Avaliação Automática de Migração em Redes Distribuídas de Conversores

Miguel Ferreira
Departamento de Sistemas de Informação, Universidade do Minho, Guimarães, Portugal
mferreira@dsi.uminho.pt

Ana Alice Baptista[1], José Carlos Ramalho[2]
Departamento de Sistemas de Informação[1]/Informática[2], Universidade do Minho, Guimarães[1]/Braga[2], Portugal
analice@dsi.uminho.pt[1], jcr@di.uminho.pt[2]

Resumo

Ao longo dos últimos anos têm vindo a ser propostas diversas estratégias que procuram solucionar o problema da preservação digital. Entre estas destacam-se a emulação, o encapsulamento e a migração. No domínio da migração, as mais recentes iniciativas introduzem redes distribuídas de conversores. Neste contexto, propõe-se uma extensão às actuais redes de migração, assente num novo conjunto de serviços, que irá permitir ao utilizador realizar as suas tarefas de preservação mais eficazmente. O novo conjunto de serviços irá ser capaz de, automaticamente, determinar a quantidade de informação perdida numa dada migração, gerar relatórios detalhados de migração para anexação à metainformação de preservação dos objectos digitais convertidos e produzir recomendações acerca dos formatos de destino ou os caminhos de migração mais adequados aos requisitos de preservação de cada utilizador.

Palavras-chave: Preservação digital, emulação, encapsulamento, migração, qualidade de migração, Web services, Arquitecturas orientadas ao serviço.

1 Introdução

Desde a invenção da escrita que existe uma preocupação incessante por parte do ser humano em preservar os artefactos que resultam dos seus processos intelectuais e criativos [Proença & Lopes 2004]. A preservação desses artefactos permite às gerações futuras compreender e contextualizar a história e cultura dos seus povos [Lee, Slattery, Lu, Tang & McCrary 2002]. Os museus, as bibliotecas e os arquivos assumem neste contexto um papel determinante responsabilizando-se pela preservação e longevidade desses artefactos.

Uma parte significativa da produção intelectual actual é produzida em suportes digitais. A simplicidade com que este tipo de material pode ser criado e disseminado através das modernas redes de comunicação, aliada à qualidade dos resultados obtidos, constitui um factor determinante na adopção de ferramentas de autoria digitais. Contudo, este tipo de material é geralmente acompanhado de um problema estrutural que coloca em risco a sua longevidade. Embora um documento digital possa ser copiado infinitas vezes sem perder qualidade, este requer um contexto tecnológico, hardware e software, que possibilite a sua apresentação de forma inteligível ao ser humano. Esta dependência tecnológica torna-o vulnerável à rápida obsolescência a que a tecnologia está sujeita.

Designa-se, assim, por preservação digital o conjunto de actividades ou processos responsáveis por garantir o acesso continuado, a longo-prazo, à informação e restante herança cultural existente em formatos digitais [Webb 2003]. Neste contexto, considera-se um objecto digital todo e qualquer objecto de informação que possa ser representado através de uma sequência de dígitos binários (*bit stream*) [Thibodeau 2002]. Esta definição é suficientemente alargada para acomodar tanto, informação que nasceu num contexto

tecnológico digital (objectos nado-digitais), como informação digital obtida a partir de suportes analógicos (objectos digitalizados). Documentos de texto, fotografias digitais, diagramas vectoriais, bases de dados, sequências de vídeo e áudio, modelos de realidade virtual, páginas Web, e jogos ou aplicações de software são apenas alguns exemplos do que pode ser considerado um objecto digital.

Este artigo está organizado da seguinte forma: na secção 1 são introduzidos alguns conceitos associados ao domínio da preservação digital; na secção 2 é apresentado o trabalho relevante que está relacionado com esta investigação; a secção 3 descreve a investigação em curso; na secção 4 são descritas as experiências a realizar e a metodologia de investigação; finalmente, na secção 5 são tecidas algumas conclusões.

