Middleware Applied to Digital Preservation: A Literature Review

Eriko Brito, Paulo Cesar Abrantes, Bruno de Freitas Barros
Recife Center for Advanced Studies and Systems (CESAR)
Recife – PE, Brazil
E-mails: eriko.brito@outlook.com, {pc.abrantes, barrosbruno}@gmail.com

Abstract — **Maintaining digital collections available for humanity use at long term is a big challenge for the areas of digital preservation, management policies of curator centers and technologies for data reproducibility. This paper performs a literature review to investigate middleware options for digital preservation, listing its main features and applications. Seven solutions were found and it was concluded that the cataloged technological bases are mature enough, which indicates an optimistic future for the digital curation area.**

Keywords—Digital Curation; Digital Preservation; Reproducibility; Middleware.

I. INTRODUCTION

Research and solutions in the digital preservation area have evolved significantly in recent decades establishing their technological, methodological and political apparatus. Brito et al. [1] point out that compared to the physical collections preservation, digital content brings an association, almost paradoxically, between a great potential risk and a great potential for protection. The potential risk is represented by the ephemerality of digital storage that can be irretrievably lost because of technical or human failure much more easily and quickly than in the case of physical representations of content. The potential for protection, in turn, is anchored in the fact that digital collections can be endlessly reproduced and stored with full fidelity and integrity.

The continuity of digital collections depends, mostly, on implementing strategies that take full advantage of the potential for protection, attempting to neutralize its inherent potential risk. However, the challenge can represent much more of a social and institutional problem than a purely technical issue, because, particularly concerning digital preservation, it depends on institutions that undergo changes of direction, mission, administration and funding sources, as Arellano defends [2].

Concomitantly, in the information technology area, distributed systems are established as an information-sharing pattern. That leads us to cloud computing that, according to Mell et al. [3], is a ubiquitous, convenient and on demand model for sharing computing resources that can be managed and made available with minimal effort.

Applying the distributed processing principles and cloud computing to the maturity scenario of digital preservation seems to be an only natural option. Wittek et al. [4] relates distributed systems to distributed digital preservation, pointing out that it can ensure the replication of digital artifact copies between geographically separated servers. It is important to say that distributed digital preservation is not only the act of ensuring the backup of digital artifacts, but also the possibility to access the data over the years and to reuse it.

Among the distributed systems and the digital preservation, there is the middleware, which, for Rocha et al. [5], is the group of components located between the operating system and the application, promoting generic services to support the execution of distributed applications. For this Literature Review (LR), the presented concept of middleware is extended to a layer situated between the business rules of the curator center and the supporting IT infrastructure of digital preservation.

The remainder of the paper is structured as follows. Section 2 presents the review planning with its goals. In Section 3, an overview of the middleware options found is presented. Section 4 will show a detailed analysis of each option and, finally, Section 5 presents conclusion and register for future work.

II. REVIEW PLANNING

This review follows the guidance of Kitchenham et al. [6] in its structure. It contains the objectives of the review, the research questions to be presented, the criteria used for the negative scope of the research, the strategy that will be used and the way it will be conducted. It is expected that in this way, other researchers can repeat the procedure according to their own definitions.

A. Objectives

The goal of this LR is to identify existing options in literature of middleware solutions used in digital preservation processes of curation centers. The result of this work can provide the discovery of challenges and trends in this research field.

B. Research Question

The research question that guides this LR is: what middleware options for digital preservation are currently available in the literature?

C. Exclusion Criteria

For the established research question, it was decided that some productions will be excluded from the scope of this LR. Specifically, scientific productions that:

- Are in proposal stage;
- Present the state of art for the research question;
- Were published before the year of 2010.

 ISBN: 978-1-61208-438-1

D. Research Strategy

The research strategy consists in establishing the premises that will be considered by the LR to achieve its goals. Thus, in this section, will be elicited the sources of the research project, base language and keywords to be used in the search engines listed at the sources. These assumptions are defined as follows:

Sources: IEEE Xplorer, Science Direct, ACM Digital Library, Compendex and Scopus.
Language: the English language will be used as reference for the LR, as it is considered the most popular in the scientific world.
Keywords: Middleware to digital preservation, Distributed systems for digital preservation, Electronic Records Archives capabilities, reproducibility.

