DEPARTAMENTO DE SISTEMAS DE INFORMAÇÃO
ESCOLA DE ENGENHARIA · UNIVERSIDADE DO MINHO

PRESERVAÇÃO DE LONGA DURAÇÃO DE INFORMAÇÃO DIGITAL NO CONTEXTO DE UM ARQUIVO HISTÓRICO

José Miguel Araújo Ferreira

TESE DE DOUTORAMENTO

Tese submetida à Escola de Engenharia da Universidade do Minho para obtenção do grau de Doutor em Tecnologias e Sistemas de Informação, na especialidade de Sociedade da Informação, sob a orientação da Professora Doutora Ana Alice Baptista e Professor Doutor José Carlos Ramalho.

Guimarães, Maio de 2009

É AUTORIZADA A REPRODUÇÃO INTEGRAL DESTA TESE APENAS PARA EFEITOS DE INVESTIGAÇÃO, MEDIANTE DECLARAÇÃO ESCRITA DO INTERESSADO, QUE A TAL SE COMPROMETE;

Universidade do Minho, _____/_____/_____

Assinatura: _____

À minha filha...

...que brinque para sempre nos confins do firmamento...

Aos meus pais...

... por estarem presentes no momento em que mais precisei deles...

AGRADECIMENTOS

Foram várias as pessoas que directa ou indirectamente contribuíram para o desenvolvimento desta tese. A essas pessoas gostaria de prestar a minha homenagem, agradecendo-lhes todo o apoio e disponibilidade concedidos ao longo destes quatro anos.

Em primeiro lugar, gostaria de agradecer à Professora Arminda Manuela Gonçalves do Departamento de Matemática para a Ciência e Tecnologia da Universidade do Minho por me ter levado pela mão através do mundo da estatística não-paramétrica; ao Professor Eduardo Severino do Departamento de Estatística e Investigação Operacional da Faculdade de Ciências da Universidade de Lisboa pela ajuda fundamental na parametrização e planeamento de experiências; ao Professor Zhou Wang do Departamento de Engenharia Electrotécnica e Computadores da Universidade de Waterloo no Canadá, pela ajuda preciosa em torno do cálculo de similaridade de imagens; ao Professor Andreas Rauber da Universidade Técnica de Viena na Áustria, pelas conversas de café em torno de preservação digital e investigação em geral; ao engenheiro Duarte Duque pela ajuda fundamental na escrita e interpretação de formalismos matemáticos; ao Doutor Pedro Gabriel Ferreira, colega, confidente e amigo, por ter desbravado o caminho espinhoso de um doutoramento dando o exemplo para tantos outros; aos alunos, agora Engenheiros, Nuno Gonçalves, Rui Rodrigues, Samuel Cordeiro, Rick Gomes, Victor da Costa Pinheiro, Ricardo Gomes de Faria por terem materializado tudo aquilo que não tive tempo de desenvolver sozinho; aos Engenheiros Rui Castro e Luís Faria por serem os melhores programadores do país e por me terem ajudado a optimizar muitos dos componentes que descrevo nesta tese; aos técnicos da Direcção-Geral de Arquivos, Dr. Francisco Barbedo, Dra. Cecília Henriques e Dr. Luís Corujo por pacientemente me terem ensinado a ver o mundo através dos olhos de um arquivista; à Marta por me ter aturado ao longo destes anos todos e por me ter ajudado a depurar o texto que dá corpo a esta tese.

Finalmente, gostaria de agradecer aos meus orientadores, sem os quais este trabalho nunca teria sido possível. À Professora Ana Alice Baptista, filósofa da computação, mulher de ideias elevadas, que tantas vezes exigiu que o meu cérebro se colocasse em bicos-de-pés; e ao Professor José Carlos Ramalho, rico em experiências, tecnólogo por excelência, com o mais apurado sentido prático da vida; a ambos, o meu mais sincero obrigado por me terem deixado crescer intelectualmente a seu lado.

A todas estas pessoas, o meu mais sincero obrigado!

Miguel Ferreira

RESUMO

PRESERVAÇÃO DE LONGA-DURAÇÃO DE INFORMAÇÃO DIGITAL NO CONTEXTO DE UM ARQUIVO HISTÓRICO

Ao longo do século XX, a humanidade assistiu à massificação generalizada das tecnologias digitais. Estas encontram-se presentes em todos os quadrantes do mundo civilizado e suportam grande parte da actividade humana. Actividades tão dispares como consultar as horas ou planear uma missão espacial a Marte são, hoje em dia, inteiramente suportadas por tecnologias digitais.

A expansão das tecnologias digitais conduziu inevitavelmente a um aumento da produção de informação digital. Este tipo de informação acarreta consigo um problema que coloca em risco a sua acessibilidade a longo-prazo. Este tipo de material, embora possa ser copiado infinitas vezes sem perder qualidade, requer a presença de um contexto tecnológico, hardware e/ou software, para que possa ser interpretado de forma inteligível por um ser humano. Esta dependência torna-o vulnerável à rápida obsolescência a que a tecnologia está sujeita, dado que nem sempre os novos desenvolvimentos garantem a compatibilidade com tecnologias precedentes.

No sentido de mitigar o problema da obsolescência tecnológica e garantir o acesso continuado à informação digital foram apontadas diversas estratégias de preservação de informação digital, como por exemplo: a emulação, a migração de formatos e o encapsulamento. Apesar dos inúmeros progressos verificados neste domínio, continua a existir um vazio assinalável no que diz respeito à automatização de estratégias de preservação. Paralelamente, questões relacionadas com a autenticidade dos materiais, a validação de estratégias de preservação e a necessidade, sempre crescente, de reduzir custos assumem particular destaque na lista de preocupações dos profissionais da ciência da informação.

Este projecto de investigação visa atenuar o conjunto de problemas previamente enumerados, dando especial ênfase à automatização de processos de preservação baseados em migração de formatos. De forma a dar resposta a esta necessidade, foi desenvolvida uma Arquitectura Orientada ao Serviço (SOA) capaz de auxiliar organizações e/ou indivíduos na implementação de intervenções de preservação. O sistema desenvolvido é constituído por um conjunto de componentes, fisicamente distribuídos, que são capazes de realizar o seguinte conjunto de actividades: executar acções de preservação baseadas em migração de formatos (conversão); determinar a quantidade de informação, propriedades significativas e funcionalidades perdidas

durante uma migração (controlo de qualidade); produzir relatórios que possam ser utilizados como metainformação de preservação e que documentam a intervenção de preservação (autenticidade); e fornecer sugestões de formatos de destino e/ou serviços de conversão que maximizem a satisfação da entidade-cliente (selecção de alternativas de migração).

O sistema desenvolvido foi avaliado no que diz respeito à sua capacidade de produzir recomendações de alternativas de migração capazes de satisfazer os requisitos de preservação manifestados por uma entidade-cliente. A avaliação incidiu ainda sobre a capacidade demonstrada pelo sistema em aferir o nível de degradação incorrido num objecto digital durante uma migração de formato, especialmente no que toca a propriedades significativas de carácter subjectivo.

O sistema foi avaliado recorrendo a colecções de teste constituídas por imagens matriciais em diferentes formatos. Os resultados obtidos demonstram que o sistema é capaz de determinar eficazmente a similaridade gráfica entre imagens, apresentando valores de correlação superiores a 0.81 entre as opiniões produzidas por avaliadores humanos e métodos automáticos de cálculo de similaridade. No que toca à capacidade do sistema em determinar o nível de deterioração da metainformação embebida nas imagens, este apresentou valores de correlação acima dos 0.96 entre os valores produzidos pela métrica de Jaccard' e os valores de referência associados à colecção de teste.

As experiências realizadas em torno do sistema de recomendação permitiram concluir que os serviços de migração sugeridos por este componente (tendo por base informação recolhida a partir de migrações anteriores) possuem um elevado nível de correlação com as recomendações ideais calculadas para cada objecto digital pertencente à colecção de teste. Os testes realizados resultaram em valores de correlação compreendidos entre 0.68 e 0.85 com um erro de precisão máximo de 34.9%.

Os principais contributos desta investigação são: a capacidade de preservar objectos digitais recorrendo a conversão de formatos sem que haja necessidade de implementar complexos sistemas de migração; a capacidade de obter relatórios detalhados sobre a migrações realizadas permitindo, assim, documentar todo o processo de preservação e deste modo assegurar a autenticidade dos materiais; e a possibilidade de comparar diferentes alternativas de migração e identificar de forma objectiva qual destas é a mais adequada para satisfazer as necessidades de uma organização.