2 Trabalho relacionado

Ao longo dos últimos anos têm vindo a ser propostas na literatura diversas estratégias com o objectivo de minimizar o problema da preservação digital. Segundo Lee et al., as diversas estratégias de preservação podem ser agrupadas em três classes fundamentais: emulação, encapsulamento e migração [Lee et al. 2002].

A emulação consiste na utilização de um software especial, designado por emulador, de forma a reproduzir o comportamento de uma plataforma de hardware e/ou software numa outra, à partida incompatível. O recurso a emuladores possibilita a interpretação dos objectos digitais no contexto tecnológico em que foram criados [Rothenberg, Commission on Preservation and Access & Council on Library and Information Resources 1999]. A grande vantagem desta abordagem está na capacidade de reproduzir com elevado grau de fidelidade a funcionalidade e apresentação do objecto original [Lee et al. 2002; Rothenberg et al. 1999]. Os recurso a emuladores está geralmente associado à preservação de objectos digitais complexos com propriedades dinâmicas e/ou interactivas como é caso do software.

A estratégia de encapsulamento consiste em preservar, juntamente com o objecto digital, toda a informação necessária e suficiente que permita o futuro desenvolvimento de conversores, visualizadores ou emuladores. Esta informação poderá consistir, por exemplo, numa descrição formal e detalhada do formato do objecto preservado. Uma iniciativa relevante neste contexto consiste na Data Format Description Language, uma linguagem XML que permite descrever qualquer tipo de formato digital [Westhead, Wen & Carroll 2003]. Raymond Lorie propõe uma alternativa a esta estratégia substituindo a especificação formal por uma aplicação de software compilada para uma máquina virtual universal, e.g. Java Virtual Machine [Lorie 2002]. Esta aplicação tem como finalidade apresentar uma visão lógica do objecto permitindo uma navegação simples pelas suas propriedades. Esta estratégia é geralmente utilizada quando os objectos digitais não serão requisitados durante longos períodos de tempo. Procura-se, deste modo, reduzir os custos de preservação, adiando ao máximo a necessidade de uma intervenção de preservação.

A migração consiste na "(…) transferência periódica de material digital de uma dada configuração de hardware/software para uma outra ou de uma geração de tecnologia para outra subsequente" [Task Force on Archiving of Digital Information, Commission on Preservation and Access & Research Libraries Group 1996].

Na migração, ao contrário das estratégias anteriormente descritas, os objectos digitais não são conservados nos seus formatos originais. Esta estratégia tem como objectivo fundamental preservar o conteúdo intelectual do objecto e não a sua estrutura. A migração recorre a software especial, designado por conversor, para transformar os objectos existentes em formatos obsoletos em objectos compatíveis com as tecnologias mais actuais. A principal vantagem desta abordagem consiste possibilidade de um utilizador comum ser capaz de interpretar os objectos

digitais sem necessitar de quaisquer artefactos adicionais para além do software já existente no seu computador pessoal. No entanto, durante uma migração algumas das propriedades que definem o objecto digital poderão não ser correctamente transferidas para o formato de destino . Isto deve-se, sobretudo, a incompatibilidades existentes entre os formatos de partida e destino ou devido a conversores mal comportados.

2.1 Migração distribuída

Os mais recentes desenvolvimentos no contexto da migração introduzem arquitecturas distribuídas de conversores. O Typed Objects Model sintetiza um sistema distribuído de conversores, suportado por uma taxionomia de tipos e formatos de objectos, que recorre a agentes mediadores para descobrir e executar conversões entre formatos [Ockerbloom 1998]. Outras iniciativas recorrem a tecnologias mais comuns como os Web Services [Walker & Thoma 2004]. Hunter e Choudhury dão um passo em frente propondo uma rede de conversores suportada por uma camada semântica que possibilita a sua invocação automática por agentes de software [Hunter & Choudhury 2004].

