system. In *HPDC workshop on "Next Generation Distributed Data Management"*. https://www.irods.org/index.php/Publications.

Webb, C. (2003). *Guidelines for the Preservation of Digital Heritage*. United Nations Educational Scientific and Cultural Organization - Information Society Division. http://unesdoc.unesco.org/images/0013/001300/130071e.pdf.