ARIADNE
Issue 48 July 2006

Main Articles

A Foundation for Automatic Digital Preservation

Miguel Ferreira, **Ana Alice Baptista** and **José Carlos Ramalho** propose a Service-Oriented Architecture to help cultural heritage institutions to accomplish automatic digital preservation.

Introduction

Efforts to archive a large amount of digital material are being developed by many cultural heritage institutions. We have evidence of this in the numerous initiatives aiming to harvest the Web [1-5] together with the impressive burgeoning of institutional repositories [6]. However, getting the material inside the archive is just the beginning for any initiative concerned with the long-term preservation of digital materials.

Digital preservation can best be described as the activity or set of activities that enable digital information to be intelligible for long periods of time. In general, digital information kept in an archival environment is expected to be readable and interpretable for periods of time much longer than the expected lifetime of the individual hardware and software components that comprise the repository system, as well as the formats in which the items of information are encoded [7][8].

Over the past decade, a vast number of preservation strategies have emerged from the various preservation projects developed, literally, all over the world [9-12]. Nonetheless, the most cited and applied preservation strategy continues to be migration [13][14], especially in contexts where non-interactive digital objects, such as images, databases or text documents, are the focus of preservation.

Migration can be described as a '(...) set of organized tasks designed to achieve the periodic transfer of digital materials from one hardware/software configuration to another or from one generation of computer technology to a subsequent generation.' [15].

The major drawback in this approach is that whenever an object is converted to a new format, some of its original properties may not be adequately transferred to the target format. This may occur due to incompatibilities between the source and target formats or because the application used to do the conversion is not capable of carrying out its tasks correctly.

Understanding Migration

Every preservation intervention involves choices. Resources are finite, often scarce. As a result, decisions have to be taken to ensure that the best possible preservation strategy is selected from the wide range of options available. These decisions depend upon a

multiplicity of factors such as: technical expertise, users' expectations, institutional budget, existing equipment and available time [16]. In this respect migration-based strategies are no different.

In order to better understand all the steps involved in a migration process, one should consider the following sequence of activities:

Selection of a Migration Strategy

Two decisions have to be made prior to any object migration: which format should be used to accommodate the properties of the original object; and which application should be used to carry out that migration. This decision-making activity constitutes the first stage of any migration process. It is in the best interests of the preserving institution to aim for the optimal combination of target format and conversion software, i.e. one that preserves the maximum number of properties of the original object at the minimum cost.

Cost should be regarded as a multi-dimensional variable. Factors such as throughput, application charges, format openness or prevalence should be considered collectively during this decision-making activity. Objective tools or frameworks especially designed to help institutions in the selection of appropriate options would greatly simplify this exceptionally complicated task.

The Conversion

The conversion work consists of the reorganisation of the information elements that comprise the digital object into the logical structures as defined by a different format [17].

From the preserver's point-of-view, carrying out a conversion usually consists in setting up a conversion application and executing it against a collection of digital objects. Some scripts may have to be developed in order to automate the whole procedure.

Evaluation of Results

After the conversion process, the resultant objects should be evaluated in order to determine the amount of data loss incurred during migration. This is accomplished by comparing the properties that comprise the source object (also known as significant properties [18][19]) with the properties of its converted counterparts. If the evaluation results are below expectations, i.e. the object's properties have degraded to an unacceptable level, a different migration alternative should be selected and the whole process reinitiated.

In most cases, the evaluation process still requires a considerable amount of manual labour. Certain subjective properties such as the disposition of graphic elements in a text document or the presence of compression artifacts in an image file are generally inspected by human experts, rendering this activity both onerous and time-consuming [20].

A Service-Oriented Architecture for Automatic Migration

At the University of Minho research is being undertaken to devise new pathways to carry out the three outlined activities in an automated fashion (i.e. selection of migration options, conversion and evaluation). Current activities are focused on the development of a Service-Oriented Architecture (SOA) [21][22] that, by combining input from different distributed applications, enables client institutions to preserve collections of digital material automatically.

It is assumed that client institutions already possess a digital repository system capable of storing, managing and providing access to the digital objects they hold. The repository system will act as the client application that benefits from the services provided by the SOA.

In order better to understand all the functions provided by the SOA one might consider the following scenario:

The National Archives of Portugal [23] are currently engaged in the development of a

digital repository system capable of preserving authentic digital objects produced by affiliated public administration institutions (Project RODA [24]). Alongside the development of the repository software comes the creation of ingest and preservation policies that will aid producers in the preparation of their material before it is submitted to the repository. This notwithstanding, the repository will expect to be confronted by objects in formats previously unencountered and which will need to undergo a process of normalisation before being deposited. In the presence of an unrecognised format, the repository system could invoke a format identification service provided by the SOA in order to obtain information about the object's format, in addition to checking its integrity. After this operation, the repository could interrogate the SOA to obtain a list of formats to which the object could be converted. Simultaneously, the repository would inform the SOA of its preservation preferences and requirements, i.e. a list of preservation-oriented requirements derived from the policies created by the senior management of the archive. A few examples of such requirements are as follows:

- Preservation interventions should be affordable and swift;
- Interventions should preserve the maximum number of significant properties of the original object;
- The interventions should not resort to formats that are dependent on the payment of royalties.