Este tipo de migração apresenta vantagens consideráveis face às estratégias mais convencionais: a utilização de Web Services permite esconder as especificidades de cada conversor e da plataforma que o suporta; a criação de serviços redundantes assegura a fiabilidade do sistema perante situações de ruptura parcial; e a existência de múltiplos caminhos de migração permite que a solução resista ao desaparecimento gradual de parte dos conversores. Este tipo de abordagem é também compatível com uma série de variantes de migração, como por exemplo, normalização e migração a-pedido. Paralelamente, a criação de uma rede global de migradores poderá conduzir a uma redução generalizada dos custos de preservação. Qualquer organização poderá rentabilizar os seus investimentos no desenvolvimento de conversores publicando-os na rede de serviços e cobrando uma taxa pela sua utilização.

2.2 Avaliação de estratégias de preservação

Apesar do número de estratégias propostas não parar de aumentar, nenhuma destas foi até agora devidamente validada ou universalmente aceite [Rauch & Rauber 2004]. A preferência por qualquer uma das alternativas exige, geralmente, que variados factores sejam tomados em consideração, como por exemplo, as características da colecção, a satisfação da comunidade de interesse ou os custos associados ao processo de preservação [Rauch & Rauber 2004].

Rauch e Rauber desenvolveram um método de avaliação capaz de comparar e seleccionar as alternativas de preservação que melhor satisfazem as necessidades individuais de cada organização/utilizador [Rauch, Pavuza, Strodl & Rauber 2005; Rauch & Rauber 2004]. O seu trabalho é baseado em conceitos de Análise de Utilidade [Weirich et al. 2001], um método originalmente desenvolvido para auxiliar a tomada de decisões em projectos complexos nos domínios da construção civil e economia. O método segue um processo composto pelas seguintes etapas:

1. Começa com a criação de uma árvore de objectivos onde diversos critérios de preservação são compilados e organizados de uma forma hierárquica.
2. Numa segunda fase são associadas unidades de medida a cada critério de preservação, e.g. milímetro, segundo, Mb/s, EURO, etc.;
3. A terceira fase consiste na selecção de um conjunto considerável de alternativas que poderão ser utilizadas para preservar a colecção de objectos digitais. Estas alternativas serão posteriormente comparadas e ordenadas de acordo com as necessidades específicas da organização;

4. No quarto passo, cada uma das alternativas é aplicada a um conjunto representativo de objectos digitais. O resultado de cada intervenção será então avaliado à luz de cada um dos critérios que constam da árvore de objectivos;
5. No quinto passo, os resultados das avaliações são transformados e normalizados em unidades numéricas comparáveis;
6. No sexto passo são atribuídos pesos percentuais a cada um dos critérios e nós que constituem a árvore de objectivos;
7. O passo sete consiste na agregação dos valores totais e parciais obtidos das experiências;
8. Finalmente, todas as alternativas são ordenadas de acordo com os pesos atribuídos a cada um dos critérios que constam da árvore de objectivos.

3 Investigação em curso

Numa secção anterior foi descrita a forma como redes distribuídas de conversores poderão contribuir para a automatização dos processos de preservação. Neste tipo de redes será expectável que diversas sequências de conversores sejam capazes de efectuar o mesmo tipo de migração. Porém, caminhos de migração distintos poderão originar objectos consideravelmente diferentes. Isto reflecte a incapacidade dos conversores de preservar fielmente a totalidade das propriedades que caracterizam os objectos digitais. Neste contexto, propõe-se uma extensão às actuais redes de migração, assente num novo conjunto de serviços, que irá permitir às organizações realizar as suas tarefas de preservação mais eficazmente. O novo conjunto de serviços irá permitir: 1) determinar a quantidade de informação perdida numa migração; 2) gerar relatórios detalhados de migração para anexação à metainformação de preservação dos objectos digitais; e 3) produzir recomendações acerca dos formatos de destino ou os caminhos de migração mais adequados às necessidades de cada organização.

Adicionalmente o sistema proposto será capaz de crescer graciosamente permitindo a inclusão de novos serviços de migração, bem como, novos indicadores de qualidade.

