The original publication is available at http://www.springerlink.com.

M. Ferreira, A. A. Baptista and J. C. Ramalho, "*An intelligent decision support system for digital preservation,*
"International Journal on Digital Libraries, 2007. Available at http://dx.doi.org/10.1007/s00799-007-0013-x

Miguel Ferreira · Ana Alice Baptista · José Carlos Ramalho

An intelligent decision support system for digital preservation

Abstract This paper describes a Service Oriented Architecture (SOA) based on Web services technology designed to assist cultural heritage institutions in the implementation of migration based preservation interventions. The proposed SOA delivers a recommendation service and a method to carry out complex format migrations. The recommendation service is supported by three evaluation components that assess the quality of every migration intervention in terms of its performance (Migration Broker), suitability of involved formats (Format Evaluator) and data loss (Object Evaluator). Throughout the paper the whole workflow between these three components is explained in detail as well as the most relevant tasks that carried out internally in each of them. The proposed system is also able to produce preservation metadata that can be used by client institutions to document preservation interventions and retain objects' authenticity. Although the primary goal of this SOA is the implementation of migration based preservation interventions, it can also be used for other purposes such as comparing file formats or evaluating the performance of conversion applications.

Keywords Digital preservation · Decision Support Systems · Migration · Service Oriented Architectures (SOA) · Web services · Authenticity · Preservation metadata · Preservation services

Miguel Ferreira
Department of Information Systems, University of Minho, Portugal
Tel.: +351 253 510 261
E-mail: mferreira@dsi.uminho.pt

Ana Alice Baptista
Department of Information Systems, University of Minho, Portugal
Tel.: +351 253 510 310
E-mail: analice@dsi.uminho.pt

José Carlos Ramalho
Department of Informatics, University of Minho, Portugal
Tel.: +351 253 604479
E-mail: jcr@di.uminho.pt

1 Introduction

Over the last decade, the research community has come up with a considerable number of strategies aiming at solving the problem of digital preservation and technological obsolescence. Among such strategies are emulation [18,42], encapsulation [45,46] and migration [9,32,43,55], as well as an assortment of variations and combinations from all of the above, e.g. normalization [25,50], migration on-request [32, 43] or Universal Virtual Computer (UVC) [29,28].

The migration strategy can best 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." [49].

Contrary to other preservation strategies, migration based approaches do not attempt to preserve digital objects in their original form. In alternative, they transform objects from near obsolete formats into more up-to-date encodings that most users will be able to interpret using common software available on their personal computers. The main disadvantage in this type of approach is that whenever a digital object is migrated there is a high probability that some of its inner properties may not be adequately transferred to the target format (i.e. some type of data loss is expected to take place) [20,21]. The reason for this phenomenon is twofold: there might be structural incompatibilities between the source and the target formats or the converter used to carry out the transformation may be incapable of performing that task correctly. Nevertheless, migration continues to be one of the most widely applied preservation strategies [25].

Whatever strategy is in place, preservation interventions usually involve choices. As resources are often limited, decisions have to be made to make sure that the best possible preservation approach is selected from a broad range of available options. These decisions generally depend of an assortment of factors such as: technical expertise of the preserving institution, the expectations of the designated community, the financial commitment, the allocated technological infrastructure and available time [40]. Migration based strategies are not different in this regard.

In general, two decisions anticipate the implementation of a migration strategy: first, one must decide upon which format should be used to accommodate the properties of the original object (i.e. choose the target format); and secondly, which conversion application should be used to carry out the corresponding transformation. This decision making activity constitutes a major step in any migration process. In general, it is at the best interest of the preserving institution that a combination of target format and conversion software is chosen which preserves the maximum number of properties of the original object at a minimum cost. Cost, however, should be regarded as a multidimensional variable. Factors such as migration throughput, application fees, format openness and prevalence should be considered collectively during the decision-making activity. Objective tools and frameworks specially designed to assist institutions in the selection of appropriate migration alternatives would greatly simplify this exceptionally complicated task.

Following the decision-making phase is the conversion process itself. Objects are passed through conversion software in order to create faithful representations of those objects in more prevalent formats. In order to retain the objects' authenticity, this process should be documented in a high level of detail using preservation metadata [3,5,7,22, 31,37].

After the conversion process, the resultant objects should be evaluated as to determine the amount of data that was lost during migration. This procedure generally consists in comparing the significant properties [22,43] of the source object with the significant properties of its converted counterparts. If the evaluation results are bellow expectations, i.e., the conversion did not maintain a minimum set of significant properties, then a different migration procedure ought to be selected and the whole process reinitiated.

In most cases, the evaluation process 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 making this activity exceptionally onerous and time consuming [39].

2 CRiB: An intelligent system to support preservation decisions

As depicted in the previous section, migrating distinct objects to a given format may not always produce satisfactory results. Each target format comprises a set of properties which may, or may not, be sufficient to accommodate the inner properties of a candidate object for migration. Moreover, distinct conversion applications may render considerably different objects. Some of the properties that constitute the original object may not be accurately transformed by the conversion application. Others may be ignored altogether.

This paper describes a Service Oriented Architecture (SOA) [35] designed to assist client institutions in the implementation of migration based preservation interventions. The proposed system works by assessing the quality of distinct conversion applications in order to produce recommendations of optimal migration options. The recommendations produced by the system also take into account the specific preservation requirements of each client institution.

Figure 1 depicts the CRiB[1] system, an architecture composed of several distributed components that work collectively to deliver a migration advisory service [12,14]. At the heart of the CRiB system is the Migration Advisor, the service responsible for producing such recommendations.