ABSTRACT

LONG-TERM PRESERVATION OF DIGITAL INFORMATION IN THE CONTEXT OF A HISTORICAL ARCHIVE

During the second half of the 20th century, mankind has passively witnessed the worldwide proliferation of digital technologies. These technologies are currently present in every aspect of today's civilized life and natively support a great deal of human activities. Distinct actions such as telling the time or planning a mission to Mars are now entirely supported by digital technologies. This growth has been accompanied by an overwhelming expansion of digital information.

Digital information has a lot of advantages over traditional analogue information. However, it carries a structural problem that may hinder its accessibility in the long run. Digital information requires the presence of a technological environment (hardware and/or software) in order to be adequately rendered for human consumption. This technological dependency makes it vulnerable to the rapid evolution of digital technologies as well as technological ruptures caused by non-retrocompatible developments.

To insure the continuous access to digital information, several strategies have been proposed: emulation, format migration, encapsulation, etc. However, there is still a great deal of work to be done in what concerns making these processes more automatic and user-friendly. Moreover, issues regarding the authenticity of digital materials have always been a concern for information science professionals.

This thesis aims at solving the previously outlined issues, focusing especially on the automation of migration-based preservation strategies. In order to accomplish this goal, we have developed a Service Oriented Architecture (SOA) specially designed to assist cultural heritage institutions in the implementation of 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, suitability of involved formats and data loss. The proposed system is also able to produce preservation metadata that can be used by any client institution to document preservation interventions and retain objects' authenticity.

The system has been evaluated in what concerns its ability to produce suggestions of migration services that maximize the preservation requirements of any given client institution. The evaluation process also focused the system's ability to determine the level of degradation imposed to a digital object during a migration process, especially in what concerns its subjective significant properties, i.e., pixel correctness and embedded metadata.

The system was evaluated using datasets of raster images encoded in several formats. The results of this research show that the proposed system is capable of effectively calculating the similarity between digital images, revealing a correlation value superior to 0.81 between automatic similarity algorithms and the mean opinions scores provided by human evaluators. In what concerns the system's ability to determine the level of degradation occurred in the image metadata, the system showed correction values above 0.96 while using a modified version of the Jaccard similarity metric.

The recommendation system showed a level of correlation of 0.68 to 0.85 (with a maximum precision of 34.9%) when suggestions based on previously executed migrations were compared with the ideal rankings of migration services calculated specifically for a given object.

The main contributions of this research are: the ability to preserve digital information using a format migration strategy without having to deploy complex migration systems; the ability to obtain detailed migration reports that document the entire preservation intervention which can be used as preservation metadata to ensure information authenticity; and the possibility of comparing and assessing different migration options and objectively choose the one that maximises the satisfaction of a client institution.

CONTEÚDO

INTRODUÇÃO 1

1.1 Estratégias de preservação digital 2
1.2 Motivação 4
1.3 Objectivos e contributos 5
1.4 Organização da tese 8

PRESERVAÇÃO DIGITAL 11

2.1 A anatomia de um objecto digital 14
2.2 O modelo de referência OAIS 18
2.3 Estratégias de preservação digital 21
2.3.1 Preservação de tecnologia 22
2.3.2 Refrescamento 23
2.3.3 Emulação 23
2.3.4 Migração/conversão 26
2.3.5 Encapsulamento 32
2.3.6 Pedra de Roseta digital 33
2.4 Directórios de formatos 34
2.5 Autenticidade 37
2.6 Metainformação de preservação 40
2.6.1 PREMIS 40
2.7 Considerações finais 44

AUTOMATIZAÇÃO DE PROCESSOS DE MIGRAÇÃO 49

3.1 Actividades inerentes a um processo de migração 50
3.1.1 Selecção de uma alternativa de migração 50
3.1.2 Conversão de materiais 51
3.1.3 Controlo de qualidade 52
3.2 Migração em ambientes distribuídos 52
3.3 Cenário de preservação 54
3.4 Serviços de preservação 56
3.4.1 Identificador de formatos 58
3.4.2 Conversores 59
3.4.3 Controlo de qualidade 60
3.4.4 Notificador de obsolescência 61
3.4.5 Seleccionador de estratégias de migração 62
3.5 Cenário revisto 65
3.6 Considerações finais 68

CRIB – PLATAFORMA DE SERVIÇOS DE PRESERVAÇÃO 71

4.1 Visão geral 72
4.2 Core preservation services 75

4.2.1 Identificação de formatos ... 76
4.2.2 Selecção de estratégias de migração ... 77
4.2.3 Migração de formatos e controlo de qualidade ... 80
4.2.4 Serviços adicionais ... 82
4.3 Service Registry ... 83
4.4 Format Identifier ... 88
4.5 Migration Broker ... 89
4.5.1 Disponibilidade ... 92
4.5.2 Estabilidade ... 92
4.5.3 Débito ... 93
4.5.4 Custo de utilização ... 94
4.5.5 Taxa de crescimento em bytes ... 95
4.5.6 Taxa de crescimento em número de ficheiros ... 96
4.6 Object Evaluator ... 96
4.6.1 Classes de objectos ... 102
4.6.2 Taxionomias de avaliação ... 105
4.6.3 Extractores de valores de propriedades ... 109
4.6.4 Funções de similaridade ... 109
4.7 Format Evaluator ... 110
4.7.1 Ganho de preservação ... 115
4.7.2 Implicação ... 116
4.7.3 Negação ... 116
4.7.4 Razão ... 117
4.8 Migration Advisor ... 118
4.8.1 Algoritmo de recomendação ... 122
4.9 Considerações finais ... 126
4.9.1 Limitações ... 128

METODOLOGIA E AVALIAÇÃO ... 133

5.1 Avaliação do Object Evaluator ... 134
5.1.1 Protocolo experimental ... 136
5.1.2 Propriedade significativa: conformidade gráfica ... 137
5.1.3 Propriedade significativa: metainformação embebida ... 154
5.2 Avaliação do Migration Advisor ... 162
5.2.1 Caracterização da colecção de teste ... 165
5.2.2 Selecção de caminhos de conversão ... 167
5.2.3 Treino e teste do sistema ... 168
5.2.4 Resultados ... 170
5.3 Considerações finais ... 171

IMPLEMENTAÇÕES DO CRIB ... 175

6.1 Planets ... 175
6.2 RODA ... 177

CONCLUSÕES E TRABALHO FUTURO ... 181

7.1 Síntese ... 181
7.2 Conclusões e discussão ... 183
7.3 Contributos ... 191
7.4 Trabalho futuro ... 193

APÊNDICES ... 197

8.1 Ferramentas de extracção de propriedades ... 197
8.1.1 Image IO ... 198
8.1.2 ExifTool 7.15 ... 198
8.1.3 Microsoft Office Word 2003 ... 199
8.1.4 OpenOffice.org Writer 2.2 ... 200
8.1.5 PDFBox ... 200
8.2 Taxionomia geral de avaliação ... 202
8.3 Funções de similaridade ... 203
8.3.1 Similaridade numérica ... 205
8.3.2 Similaridade vectorial ... 206
8.3.3 Similaridade textual ... 207
8.3.4 Similaridade entre conjuntos ... 210
8.3.5 Similaridade de XML ... 211
8.3.6 Similaridade gráfica ... 212
8.4 Teste não-paramétrico de Wilcoxon ... 218
8.5 Validação cruzada ... 219
8.6 Licença de uso e distribuição do CRiB ... 221

ANEXOS ... 223

9.1 Interpretação de valores-P ... 223

LISTA DE FIGURAS

Figura *Página*

Figura 1 – Cassete de vídeo Betamax. 12
Figura 2 – Cassete de vídeo VHS. 12
Figura 3 – Disquete de 3.5 polegadas. 13
Figura 4 – Leitor de disquetes de 3.5 polegadas. 13
Figura 5 – Cadeia de interpretação desde o nível físico até ao nível conceptual. 16
Figura 6 – Diferentes níveis de abstracção de um objecto digital. 16
Figura 7 – Objecto digital observado a diferentes níveis de abstracção 18
Figura 8 – Modelo de referência Open Archival Information System (OAIS). 19
Figura 9 – Classificação das diferentes estratégias de preservação digital 22
Figura 10 – Exemplo de um cenário de emulação. 25
Figura 11 – Degradação do objecto digital ao longo de sucessivas migrações. 30
Figura 12 – Migração a-pedido. 30
Figura 13 – Migração distribuída baseada em Serviços Web. 31
Figura 14 – Pedra de Roseta. 33
Figura 15 – Verificação da qualidade de uma migração através de canonização 40
Figura 16 – Entidades presentes no Dicionário de Dados PREMIS. 41
Figura 17 – Diferentes representações para a mesma entidade intelectual 42
Figura 18 – Arquitectura de um sistema de preservação. 57
Figura 19 – Exemplo de árvore-objectivo. 63
Figura 20 – Processo de selecção de estratégias de preservação. 65
Figura 21 – Arquitectura geral da plataforma CRiB. 73
Figura 22 – Interface do componente Core Preservation Services 76
Figura 23 – Diagrama de classes das mensagens trocadas pelo CRiB 76
Figura 24 – Diagrama de sequência da identificação de formatos. 77
Figura 25 – Diagrama de classes de uma representação. 77
Figura 26 – Diagrama de sequência relativo à selecção de uma alternativa de migração 78
Figura 27 – Mensagens envolvidas na selecção de uma alternativa de migração. 79
Figura 28 – Diagrama de sequência do processo de conversão. 80