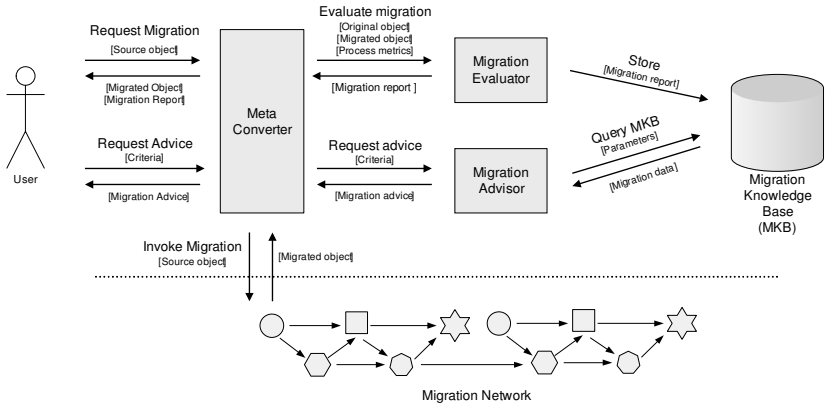

Figura 1 – Arquitectura do sistema distribuído de migração.

A Figura 1 descreve os principais componentes que constituem o sistema proposto. O componente Metaconverter constitui a interface entre o utilizador e o sistema de migração. Este é responsável pela agregação de informação sobre todas as redes de migração e seus serviços de conversão (i.e. service registry), identificação de conversões transitivas nessas redes, invocação

de conversores e registo de informação sobre o processo de conversão (e.g. custo, duração, confiabilidade, etc.).

O Migration Evaluator é responsável pela detecção automática de perdas de informação nas características essenciais dos objectos digitais convertidos e combinar esta informação com a recolhida pelo Metaconverter acerca dos processos de migração de forma a gerar relatórios de migração para cada caminho de migração. Estes relatórios serão acumulados na Migration Knowledge Base e simultaneamente devolvidos ao utilizador para que este possa aferir a qualidade da respectiva migração. Estes relatórios serão baseados no PREMIS Data Dictionary for Preservation Metadata [Caplan et al. 2005] de forma a facilitar a sua inclusão na metainformação de preservação. Os objectos convertidos serão comparados segundo múltiplos critérios. Rauch e Rauber identificaram, para algumas classes de objectos digitais, um conjunto alargado de propriedades significativas que deverão ser consideradas: documentos de texto (63 critérios identificados), objectos de áudio (136 critérios), e sequências de vídeo (324 critérios) [Rauch et al. 2005; Rauch & Rauber 2004]. O sistema proposto deverá ser capaz de responder a questões como: quanto conteúdo foi devidamente preservado?; que elementos relacionados com a apresentação do objecto não foram devidamente convertidos (e.g. fontes, cores, frequência de amostragem)?; em que medida o objecto convertido é maior que o objecto de partida?; qual o nível de suporte/abertura/estabilidade do formato de destino, etc.

O componente Migration Advisor será responsável por recomendar os formatos de destino ou os caminhos de migração mais adequados às necessidades de cada utilizador. As recomendações fornecidas pelo sistema serão geradas confrontando os requisitos de cada utilizador com a informação armazenada, ao longo do tempo, na Migration Knowledge Base.

4 Metodologia e experiências propostas

Neste trabalho propõe-se desenvolver um sistema distribuído capaz de assistir organizações ou indivíduos na selecção e execução de intervenções de preservação baseadas em migração. As duas questões de investigação centrais neste trabalho são: 1) Será exequível especificar e desenvolver um sistema capaz de automaticamente detectar a quantidade de informação perdida numa migração e com isso gerar relatórios de migração em formatos apropriados que facilitem a sua anexação à metainformação de preservação de um objecto digital? 2) Será exequível especificar e desenvolver um sistema que, baseado na informação recolhida ao longo do tempo, seja capaz de produzir recomendações sobre quais os formatos ou caminhos de migração mais adequados às necessidades de uma organização/utilizador?