E. Review Conduction

This paper was planned and produced in May 2015 in response to the approval requirements of the discipline of Systems Interoperability of the Professional Master's degree in Software Engineering at CESAR.EDU. The sources were found by using search strings formed by logical combinations of keywords presented as follows:

("digital *preservation"* OR "digital *curation"* OR *reproducibility* OR *cloud* OR "*distributed systems*") AND *middleware*

The collection of articles was obtained by reading the abstracts, guided by the research question and exclusion criteria presented. As a result of this approach, there were seven relevant papers to the theme of this LR, which are presented in Section 3.

In summary, the results obtained from the data sources were:

a) IEEE Xplorer returned 37 works, out of which 3 were considered aligned to the research question.
b) ACM Digital Library returned 47 results, some of them also found in IEEE Xplorer, and 4 of them were more relevant to this LR.
c) In Science Direct 20 results were located.
d) In Scopus, 7 and,
e) In Compendex, 5 results.

Results **c**, **d** and **e** were classified as outside of the LR scope either because they match the exclusion criteria or because the abstracts were not aligned with the research question.

III. Overview of Selected Options

This section is dedicated to present the summarized results that are the most relevant to this LR. The following seven subsections are identified by the titles of the papers. They describe relevant aspects of each result and its purpose.

A. Digital Preservation in Grids and Clouds: A Middleware Approach

Digital preservation can be seen as an effort to retain, as long as necessary, digital material for future use on research, consultations or any other form of knowledge management.

Several of the current digital preservation systems are backed in the computational grid technology, but the advent of cloud computing and its potential has become a strong and attractive possibility [4].

Placed in the business layer of the SHAMAN model, the proposal made by Peter Wittek and Sandor Daranyi suggests a middleware which is flexible enough to enable a quick and transparent switching between cloud computing and grid computing, in compliance with business rules and requirements of the entity that needs to preserve its digital collection [7]. Figure 1 shows the proposed architecture that includes, on the left, an archive layer governed by a set of pre-established policies and, on the right, computational scalability components in clouds or grids.

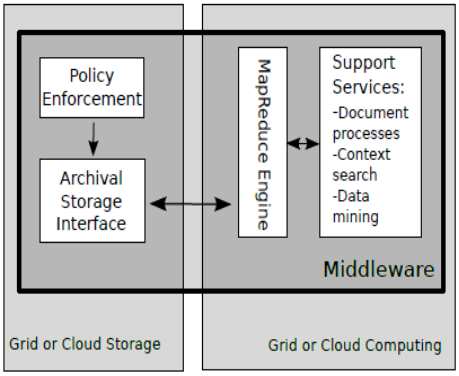

Figure 1. Overview of the Peter and Sándor's proposal

The considerations of the authors suggest that small businesses can be the biggest beneficiaries of the switching flexibility, by replacing servers or grid acquisitions with service level agreements with computing service providers in the clouds.

B. Content server system architecture for providing differentiated levels of service in a Digital Preservation Cloud

Quyen L. Nguyen and Alla Lake [8] write about storage challenges and resulting preservation of the rapidly growing volume of digital records and the need for some companies and industries to stick to directives such as Sarbanes-Oxley Act. It is important to note that digital preservation covers the need of keeping information available and accessible regardless of the hardware and features that originated it.

In this context, and to the authors, preservation in the clouds is a simple and economical option for models such as the Open Archive Information System (OAIS), as seen in Figure 2, which requires great engineering effort and planning.

 ISBN: 978-1-61208-438-1

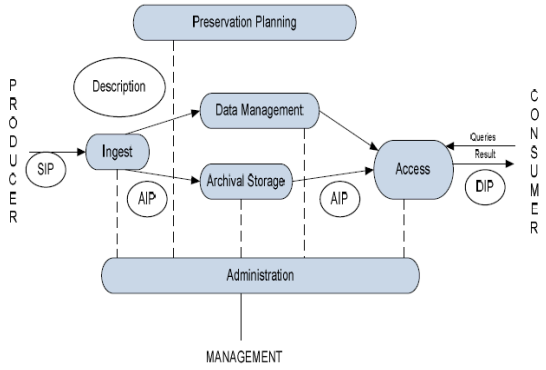

Figure 2. The OAIS Model

In the LDPaaS proposal, the Ingest layers, Preservation, Archival Storage and Access are abstracted and offered as individual services in the clouds. With this arrangement the proposal provides differentiated service levels for the various needs of long term digital preservation.