The SOA would then address all of these criteria with information previously acquired about the behaviour and quality of all accessible conversion applications and would then produce a ranked list of optimal migration options. The repository system could then select the most suitable one from this list and request the SOA to carry out the corresponding migration.

After the conversion process, the repository system would receive a new digital object (better yet, a new digital representation of the source digital object) and a migration report stating the amount of data lost in that migration. This report could then be merged with the preservation metadata already maintained by the repository in order to document the preservation intervention and sustain the object's authenticity. On a regular basis, the repository would consult with a notification service to determine if any of the formats it holds are at risk of becoming obsolete. When a format falls into that condition, a new migration process is triggered.

A close examination of the outlined scenario enables us to identify the following services:

- A format identification service that also checks the integrity of digital objects;
- A service that produces recommendations of optimal migration options (selection of a migration option);
- A service to carry out format migrations (the conversion);
- A service to determine the amount of data loss resulting from a migration (evaluation of results);
- A service that provides information about the formats that are at risk of becoming obsolete.

The general architecture of the proposed SOA is depicted in Figure 1. This design does not intend to be prescriptive or limiting in any way. The goal is to provide a framework for discussion by pointing out the fundamental elements that should be present in such a system. Several interesting and competing research projects are presented as promising candidates to implement some of these elements. We recognise of course that many other initiatives and solutions might also exist outside the scope of our work or this article.

