Automatic Evaluation of Migration Quality in Distributed Networks of Converters

Miguel Ferreira

Department of Information Systems,
University of Minho, Guimarães, Portugal
mferreira@dsi.uminho.pt

Abstract. Migration has always played an important role in digital preservation. The most recent developments in this context introduce networks of remotely distributed converters. In this paper we propose an extension to current migration networks to enable users to: a) determine to what extent have the essential characteristics of digital objects been preserved during a migration; b) generate detailed migration reports for inclusion in the objects' preservation metadata; and c) provide advice on the available migration paths or target formats that will best suit users' requirements.

1 Introduction

A significant part of current intellectual output is being created in some sort of digital form. The simplicity by which this type of material can be created and disseminated over current computer networks, combined with the high quality of results are significant factors that lead to the massive adoption of digital authoring tools. However, digital material suffers from a structural problem that puts its longevity at risk. Although a digital document may be copied an infinite number of times without loosing quality, it requires a technological environment in order to transform its computational representation (i.e. bit stream) into something that human beings can understand. This dependency on technology makes this type of material vulnerable to the rapid and unpredictable obsolescence that affects all technological artifacts [1].

The usable lifetime of a digital document is, in general, several times lower than that of a similar document stored in an analogue medium. A paper document, for instance, can be preserved for more than 100 years and a microfilm, if stored in appropriate conditions, will still be readable in 500 years time [2]. On the other hand, the life expectancy of a modern technological platform may be as short as 5 years [3-6].

Digital preservation is here defined as the set of processes and activities that ensure the continued access to information and all kinds of cultural heritage existing in digital formats [7]. In this context, a digital object is "(…) an information object, of any type of information or any format, that is expressed in digital form" [8]. This definition is general enough to accommodate both information that was created within a technological environment (born-digital objects) and information that was obtained

This work was supported by the FCT under the grant SFRH/BD/17334/2004.

from analogue sources (digitalised objects). Text documents, digital photographs, vector graphics, databases, video/audio sequences, virtual reality models, Web pages and computer games or software applications, are a just few examples of what may be considered digital objects.

This paper is organised as follows: section 2 provides some background on the most prominent digital preservation strategies; section 3 describes the proposed research where a special emphasis is put on the research questions, methodology and expected contributions; section 4 outlines a few points for discussion during the doctoral consortium; section 5 summarises the proposed research.

2 Related work

Over the last decade, a considerable number of strategies have been proposed aiming to solve the problem of digital preservation and technological obsolescence. According to Lee et al. the available strategies may be aggregated into 3 major categories: emulation, encapsulation and migration[1] [9].

2.1 Preservation strategies

The emulation strategy consists in the use of special software to reproduce the behaviour of a given hardware/software platform in a different technological environment [10]. This strategy allows the user to interpret digital objects by running the same software that was used in their creation [9, 10]. Emulation plays an important role in preservation, especially in the interpretation of highly dynamic and/or interactive objects, such as software [11].

The encapsulation strategy aspires to solve the problem of digital preservation by storing together with the object, enough information about how it should be interpreted [12]. This information may consist of a formal specification of the object's format which will enable the future development of viewers, emulators or converters. Lorie argues that this formal specification should be expressed in some sort of machine language compiled for a virtual machine [13, 14]. If the virtual machine is sufficiently well documented, it should be possible to recreate it in a future platform. The encapsulation strategy is usually applied to collections of objects which are expected to remain unused for long periods of time. Using encapsulation on such collections may reduce the number of preservation interventions leading to a generalised reduction in preservation costs.

The migration strategy consists of 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." [3]. Contrary to other preservation strategies, migration does not intend to maintain the digital object in its original format. Alternatively, it

[1] In this paper, the term *conversion* will be used to address a one step transformation between two distinct formats and the term *migration* to describe transformations that involve multiple sequential conversions.

converts the object from a near obsolete format to a format that users are able to interpret with their up-to-date computers. The main disadvantage of this approach relies on the fact that when an object is migrated, some of its properties may not be completely or adequately transferred to the target format, i.e. some sort of data loss may be experienced. The reason for this is twofold: there may be structural incompatibilities between the source and the target formats or the converter may be faulty and therefore incapable of performing its task appropriately.

There are a considerable number of migration variants. The most relevant ones comprise normalisation, migration on request and distributed migration.

Normalisation consists in the conversion of digital objects from a wide range of distinct formats to a smaller and more manageable set of formats. The rationale behind it is based on the idea that by reducing the number of formats in custody, subsequent preservation interventions will result simplified as the same strategy may be applied to a higher number of objects. Normalisation is usually performed by the digital archive during the ingest process of new material or by the material's producer before any archival process is even initiated [6, 15].