Figura 29 – Diagrama de classes associadas ao processo de conversão..... 81
Figura 30 – Outros métodos disponibilizados pelo CRiB..... 83
Figura 31 – Relações entre entidades que descrevem um serviço de conversão..... 84
Figura 32 – Arquitectura detalhada do Service Registry..... 86
Figura 33 – Métodos disponibilizados pelo Service Registry..... 87
Figura 34 – Métodos disponibilizados pelo Format Identifier..... 88
Figura 35 – Arquitectura detalhada do Migration Broker..... 89
Figura 36 – Métodos disponibilizados pelo Migration Broker..... 90
Figura 37 – Mensagens trocadas pelo Migration Broker..... 90
Figura 38 – Interface comum a todos os serviços de conversão..... 91
Figura 39 – Caminho de migração com baixa estabilidade..... 93
Figura 40 – Cálculo do tempo de migração..... 94
Figura 41 – Cálculo do custo de utilização de uma migração composta..... 94
Figura 42 – Arquitectura detalhada do Object Evaluator..... 98
Figura 43 – Arquitectura detalhada do comparador de objectos conceptuais..... 99
Figura 44 – Métodos disponibilizados pelo Object Evaluator..... 101
Figura 45 – Mensagens trocadas pelo Object Evaluator..... 102
Figura 46 – Taxionomia de avaliação de imagens matriciais..... 106
Figura 47 – Taxionomia de avaliação de documentos de texto..... 108
Figura 48 – Arquitectura do Format Evaluator..... 111
Figura 49 – Diagrama de classes associadas ao Format Evaluator..... 112
Figura 50 – Cálculo do benefício de migração..... 114
Figura 51 - Diagrama de sequência do processo de recomendação..... 120
Figura 52 – Arquitectura do Migration Advisor..... 121
Figura 53 – Diagrama de classes e mensagens trocadas pelo Migration Advisor..... 122
Figura 54 – Arquitectura geral do motor de recomendação..... 122
Figura 55 – Cálculo de pontuação de um caminho de migração..... 123
Figura 56 – Exemplo de normalização de taxionomia pesada segundo uma escala Likert de 1 a 5..... 123
Figura 57 – Agregação de resultados e cálculo de pontuação..... 126
Figura 58 – *Screenshot* da aplicação utilizada para comparar imagens..... 144
Figura 59 – Projecções de MOS com (a) RMSE, (b) UQI, (c) SSIM e (d) CBM..... 149

Figura 60 – Conjunto de imagens com RMSE≈0.96 e valores de UQI, SSIM e CBM distintos..... 153
Figura 61 – Teste do sistema de recomendação..... 165
Figura 62 – Plato e os serviços de migração do CRiB..... 177
Figura 63 – Interface gráfica do Repositório de Objectos Digitais Autênticos..... 178
Figura 64 – Taxionomia geral de avaliação..... 202
Figura 65 – Algoritmo da distância de Levenshtein..... 208
Figura 66 – Definição formal de imagem matricial..... 212
Figura 67 – Classes de métricas de similaridade gráfica..... 213
Figura 68 – Detecção de (1) contornos, (2) texturas e (3) regiões planas usando uma máscara de Sobel..... 217
Figura 69 – Diagrama de processamento da métrica CBM..... 218
Figura 70 – Exemplo do método de validação cruzada com 4 dobras..... 221

LISTA DE TABELAS

Tabela *Página*

Tabela 1 – Possíveis estratégias de preservação por nível de abstracção. 47
Tabela 2 – Elementos de metainformação sobre a organização que desenvolveu o serviço de conversão. 85
Tabela 3 – Elementos de metainformação que descrevem serviços de conversão 85
Tabela 4 – Elementos de metainformação que descrevem a localização do serviço 86
Tabela 5 – Elementos de metainformação que descrevem os contactos de uma organização. 86
Tabela 6 – Exemplo de uma taxionomia de avaliação de documentos de texto 100
Tabela 7 – Exemplo de uma taxionomia de avaliação de objectos áudio 100
Tabela 8 – Formatos suportados pelo CRiB. 102
Tabela 9 – Propriedades associadas a imagens matriciais. 107
Tabela 10 – Propriedades associadas a documentos de texto 109
Tabela 11 – Características técnicas avaliadas pelo Format Evaluator 113
Tabela 12 – Cálculo da função Gain 115
Tabela 13 – Cálculo da função Implication. 116
Tabela 14 – Cálculo da função Not. 117
Tabela 15 – Cálculo de desempenho médio de um caminho de migração. 124
Tabela 16 – Normalização de desempenho médio de um caminho de migração 125
Tabela 17 – Avaliações produzidas por intervenientes humanos. 145
Tabela 18 – MOS e desvio-padrão após remoção de valores discrepantes 147
Tabela 19 – Avaliações produzidas pelos algoritmos RMSE, UQI, SSIM e CBM 148
Tabela 20 – Valores de similaridade ajustados aos valores de MOS. 150
Tabela 21 – Desempenho dos vários algoritmos de cálculo de similaridade de imagem. 151
Tabela 22 – Tipos de falhas na metainformação embebida que poderão ocorrer durante uma conversão de formatos 155
Tabela 23 – Colecção de teste utilizada na experiência com metainformação embebida 156
Tabela 24 – Resultados produzidos pelos métodos XML Diff e Jaccard 158
Tabela 25 – Desempenho dos dois métodos de cálculo de similaridade de metainformação embebida. 159

Tabela 26 – Resultados produzidos pelo método de Jaccard modificado..... 161
Tabela 27 – Desempenho dos dois métodos de cálculo de similaridade de metainformação embebida. 162
Tabela 28 – Descrição das colecções de imagens utilizadas na avaliação do componente Migration Advisor..... 166
Tabela 29 – Caminhos de conversão utilizados na avaliação do Migration Advisor..... 167
Tabela 30 – Dados relativos ao treino e teste do componente Migration Advisor. 168
Tabela 31 – Resultados da validação cruzada efectuada ao Migration Advisor..... 170
Tabela 32 – Características da nova colecção de teste de cardinalidade 10..... 171
Tabela 33 – Resultados da validação cruzada efectuada ao Migration Advisor com a nova colecção de teste de cardinalidade 10..... 171
Tabela 34 – Propriedades extraídas e formatos suportados pela biblioteca Java Image I/O.. 198
Tabela 35 – Propriedades extraídas e formatos suportados pela ferramenta ExifTool..... 199
Tabela 36 – Propriedades extraídas pela ferramenta Microsoft Office Word 2003..... 199
Tabela 37 – Propriedades extraídas pela ferramenta OpenOffice.org Writer 2.2..... 200
Tabela 38 – Propriedades extraídas pela ferramenta PDFBox..... 201
Tabela 39 – Métricas utilizadas para comparar imagens matriciais..... 204
Tabela 40 – Métricas utilizadas para comparar documentos de texto..... 205
Tabela 41 – Resultados da aplicação do teste de Wilcoxon para comparação de médias..... 219

LISTA DE EQUAÇÕES

Equação *Página*

Equação 1 – *Mean Opinion Score* (MOS). 146
Equação 2 – Taxa de valores não-discrepantes. 151
Equação 3 – Coeficiente de Similaridade de Jaccard. 157
Equação 4 – Exemplo da aplicação do coeficiente de Jaccard. 160
Equação 5 – Definição da função *first* e versão modificada do método de Jaccard. 161
Equação 6 – Número de conversões mediante o tamanho da colecção de teste 168
Equação 7 – Relação entre similaridade e distância. 203