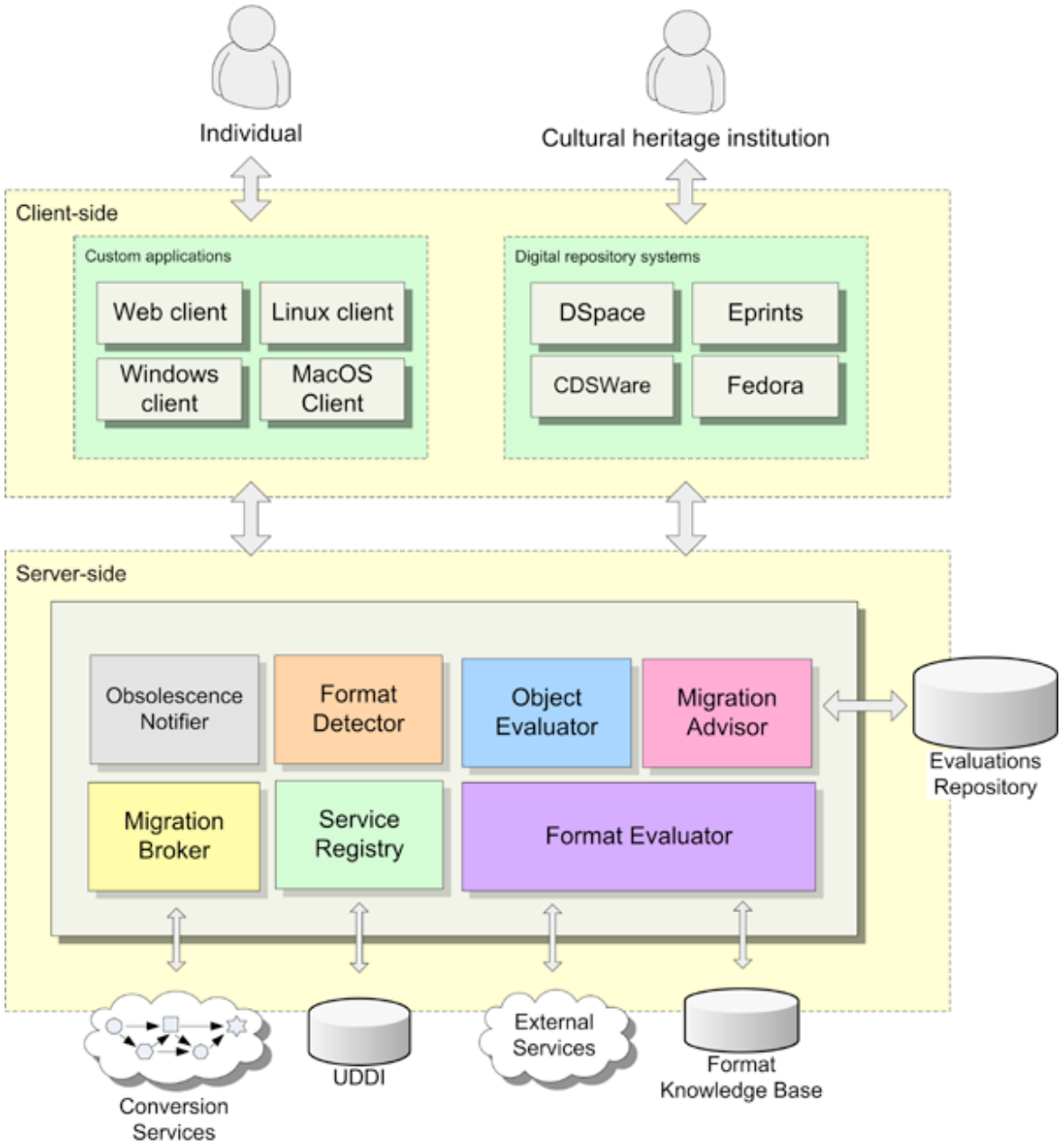

Figure 1: General architecture for a SOA capable of delivering automatic digital preservation.

The figure is divided into two major sections: the client and the server-side. The client-side depicts a few examples of applications that may use the services provided by the SOA. Among these are: digital repository systems like DSpace [25], Fedora [26] or Eprints [27], and custom applications developed by individual users.

It is important to point out that any application capable of invoking a Web service may make use of the proposed SOA.

On the server-side are depicted the chief components comprising this framework. Each of these components is actually an independent application with distinctive roles and responsibilities that co-operate with each other by exchanging messages. This approach makes it possible for each component to be governed by a different organisation and facilitates the distribution of workload.

Obsolescence Notifier

The first of these components is the Obsolescence Notifier, a service responsible for

raising awareness among client institutions of the file formats that are at risk of becoming obsolete. This service should to be consulted regularly by client institutions in order to determine if the objects in their custody are close to becoming unreadable to their designated community.

Several resources are available that could be used to support such a service. A few examples are as follows:

- The report 'Risk Management of Digital Information: A File Format Investigation' by Lawrence, et al [17], is a study on the impact of migration on file integrity and can provide some guidance in assessing the risk involved in keeping certain file formats.
- The INFORM Methodology [28] is an approach for measuring the durability of digital formats.
- The service could also be supported by a group of human experts responsible for monitoring consumer trends and emerging technologies. Several institutions already perform this type of work, although in a more generalised fashion, e.g. DigiCULT with its annual Technology Watch Reports [29] and the Digital Preservation Coalition with its monthly themed Technology Watch Reports [30].

Format Detector

The Format Detector, as the name suggests, is a service capable of identifying the underlying encoding of a digital object. The client institution should be able to monitor, migrate and validate the integrity of digital objects without human intervention and this service is indispensable in accomplishing that goal. Furthermore, it enables digital formats to be identified according to the naming scheme used by other components that comprise the proposed SOA (e.g. the Migration Broker).

The following applications are potential candidates for supporting such a service:

- JHOVE, jointly developed by JSTOR and the Harvard University Library, is an application specifically designed to identify and characterise digital formats. In fact, JHOVE is more of a technical metadata extractor than a simple format identifier. At the moment, JHOVE is capable of supporting 11 different file formats [31].
- Droid, on the other hand, was developed exclusively to identify digital formats. It was developed by the National Archives of the United Kingdom, the creators of the PRONOM format registry [32][33], and currently supports hundreds of different file formats [34].

Some institutions and initiatives have been developing services capable of carrying out format migrations [35-40]. Such initiatives rely on a common set of communication protocols to support the discovery and invocation of conversion procedures. Any conventional application may also be used as a service if an appropriate application wrapper is developed [22], i.e. a small piece of software that acts as the intermediary between the application and communication protocol.

In this type of approach, a client application is used to send out a digital object to a remote procedure that, after unpackaging the received message, converts the embedded object and returns the result back to the client.

Standard protocols, such as the ones that accompany Web services technology [41], may play an important role in this domain due to their open-standard and platform-independent characteristics.

A distributed approach to migration introduces some appealing properties:

- the use of Web services hides the complexity of the conversion software that is being used under the hood and promotes interoperability by cloaking the peculiarities of the supporting hardware and operating system;
- combining conversion services enables new migrations to be performed and makes this solution capable of coping with the gradual disappearance of converters [22];
- the development of redundant services ensures that the migration network remains functional during situations of partial breakdown.

- this approach is compatible with several variants of migration, such as normalisation [13][42-47] and migration on-request [10].

However, requiring the presence of a computer network to carry out format migrations hardly seems reasonable in a preservation context. This type of reliance on technology is generally very undesirable. However, digital preservation is a global problem. A distributed approach may very well prove to be an effective way to handle the intricacies of preservation as it allows institutions worldwide to share their solutions and co-operate in the network of services.

Service Registry

The Service Registry component is responsible for managing information about existing conversion services. It stores metadata about its producer/developer (e.g. name, description and contact), about the service itself (e.g. name, description, the source/target formats that it is capable of handling, cost of invocation, etc.) and information on how the service should be invoked by a client application (i.e. its access point).