Migration on request aims at reducing the data loss problem usually associated with migration strategies by applying all the necessary transformations to the original objects instead of converting any of the derivatives that have resulted from previous migrations [16]. In this strategy the focus is not on the preservation of the digital object but in making sure that the migration tool remains usable over time [9].

2.2 Distributed migration

The most recent endeavours in the field of migration propose architectures of remotely distributed converters. The Typed Objects Model proposed by Ocklerbloom describes a network of mediator agents that assist users in the discovery and execution of object conversions [17]. Other initiatives explore similar concepts but resorting to more widely spread technologies. The Lister Hill National Center for Biomedical Communications has developed a Web Service which enables anyone to convert objects from fifty different formats to PDF [18]. Hunter and Choudhury propose a distributed network of Web Services that provide advanced migration services to users [19]. This network is supported by a semantic layer which enables software clients to autonomously find and invoke desired migrations.

This type of migration introduces several advantages over the more traditional solutions: the use of Web Services hides the complexity of specific converters as well as the peculiarities of the technological platform that supports them; the development of redundant services insures that the network remains functional during situations of partial break down; and the existence of multiple migration paths enables this solution to cope with the gradual disappearing of converters. Moreover, this approach is compatible with several variants of migration, such as migration on request and normalisation.

The creation of a global network of converters may also result in a generalised reduction in preservation costs. Any willing organisation will be able to gain profit over its investments by publishing their own converters on the migration network and charging a small fee for its utilisation.

2.3 Evaluation of preservation strategies

Although the number of proposed strategies for digital preservation continues to grow none of them has been adequately proved or universally accepted [20]. A preserver's decision for or against a specific preservation strategy is usually accompanied by certain amount of discomfort. Unfortunately, his/her decision is not only conditioned by the lack of universal acceptance. Several factors such as the satisfaction of the designated community, the characteristics of the collection or the costs involved in the preservation need to be taken into account [20].

Rauch and Rauber have built a conceptual framework which allows preservers to evaluate, compare and select preservation strategies according to their individual requirements [20-22]. Their work is based on the ideas of Utility Analysis [23], which was originally developed for infrastructure and economic research projects. The framework follows a process which is composed by the following steps:

1) It begins with the creation of an objective tree where many different preservation-aware criteria are assembled and structured in a hierarchical way. These criteria usually fall into one of the following top level categories: file characteristics, preservation processes and preservation costs (Fig. 1);
2) In step two, units of measurement such as millimetre, second or EURO are assigned to each of the criteria;
3) Step three consists in bringing together a reasonable number of alternatives that could be used to preserve the collection of digital objects. These alternatives will later be compared and ranked according to the users' preservation objectives;
4) In step four each of the selected preservation alternatives is executed against a representative set of objects. The outcome of this process is evaluated according to all the criteria and recorded for later use;
5) In step five the measured values are transformed into comparable numeric units;
6) In step six the preserver weights each of the criteria in the objective tree according to his/her requirements;
7) Step seven consists in the aggregation of partial and total values resultant from the experiments;
8) Finally, in step eight, all the alternatives are ranked and the preserver is able to identify the optimal solution for his/her preservation problem.

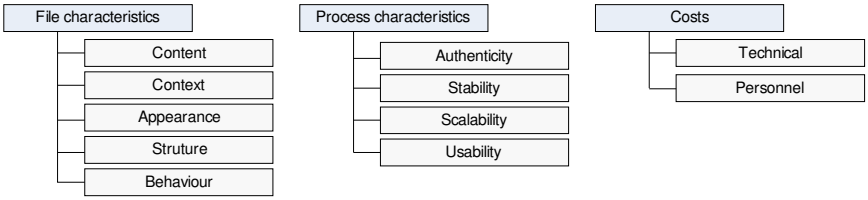

Fig. 1 - Top two levels of a generic objective tree.

3 Proposed research

In a previous section it was described how networks of distributed converters may play an important role in the automation of preservation processes. In such networks it is likely to exist different sequences of converters that originate objects in the same target format. However, distinct migration paths are likely to produce considerably different objects (Fig. 2). Moreover, different users are likely to have different expectations regarding the results of a preservation intervention. For instance, some users may be interested in preserving most of the characteristics of a digital object, while others, with a more limited budget, may have to compromise and give up on some of the characteristics in order to reduce preservation costs.