C. *Biopolis, long term preservation of digital user content*

The purpose of the Biopolis is that the digital contents of its users can be replicated in the clouds, commercially or not. With regard to copyright, the design ensures the ultimate relationship between the author and the maintainer of the digital material.

Differently from systems like Google Panoramio and Flirk, in which, as Sardis et al. [10] state, time stamps are not present, the Biopolis project supports time attribute and is prepared to offer, in the coming years, scalable storage, preservation and organization through libraries and semantic searches if needed, stamping data with its registration time.

Through the Internet, the users can add their digital content via the web interface of the Biopolis system, by logging the geographic position of upload and copyrights for appropriate action. After the content for retention time setting is filtered, storage, procurement, distribution, preservation, recovery and reuse options are provided.

In terms of middleware, the Biopolis project has levels with its own API functions, which are used as clients in a common runtime environment providing scalability, high availability and routing messages. Figure 3 shows the middleware components.

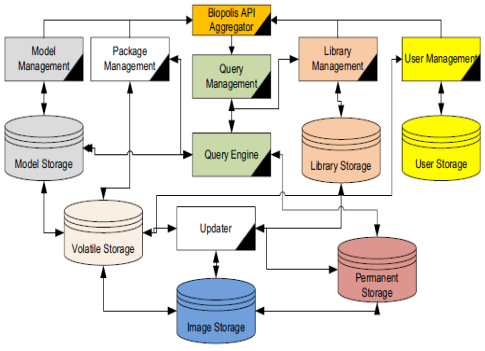

Figure 3. Biopolis Middleware

D. *PDS cloud: Long term digital preservation in the cloud*

The Preservation DataStores (PDS) proposal is a preservation cloud based on the OAIS model, developed by Ccsds [9] as an infrastructure component of the European Union ENSURE. It employs multiple heterogeneous providers and embodies the concept of object-informational preservation.

The authors conducted a gap analysis concluding that "just throwing" the data in a cloud is not the solution for preservation repositories. Instead, a more professional approach is expected. The main features of PDS Cloud includes: a) multiple clouds storage support, b) enhancement of future understandability of content by supporting data access using cloud based virtual appliances and, c) advanced services based on the OAIS model.

As shown in Figure 4, the PDS Cloud architecture is implemented as a middleware, composed by a broker that OAIS interconnects between multiple entities and the cloud. On the front-end, PDS Cloud exposes to the client a set of OAIS-based preservation services such as ingest, access, delete and preservation actions on OAIS Archival Information Packages (AIPs).

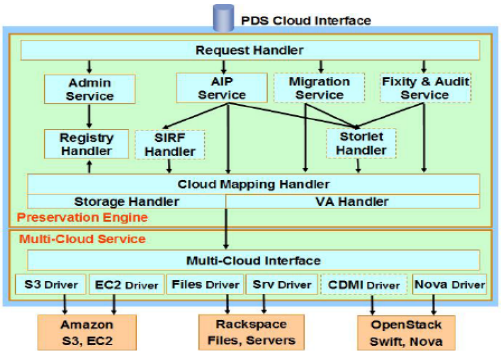

Figure 4. PDS Cloud Architecture

On the back-end, it leverages heterogeneous storage and computes cloud platforms from different vendors. AIPs may be stored on multiple clouds simultaneously to exploit different storage cloud capabilities and pricing structures, and to increase data survivability.

In the conclusions, Rabinovici-Cohen et al. [11] point out that the main purpose of PDS Cloud is to keep the responsiveness of long term digital material by adhering to the changes of technological scenarios.

E. *Rule-based curation and preservation of data: A data grid approach using iRODS*

In this paper, Hedges et al. [12] present the implementation of a data management layer to support a system of preservation research data. For data storage, the authors suggest the use of the e-Science technology and grid computing middleware, presenting how integrated Rule-Oriented Data Management System (iRODS) can be used to implement complex strategies of digital preservation.

The contextualization involves the consideration that the Storage Resource Broker (SRB) developed by the San Diego

Supercomputer Center (SDSC) is the most widely used data management middleware for digital preservation. The iRODS, open source, is presented as its successor, also developed by SDSC, with significant evolution especially in the political representation of capacity in terms of rules.