It is important that the Service Registry is populated with rich metadata. Much of the information delivered to end-users after a conversion will be obtained from this data source. This information can be used to document the preservation intervention as it outlines all the components that took part in the migration process and describes the outcome of the event in terms of data loss and object degradation (see Object Evaluator). This migration report constitutes what PREMIS refers to as an Event Entity [48].

One of the major advantages of using Web services in this context is in the capacity to combine tens or hundreds of conversion services to create new migration operations. However to accomplish this, each conversion service should respect a well-defined interface that establishes the arguments that each conversion service should be capable of handling. Although this interface is essential to produce service compositions, it is not sufficient on its own. Each conversion service must be described with source and target format metadata elements whose values are obtained from a controlled vocabulary. This is fundamental to enable the computation of the migration network (i.e. all possible migration paths between two given formats).

Several initiatives are considered suitable candidates to provide that controlled vocabulary:

1. The PRONOM registry, an initiative from the National Archives of the UK aims at building a registry of information about every existing file format [32][33];
2. The Digital Formats Web site, created by the Library of Congress aims at providing information about digital content formats [49][50];
3. The Global Digital Format Registry intends to provide sustainable distributed services to store, discover and deliver representation information about digital formats [51][52];
4. Representation Information Registry/Repository is an OAIS representation information registry for digital data and is currently being developed by the Digital Curation Centre [53].

Migration Broker

The Migration Broker is responsible for carrying out object migrations. In practice, this component is responsible for making sure that composite conversions are performed atomically from the point of view of the client application and the rest of the SOA components. Additionally, this component is responsible for recording the performance of each migration service. The results of these measurements are stored in the Evaluations Repository, a knowledge base that supports the recommendation system (see Migration Advisor).

A prototype of the proposed SOA is currently being devised at the University of Minho and is presently capable of measuring the following process-related criteria:

- Availability, i.e. the probability of a service being operational at the time of invocation.
- Stability, i.e. the capacity of a service to carry out what it purports to do.
- Throughput, i.e. the amount of work that the service is capable of doing per time unit. The workload is determined by the size of the object to be converted.
- Cost, i.e. the amount that a client will have to pay to use the service on one occasion. The cost of a composite migration is the sum of each individual cost.
- Outcome size, i.e. the size in bytes of the resulting object when compared with the original.
- Outcome file count, i.e. the number of files in the resulting representation.

Format Evaluator

The Format Evaluator provides information about the current status of file formats. This information enables the Migration Advisor to determine which formats are better suited to accommodate the properties of source objects by looking at the characteristics of each pair of formats. This service is supported by a data store containing facts about formats (i.e. Format Knowledge Base), but could also exploit external sources of information such as the PRONOM registry or Google Trends [54], to determine automatically a format's prevalence and usage.

The current prototype is capable of determining the potential gain (in terms of preservation) that one might obtain in converting an object from its original format to a new one by considering the following set of criteria:

- Market share, i.e. whether the format is widely accepted or simply a niche format. Market share is also known as 'adoption'. Adoption refers to the degree to which the format is already used by the primary creators, disseminators, or users of information resources;
- Support level, i.e. whether the creator of the format provides good technical support on the format;
- Is standard, i.e. whether the format has been published by a standards organisation;
- Open specification, i.e. whether specification can be independently inspected.
- Supports compression, i.e. whether the format supports any type of compression.
- Lossy compression only, i.e. whether the format only supports a lossy type of compression.
- Supports transparency, i.e. whether the format supports transparency features.
- Embedded metadata, i.e. whether the format may contain embedded metadata.
- Royalty-free, i.e. whether royalties or licence fees are payable.
- Open source, i.e. whether there are decoders whose source can be independently inspected.
- Backwardly compatible, i.e. whether revisions have support for previous versions.
- Documentation level, i.e. whether the format specification is well documented.
- Competing formats available, i.e. whether competing or similar formats exist.
- DRM support, i.e. whether DRM (Digital Rights Management), encryption or digital signatures can be used.
- Update frequency, i.e. whether revisions happen so fast that the archive cannot keep up with demand.
- Supports custom extensions, i.e. whether extensions, such as executable sections or narrowly supported features, can be added to the format.
- Life time, i.e. how many years have passed since the format has been officially released.
- Transparent decoding, i.e. the degree to which the digital representation is open to direct analysis with basic tools, including human readability using a text-only editor.
- Reader single producer, i.e. whether the reader/viewer is produced by a single entity.
- Single reader, i.e. whether the format can only be read by one piece of software.
- Open source reader, i.e. whether the source-code of the reader software can be independently inspected.
- Multiplatform reader, i.e. whether the reader software can be run on various platforms (e.g. operating systems or hardware).

All of these criteria are being considered by the Migration Advisor to rank all the available

migration options.

This component is capable of measuring the preservation gain of performing a certain transformation. For example, if the target format is royalty-free while the source format is not, there is a preservation gain associated with that transformation. On the other hand, if the target format only supports a lossy type of compression, while the source format is not compressed at all, there is a potential risk of losing important information in the process.