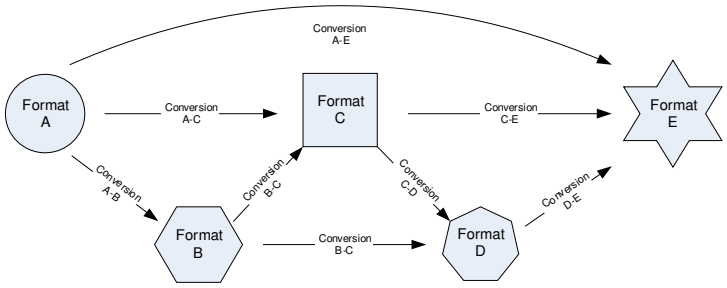

Fig. 2 - Multiple migration paths between two distinct object formats. The different geometric shapes intend to represent changes in the object's characteristics after each conversion.

In this context, we propose an extension to current migration networks based on a new set of services which will help users perform their preservation tasks more effectively by:

a) Determining the amount of data loss occurred in a migration;
b) Generating detailed migration reports in forms appropriate to be appended to the object's preservation metadata;
c) Providing recommendations on which migration paths or target formats will best fulfil users' preservation requirements.

In addition, the system will be able to scale smoothly by supporting the inclusion of new quality indicators and new migration services.

Fig. 3 depicts the top level components of the proposed system. The Metaconverter component constitutes the interface between the user and the proposed system. It is responsible for the aggregation of information about available migration networks (service registry), identification of transitive conversions on those networks, invocation of appropriate converters and the registration of information about performed migrations (e.g. cost, time, reliability).

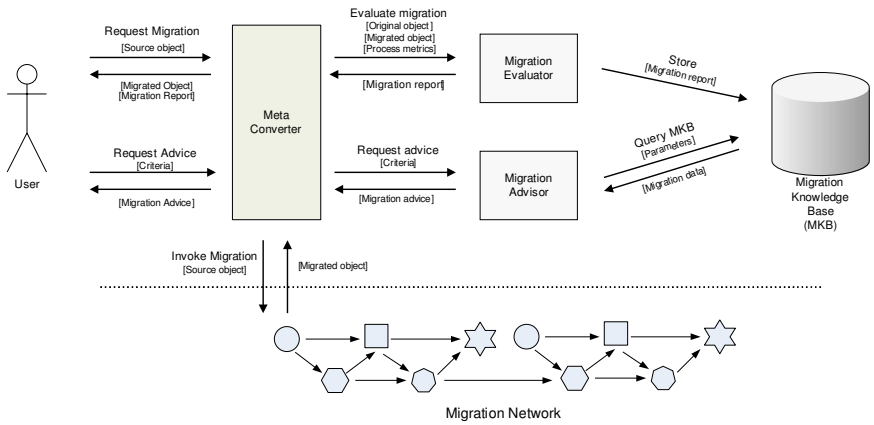

Fig. 3 - Extension to current migration networks.

The Migration Evaluator is responsible for automatically detecting differences in the essential characteristics of objects and its converted counterparts. Additionally, the Evaluator will consider the information about the conversion processes produced by the Metaconverter and combine both to generate quality reports on each migration path. These reports will be processed and stored in the Migration Knowledge Base. They will also be delivered to users, allowing them to acknowledge the quality of a specific migration and to adequately document the preservation intervention. At the moment we expect to base these reports on the PREMIS Data Dictionary for Preservation Metadata [24].

The digital objects and their converted versions will be compared and evaluated according to multiple criteria. Rauch and Rauber [20-22] have identified several quality criteria for at least three types of digital objects: e-journal documents (63 criteria), audio files (136 criteria) and video sequences (324 criteria). Examples of such criteria may be expressed in the form of questions: how much of the content has been preserved?; which appearance items were not effectively preserved in the conversion (e.g. font style, colour, frame rate, bit rate)?; is the target object bigger than the source object?; what is the stability/support/openness of the target format?; etc.

The Migration Adviser is responsible for generating appropriate advice on the migration paths or destination formats that will best suit the preservation requirements of the user. It accomplishes this by confronting the users' set of criteria with the information collected overtime stored in the Migration Knowledge Base.

3.1 Research questions

We claim that is possible to develop a system to assist users in the selection and execution of preservation interventions based on migration strategies. The proposed system will be able to operate autonomously and integrate seamlessly with current and future distributed migration networks. The system will be available as a Web

Service enabling any digital repository system to expand its functionalities by remotely invoking the service.