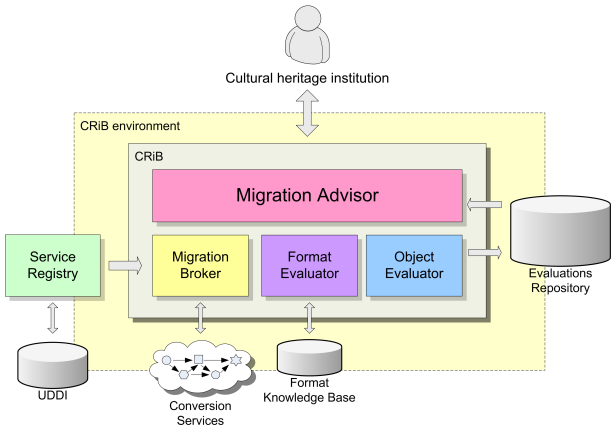


Fig. 1 The general architecture of the CRiB system.

In order to generate an appropriate recommendation, the Migration Advisor resorts to the Evaluations Repository (Figure 1), a database containing quality measurements collected over time by three evaluation components: the Migration Broker, the Format Evaluator and the Object Evaluator.

Client institutions may state their individual preservation needs by assigning weights to each of the evaluation criteria that the system is capable of handling. When a recommendation is requested, the collected evaluations will be confronted with the requirements outlined by the client institution and a ranked list of migration options will be produced. Client institutions may then take advantage of the system to carry out the suggested migration procedures.

In order to rank all the migration possibilities, the Evaluations Repository must first be populated with data. This is generally called training and basically consists in requesting the system to convert a large quantity of digital objects of different "shapes" and "sizes" using all reachable converters. This operation forces all evaluators to produce reports that will nourish the Evaluations Repository. The system keeps doing these evaluations even when it is in production. The system is regarded as intelligent because it learns with every migration executed as opposed of being pre-programmed with human defined rules.

Conversion applications are integrated with the CRiB system by means of application wrappers [14], i.e. small

[1] CRiB stands for Conversion and Recommendation of Digital Object Formats.

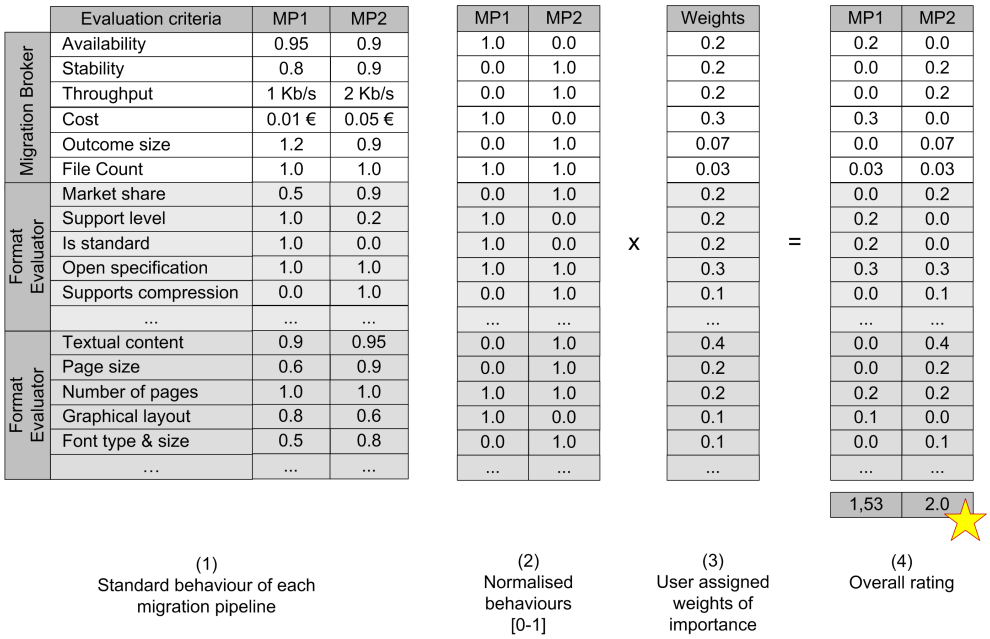


Fig. 2 Steps involved in the ranking of migration alternatives. MP1 and MP2 represent two distinct migration pipelines to convert text documents.

software layers that enable applications to be remotely invoked as Web Services.

The computation of the ranked list of alternatives is based on the same principles of the evaluation framework described by Rauch and Rauber in [38–41,54]. The process within the CRiB is orchestrated as follows:

1. For each conversion application, a standard or average behaviour is calculated taking into account all the evaluation criteria supported by the system. 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 standard behaviour of each criterion is then normalised into a comparable scale of zero to one (Figure 2, step 2). The highest measurements assume the value of one whilst the lowest are normalised to zero. 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 application cost criterion, for example, is always inverted before the normalisation step, as higher values of cost correspond to a lower preservation performance;
3. The client institution is then asked to assign weights to each evaluation criteria according to their perception of importance. These weights are then multiplied by the values calculated in the previous step (Figure 2, step 3);
4. The overall score for a given converter is obtained by summing up all the ensuing values. The most apt migration option is the one that attains the highest score (Figure 2, step 4). The resulting scale ranges from zero to infinite. The upper bound depends solely on the number of criteria being considered during the evaluation process.

The following sections describe the inner workings of each of the evaluation components.