The criteria present in this evaluation taxonomy were assembled from various bibliographic sources such as [28][49][55]. Groups of format experts and digital curators may also contribute with additional criteria to enrich the evaluation taxonomy.

Object Evaluator

The Object Evaluator is in charge of judging the quality of the migration outcome. It accomplishes this by comparing objects submitted for migration with their converted counterparts. Again, these evaluations will be performed according to a range of criteria. These criteria, known in this context as significant properties, constitute the set of attributes of an object that should be kept intact during a preservation intervention [19]. They constitute the array of attributes that characterise an object as a unique intellectual entity, independently of the encoding used to represent it. The Bible for example, may exist in many different formats and media, e.g. ASCII text, Portable Document Format, written on paper or carved on stone, and still be regarded as the Holy Bible. Considering text documents as an example, some significant properties could be: the number of characters, the order of those characters, the page size, the number of pages, the graphical layout, the font type and size.

Migration Advisor

The Migration Advisor is responsible for producing suggestions of migration alternatives. In reality this component acts as a decision support centre for client institutions and is capable of determining the best possible choice within a wide range of options. It accomplishes this by confronting the preservation requirements outlined by client institutions with the accumulated knowledge about the behaviour of each accessible migration path.

The behaviour of each migration path is determined by taking into consideration the sets of criteria previously described: conversion performance, status of the formats involved and data loss (handled respectively by the Migration Broker, Format Evaluator and Object Evaluator).

Evaluations Repository

In order to generate an appropriate recommendation, the Migration Advisor resorts to the Evaluations Repository, a database containing all the measurements taken by the Object Evaluator, the Format Evaluator and the Migration Broker. Averaging these readings provides a general idea about the behaviour of each migration path.

Different institutions will have distinct preservation needs. They will be able to state their individual requirements by weighting the importance of each criteria handled by the system. This enables the system to rank all the alternatives according to their level of aptness to resolve the preservation problem of the client institution.

In order to rank all possible options, the Evaluations Repository must be populated with data. This is generally called training and basically consists in requesting the SOA to convert a large set of digital objects in different forms and sizes using all possible migration paths. This operation forces all the evaluators to produce reports that will be used by the Migration Advisor to compute an appropriate recommendation.

The Migration Advisor uses the same principles as the evaluation framework described by Rauch and Rauber [16][20][55-57]. The process within the SOA is orchestrated as follows:

1. For each migration path, a standard or average behaviour is determined for all of

the evaluation criteria. This task is performed by the Migration Advisor whenever a suggestion is requested. It is assumed that the Evaluations Repository has already been populated with data (Figure 2, step 1);

2. The average behaviour per criterion of a given migration path is then normalised into a comparable scale of 0 to 1 (Figure 2, step 2). The highest measurements assume the value of 1 whilst the lowest are normalised to a 0. All other values are spread between these two figures. It is important to point out that evaluations always produce positive preservation results, i.e. high values correspond to a better preservation performance. The cost criterion, for example, must be inverted before the normalisation step, as higher values of cost correspond to lower preservation performances.
3. The client institution is then asked to assign weights to each evaluation criteria according to its perception of importance. These weights are then multiplied by the normalised values calculated in the previous step (Figure 2, step 3);
4. The overall score of a migration path is obtained by summing up all the ensuing values. The most appropriate migration option is the one that attains the highest score (Figure 2, step 4).

	Evaluation criteria	MP1	MP2
Migration Broker	Availability	0.95	0.9
	Stability	0.8	0.9
	Throughput	1 Kb/s	2 Kb/s
	Cost	0.01 €	0.05 €
	Outcome size	1.2	0.9
	File Count	1.0	1.0
Format Evaluator	Market share	0.5	0.9
	Support level	1.0	0.2
	Is standard	1.0	0.0
	Open specification	1.0	1.0
	Supports compression	0.0	1.0
	...	...	...
Format Evaluator	Textual content	0.9	0.95
	Page size	0.6	0.9
	Number of pages	1.0	1.0
	Graphical layout	0.8	0.6
	Font type & size	0.5	0.8
	...	...	...

MP1	MP2
1.0	0.0
0.0	1.0
0.0	1.0
1.0	0.0
0.0	1.0
1.0	1.0
0.0	1.0
1.0	0.0
1.0	0.0
1.0	1.0
0.0	1.0
...	...
0.0	1.0
0.0	1.0
1.0	1.0
1.0	0.0
0.0	1.0
...	...

x

Weights
0.2
0.2
0.2
0.3
0.07
0.03
0.2
0.2
0.2
0.3
0.1
...
0.4
0.2
0.2
0.1
0.1
...