LISTA DE FÓRMULAS

Fórmula *Página*

Fórmula 1 – Disponibilidade. 92
Fórmula 2 – Estabilidade. 93
Fórmula 3 – Débito de conversão 93
Fórmula 4 – Taxa de crescimento em bytes de representações convertidas 95
Fórmula 5 – Taxa de crescimento em número de ficheiros. 96
Fórmula 6 – Ratio. 117
Fórmula 7 – Exemplo de aplicação da função Ratio. 118
Fórmula 8 – Normalização de pesos. 124
Fórmula 9 – Normalização de vectores de desempenho 125
Fórmula 10 – Definição matemática de distância. 203
Fórmula 11 – Definição matemática de similaridade. 204
Fórmula 12 – Distância proporcional 205
Fórmula 13 – Similaridade proporcional. 206
Fórmula 14 – Similaridade euclidiana. 206
Fórmula 15 – Igualdade textual relaxada. 209
Fórmula 16 – Métrica de comparação de cadeias de caracteres de Jaro. 209
Fórmula 17 – Similaridade de Jaro-Winkler. 210
Fórmula 18 – Coeficiente de similaridade de Jaccard 211
Fórmula 19 – Função *first*. 211
Fórmula 20 – Coeficiente de similaridade de Jaccard modificado. 211
Fórmula 21 – Normalized Root Mean Squared Error (NRMSE). 214
Fórmula 22 – Universal Image Quality Index (UQI) de uma componente de cor. 215
Fórmula 23 – Fórmulas auxiliares ao cálculo de UQI. 215
Fórmula 24 – Valor global de UQI. 215
Fórmula 25 – Fórmulas auxiliares ao cálculo de SSIM. 216
Fórmula 26 – Structural Similarity (SSIM) de uma componente de cor. 216
Fórmula 27 – Valor de SSIM que combina as quatro componentes de cor. 216

Fórmula 28 – Valor global de SSIM que combina os valores de SSIM das *M* janelas amostradas.....217
Fórmula 29 – Diferença entre as avaliações subjectivas e os valores objectivos.....219
Fórmula 30 – Formulação de hipóteses.....219

GLOSSÁRIO

Arquitectura Orientada ao Serviço. Arquitectura de software onde vários componentes disponibilizam recursos computacionais aos restantes participantes da rede sob a forma de serviços independentes, invocáveis de forma normalizada através de um protocolo comum (ver Serviço *Web*).

Arquivo. Organização responsável por gerir, descrever, armazenar e garantir acesso a informação.

ASCII. American Standard Code for Information Interchange. Conjunto de códigos capaz de representar letras, dígitos e outros símbolos, amplamente utilizado por computadores na troca e armazenamento de informação textual.

Autenticação. Processo responsável por assegurar que um utilizador, serviço ou recurso é exactamente aquele que se propõe ser (i.e., comprovação de identificação).

CD-ROM. Compact Disc Read-Only Memory. Suporte físico de armazenamento baseado em tecnologia óptica.

Comunidade de interesse. Conjunto identificável de consumidores de informação de um dado repositório ou arquivo.

Conversão. Ver Migração.

Digitalização. Processo responsável pela transformação de informação analógica em informação digital.

Disco rígido. Suporte de armazenamento de informação digital baseado em tecnologia magnética.

DVD. Digital Versatile Disk. Suporte físico de armazenamento baseado em tecnologia óptica. Fisionomicamente semelhante a um CD-ROM mas com uma capacidade de armazenamento várias vezes superior.

Emulador. Software capaz de reproduzir o comportamento de uma plataforma de hardware e/ou software numa outra que de outro modo seria incompatível.

Encapsulamento. Preservar, juntamente com um objecto digital, toda a informação necessária e suficiente para permitir o futuro desenvolvimento de conversores, visualizadores ou emuladores que garantam o acesso à informação veiculada. Esta informação poderá consistir, por exemplo, numa descrição formal e detalhada do formato do objecto preservado.

Estratégia de preservação digital. Abordagem técnica que garante o acesso continuado à informação existente em formatos digitais (ver Migração, Emulador ou Encapsulamento).

Flash-drive. Dispositivo que combina uma memória flash com uma interface USB, vulgarmente utilizado para armazenar informação digital. Este dispositivo é também vulgarmente conhecido por *pen-drive*.

GIF. Graphics Interchange Format. Formato matricial para representação de imagens digitais.

Ingestão. Processo ou componente responsável pela recepção de material de arquivo.

Internet. Rede global de comunicação baseada no protocolo TCP/IP.

Java. Linguagem de programação orientada ao objecto desenvolvida na década de 90. Contrariamente às linguagens de programação convencionais, que são compiladas para código nativo, a linguagem Java é compilada para *bytecode*, ou seja, código que é executado por uma máquina virtual.

JPEG. Joint Photographic Experts Group. Formato matricial para representação de imagens digitais.

Material digital. Conjunto de informação ou objectos digitais.

Metadados. Ver Metainformação.

Metainformação. Informação utilizada para descrever um determinado objecto ou recurso.

Migração. Transferência periódica de material digital de uma configuração de hardware/software para outra, ou de uma geração de tecnologia para outra subsequente.

Objecto digital. Todo e qualquer objecto de informação que possa ser representado através de uma sequência de dígitos binários (*bitstream*). 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.

Objecto nado-digital. Objecto criado recorrendo apenas a ferramentas ou processos digitais, ou seja, objecto digital que não passou por um processo de digitalização.

PDF. Portable Document Format. Formato digital vulgarmente utilizado para representar documentos de texto com formatação e estrutura.

Pixel. Abreviatura de *picture element*. O mais pequeno elemento de informação visual que faz parte de uma imagem digital.

PNG. Portable Network Graphics. Formato matricial para representação de imagens digitais.

Preservação digital. Conjunto de actividades ou processos responsáveis por garantir o acesso continuado e a longo-prazo a informação e restante património cultural existente em formatos digitais.

Propriedade significativa. Característica técnica ou atributo que caracteriza um objecto digital considerada relevante para efeitos de preservação.

Refrescamento. Processo que consiste na cópia de informação de um suporte físico de armazenamento para outro do mesmo tipo.

Repositório digital. Sistema de informação responsável por gerir e armazenar informação digital.

Service Oriented Architecture (SOA). Ver Arquitectura Orientada ao Serviço.

TARGA. Truevision TGA. Formato matricial utilizado para representar imagens digitais.

TIFF. Tagged Image File Format. Formato matricial vulgarmente utilizado para representar imagens digitais.

Web Service. Forma de trocar informação onde são utilizados protocolos de ligação e formatos de mensagens normalizados baseados em XML/SOAP. De modo a facilitar a descoberta de serviços, estes são geralmente publicados em directórios, vulgarmente designados por UDDI (Universal Description, Discovery and Integration).

SIGLAS E ACRÓNIMOS

AHDS. Arts and Humanities Data Service.

ASCII. American Standard Code for Information Interchange.

CBM. Content-Based Image Quality Metric

CD-ROM. Compact Disc Read-Only Memory.

CRiB. Conversion and Recommendation of Digital Object Formats.

DGARQ. Direcção-Geral de Arquivos

DVD. Digital Versatile Disk.

Exif. Exchangeable image file format

GIF. Graphics Interchange Format.

HTTPS. Hypertext Transfer Protocol over Secure Socket Layer

IIM. Information Interchange Model

IPTC. International Press Telecommunications Council

ITU. International Telecommunication Union

JPEG. Joint Photographic Experts Group.

KFCV. k-fold cross-validation

MAE. Mean Absolute Error

MOS. Mean Opinion Score

MSE. Mean Squared Error

NMSE. Normalized Mean Squared Error

OAIS. Open Archival Information System.

PDF. Portable Document Format.

Planets. Preservation and Long-term Access through Networked Services

Planets. Preservation And Long-Term Access Through Networked Services.

PNG. Portable Network Graphics.

PREMIS. Preservation Metadata: Implementation Strategies.

RMSE. Root Mean Squared Error

ROAR. Registry of Open Access Repositories

RODA. Repositório de Objectos Digitais Autênticos

SOA. Service Oriented Architecture.

SSIM. Structured Similarity

TIFF. Tagged Image File Format.

TOM. Typed Objects Model.

UNO. Universal Network Object.

UQI. Universal Image Quality Index.

URL. Uniform Resource Locator

VHS. Video Home System

WS-BPEL. Web services Business Process Execution Language

XMP. Extensible Metadata Platform

Capítulo 1
Introdução

Ao longo da segunda metade do século XX, a humanidade assistiu à massificação generalizada das tecnologias digitais. Estas encontram-se presentes em todos os quadrantes do mundo civilizado e suportam grande parte da actividade humana. Actividades tão dispares como consultar as horas ou planear uma missão espacial a Marte são, hoje em dia, inteiramente suportadas por tecnologias desta natureza. Esta expansão foi desde logo acompanhada por um aumento da produção de informação digital.

Um estudo realizado pela consultora IDC[1] revela que a produção de informação digital tem vindo a sofrer um aumento com características exponenciais. Em 2007, o universo digital foi estimado em 281 Exabytes de informação (i.e., mil milhões de Gigabytes), ou seja, cerca de 45 Gigabytes por cada pessoa existente no planeta. Em 5 anos, prevê-se que esse valor seja 10 vezes superior (Gantz et al., 2008).