2.1 Migration Broker

The Migration Broker component is responsible for carrying out format migrations as well as making sure that composite conversions (i.e. conversions composed by a sequence of transformations performed by distinct conversion applications) are performed atomically from the system's point of view, i.e., composite converters are handled in the same way as single converters. Additionally, this component is responsible for measuring the performance of each conversion option (single-step or composite). The performance is determined according to the criteria outlined in Table 1.

To support the discovery of conversion services, the Migration Broker resorts an additional component, the Service Registry. This component is responsible for managing information about all conversion applications known to the system. The information stored in the Registry is composed by the name, description and contact of the converter's developer; the description and cost of execution of each conversion service and information describing how it may be invoked by a client application (i.e. its access point).

Moreover, each conversion service is described by a pair of source and target format descriptors. It is critical that the values used in these metadata elements are obtained from a controlled vocabulary in order to facilitate the computation of composite migrations. Our prototype currently uses

Table 1 Process-related evaluation criteria.

Criterion name	Description
Availability	The probability of a service being operational at the time of invocation.
Stability	The capacity of a service to carry out what it purports to do.
Throughput	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	The amount of economic units that a client must pay in order to use the service a single time. The cost of a composite migration is the sum of the costs of each individual converter.
Outcome size	The size in bytes of the resulting object when compared with the original.
Outcome file count	The number of files in the resulting representation in relation to the original one.

the PRONOM registry [6,51] for this purpose. This decision was supported by the fact that the PRONOM appeared to be the most advanced initiative in this domain and the only one that explicitly stated the creation of services for consultation of its data store as part of its short-term objectives. Additionally considered options consisted in the Global Digital Format Registry [1,19], the Representation Information Registry/Repository [8] and MIME Media Types [15]. All of those were discarded due to lack of documentation, semantic precision or available tools.

It is important to note that the Service Registry is currently supported by an Universal Description, Discovery and Integration server (UDDI) [34]. The UDDI was initially considered due to its ability to store, search and publish Web service's metadata.

2.2 Format Evaluator

The recommendation process is additionally supported by information produced by a Format Evaluator. This component delivers information about the current status of file formats known and supported by the system. This information enables the Migration Advisor to determine which formats are better candidates to accommodate the properties of source objects by exclusively looking at the characteristics of each pair of formats. If, for instance, a certain target format is royalty-free whilst the source format is not, then one might consider that there will be an improvement in terms of preservation status if one were to convert an object from its source format to this particular royalty-free target format. On the other hand, if the target format exclusively supports a lossy type of compression while the source format is not compressed at all, then there will be a potential risk of loosing relevant information. The transformation that corresponds to the second example should score considerably lower in terms of its suitability for preservation.

The Format Evaluator is currently being supported by a data store of known facts about file formats, i.e. the Format Knowledge Base. In the future, this component could resort to additional sources of information such as format registries or services provided by other institutions. For example, Google Trends [16] could be used to determine how a format's popularity and prevalence has evolved over time.

The current prototype is capable of determining the potential gain in converting an object from its original format to a novel one by considering the criteria depicted in Table 2. These criteria were collected from a range of bibliographic sources such as [26,41,48]. Format experts and digital curators could also contribute with additional criteria to enrich the evaluation process.

The evaluations provided by this component are somewhat different than the ones produced by the Migration Broker. The latter focus mostly on objective criteria whose values derive directly from measurements taken during the conversion process (e.g. how much time it took to perform a transformation). The Format Evaluator on the other hand is supported by a data store (i.e. Formats Knowledge Base) of previously assembled facts about the formats that system is capable of handling. After a format transformation, the potential preservation gain is determined by applying specific comparison functions to the criteria values collected from this data store (see the specification of some comparison functions in Figure 3).

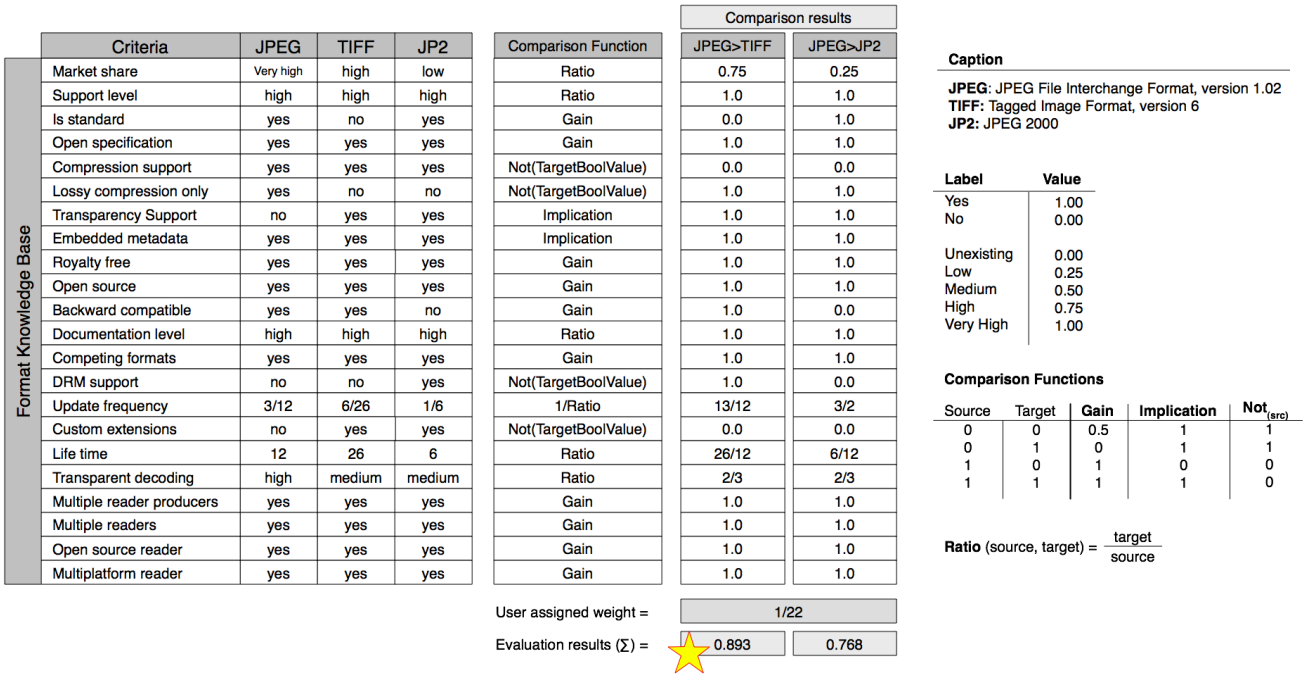
Consider, for example, an institution that wants to preserve a collection of high quality JPEG 1.02 files that have resulted from a previously undertaken digitalization process. This institution wants to know which format is most suitable for preserving those files.