=

MP1	MP2
0.2	0.0
0.0	0.2
0.0	0.2
0.3	0.0
0.0	0.07
0.03	0.03
0.0	0.2
0.2	0.0
0.2	0.0
0.3	0.3
0.0	0.1
...	...
0.0	0.4
0.0	0.2
0.2	0.2
0.1	0.0
0.0	0.1
...	...
1,53	2.0

(1) Standard behaviour of each migration path

(2) Normalised behaviours [0-1]

(3) User assigned weights of importance

(4) Overall rating

Figure 2: Steps involved in the ranking migration alternatives. MP1 and MP2 represent two different migration paths to convert text documents. (This table is displayed in an alternative format.)

Evaluation

A prototype implementing the concepts described in this paper is currently being devised at the University of Minho and is expected to be fully operational by the end of 2006 [22][58]. The purpose of this prototype is to evaluate the suitability of the proposed architecture and the precision of the recommendation system. Precision will be assessed using cross-validation techniques.

Still images are generally represented by simple structures and, for that reason, are being used to guide the development of the prototype. More complex objects, like text documents produced by word-processing software, will be considered afterwards in order to assess their effectiveness in handling more subjective criteria, such as appearance or text layout.

The current prototype fully implements the following components:

- a Service Registry, being supported by an UDDI server.
- a Migration Network with 91 registered converters.
- a Migration Broker, fully capable of combining migration services and recording their performance;
- a Format Detector, being supported by Droid [34];
- a Format Evaluator, capable of comparing several file formats according to the criteria outlined in the previous section;
- a Migration Advisor, fully capable of computing recommendations based on the information collected in the Evaluations Repository.

Adding new evaluation criteria to the prototype is as easy as updating a configuration file. The real complexity relies on creation of new criterion evaluators. Once developed, these evaluators can be placed in the servers file system (to be loaded during system's bootstrap) or remotely invoked.

New conversion services can be attached to the system by simply adding them to the Service Registry. Training the Migration Advisor to recognise new conversion services is, of course, essential.

Throughout this article, Web services have been presented as a promising technology to support the proposed SOA. However, many other protocols exist which could be used to implement these ideas. Different technologies could even be utilised together by means of gateways or proxies. For example, a gateway is currently being developed to enable converters provided by the TOM Conversion Service [35][59], a technology that uses a non-standard communication protocol, to be used by our prototype.

Conclusions and Future Work

This article describes the set of components that are necessary to build a Service-Oriented Architecture (SOA) to enable cultural heritage institutions to carry out digital preservation with minimum human intervention.

The proposed SOA enables institutions to co-operate in the establishment of a global advisory service that, among other things, will be capable of producing recommendations of optimal migration options, perform format migrations and thoroughly document preservation interventions by generating appropriate preservation metadata.

The proposed SOA could also be used as an objective tool for comparing file formats and conversion software. It could be used to provide an on-demand migration service, i.e. a service capable of converting objects from their archival configurations to formats more suitable for dissemination; as well as a normalisation procedure for ingest work.

Although a prototype for this SOA is still under development, some conclusions can already be drawn. The set of digital objects used to train the recommendation system should be as heterogeneous as possible in terms of shape and size, and should contain at least a couple of thousand objects. Small or homogeneous object sets generate very imprecise recommendations due to overfitting in the learning process [60].

Further research could be conducted to detect patterns in the user-assigned weights. Such patterns would represent user profiles and would enable the recommendation process to be automated one step further.

The proposed SOA could also contribute to fostering new lines of research such as the improvement, or the development, of comparison algorithms for different classes of objects, e.g. image, text, audio, video or datasets. Comparators such as these are necessary to develop a general purpose Object Evaluator. Further work should also be conducted to devise a general evaluation taxonomy for several classes of digital objects.

Acknowledgments

The work reported in this article has been funded by the FCT (Fundação para a Ciência e a Tecnologia, Portugal) under the grant SFRH/BD/17334/2004.