The central research questions in this work are as follows:

- Is it feasible to design and implement a system to automatically determine the amount of data loss occurred in a migration as well as generating migration reports in appropriate forms to be appended to the object's preservation metadata?
- Is it feasible to design and implement a system which, based on previously collected information about the quality of each migration path on a network of converters, is able to help users choose the most appropriate migration path or target format to solve their preservation problem?

These research questions are closely related to the Migration Evaluator and the Migration Advisor components of the system described earlier.

3.2 Research methodology

We intend to prove the concepts expressed in our research questions (automatic quantification of data loss and automatic recommendation of appropriate migration strategies) by developing a prototype of the described system and by running a controlled set of experiments in order to empirically attest its viability.

3.3 Proposed experiments

Rauch and Rauber have conducted several experiments to assess the effectiveness of their framework for evaluating preservation strategies [20-22]. These experiments were conducted at different institutions which had collections of objects that they wanted to preserve. Using their evaluation framework Rauch and Rauber were able to compare several distinct preservation approaches and rank them according to the specific preservation objectives of each institution. Unfortunately, the collections of objects used in these experiments are not publicly available; therefore, we intend to reproduce some of these experiments in order to collect useful empirical data which will enable us to validate our system.

We expect to validate the concept of automatic quantification of data loss occurred in a migration by comparing the evaluation reports produced by the Migration Evaluator component with analogous work carried out by a group of human experts. The experiment consists in assembling a collection of objects, migrate these objects to different formats using distinct migration paths and have a group of experts look at the outcome in order to evaluate the amount of data loss occurred in each of the essential characteristics of the migrated objects. Human evaluations will them be compared with the evaluations performed by the Migration Evaluator component.

The concept of automatic recommendation of appropriate migration strategies will be validated by applying the evaluation framework proposed by Rauch and Rauber to

a collection of objects and a set of available migrations and comparing its results with the recommendations produced by the Migration Advisor.

3.4 Key contributions

The main contributions that may result from this research are:

- **For individual preservers, digital archives and libraries:** the ability to invoke remote conversion services, receive detailed migration reports for inclusion in the preservation metadata, compare different migration options and obtain advice on which migration paths or destination formats are best suited for their specific preservation requirements;

- **For designers and programmers of converters:** the possibility of publishing and selling their conversion applications. The converters have to be externalised as Web Services and registered in the Metaconverter. Afterwards, the system will automatically test and rank all new conversion services. These results will be accumulated in the Migration Knowledge Base in order to generate advice on the best available conversions.
- **For metadata schema creators and users:** sharing experiences on practical utilisation of recently created metadata schemas may contribute to an increased adoption by the user community. It may also help creators to improve future versions of those schemas and accelerate the development of XML bindings.

4 Discussion

By attending this Doctoral Consortium we expect to obtain feedback from senior researchers on the following topics:

- **Relevance**: is this research considered relevant from an international perspective?

- **Research methodology:** is the research methodology adequate to validate the proposed research?

- **Architecture:** does the proposed architecture appears to be technically sound?

- **Services description:** what technology do you recommend to implement the service registry and migration network description (e.g. OWL, Topics Maps, RDF, relational database)?

- **Format registry:** which format registry standard do you recommend to uniquely identify a digital object's format (e.g. MIME types, TOM type descriptors, Global Digital Format Registry, etc.)?

- **Preservation metadata standard:** which preservation metadata schema do you recommend to accommodate the migration evaluation reports produced by the Migration Evaluator component (e.g. PREMIS data dictionary)?

5 Conclusions

In this paper we propose a system which will enable users to select optimal migrations or target formats from a distributed network of conversion services. It accomplishes this, by automatically comparing, according to multiple criteria, the converted objects with their original counterparts and by recording information about the migration process it self. The results of these evaluations will be accumulated in a knowledge base enabling the calculation of recommendations about which migrations are more capable of fulfilling the preservation requirements of the user.

The documentation of preservation interventions is a fundamental tool for the assessment of authenticity of the preserved objects. For that reason, the system will be able to generate detailed migration reports in appropriate forms for inclusion in the objects' preservation metadata.