The Format Evaluator could be consulted in order determine which format would offer the most appealing set of preservation features. Figure 3 depicts the facts available in our current version of the Format Knowledge Base for the formats JPEG 1.02, TIFF 6 and JPEG 2000. The figure also depicts the overall preservation gain that one would obtain if any of the JPEG 1.02 files were to be transformed to either of the other formats (assuming that the client institution gives equal importance to each of the considered criteria). Figure 3 also outlines the inner workings of the functions used to compare each of the evaluation criteria.

One should also notice in Figure 3 that TIFF 6 has objectively proven to be more apt for preservation than JPEG 2000, fundamentally because it is presented as a more prevalent and stable format (i.e. has a higher market share and life time).

2.3 Object Evaluator

The third evaluation component is the Object Evaluator. This component is responsible for judging the quality of a migration outcome by comparing the objects submitted to migration with its converted counterparts. Again, these evaluations



Format Knowledge Base

Criteria	JPEG	TIFF	JP2	Comparison Function	Comparison results: JPEG>TIFF	Comparison results: JPEG>JP2
Market share	Very high	high	low	Ratio	0.75	0.25
Support level	high	high	high	Ratio	1.0	1.0
Is standard	yes	no	yes	Gain	0.0	1.0
Open specification	yes	yes	yes	Gain	1.0	1.0
Compression support	yes	yes	yes	Not(TargetBoolValue)	0.0	0.0
Lossy compression only	yes	no	no	Not(TargetBoolValue)	1.0	1.0
Transparency Support	no	yes	yes	Implication	1.0	1.0
Embedded metadata	yes	yes	yes	Implication	1.0	1.0
Royalty free	yes	yes	yes	Gain	1.0	1.0
Open source	yes	yes	yes	Gain	1.0	1.0
Backward compatible	yes	yes	no	Gain	1.0	0.0
Documentation level	high	high	high	Ratio	1.0	1.0
Competing formats	yes	yes	yes	Gain	1.0	1.0
DRM support	no	no	yes	Not(TargetBoolValue)	1.0	0.0
Update frequency	3/12	6/26	1/6	1/Ratio	13/12	3/2
Custom extensions	no	yes	yes	Not(TargetBoolValue)	0.0	0.0
Life time	12	26	6	Ratio	26/12	6/12
Transparent decoding	high	medium	medium	Ratio	2/3	2/3
Multiple reader producers	yes	yes	yes	Gain	1.0	1.0
Multiple readers	yes	yes	yes	Gain	1.0	1.0
Open source reader	yes	yes	yes	Gain	1.0	1.0
Multiplatform reader	yes	yes	yes	Gain	1.0	1.0
User assigned weight =					1/22	
Evaluation results (Σ) =					0.893	0.768

Caption

JPEG: JPEG File Interchange Format, version 1.02
TIFF: Tagged Image Format, version 6
JP2: JPEG 2000

Label	Value
Yes	1.00
No	0.00
Unexisting	0.00
Low	0.25
Medium	0.50
High	0.75
Very High	1.00

Comparison Functions

Source	Target	Gain	Implication	$Not_{(src)}$
0	0	0.5	1	1
0	1	0	1	1
1	0	1	0	0
1	1	1	1	0

Ratio (source, target) = $\frac{\text{target}}{\text{source}}$

Fig. 3 The inner workings of the Format Evaluator.

will be performed according to multiple criteria. These criteria, known in this context as significant properties, constitute the set of attributes that should be maintained intact during the preservation intervention [43]. They represent the array of properties that characterise an object as a distinctive intellectual entity independently of the encoding that is being used to represent it.

A different set of significant properties must be compiled for each class of digital objects. Take text documents, for example. One might select properties such as the number of characters in the document, the order of those characters, the page size and the graphical layout. Some of these properties are not applicable to other classes of digital objects, e.g. still images.

One should note that significant properties are not well defined for the majority of object classes. Previous work by Rauch and Rauber [38–41] have shown how complex the process of collecting significant properties can be for some types of digital objects. This process usually entails a careful study of the technical characteristics of the formats within a class of digital objects as well as a perceptive analysis of the features that compose sample objects in that domain.

The Arts and Humanities Data Service [2] and the Library of Congress [26] have been publishing technical reports on distinct classes of digital objects. Within those reports it is possible to find a considerable number of significant properties that one might regard as relevant for evaluation.