Serão certamente variadas as razões que conduziram à adopção massificada de ferramentas digitais. No entanto, a qualidade dos produtos resultantes da sua exploração, aliada à facilidade da sua disseminação, foram factores importantes que explicam a adopção generalizada deste

[1] http://www.idc.com

tipo de ferramentas e o aumento crescente de informação digital (Teixeira, Ferreira, & Verhaegh, 2003).

Apesar das inúmeras vantagens que decorrem da utilização de informação digital, é importante realçar que esta é acompanhada de um problema estrutural que coloca em risco a sua longevidade. Este tipo de material, embora possa ser copiado infinitas vezes sem perder qualidade, requer a presença de um contexto tecnológico, hardware e/ou software, para que possa ser apresentado de forma inteligível a um ser humano. Esta dependência tecnológica torna-o particularmente vulnerável à rápida obsolescência a que a tecnologia está sujeita (Chen, 2001).

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 património cultural existente em formatos digitais (Webb, 2003). Neste contexto, designa-se por *objecto digital* todo e qualquer objecto de informação que possa ser representado através de uma sequência de dígitos binários[2] (Thibodeau, 2002). Documentos de texto, fotografias digitais, diagramas vectoriais, bases de dados, sequências de vídeo e áudio, modelos de realidade virtual, páginas Web, jogos e aplicações de software são apenas alguns exemplos do que pode ser considerado um objecto digital.

1.1 Estratégias de preservação digital

Ao longo dos últimos anos têm vindo a ser propostas diversas estratégias no sentido de minimizar o impacto da obsolescência tecnológica no acesso à informação digital. Segundo Lee et al., as várias estratégias de preservação de informação digital podem ser agrupadas em três classes fundamentais: emulação, encapsulamento e migração (Lee, Slattery, Lu, Tang, & McCrary, 2002).

A *emulação* consiste na utilização de um software especial, designado *emulador*, capaz de 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 num ambiente tecnológico semelhante àquele em que foram criados, ainda que tratando-se de um ambiente virtual (Rothenberg, Commission on Preservation and Access, & Council on Library and Information Resources, 1999). A grande vantagem desta abordagem está na

[2] Esta definição é suficientemente lata para acomodar tanto, informação que nasceu num contexto tecnológico digital (objectos nado-digitais), como informação digital obtida a partir de suportes analógicos (objectos digitalizados).

capacidade de reproduzir com elevado grau de fidelidade a funcionalidade e apresentação do objecto original (Lee et al., 2002; Rothenberg et al., 1999). O recurso a emuladores está geralmente associado à preservação de objectos digitais complexos[3] dotados de propriedades dinâmicas e/ou interactivas como é caso das aplicações de software.

A estratégia de encapsulamento consiste em preservar, juntamente com o objecto digital, toda a informação necessária e suficiente para suportar o futuro desenvolvimento de conversores, visualizadores ou emuladores. Esta informação poderá consistir, por exemplo, numa especificação formal e detalhada do formato associado ao objecto preservado. Raymond Lorie propõe uma variante desta estratégia onde esta especificação formal é substituída por uma aplicação de software compilada para uma máquina virtual universal, e.g. Java Virtual Machine (Raimond A. Lorie, 2002). Esta aplicação tem como finalidade apresentar uma visão lógica do objecto, possibilitando desta forma, uma navegação simples através das suas propriedades.

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). Num contexto de 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 estrutura utilizada para o representar. A migração recorre a software de conversão para transformar os objectos codificados em formatos obsoletos em objectos cujos formatos são compatíveis com as plataformas tecnológicas mais actuais. A principal vantagem desta abordagem consiste na possibilidade de um utilizador convencional ser capaz de interpretar os objectos digitais preservados sem necessidade de artefactos adicionais para além do software existente no seu computador pessoal. No entanto, a aplicação desta estratégia pode resultar na perda de propriedades essenciais do objecto digital. Isto deve-se, sobretudo, a incompatibilidades existentes entre os formatos de partida e chegada ou à utilização de conversores incapazes de realizar devidamente as tarefas a que se propõem.

Apesar da existência de diversas estratégias de preservação digital, nenhuma delas foi até ao momento devidamente validada ou universalmente aceite (Rauch & Rauber, 2004). A escolha

[3] Objectos digitais complexos são geralmente constituídos por vários subcomponentes que poderão, inclusivamente, estar distribuídos por vários nós de processamento, i.e., servidores. Um exemplo deste tipo de objectos são páginas Web que são constituídas por texto, imagens, vídeos, ligações a outras páginas, etc.

de qualquer uma das alternativas expostas necessita geralmente que diversos factores sejam tomados em consideração, como por exemplo: as características da colecção que se pretende preservar, a satisfação dos potenciais utilizadores da informação ou os custos associados ao processo de preservação (Rauch & Rauber, 2004).

Rauch e Rauber desenvolveram um método de avaliação capaz de comparar e seleccionar alternativas de preservação de acordo com as necessidades individuais de cada organização preservadora (Rauch, Pavuza, Strodl, & Rauber, 2005; Rauch & Rauber, 2004). O seu trabalho é baseado em conceitos de Análise de Utilidade, um método originalmente desenvolvido para auxílio à tomada de decisão em projectos complexos nos domínios da construção civil e economia (Weirich et al., 2001). O método resulta na ordenação de várias alternativas de preservação de acordo com os requisitos específicos manifestados pela entidade preservadora, facilitando deste modo a identificação da estratégia e parâmetros associados mais adequados num dado contexto organizacional.

1.2 Motivação

O problema geral da obsolescência tecnológica afecta todos aqueles que lidam com informação digital. Afecta indivíduos que acumulam toda uma vida de memórias materializadas em fotografias, músicas e filmes codificados em formatos digitais (Teixeira et al., 2003). Afecta organizações que produzem no seu dia-a-dia grandes volumes de informação, muita desta vital para o exercício da sua actividade. Afecta as instituições de índole cultural, como arquivos, bibliotecas e museus, onde se começam a dar os primeiros passos na incorporação de artefactos digitais com elevado valor patrimonial ou com imposições legais que determinam a sua retenção e preservação a longo-prazo.

Mesmo aqueles que não manipulam directamente informação digital dependem desta no seu dia-a-dia. A televisão que chega a suas casas é suportada por formatos digitais, assim como a música que consomem, os seus registos fiscais e financeiros, as fotografias e *outdoors* que vêem na rua, as conversas ao telemóvel, os raios-X e registos clínicos mantidos pelo seu médico de família, até mesmo os livros que lêem confortavelmente na praia ou no sofá já existiram, de uma forma ou outra, em formato digital.

É também de realçar o aumento de informação científica publicada em formatos digitais, assim como o aumento do número de repositórios responsáveis pela sua conservação (Brody, 2005; Ferreira, Saraiva, Rodrigues, & Baptista, 2008). Este facto reforça a ideia de que as questões relacionadas com a preservação de informação digital deverão ser encaradas de uma forma

concertada sendo necessário a elaboração de mecanismos que facilitem, sistematizem e validem os processos que lhe são inerentes (C. A. Lynch, 2003).

Apesar dos progressos sentidos no domínio da preservação digital, continua a existir um vazio assinalável no que diz respeito à automatização de estratégias de preservação (Ross & Hedstrom, 2005). Em paralelo, questões relacionadas com a autenticidade dos objectos digitais, a validação das actuais estratégias de preservação e a necessidade, sempre crescente, de reduzir os custos da sua implementação assumem particular destaque na lista de preocupações dos profissionais envolvidos em processos de preservação de materiais digitais (Ross & Hedstrom, 2005). Este trabalho de investigação visa contribuir para a solução destes problemas, dando especial ênfase à automatização dos processos de preservação baseados em migração.

Neste projecto propõe-se modelar e desenvolver uma arquitectura de software baseada em serviços, capaz de assistir organizações e indivíduos na selecção e execução de estratégias de preservação baseadas em migração. A arquitectura proposta incorporará também mecanismos de controlo de qualidade que garantam, de forma prática e eficaz, a autenticidade dos materiais preservados. Adicionalmente, as estratégias de migração sugeridas pela arquitectura procurarão maximizar a satisfação dos utilizadores da informação preservada e ir ao encontro das políticas de preservação definidas pela entidade preservadora.

1.3 Objectivos e contributos

Neste projecto de doutoramento procurou-se identificar e desenvolver o conjunto mínimo de serviços que facilitassem a implementação transversal de estratégias de preservação digital baseadas em migração, garantindo o cumprimento dos seguintes requisitos:

- A execução da estratégia de migração deve prescindir de intervenção humana;
- Os objectos preservados deverão permanecer autênticos[4] independentemente das intervenções de preservação a que forem sujeitos;
- As acções de preservação exercidas sobre os objectos deverão maximizar a satisfação dos seus potenciais utilizadores e as políticas da entidade responsável pela conservação dos mesmos.