This feature of the iRODS middleware allowed the authors to explore two key points:

- preservation actions taken when a digital resource is ingested into an iRODS data grid;
- post-ingest management of the integrity and authenticity of curated digital resources.

In its conclusion, the paper exalted the iRODS skills, such as the flexibility to implement the rules in a sequence of actions to be executed in particular contexts or when certain events like the ingest of the file into the grid, or a timer occur.

F. New Roles for New Times: Digital Curation for Preservation

The work of the authors was to examine tools and techniques used to automate the exchange of significant data volumes between MetaArchieve Cooperative, which uses "Lots of Copies Keep Stuff Safe" (LOCKSS) and the Chronopolis preservation system, which uses the Storage Resource Broker (SRB). It is expected that this work enable the use of preservation systems to share data between these two preservation networks in the United States of America [13].

Staff from the MetaArchive and Chronopolis are investigating technologies that allow data exchange between LOCKSS and iRODS, based on real-world, practical implementations within MetaArchive, Chronopolis and CDL. Three methods of exchange are being developed and evaluated.

The first method uses BagIt and related technologies to package and transfer large collections efficiently and reliably. A second more, sophisticated tool, allows the user to identify an existing BagIt bag, transfer its contents, ingest these objects into a storage zone and perform quality assurance testing for the whole process. The third proposed method utilizes a LOCKSS plugin. LOCKSS has its own technical architecture that can be enhanced by the use of custom-created, XML-based plugins, which allows data to be manipulated according to defined rules. They will create plugins that will allow the LOCKSS system to interact with an iRODS system.

In its conclusion, the paper emphasizes the importance of the project that will integrate two major digital data preservation platforms and, by doing so, improve multi-disciplinary and multi-institutional scientific exploration that is highly data-driven. An integrated LOCKSS/iRODS infrastructure will better support the growing diversity in formats, visualization, and analytical tools that empower researchers to utilize information and data more effectively.

G. Semantic Middleware for E-science Knowledge Spaces

Futrelle et al. [14] present in his paper a middleware called Tupelo, which implements Knowledge Spaces. It enables scientists to find, use, relate and discuss data and metadata work in a distributed environment. Its construction is based on a combination of semantic web technology for data management and workflow. The main benefit of Tupelo middleware is the simplification of interoperability by providing the Knowledge Space view of heterogeneous resources distributed in institutional repositories.

In architecture terms, Tupelo is based on an abstraction called "context", which represents a kind of semantic view of distributed resources. Context implementations are responsible for performing as "operators", which are atomic descriptions of requests to either retrieve or modify the contents of a context. Two primary kinds of operations are provided:

1. Metadata operations, including asserting and retracting statements (i.e., RDF statements) and searching for statements that match a query; and
2. Data operations, including reading, writing, and deleting binary large objects (BLOB's), each of which is identified with a URI.

In its closing remarks, the paper suggests that Tupelo's interoperability-based architecture allows it to be used to connect, without replacing or displacing, existing software stacks to add context and help integrate the heterogeneous aspects of large-scale scientific work, including observation, analysis, organization, and publication.

Another significant consideration was related to reducing the development effort required to support scientific domains, allowing an active view of scientific work with strong guarantee of reusability, based on explicit semantics and declarative descriptions of analytic processes, opening new opportunities for more effectively disseminating and preserving the fruits of ongoing, evolving scientific discovery.

IV. CONCLUSION

This paper presented a literature review of middleware available for the area of digital preservation. The main objective was to list options available for this context and to present their main characteristics and applicability.

It was possible to observe the complexity of the long-term digital preservation process, and that this is still an evolving model. It was contextualized the term middleware as a layer placed between the business rules of the curator center and the supporting IT infrastructure of digital preservation.

Thus, it was possible to map the following results as the most cited middleware related to digital curation and preservation:

a) The Storage Resource Broker (SRB), developed by the San Diego Supercomputer Center (SDSC);

 ISBN: 978-1-61208-438-1

b) iRODS, open-source, presented as its successor and also developed by SDSC, and the;
c) LOCKSS Program as an open-source, library-led digital preservation system built on the principle that "lots of copies keep stuff safe".

Table 1 shows an overview of the solutions found, being organized by the name of the paper, the model used as the base of the solution, the use of grids or clouds and its main characteristics.