The University of Minho is currently collaborating with the Portuguese National Archives[2] in the development of a digital repository capable of preserving authentic digital objects produced by affiliated public administration institutions – the RODA project [36]. During the planning stages of this project, several meetings were held in order to devise a general taxonomy of significant properties for the three classes of digital objects that the repository is expected to handle, i.e. still images, text documents and relational databases.

Work is still underway in order to create a consistent set of significant properties for relational databases as little or no information is available regarding this specific subject. Nonetheless, a considerable list of criteria has already been assembled for the remaining classes (see Table 3 and Table 4). Some of these criteria were suggested by team members during project meetings while others were obtained from sources of information such as [10,11,26,41].

Most criteria outlined in Table 3 and Table 4 are considered fairly objective and easy to evaluate. This means that detecting migration-induced changes in those properties is a rather straightforward task. However, some of these criteria require more delicate approaches. Subjective criteria such as pixel correctness, character correctness, page layout or metadata are supported by more complex data types such as pixel data, textual information or XML. Consequently, special comparison functions must be devised in order to objectively compare those properties.

[2] Instituto dos Arquivos Nacionais/Torre do Tombo

A preliminary study on similarity algorithms for these types of data has already been initiated. So far, we have managed to collect an assortment of promising algorithms for comparing pixel data [27,44,47,52,53,60] and textual information [4,23,24,33,59]. Present work is mostly focused on the test and selection of the most appropriate similarity algorithm for each of these types of data.

The internal architecture of the Object Evaluator is depicted in Figure 4. As stated before, this component is responsible for measuring how objects have been modified by migration software. The output of this component is an evaluation report that, for each significant property, outlines how respectful the converted object is to the original (in a scale of zero to one).

In order to compare objects in different formats one must resort to format parsers, i.e. special software capable of interpreting the structure of a given file format. Once a digital object has been parsed, one must extract each of the significant properties defined in the evaluation taxonomy for that particular class of objects. The collection of extracted properties could be regarded as the canonical (or conceptual) object [30] since information relevant to that object is now free from all technological peculiarities. The last stage in the evaluation process is the application of a set of similarity functions capable of detecting how much the information has changed in relation to the original.

Some subjective criteria can not be evaluated solely by looking at properties explicitly defined in the file format. Some of these criteria take into to consideration an assortment of other properties while others require them to be preprocessed before similarity functions can be applied. Consider the *page layout* criterion in a text document. In order to faithfully compare the graphical layout of two text documents (independently of their encoding), one must first break the documents into a sequence of pages, convert each page into an pixel-based representation with a common resolution (similar to printing the document to an image file instead of a printer) and then apply an image similarity function to each of the resultant images. The resulting similarity ratio must take into account the overall evaluation of all the pages in the document.

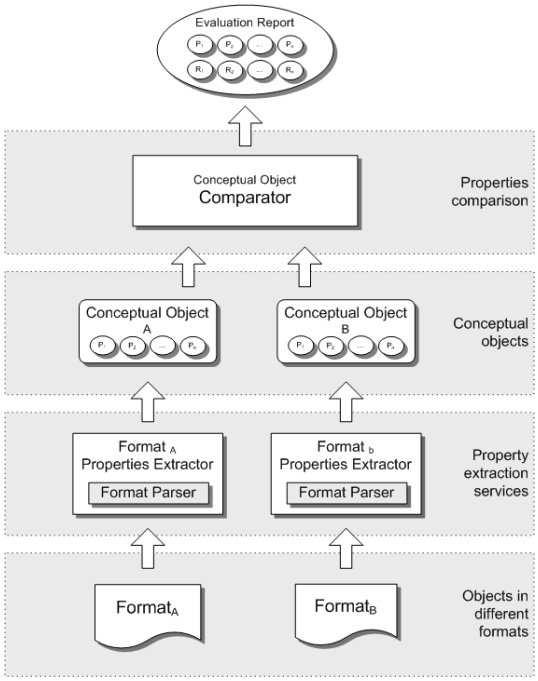


Fig. 4 Internal architecture of the Object Evaluator.

3 Conclusions

This paper proposes a software architecture based on distributed services that may help institutions to carry out migration based preservation interventions. The proposed system enables institutions to cooperate in the edification of a global advisory service that, among other things, is capable of producing recommendations of optimal migration alternatives; perform format migrations (resorting to converter composition whenever necessary); and thoroughly document preservation interventions by generating reports in the form of metadata (in what PREMIS refers to as an Event Entity [37]).

When a client institution requests a recommendation, the system computes an optimal migration option by confronting the preservation requirements outlined by the client institution with the evaluation criteria stored overtime in the Evaluations Repository. After this operation, the client institution may take advantage of the system to carry out the corresponding conversion. In the end of each conversion, the client institution will receive a novel representation of the submitted object and a metadata report that fully describes the procedures undertaken (i.e. the event) and the results of the conversion (i.e. the outcome). This report can then be embedded with the preservation metadata that accompanies the object within the archival environment in order to retain its authenticity.

The current prototype is being supported by Web services technology [17] as it appears to be well-suited for supporting the development of SOAs and due to their open-standard and platform-independent characteristics. It is important to point out that many other protocols could equally be used to implement these ideas. Distinct Remote Procedure Call (RPC) technologies could even be combined and used together with the CRiB as long as gateways or proxies are implemented.