[4] Ver discussão sobre Autenticidade na secção 2.5 na página 37.

É importante referir que a implementação de uma estratégia de migração pressupõe a realização de um conjunto mínimo de actividades, nomeadamente, a selecção de uma estratégia de migração de entre um conjunto alargado de opções disponíveis, a conversão dos materiais e a avaliação dos resultados obtidos numa perspectiva de controlo de qualidade.

A questão de investigação que norteou este trabalho foi:

> *Qual o conjunto de serviços que permite implementar, de forma transversal e automática, todos os processos inerentes à migração de objectos digitais num contexto de preservação digital, sem que haja prejuízo da sua autenticidade?*

De forma a dar resposta à questão de investigação apresentada, foi desenvolvida uma Arquitectura Orientada ao Serviço (SOA) capaz de auxiliar organizações e indivíduos na implementação de intervenções de preservação baseadas em migração. A arquitectura desenvolvida é constituída por um conjunto de componentes fisicamente distribuídos que permitem realizar o seguinte conjunto de actividades:

- Oferecer um conjunto alargado de serviços de migração de formatos (conversão);
- Disponibilizar um mecanismo de controlo de qualidade baseado em critérios pré-estabelecidos que permite aferir o nível de serviço prestado por cada conversor (controlo de qualidade);
- Produzir relatórios de migração que possam ser utilizados para documentar a intervenção de preservação (autenticidade);
- Fornecer sugestões de formatos de destino e/ou conversores que maximizem a satisfação da organização ou individuo (selecção).

A arquitectura que resultou deste trabalho permite a qualquer entidade-cliente realizar o conjunto de actividades previamente enunciado, bastando para tal invocar remotamente os serviços disponibilizados.

Os principais contributos desta investigação são:

Para entidades carentes de preservação digital

- A capacidade de preservar os seus objectos digitais através da conversão de formatos sem que haja necessidade de implementar localmente complexos sistemas de migração;

- A capacidade de obter relatórios detalhados sobre a migração realizada permitindo, assim, documentar todo o processo de preservação e deste modo assegurar a autenticidade dos materiais;

- A possibilidade de comparar diferentes alternativas de migração e identificar de forma objectiva qual destas é a mais adequada para satisfazer as suas necessidades organizacionais;

- A possibilidade de determinar, para cada objecto digital, o conjunto de propriedades significativas que não foram devidamente preservadas no processo de migração.

Para a indústria de software

- A possibilidade de vender as suas aplicações de conversão através da plataforma de serviços desenvolvida;

- A capacidade de avaliar de forma objectiva a qualidade das suas aplicações através de uma plataforma provida de dezenas de critérios de avaliação;

- A possibilidade de comparar o desempenho das suas aplicações com o desempenho de centenas de outras numa arena imparcial que favorece a concorrência;

- Um modelo de avaliação de mecanismos de migração/exportação que poderá vir a ser implementado por aplicações de software com suporte para múltiplos formatos e que permite ao utilizador identificar os formatos mais adequados para armazenar objectos produzidos nessa aplicação.

Para investigadores em preservação

- A identificação e caracterização dos diferentes serviços e componentes funcionais que possibilitam a implementação de estratégias de preservação baseadas em migração sem que haja prejuízo da autenticidade dos materiais;

- A recolha e desenvolvimento de funções de similaridade adequadas a diferentes tipos de dados que permitem aferir de forma objectiva a degradação incorrida ao nível das propriedades significativas dos objectos digitais devido à migração de formatos;

- O acesso a um modelo de arquitectura e respectiva implementação capaz de avaliar o desempenho de uma migração segundo múltiplos critérios, nomeadamente: performance operacional, aptidão dos formatos envolvidos e quantificação da informação perdida durante uma intervenção de preservação.

É importante referir que todas as experiências de validação realizadas em torno da arquitectura proposta tiveram como base objectos pertencentes à classe imagens matriciais.

1.4 Organização da tese

Este documento está organizado em 7 capítulos:

O primeiro capítulo, Introdução, apresenta uma visão geral sobre a investigação desenvolvida. É efectuada uma introdução à temática da preservação digital onde são descritas, sucintamente, as principais estratégias de preservação propostas pela comunidade científica. Neste capítulo são ainda apresentadas as motivações que conduziram ao desenvolvimento desta tese, a questão de investigação que a norteou e os contributos que dela resultaram.

O segundo capítulo descreve todo o trabalho que serviu de base à investigação realizada, i.e., o estado da arte. Nele são abordados temas como: o conceito de objecto digital, de preservação digital, o modelo de referência OAIS, estratégias de preservação digital, directórios de formatos, critérios para a autenticidade, metainformação de preservação e modelos de avaliação de estratégias de preservação.

O capítulo seguinte consiste num enquadramento teórico que facilita a compreensão das diferentes fases inerentes a um processo de migração. Este capítulo apresenta um cenário de preservação que permite identificar algumas das principais dificuldades com que um profissional da área se debate, servindo de ponto de partida para a identificação de um conjunto de serviços considerados indispensáveis para que seja possível automatizar processos de preservação baseados em migração. É ainda descrito em detalhe um conjunto de ferramentas que permitem implementar os serviços de preservação previamente identificados.

O quarto capítulo introduz o CRiB, a plataforma de serviços de preservação proposta nesta tese. Partindo de uma visão geral da sua arquitectura, são apresentados todos os componentes lógicos que a constituem, bem como todos os detalhes da sua implementação. É ainda neste capítulo que são apresentadas as taxionomias de avaliação utilizadas durante o processo de controlo de qualidade e recomendação de estratégias de preservação.

O quinto capítulo é dedicado à metodologia de validação dos componentes desenvolvidos. Nele são apresentadas todas as experiências realizadas em torno da plataforma e que demonstram a sua adequabilidade aos fins a que se propõe.

O sexto capítulo apresenta e descreve dois projectos com relevância nacional e internacional que adoptaram a plataforma de serviços que irá ser apresentada ao longo desta tese.

A tese termina com um conjunto de considerações finais, contributos e apontamentos sobre trabalho futuro.

Capítulo 2
Preservação digital

Desde a invenção da escrita que existe uma manifesta preocupação em torno da preservação de artefactos que resultam de processos intelectuais e criativos do ser humano (Proença & Lopes, 2004). A preservação desses artefactos permite que gerações futuras sejam capazes de compreender e contextualizar a história e a cultura dos seus povos (Lee et al., 2002). Os museus, as bibliotecas e os arquivos têm assumido, neste contexto, um papel determinante responsabilizando-se pela sua preservação e conservação.

Nos dias que correm, uma parte significativa da produção intelectual é realizada com o auxílio de ferramentas digitais. A simplicidade com que o material digital pode ser criado e disseminado através das modernas redes de comunicação e a qualidade dos resultados obtidos são factores determinantes na adopção deste tipo de ferramentas.

Apesar das inúmeras vantagens inerentes à sua utilização, o material digital acarreta um problema estrutural que coloca em risco a sua longevidade. Embora um documento digital possa ser copiado infinitas vezes sem qualquer perda de qualidade, este exige a presença de um contexto tecnológico para que possa ser interpretado por um ser humano. Esta dependência tecnológica torna-o particularmente vulnerável à rápida obsolescência a que geralmente a tecnologia está sujeita (Ferreira, Baptista, & Ramalho, 2005).

O curso da história tem revelado inúmeros exemplos fatídicos de obsolescência tecnológica. Na década de 70, a multinacional japonesa Sony introduziu um formato de vídeo designado Betamax (Figura 1). Comparativamente ao comum VHS[5] (Figura 2), a cassete Betamax era de menores dimensões e oferecia uma qualidade de imagem superior. O pico da sua popularidade foi atingido em 1983 quando cerca de um terço do mercado de vídeo doméstico era dominado por este formato (IEEE History Center; Nayak & Ketteringham, 1994; Shiraishi, 1985).



Figura 1 – Cassete de vídeo Betamax.

Apesar do seu sucesso comercial, o facto de a Sony não facilitar o licenciamento de produção a terceiros foi decisivo para que ocorresse uma viragem radical no mercado dos pequenos electrodomésticos e os consumidores adoptassem massivamente o formato VHS. Em pouco tempo, o formato Betamax desapareceu do mercado europeu e norte-americano, sendo hoje em dia muito difícil encontrar um dispositivo capaz de apresentar o conteúdo armazenado numa dessas cassetes (Nayak & Ketteringham, 1994).



Figura 2 – Cassete de vídeo VHS.