TABLE 1. SOLUTIONS OVERVIEW

Paper Title	Base model	Grid	Cloud	Main Characteristic
Digital Preservation in Grids and Clouds: A Middleware Approach	SHAMAN	Yes	Yes	Allows a transparent switching between cloud computing and grid computing
Content server system architecture for providing differentiated levels of service in a Digital Preservation Cloud	OAIS	No	Yes	Abstracts the layers and offers them as individual services in the clouds.
Biopolis, long term preservation of digital user content	None	No	Yes	Provides scalability, high availability and routing messages.
PDS cloud: Long term digital preservation in the cloud	OAIS	No	Yes	Supports multiple clouds storage and provides data access using cloud based virtual appliances
Rule-based curation and preservation of data: A data grid approach using iRODS	SRB	Yes	No	Implements a rule-oriented data management layer to support a system of preservation research data
New Roles for New Times: Digital Curation for Preservation	LOCKSS and iRODS	Yes	Yes	Allows data exchange between LOCKSS system and iRODS
Semantic Middleware for E-science Knowledge Spaces	None	Yes	No	Simplifies the interoperability by providing the Knowledge Space view of heterogeneous resources distributed in institutional repositories

Although Brito et al. [1] claim that the digital preservation area is still in the early stages of its formation and that the technological, methodological and political apparatus to preserve digital information is still being built, it was found mature middleware options related to digital preservation and access to long term information.

After this research, it was realized that information security applied to digital curation is an area that can be explored, so as future work is suggested a review of the literature to list the existing solutions.

ACKNOWLEDGMENT

To the CESAR.EDU teaching staff that contributed with methodological guidance for the development of this paper. We also appreciate the patience and dedication of our families that unconditionally supported the work that has been done so far.

REFERENCES

[1] E. Brito, R. Costa, A. Duarte, P. De Pós-graduação, and I. Ppgi, "The adoption of model canvas in data management plans for digital curation in research projects," 2012.

[2] M. Arellano, "Digital Preservation Criteria of Scientific Information," Brazil: University of Brasilia, 2008, p. 50.

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing Recommendations of the National Institute of Standards and Technology," Natl. Inst. Stand. Technol. Inf. Technol. Lab., vol. 145, 2011, p. 7.

[4] P. Wittek and S. Darányi, "Digital Preservation in Grids and Clouds: A Middleware Approach," J. Grid Comput., vol. 10, no. 1, 2012, pp. 133–149.

[5] V. H. Rocha, F. S. Ferraz, H. N. De Souza, and C. A. G. Ferraz, "ME-DiTV : A middleware extension for digital TV," International Conference on Software Engineering Advances, 2012, pp. 673-677.

[6] B. Kitchenham, et al., "Systematic literature reviews in software engineering – A tertiary study," Inf. Softw. Technol., vol. 52, no. 8, Aug. 2010, pp. 792–805.

[7] P. Innocenti, et al., "Assessing digital preservation frameworks: the approach of the SHAMAN project,", 2009, pp. 412–416.

[8] Q. L. Nguyen and A. Lake, "Content server system architecture for providing differentiated levels of service in a Digital Preservation Cloud," Proc. IEEE 4th Int. Conf. Cloud Computing, 2011, pp. 557–564.

[9] Ccsds, "Reference Model for an Open Archival Information System (OAIS)," Forsp. Data Syst., no. January, 2002, pp. 1–148.

[10] E. Sardis, A. Doulamis, V. Anagnostopoulos, and T. Varvarigou, "Biopolis, long term preservation of digital user content," IEEE 10th Int. Conf. on e-Business Engineering, Sept. 2013, pp. 478-483.

[11] S. Rabinovici-Cohen, J. Marberg, K. Nagin, and D. Pease, "PDS cloud: Long term digital preservation in the cloud," Proc. IEEE Int. Conf. Cloud Eng. IC2E, 2013, pp. 38–45.

[12] M. Hedges, A. Hasan, and T. Blanke, "Management and preservation of research data with iRODS," in Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience - CIMS '07, 2007, p. 17.

[13] T. Walters and K. Skinner, "New Roles for New Times: Digital Curation for Preservation," Washington: Association of Research Libraries, 2011.

[14] J. Futrelle, et. al., "Semantic middleware for e-Science knowledge spaces," Concurr. Comput. Pract. Exp., vol. 23, no. 17, 2011, pp.2107-2117

 ISBN: 978-1-61208-438-1