References

1. PANDORA - Australia's Web Archive http://pandora.nla.gov.au/
2. UK Web Archiving Consortium http://www.webarchive.org.uk/
3. The Wayback Machine http://www.archive.org/
4. Austrian On-Line Archive http://www.ifs.tuwien.ac.at/~aola/
5. Kulturarw3 - Long time preservation of electronic documents http://www.kb.se/kw3/ENG/
6. Growth of Institutional Archives over Time http://archives.eprints.org/index.php?action=analysis
7. D. S. H. Rosenthal, T. Robertson, T. Lipkis, V. Reich, and S. Morabito, "Requirements for Digital Preservation Systems", *D-Lib Magazine* 11 (11), 2005 http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html
8. L. S. Lin, C. K. Ramaiah, and P. K. Wal, "Problems in the preservation of electronic records", Library Review 52 (3), pp. 117-125, 2003.
9. A. R. Heminger and S. B. Robertson, "Digital Rosetta Stone: A Conceptual Model for Maintaining Long-Term Access to Digital Documents", presented at 6th DELOS Workshop, Tomar, Portugal, 1998.
10. P. Mellor, P. Wheatley, and D. M. Sergeant, "Migration on Request, a Practical Technique for Preservation", presented at ECDL '02: 6th European Conference on Research and Advanced Technology for Digital Libraries, London, UK, 2002.
11. S. Granger, "Emulation as a Digital Preservation Strategy", *D-Lib Magazine* 6 (10), 2000 http://www.dlib.org/dlib/october00/granger/10granger.html
12. R. A. Lorie, "A Methodology and System for Preserving Digital Data", presented at Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02), Portland, Oregon, 2002.
13. K.-H. Lee, O. Slattery, R. Lu, X. Tang, and V. McCrary, "The State of the Art and Practice in Digital Preservation", Journal of Research of the National Institute of Standards and Technology 107 (1), pp. 93-106, 2002.
14. P. Wheatley, "Migration: a Camileon discussion paper", *Ariadne* issue 29, September 2001 http://www.ariadne.ac.uk/issue29/camileon/
15. Task Force on Archiving of Digital Information, Commission on Preservation and Access, and Research Libraries Group, Preserving digital information: report of the Task Force on Archiving of Digital Information. Washington, D.C.: Commission on Preservation and Access, 1996.
16. C. Rauch and A. Rauber, "Preserving Digital Media: Towards a Preservation Solution Evaluation Metric", presented at International Conference on Asian Digital Libraries, Shanghai, China, 2004.
17. G. W. Lawrence, W. R. Kehoe, O. Y. Rieger, W. H. Walters, and A. R. Kenney, "Risk Management of Digital Information: A file format investigation", Council on Library and Information Resources, Washington, DC 2000.
18. H. Hofman, "Can Bits and Bytes be Authentic? Preserving the Authenticity of Digital Objects", presented at International Federation of Library Associations Conference, Glasgow, 2002.
19. A. Rusbridge, "Migration on Request", University of Edinburgh - Division of Informatics, 4th Year Project Report 2003.
20. C. Rauch, F. Pavuza, S. Strodl, and A. Rauber, "Evaluating preservation strategies for audio and video files", presented at DELOS Digital Repositories Workshop, Heraklion, Crete, 2005.
21. OASIS SOA Reference Model TC, "OASIS Reference Model for Service Oriented Architectures (Working Draft 10)", OASIS 2005.
22. M. Ferreira, A. A. Baptista, and J. C. Ramalho, "CRiB: A service oriented architecture for digital preservation outsourcing", presented at XATA - XML: Aplicações e Tecnologias Associadas, Portalegre, Portugal, 2006.
23. Instituto dos Arquivos Nacionais/Torre do Tombo Web site http://www.iantt.pt/
24. RODA (Repositório de Objectos Digitais Autênticos) Web site http://roda.iantt.pt/
25. DSpace Web site http://www.dspace.org/
26. Fedora Web site http://www.fedora.info/
27. EPrints Web site http://www.eprints.org/
28. A. Stanescu, "Assessing the Durability of Formats in a Digital Preservation Environment", *D-Lib Magazine* 10 (11), 2004.
29. Technology Watch Reports http://www.digicult.info/
30. Technology Watch Reports http://www.dpconline.org/graphics/reports/
31. JHove - JSTOR/Harvard Object Validation Environment http://hul.harvard.edu/jhove/
32. J. Darlington, "PRONOM - A Practical Online Compendium of File Formats", RLG