Um exemplo mais recente de obsolescência tecnológica, desta vez no domínio digital, reporta-se ao uso das populares disquetes de 3.5 polegadas (Figura 3). Em Março de 2003, o fabricante Dell Computer Corporation anunciou que os seus computadores deixariam de integrar

[5] Video Home System.

dispositivos de leitura para este tipo de suporte (Figura 4). Vários fabricantes seguiram de imediato o seu exemplo (Kenney, McGovern, Entlich, Kehoe, & Olsen, 2003).



Figura 3 – Disquete de 3.5 polegadas.

Actualmente, é ainda possível adquirir dispositivos capazes de ler disquetes de 3.5 polegadas. No entanto, o mercado inclina-se rapidamente para o uso de DVD e *flash-drives*.



Figura 4 – Leitor de disquetes de 3.5 polegadas.

É importante salientar que a obsolescência tecnológica não se manifesta apenas ao nível dos suportes físicos. Toda a informação digital tem necessariamente de respeitar as regras lógicas de um formato. Isto permite às aplicações de software abrir e processar adequadamente a informação armazenada. À medida que o software vai evoluindo, também os formatos por ele suportados vão sendo alvo de actualização.

É bastante comum encontrar aplicações de software capazes de carregar os ficheiros produzidos por versões anteriores da mesma aplicação. No entanto, essa capacidade raramente vai para além das duas versões precedentes (Kenney et al., 2003).

No mundo actual, onde cada vez mais organizações dependem da informação digital que produzem, torna-se premente a implementação de técnicas e de políticas concertadas que vão no sentido de garantir a perenidade e a acessibilidade a este tipo de informação.

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 património cultural existente em formatos digitais (Webb, 2003). A preservação digital consiste na capacidade de garantir que a informação digital permanece acessível, interpretável e

autêntica na presença de uma plataforma tecnológica diferente daquela que fora inicialmente utilizada no momento da sua criação.

Foram muitas as iniciativas que ajudaram a construir a base de conhecimento que hoje suporta o domínio científico da preservação digital. Dessas iniciativas resultaram ideias, conceitos e estratégias que levaram à discussão e ao reconhecimento universal deste problema. Neste capítulo pretende-se descrever as mais relevantes iniciativas no domínio da preservação digital, bem como contextualizar os principais conceitos que orientam a linha de pensamento que alicerça esta tese.

Este capítulo está organizado da seguinte forma: a secção 2.1 introduz o conceito de objecto digital; na secção 2.2 é introduzida alguma da terminologia que será utilizada ao longo da tese socorrendo-se para tal do modelo de referência OAIS; na secção 2.3 são descritas as principais estratégias de preservação apontadas pela comunidade científica; na secção 2.4 são descritas as iniciativas mais relevantes no domínio dos directórios de formatos; a secção 2.5 aborda questões relacionadas com autenticidade e introduz o conceito de propriedade significativa realçando a sua importância na elaboração de políticas de preservação; a secção 2.6 explora a importância da utilização de normas de metainformação como forma de dar suporte às actividades de preservação digital, dando especial ênfase ao dicionário de dados PREMIS; o capítulo termina, na secção 2.7, com um sumário e uma reflexão final que têm como objectivo relacionar os diferentes conceitos e iniciativas apresentadas ao longo do capítulo.

2.1 A anatomia de um objecto digital

Um objecto digital define-se como todo e qualquer objecto de informação que possa ser representado através de uma sequência de dígitos binários[6] (Thibodeau, 2002). Esta definição é suficientemente abrangente para acomodar tanto informação nascida num contexto tecnológico digital (objectos nado-digitais) como informação digital obtida a partir de suportes analógicos (objectos digitalizados).

Alguns exemplos elucidativos de objectos digitais são: documentos de texto, fotografias digitais, diagramas vectoriais, bases de dados, sequências de áudio/vídeo, modelos de realidade virtual, páginas *Web* e aplicações de software.

[6] Do inglês *bit stream*.

De modo a promover a compreensão e o enquadramento das diferentes estratégias de preservação que serão descritas ao longo deste capítulo, torna-se fundamental considerar e analisar os diferentes níveis a que os objectos digitais podem ser interpretados.

Para que um ser humano seja capaz de decifrar um objecto digital, há um conjunto de transformações que deverão ocorrer. Um objecto digital começa por ser um objecto físico, i.e., um conjunto de símbolos ou sinais inscritos num suporte físico (e.g. disco rígido, CD, DVD, disquete, memória-*flash*, etc.).

O suporte físico define o domínio dos símbolos a utilizar. Considere-se o seguinte exemplo: uma fotografia digital pode ser inscrita numa vasta gama de suportes físicos, no entanto, os símbolos ou sinais físicos utilizados para a representar num CD-ROM diferem substancialmente dos utilizados para a representar num disco rígido (Thibodeau, 2002). No primeiro exemplo, os símbolos utilizados são essencialmente pequenos orifícios reflectores dispostos em espiral sobre uma base de policarbonato. No segundo, são utilizados padrões magnéticos sobre um prato metálico. Independentemente do suporte utilizado, a fotografia é exactamente a mesma.

O objecto físico constitui aquilo que, geralmente, o hardware é capaz de interpretar (Figura 6). O hardware assume aqui a responsabilidade de transformar os símbolos inscritos no suporte físico num conjunto de dados que o software é capaz de manipular. Esse conjunto de dados encontra-se organizado segundo as regras decretadas pelo software utilizado na criação do objecto digital. Essas regras ou estruturas de dados constituem aquilo que vulgarmente se designa por formato de um objecto digital (Thibodeau, 2002). Essas estruturas constituem o nível de abstracção lógico ou sintáctico do objecto digital.

O software assume a responsabilidade de preparar o objecto lógico para que este possa ser devidamente apresentado a um receptor humano. Nesta fase, os sinais digitais manipulados no interior do computador são transformados em sinais analógicos que serão veiculados até ao receptor humano através de um periférico de saída (Figura 5).

A imagem que posteriormente se forma na mente do receptor constitui o que vulgarmente se designa por um objecto conceptual ou objecto semântico (Figura 6).

Os objectos semânticos assumem formas ou concepções familiares aos seres humanos, i.e., formas que existem no mundo real e que lhes são conhecidas, como livros, filmes ou fotografias. Do ponto de vista do ser humano, o objecto conceptual constitui aquilo que deve ser preservado.

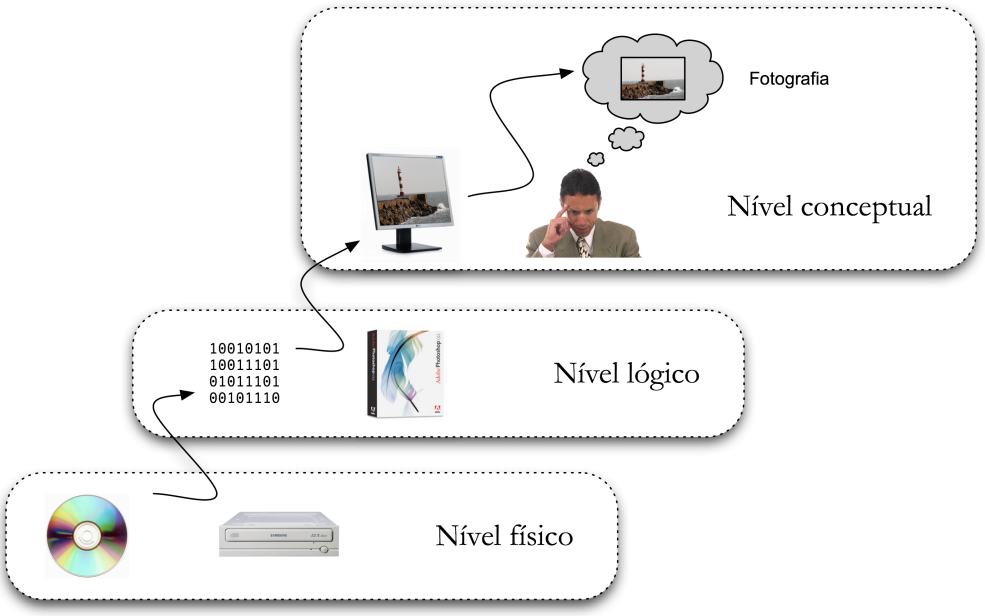


Figura 5 – Cadeia de interpretação desde o nível físico até ao nível conceptual.

Não obstante, cada ser humano acaba por fazer uma interpretação individual do objecto recebido. Essa interpretação será aqui designada por *objecto experimentado* (Figura 6). Apesar de teoricamente ser possível captar e preservar o objecto experimentado, nenhuma das estratégias de preservação apresentadas ao longo deste capítulo irão abordar seriamente esta questão.