A prototype for the proposed SOA is currently being developed at the University of Minho. Present work is mostly focused on the implementation of similarity functions for subjective criteria in the Object Evaluator. All other components are already built and functional. A Web interface called Migration Workbench [13] has been developed and published which allows users to convert digital objects from and to many different file formats as well as assess the re-

sults of the evaluations performed by the system. Once the Object Evaluator is complete and tested, the suggestion service will also be made available to the general public.

It is important to stress that for evaluation purposes the prototype will only be capable of producing suggestions and converting objects in a limit number of file formats, i.e. just a few formats belonging to the still images and text documents classes.

Some parallel contributions are also expected from this research. When concluded, this SOA will constitute an objective tool for comparing file formats and conversion applications. This work could also contribute to foster new lines of research such as the development or improving of similarity functions for different types of information, e.g. image, text, audio, video or datasets. These similarity functions are necessary to develop a general purpose Object Evaluator capable of handling all types of object classes.

Acknowledgements The work reported in this paper has been funded by the FCT under the grant SFRH/BD/17334/2004.

References

1. Abrams, S.L., Seaman, D.: Towards a global digital format registry. In: World Library and Information Congress: 69th IFLA General Conference and Council (2003)
2. Arts and Humanities Data Service: Ahds repository policies and procedures (2006). URL http://ahds.ac.uk/preservation/ahds-preservation-documents.htm
3. Authenticity Task Force: Requirements for assessing and maintaining the authenticity of electronic records. Tech. rep., InterPARES Project (2002)
4. Cohen, W.W., Ravikumar, P., Fienberg, S.E.: A comparison of string distance metrics for name-matching tasks. In: S. Kambhampati, C.A. Knoblock (eds.) Information Integration on the Web (IIWeb). Acapulco, Mexico (2003)
5. Cullen, C.T., Hirtle, P.B., Levy, D., Lynch, C.A., Rothenberg, J.: Authenticity in a Digital Environment. Council on Library and Information Resources, Washington, DC (2000)
6. Darlington, J.: Pronom - a practical online compendium of file formats. RLG DigiNews **7**(5) (2003)
7. Diessen, R.J.v., Werf-Davelaar, T.v.d.: Authenticity in a digital environment. Report 2, Koninklijke Bibliotheek and IBM (2002)
8. Digital Curation Centre: Oais representation information registry/repository (2006). URL http://dev.dcc.ac.uk/twiki/bin/view/Main/DCCRegRepV04
9. Digital Preservation Testbed: Migration: Context and current status. White paper, National Archives and the Ministry of the Interior and Kingdom Relations (2001)
10. Eadie, M.: Preservation handbook - binary text/word processor documents. Tech. rep., Arts and Humanities Data Service (2005)
11. Eadie, M.: Preservation handbook - bitmap (raster) images. Tech. rep., Arts and Humanities Data Service (2005)
12. Ferreira, M.: Crib - conversion and recommendation of digital object formats web site (2006). URL http://crib.dsi.uminho.pt
13. Ferreira, M.: Crib: Migration workbench (2006). URL http://digitarq.di.uminho.pt/MigrationWorkbench
14. Ferreira, M., Baptista, A.A., Ramalho, J.C.: Crib: A service oriented architecture for digital preservation outsourcing. In: J.C. Ramalho, J.C. Lopes, A. Simões (eds.) XATA - XML: Aplicações e Tecnologias Associadas. Portalegre, Portugal (2006)
15. Freed, N., Borenstein, N.: Multipurpose internet mail extensions (mime) part two: Media types. RFC 2046, Network Working Group (1996)
16. Google: Google trends (2006). URL http://www.google.com/trends
17. Graham, S., Simeonov, S., Boubez, T., Davis, D., Daniels, G., Nakamura, Y., Neyama, R.: Building Web Services with Java: Making Sense of XML, SOAP, WSDL and UDDI. Sams Publishing (2002)
18. Granger, S.: Emulation as a digital preservation strategy. D-Lib Magazine **6**(10) (2000)
19. Harvard University Library: Global digital format registry (2002). URL http://hul.harvard.edu/gdfr
20. Hedstrom, M.: Digital preservation: Problems and prospects. Digital Library Network (DLnet) **20** (2001)
21. Heslop, H., Davis, S., Wilson, A.: An approach to the preservation of digital records (2002)
22. Hofman, H.: Can bits and bytes be authentic? preserving the authenticity of digital objects. In: International Federation of Library Associations Conference. Glasgow (2002)
23. Jaro, M.A.: Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. Journal of the American Statistical Association **84**, 414–420 (1989)
24. Jaro, M.A.: Probabilistic linkage of large public health data files. Statistics in Medicine **14**, 491–498 (1995)
25. Lee, K.H., Slattery, O., Lu, R., Tang, X., McCrary, V.: The state of the art and practice in digital preservation. Journal of Research of the National Institute of Standards and Technology **107**(1), 93–106 (2002)
26. Library of Congress: Digital formats web site (2004). URL http://www.digitalpreservation.gov/formats
27. Lorenzetto, G.P., Kovesi, P.: A phase based image comparison technique. In: Fifth International Conference on Digital Image Computing, Techniques, and Applications. Perth, Australia (1999)
28. Lorie, R.A.: Long term preservation of digital information. In: First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01). ACM, Roanoke, Virginia, USA (2001)
29. Lorie, R.A.: A methodology and system for preserving digital data. In: Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02), pp. 312–319. New York: ACM Press, Portland, Oregon (2002)
30. Lynch, C.: Canonicalization: A fundamental tool to facilitate preservation and management of digital information. D-Lib Magazine **5**(9) (1999)
31. MacNeil, H., Wei, C., Duranti, L., Gilliland-Swetland, A., Guercio, M., Hackett, Y., Hamidzadeh, B., Iacovino, L., Lee, B., McKemmish, S., Roeder, J., Ross, S., Wan, W.k., Xiu, Z.Z.: Authenticity task force report. Tech. rep., InterPARES Project (2001)
32. Mellor, P., Wheatley, P., Sergeant, D.M.: Migration on request, a practical technique for preservation. In: M. Agosti, M.C. Thanos (eds.) ECDL '02: 6th European Conference on Research and Advanced Technology for Digital Libraries, pp. 516–526. Springer-Verlag, London, UK (2002)
33. Monge, A.E., Elkan, C.: The field matching problem: Algorithms and applications. In: Second International Conference on Knowledge Discovery and Data Mining, pp. 267–270. Portland, USA (1996)
34. OASIS: Universal description, discovery and integration (uddi) (2005). URL http://www.uddi.org
35. OASIS SOA Reference Model TC: Oasis reference model for service oriented architectures (working draft 10). Tech. rep., Organization for the Advancement of Structured Information Standards (OAIS) (2005)
36. Portuguese National Archives (Instituto dos Arquivos Nacionais/Torre do Tombo) and University of Minho: Roda (repositório de objectos digitais autênticos) web site (2006). URL http://roda.iantt.pt
37. PREMIS Working Group: Data dictionary for preservation metadata: final report of the premis working group. Final report, OCLC Online Computer Library Center and Research Libraries Group (2005)
38. Rauch, C.: Preserving digital entities - a framework for choosing and testing preservation strategies. Master thesis, Vienna University of Technology (2004)