DigiNews 7 (5), 2003.
33. PRONOM - The file format registry http://www.nationalarchives.gov.uk/pronom/
34. UK National Archives, "Droid: Digital Record Object Identification," 1.0 ed. Surrey: UK National Archives, 2005.
35. J. M. Ockerbloom, "Mediating Among Diverse Data Formats," in School of Computer Science. Pittsburg: Carnegie Mellon University, 1998, pp. 164.
36. F. L. Walker and G. R. Thoma, "A Web-Based Paradigm for File Migration", presented at IS&T's 2004 Archiving Conference, San Antonio, Texas, USA, 2004.
37. J. Hunter and S. Choudhury, "A Semi-Automated Digital Preservation System based on Semantic Web Services", presented at Joint ACM/IEEE Conference on Digital Libraries (JCDL'04), 2004.
38. Preservation webservices Architecture for Newmedia and Interactive Collections (PANIC) http://metadata.net/panic/
39. J. M. Wing and J. Ockerbloom, "Respectful Type Converters", IEEE Transactions on Software Engeneering 26 (7), 2000.
40. J. Hunter and S. Choudhury, "PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services", International Journal on Digital Libraries 6 (2), pp. 174-183, 2006.
41. S. Graham, S. Simeonov, T. Boubez, D. Davis, G. Daniels, Y. Nakamura, and R. Neyama, *Building Web Services with Java: Making Sense of XML, SOAP, WSDL and UDDI*: Sams Publishing, 2002.
42. K. Thibodeau, "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years", presented at The State of Digital Preservation: An International Perspective, Washington D.C., 2002.
43. H. Hofman, "How to keep digital records understandable and usable through time?" presented at Long-Term Preservation of Electronic Records, Paris, France, 2001.
44. H. Heslop, S. Davis, and A. Wilson, "An Approach to the Preservation of Digital Records." Camberra, Australia: National Archives of Australia, 2002.
45. A. G. Howel, "Preserving Digital Information: Challenges and Solutions", Cooperative Action by Victorian Academic Libraries, Victorian university libraries, State Library of Victoria 2004.
46. G. Hodge and E. Frangakis, "Digital Preservation and Permanent Access to Scientific Information: The State of the Practice", International Council for Scientific and Technical Information & CENDI, Report 2004-3: Rev. 05/04, 2004.
47. M. Hedstrom, "Digital Preservation: A time bomb for digital libraries", Computers and the Humanities 31, pp. 189-202, 1998.
48. PREMIS Working Group, "Data dictionary for preservation metadata: final report of the PREMIS Working Group", OCLC Online Computer Library Center & Research Libraries Group, Dublin, Ohio, USA, Final report 2005.
49. Digital Formats Web site http://www.digitalpreservation.gov/formats/
50. C. R. Arms and C. Fleischhauer, "Digital Formats: Factors for Sustainability, Functionality, and Quality", presented at IS&T Archiving Conference, Washington, D.C., USA, 2005.
51. S. L. Abrams and D. Seaman, "Towards a global digital format registry", presented at World Library and Information Congress: 69th IFLA General Conference and Council, 2003.
52. Global Digital Format Registry http://hul.harvard.edu/gdfr/
53. OAIS Representation Information Registry/Repository http://dev.dcc.ac.uk/twiki/bin/view/Main/DCCRegRepV04
54. Google Trends http://www.google.com/trends
55. C. Rauch, A. Rauber, H. Hofman, J. Bogaarts, R. Vedegem, F. Pavuza, J. Ahmer, and M. Kaiser, "A Framework for Documenting the Behaviour and Funcionality of Digital Objects and Preservation Strategies", DELOS Network of Excellence, Glasgow 2005.
56. C. Rauch, "Preserving Digital Entities - A Framework for Choosing and Testing Preservation Strategies," in Institute for Software Technology and Interactive Systems. Vienna: Vienna University of Technology, 2004.
57. P. Weirich, B. Skyrms, E. W. Adams, K. Binmore, J. Butterfield, P. Diaconis, and W. L. Harper, *Decision Space: Multidimensional Utility Analysis*. Cambridge, 2001.
58. CRiB - Conversion and Recommendation of Digital Object Formats Web site http://crib.dsi.uminho.pt/
59. TOM Conversion Service http://tom.library.upenn.edu/convert/
60. I. V. Tetko, D. J. Livingstone, and A. I. Luik, "Neural network studies. 1.

Comparison of overfitting and overtraining", Journal of Chemical Information and Computer Sciences 35, pp. 826-833, 1995.

Author Details

Miguel Ferreira
PhD student
Department of Information Systems
University of Minho
4800-058 Azurém-Guimarães
Portugal

Email: mferreira@dsi.uminho.pt
Web site: http://www.dsi.uminho.pt/~ferreira

Graduated as a Systems and Informatics Engineer, has worked as a consultant at the Arquivo Distrital do Porto (Oporto's Archive) and as a researcher at the University of Minho. Since 2003 has been publishing in field of digital archives/libraries and preservation. Currently, is developing work as a PhD student and coordinating several research projects at the Arquivo Distrital do Porto and the Portuguese National Archives (Instituto dos Arquivos Nacionais/Torre do Tombo).

Ana Alice Baptista
Auxiliary professor
Department of Information Systems
University of Minho
4800-058 Guimarães
Portugal

Email: analice@dsi.uminho.pt
Web site: http://www.dsi.uminho.pt/~analice

Auxiliary Professor at the Department of Information Systems of University of Minho, Ana has been publishing in the areas of Knowledge Society, Scholarly Communication, Information Access & Retrieval and Semantic Web. She is also interested in the social aspects of the Internet, primarily on its impacts on scholarly communication.

José Carlos Ramalho
Auxiliary Professor
Computer Science Department
University of Minho
4710-057 Braga
Portugal

Email: jcr@di.uminho.pt
Web site: http://www.di.uminho.pt/~jcr

Auxiliary Professor at the Computer Science Department of the University of Minho, has a Masters on 'Compiler Construction' and a PhD on the subject 'Document Semantics and Processing''. Has been managing projects and publishing in the field of Markup Languages since 1995.

Return to top

Article Title: "A Foundation for Automatic Digital Preservation"
Author: Miguel Ferreira, Ana Alice Baptista and José Carlos Ramalho
Publication Date: 30-July-2006 Publication: Ariadne Issue 48
Originating URL: http://www.ariadne.ac.uk/issue48/ferreira-et-al/
 File last modified: Friday, 25-Aug-2006 13:56:14 BST

Ariadne is published every three months by UKOLN. UKOLN is funded by MLA the Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.