Pretende-se validar os conceitos "quantificação automática de informação perdida numa migração" e "recomendação automática de estratégias de migração" desenvolvendo um protótipo do sistema descrito e realizando uma série de experiências de forma a atestar empiricamente a sua viabilidade (i.e. prova de conceito). Para tal, pretende-se reproduzir as experiências realizadas por Rauch e Rauber de forma a reunir dados empíricos que possam ser comparados com os resultados produzidos automaticamente pelo sistema proposto. Nas experiências de Rauch e Rauber a determinação da informação perdida em cada migração é realizada manualmente por um grupo de especialistas. A sugestão de estratégias de preservação é obtida recorrendo ao seu método de avaliação de estratégias de preservação.

5 Conclusão

Neste artigo é proposto um sistema baseado numa arquitectura orientada ao serviço capaz de assistir organizações e indivíduos a desempenhar mais eficazmente as suas tarefas de preservação. O sistema proposto tem como objectivo desempenhar as seguintes funções: 1)

determinar, segundo múltiplos critérios, a quantidade de informação perdida durante um migração; 2) gerar relatórios de qualidade associados a cada migração em formatos adequados que facilitem a sua inclusão na metainformação de preservação associada aos objectos digitais; e 3) fornecer sugestões acerca dos formatos de destino e os caminhos de migração que melhor satisfazem as necessidades de preservação de cada organização/utilizador.

6 Referências

Caplan, P., Guenther, R., Dale, R., Lavoie, B., Barnum, G., Blair, C., et al. (2005). *Data Dictionary for Preservation Metadata* (Final report): PREMIS Working Group (OCLC/RLG).

Hunter, J. & Choudhury, S. (2004). *A Semi-Automated Digital Preservation System based on Semantic Web Services.* Paper presented at the Joint ACM/IEEE Conference on Digital Libraries (JCDL'04).

Lee, K.-H., Slattery, O., Lu, R., Tang, X. & McCrary, V. (2002). 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.

Lorie, R. A. (2002, July 13-17 2002). *A Methodology and System for Preserving Digital Data.* Paper presented at the Second ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'02), Portland, Oregon.

Ockerbloom, J. M. (1998). *Mediating Among Diverse Data Formats.* Unpublished PhD Thesis, Carnegie Mellon University, Pittsburg.

Proença, A. & Lopes, S. (2004). *Digital Preservation* (Monography). Covilhã: Departamento de Informática da Universidade da Beira Interior.

Rauch, C., Pavuza, F., Strodl, S. & Rauber, A. (2005). *Evaluating preservation strategies for audio and video files.* Paper presented at the DELOS Digital Repositories Workshop, Heraklion, Crete.

Rauch, C. & Rauber, A. (2004). *Preserving Digital Media: Towards a Preservation Solution Evaluation Metric.* Paper presented at the International Conference on Asian Digital Libraries, Shanghai, China.

Rothenberg, J., Commission on Preservation and Access & Council on Library and Information Resources. (1999). *Avoiding technological quicksand: finding a viable technical foundation for digital preservation: a report to the Council on Library and Information Resources*. Washington, DC: Council on Library and Information Resources.

Task Force on Archiving of Digital Information, Commission on Preservation and Access & Research Libraries Group. (1996). *Preserving digital information: report of the Task Force on Archiving of Digital Information*. Washington, D.C.: Commission on Preservation and Access.

Thibodeau, K. (2002). *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years.* Paper presented at the The State of Digital Preservation: An International Perspective, Washington D.C.

Walker, F. L. & Thoma, G. R. (2004). *A Web-Based Paradigm for File Migration.* Paper presented at the IS&T's 2004 Archiving Conference, San Antonio, Texas, USA.

Webb, C. (2003). *Guidelines for the Preservation of Digital Heritage*: United Nations Educational Scientific and Cultural Organization - Information Society Division.

Weirich, P., Skyrms, B., Adams, E. W., Binmore, K., Butterfield, J., Diaconis, P., et al. (2001). *Decision Space: Multidimensional Utility Analysis*. Cambridge.

Westhead, M., Wen, T. & Carroll, R. (2003). *Describing Data on the Grid.* Paper presented at the 4th International Workshop on Grid Computing (GRID 2003), Phoenix, USA.