39. Rauch, C., Pavuza, F., Strodl, S., Rauber, A.: Evaluating preservation strategies for audio and video files. In: DELOS Digital Repositories Workshop. Heraklion, Crete (2005)
40. Rauch, C., Rauber, A.: Preserving digital media: Towards a preservation solution evaluation metric. In: Z. Chen, H. Chen, Q. Miao, Y. Fu, E.A. Fox, E.P. Lim (eds.) International Conference on Asian Digital Libraries, vol. 3334, pp. 203–212. Springer, Shanghai, China (2004)
41. Rauch, C., Rauber, A., Hofman, H., Bogaarts, J., Vedegem, R., Pavuza, F., Ahmer, J., Kaiser, M.: A framework for documenting the behaviour and funcionality of digital objects and preservation strategies. Tech. rep., DELOS Network of Excellence (2005)
42. Rothenberg, J., Commission on Preservation and Access and Council on Library and Information Resources: Avoiding technological quicksand: finding a viable technical foundation for digital preservation: a report to the Council on Library and Information Resources. Council on Library and Information Resources, Washington, DC (1999). By Jeff Rothenberg. ill. ; 28 cm. "January 1999" "Commission on Preservation and Access, Digital Libraries"–Cover.
43. Rusbridge, A.: Migration on request. 4th year project report, University of Edinburgh - Division of Informatics (2003)
44. Rushmeier, H., Ward, G., Piatko, C., Sanders, P., Rust, B.: Comparing real and synthetic images: Some ideas about metrics. In: Eurographics Workshop on Rendering Techniques. Springer, Dublin, Ireland (1995)
45. Shepard, T., MacCarn, D.: The universal preservation format: Background and fundamentals. In: Sixth DELOS Workshop. Tomar, Portugal (1998)
46. Shepard, T., MacCarn, D.: The universal preservation format: A recommended practice for archiving media and electronic records. Tech. rep., WGBH Educational Foundation (1999)
47. Shrestha, B., O'Hara, C.G., Younan, N.H.: Jpeg2000: Image quality metrics. In: American Society for Photogrammetry and Remote Sensing. Baltimore, USA (2005)
48. Stanescu, A.: Assessing the durability of formats in a digital preservation environment. D-Lib Magazine **10**(11) (2004)
49. Task Force on Archiving of Digital Information and Commission on Preservation and Access and Research Libraries Group: Preserving digital information: report of the Task Force on Archiving of Digital Information. Commission on Preservation and Access, Washington, D.C. (1996). Commissioned by the Commission on Preservation and Access and the Research Libraries Group. 28 cm. "May 1, 1996."
50. Thibodeau, K.: Overview of technological approaches to digital preservation and challenges in coming years. In: C.o.L. Resources, Information (eds.) The State of Digital Preservation: An International Perspective. Documentation Abstracts, Inc. - Institutes for Information Science, Washington D.C. (2002)
51. UK National Archives: Pronom - the file format registry (2002). URL http://www.nationalarchives.gov.uk/pronom
52. Wang, L.W., Zhang, Y., Feng, J.F.: On the euclidean distance of images. Ieee Transactions on Pattern Analysis and Machine Intelligence **27**(8), 1334–1339 (2005). 934HW Times Cited:0 Cited References Count:18
53. Wang, Z., Bovik, A.C.: A universal image quality index. Ieee Signal Processing Letters **9**(3), 81–84 (2002). 543AR Times Cited:70 Cited References Count:4
54. Weirich, P., Skyrms, B., Adams, E.W., Binmore, K., Butterfield, J., Diaconis, P., Harper, W.L.: Decision Space: Multidimensional Utility Analysis. Cambridge University Press, Cambridge (2001)
55. Wheatley, P.: Migration: a camileon discussion paper. Ariadne **29** (2001)
56. Wikipedia contributors: Color depth. URL http://en.wikipedia.org/w/index.php?title=Color_depth
57. Wikipedia contributors: Color space. URL http://en.wikipedia.org/w/index.php?title=Color_space
58. Wikipedia contributors: Image compression. URL http://en.wikipedia.org/w/index.php?title=Image_compression
59. Winkler, W.E.: The state of record linkage and current research problems. Tech. rep., U.S. Bureau of the Census (1999)
60. Zhou, H., Chen, M., Webster, M.F.: Comparative evaluation of visualization and experimental results using image comparison metrics. In: IEEE Visualization. Boston, USA (2002)

Table 2 Format-related evaluation criteria.