References

1. Chen, S.-S., *The Paradox of Digital Preservation.* IEEE Computer, 2001. **34**(3): p. 24-28.
2. Burkel, R., *The Role of Microfilm in Information Management.* Information Management Journal, 2003. **37**(1): p. 58-65.
3. 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*. 1996, Washington, D.C.: Commission on Preservation and Access. 59.
4. Conway, P., *Digitizing Preservation: paper and microfilm go electronic.* Library Journal, 1994. **119**: p. 42-45.
5. Graham, P. *Preserving the Digital Library*. in *Long term preservation of electronic materials - a JISC/British Library Workshop*. 1995. University of Warwick: British Library Research and Development Department.
6. Hedstrom, M., *Digital Preservation: A time bomb for digital libraries.* Computers and the Humanities, 1998. **31**: p. 189-202.
7. Webb, C., *Guidelines for the Preservation of Digital Heritage*. 2003, United Nations Educational Scientific and Cultural Organization - Information Society Division.
8. Thibodeau, K. *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years*. in *The State of Digital Preservation: An International Perspective*. 2002. Washington D.C.: Documentation Abstracts, Inc. - Institutes for Information Science.
9. Lee, K.-H., et al., *The State of the Art and Practice in Digital Preservation.* Journal of Research of the National Institute of Standards and Technology, 2002. **107**(1): p. 93-106.

10. 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*. 1999, Washington, DC: Council on Library and Information Resources. vi, 35 p.
11. Woodyard, D. *Digital Preservation: The Australian Experience*. in *Third Conference Digital Library: Positioning the Fountain of Kowledge*. 2000. Malaysia.
12. Digital Preservation Testbed, *Migration: Context and Current Status*. 2001: The Hague.
13. Lorie, R.A. *Long Term Preservation of Digital Information*. in *First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01)*. 2001. Roanoke, Virginia, USA: ACM.
14. Lorie, R.A. *A Methodology and System for Preserving Digital Data*. in *Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02)*. 2002. Portland, Oregon: New York: ACM Press.
15. Hodge, G. and E. Frangakis, *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice*. 2004, International Council for Scientific and Technical Information & CENDI.
16. Mellor, P., P. Wheatley, and D.M. Sergeant. *Migration on Request, a Practical Technique for Preservation*. in *ECDL '02: 6th European Conference on Research and Advanced Technology for Digital Libraries*. 2002. London, UK: Springer-Verlag.
17. Ockerbloom, J.M., *Mediating Among Diverse Data Formats*, in *School of Computer Science*. 1998, Carnegie Mellon University: Pittsburg. p. 164.
18. Walker, F.L. and G.R. Thoma. *A Web-Based Paradigm for File Migration*. in *IS&T's 2004 Archiving Conference*. 2004. San Antonio, Texas, USA.
19. Hunter, J. and S. Choudhury. *A Semi-Automated Digital Preservation System based on Semantic Web Services*. in *Joint ACM/IEEE Conference on Digital Libraries (JCDL'04)*. 2004: ACM.
20. Rauch, C. and A. Rauber. *Preserving Digital Media: Towards a Preservation Solution Evaluation Metric*. in *International Conference on Asian Digital Libraries*. 2004. Shanghai, China: Springer.
21. Rauch, C., et al. *Evaluating preservation strategies for audio and video files*. in *DELOS Digital Repositories Workshop*. 2005. Heraklion, Crete.
22. Rauch, C., *Preserving Digital Entities - A Framework for Choosing and Testing Preservation Strategies*, in *Institute for Software Technology and Interactive Systems*. 2004, Vienna University of Technology: Vienna.
23. Weirich, P., et al., *Decision Space: Multidimensional Utility Analysis*, ed. Cambridge University Press. 2001, Cambridge.
24. Caplan, P., et al., *Data Dictionary for Preservation Metadata*. 2005, PREMIS Working Group (OCLC/RLG).

Appendix: Advisors' statement

Miguel Ferreira is currently working on his Ph.D. project within the "Information Society Research Group". His work is being supervised by Ana Alice Baptista from the Information Systems Department and José Carlos Ramalho from the Computer Science Department.

Before his Ph.D., Miguel has worked with José Carlos Ramalho in the design and implementation of the DigitArq system: a system to support digital archives (creation and management of finding aids) which is currently being deployed in dozens of public institutions in Portugal. This project has recently been awarded with the Information Society Prize.

After the DigitArq project he started working with Professor Ana Alice Baptista as a researcher in the field of Open Access Repositories. During this time he developed several add-ons for the DSpace platform, some of which are being integrated in the official distribution.

The experience and knowledge acquired in the DigitArq and DSpace projects enabled him to create an interesting proposal for a Ph.D. project in the area of Digital Preservation. At the moment Miguel has a sound knowledge of existing solutions and approaches to Digital Preservation. His work will follow the implementation of the system described in this paper.

The research group where he is working has a wide range of research interests, covering works from computer science to social science. At the moment Miguel is part of a team of four elements who share doubts and knowledge as they pursue their research goals.