Criterion name	Description
Market share	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. A high level of adoption is better for preservation purposes.
Support level	The level of technical support on the format given by its official creator. A high level of support is preferred in a preservation context.
Is standard	Whether the format has been published by an official standards organisation. Standard formats are preferred over non-standard ones.
Open specification	Whether format specification can be independently inspected. Open formats are highly recommended in preservation contexts.
Compression support	Whether the format supports any type of compression. Uncompressed formats are generally advocated by the community.
Lossy compression only	Whether the format exclusively supports a lossy type of compression. Lossy compression schemes are highly inadvisable.
Transparency support	Whether the format offers transparency features. This criterion is somewhat specific to certain types of formats (e.g. raster images). If the source format contains transparency features then the target format should be able to accommodate those items.
Embedded metadata	Whether the format contains embedded metadata. The target format should be able to accommodate the source format's embedded metadata.
Royalty free	Whether royalties or license fees have to be requested in order to use or produce the format. Royalty free formats are preferred.
Open source	Whether there are decoders whose source can be independently inspected. The existence of open source decoding software is highly recommended.
Backward compatible	Whether revisions have support for previous versions. Backward compatibility is a desirable feature.
Documentation level	Whether the format specification is well documented. The system favours well documented formats.
Competing formats	Whether competing or similar formats exist. The existence of competing formats makes a format attractive for preservation as information may hereafter be more easily converted.
DRM support	Whether DRM (Digital Rights Management), encryption or digital signatures can be used. Any type of functionality which may hinder access to information is considered inadvisable.
Update frequency	How often a format has been revised since its official release. This criterion is determined according to the following formula: no. releases / (current year - release year). Stable formats are preferred. If revisions happen very often, the archive may not be able keep up.
Supports custom extensions	Whether extensions, such as executable sections or narrowly supported features can be added to the format. Formats that support such features are inadvisable.
Life time	How many years have passed since the format has been officially released. Long lasting formats are commonly preferred over young unestablished formats.
Transparent decoding	The degree to which the digital representation is open to direct analysis using basic tools, e.g. human readability resorting to a text-only editor. Formats that can be easily inspected and/or interpreted are preferred.
Multiple reader producers	Whether readers/viewers are produced by various entities. For preservation purposes one should not rely on readers produced by a single entity.
Multiple readers	Whether the format can be rendered by various pieces of software. For preservation purposes one should not rely on formats that can only be viewed using a specific reader.
Open source reader	Whether the source-code of the reader software can be independently inspected. The existence of open source readers/viewers is a highly desirable feature.
Multiplatform reader	Whether the reader software can be run or has versions for several different platforms (e.g. operating systems or hardware). The existence rendering software for concurrent platforms is a highly desirable feature in a preservation context.

Table 3 Taxonomy of significant properties for still images.

Criterion name	Description
appearance::resolution::width	The width of the digital image measured in pixels.
appearance::resolution::height	The height of the digital image measured in pixels.
appearance::colour::model	Abstract mathematical model describing the way colors can be represented as tuples of numbers, typically as three or four values or color components (e.g. RGB, sRGB, HSL, HSV, YUV, CMYK) [57].
appearance::colour::depth	The number of bits used to represent the color of a single pixel in a bitmapped image or video frame buffer [56].
content::completeness::pixel_correctness	How well pixels are respectful to the original image. In cases of multiple page images, the comparison is performed by page and an overall similarity value is calculated by averaging the whole page set results.
content::completeness::page_count	The number of pages that constitute the image.
context::metadata	Some image formats embed metadata. This criterion intends to measure how much of that metadata has been preserved.
structure::compression::method	Image compression can be lossy or lossless. Examples of lossless image compression methods are: run-length encoding, entropy coding and adaptive dictionary algorithms such as LZW. Examples of methods for lossy compression are: reducing the color space to the most common colors in the image, chroma subsampling, transform coding and Fractal compression [58].
structure::compression::level	The level compression used in the object. The value zero is used when lossless compression methods are in place.

Table 4 Taxonomy of significant properties for text documents.

Criterion name	Description
appearance::page::width	The width of the document measured in millimeters in relation to the original document.
appearance::page::height	The height of the document measured in millimeters.
appearance::page::layout	How similar the layout of the text-document is in relation to the original.
appearance::page::margins::left	The size of the left margin of the document.
appearance::page::margins::right	The size of the right margin of the document.
appearance::page::margins::top	The size of the top margin of the document.
appearance::page::margins::bottom	The size of the bottom margin of the document.
appearance::page::style::background color	The predominant background color of the document.
appearance::page::style::font faces	Collection of fonts used throughout the document.
content::completeness::character correctness	How well text characters are respectful to the original document. In cases of multiple page documents, the comparison should be performed by page and an overall similarity value calculated by averaging the whole page set results.
content::completeness::page count	The number of pages that constitute the text document.
content::completeness::image count	The number of images embedded in the text document.
context::metadata	Some text document formats carry embedded metadata. This criterion is expected to determine how much of that metadata has been preserved.