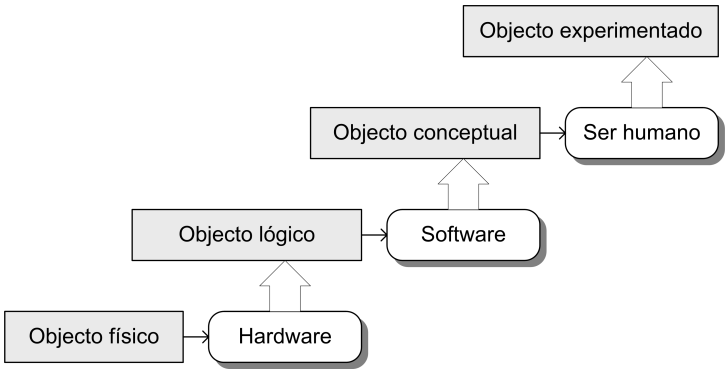


Figura 6 – Diferentes níveis de abstracção de um objecto digital.

De modo análogo, quando um ser humano assume o papel de emissor (ou produtor de informação), este mesmo conjunto de transformações é realizado, mas em sentido reverso. Nesta situação, o objecto conceptual que ganhou forma no cérebro do emissor é codificado numa linguagem passível de ser comunicada (e.g. linguagem verbal, linguagem gráfica, música, etc.). Essa linguagem poderá então ser transmitida a um receptor ou armazenada num suporte físico adequado à sua retenção, passando inevitavelmente por um processo intermédio de codificação que permite transformar a linguagem "humana" em códigos passíveis de serem processados por um computador ou outro qualquer dispositivo digital.

Numa situação ideal, o objecto conceptual formado na mente do emissor será igual ao objecto conceptual concebido pelo receptor. Somente nessa situação a comunicação poderá ser considerada perfeita.

A preservação digital é responsável por garantir que a comunicação entre um emissor e um receptor é possível, não só através do espaço, mas também através do tempo. Trata-se também de um problema de interoperabilidade, não entre sistemas contemporâneos, mas entre sistemas de épocas distintas.

Para que a preservação de um objecto digital seja possível, é necessário assegurar que todos os níveis de abstracção anteriormente descritos (i.e., físico e lógico) se mantenham acessíveis e interpretáveis. Se a cadeia de interpretação que permite elevar um objecto desde o nível físico até ao nível conceptual for interrompida, a comunicação deixa de ser possível e o objecto perder-se-á para sempre (Oltmans, Diessen, & Wijngaarden, 2004; Werf, 2002).

Segundo uma outra perspectiva, um dado objecto conceptual pode ser representado de diversas formas, ou seja, este pode ser codificado em diferentes formatos lógicos, e cada um destes ser inscrito em vários suportes físicos sem qualquer prejuízo da mensagem veiculada (Hofman, 2002a). Voltando ao exemplo anterior, é possível conceber que uma fotografia digital possa ser codificada em diversos formatos distintos, como TIFF, JPEG ou PNG, e cada um destes possa ser armazenado em diferentes suportes físicos distintos, e.g. DVD, disco rígido, memória-*flash*, cartões perfurados, entre outros. (Figura 7).

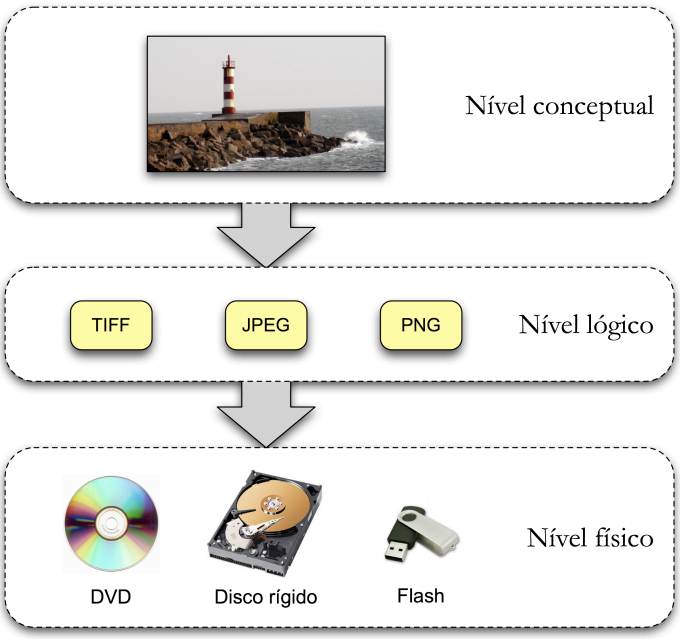


Figura 7 – Objecto digital observado a diferentes níveis de abstracção.

O enquadramento do conceito objecto digital sob uma perspectiva semiótica, i.e., recorrendo a diferentes níveis de abstracção, facilita a compreensão das diversas estratégias de preservação que serão apresentadas ao longo deste capítulo.

2.2 O modelo de referência OAIS

Em 1990, o Consultative Comitee for Space Data Systems (CCSDS) iniciou um esforço conjunto com a International Organization for Standardization (ISO) com o objectivo de desenvolver um conjunto de normas capazes de regular o armazenamento a longo-prazo de informação digital produzida no âmbito de missões espaciais.

Deste esforço nasceu o modelo de referência OAIS (Open Archival Information System), um modelo conceptual que visa identificar os componentes funcionais que deverão fazer parte de um sistema de informação dedicado à preservação digital, bem como as suas interfaces internas e externas e os objectos de informação trocados no seu interior (Consultative Committee for Space Data Systems, 2002; B. F. Lavoie, 2004). O modelo foi aprovado como uma norma internacional ISO em 2003 – ISO Standard 14721:2003 (Consultative Committee for Space Data Systems, 2002).

Um dos contributos mais notáveis desta iniciativa foi a definição de uma terminologia própria que viria a facilitar a comunicação entre os diversos intervenientes envolvidos na preservação de objectos digitais (Saramago, 2004). É importante referir que ao longo desta tese a terminologia utilizada segue o modelo de referência OAIS.

A Figura 8 ilustra os diferentes componentes funcionais, bem como os pacotes de informação trocados no interior de um repositório digital compatível com o modelo de referência OAIS.

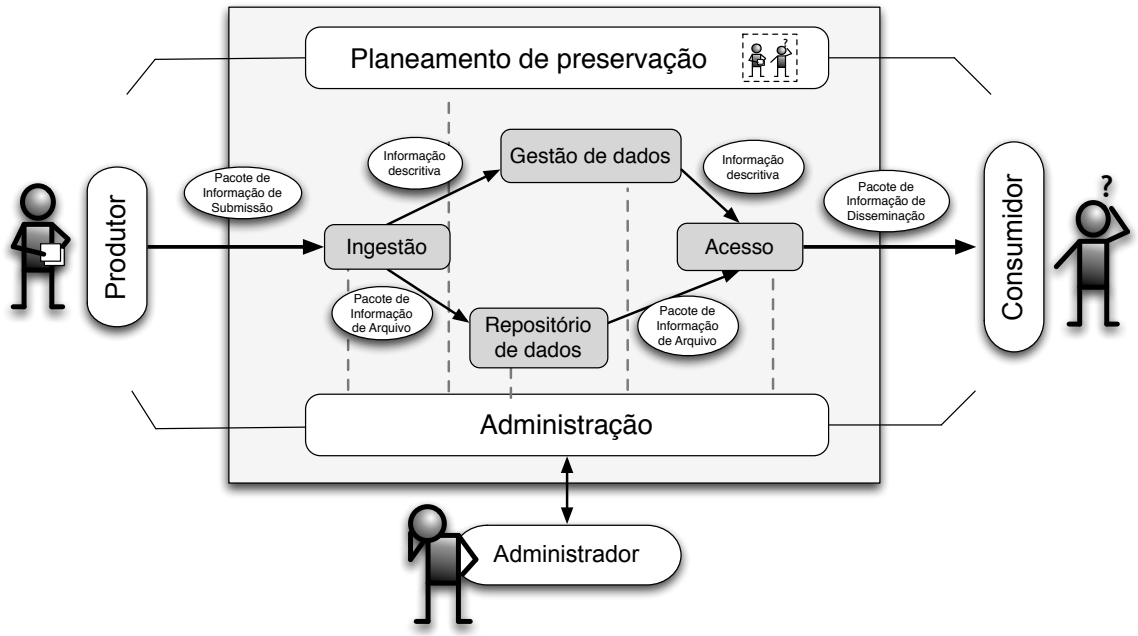


Figura 8 – Modelo de referência Open Archival Information System (OAIS).

O Produtor deverá ser entendido como a entidade externa ao repositório que se responsabiliza pela submissão de novo material no repositório. Este material é aqui representado pelo Pacote de Informação de Submissão[7] (SIP).

Durante o processo de submissão, designado neste contexto por Ingestão, o repositório é responsável por verificar a integridade da informação recebida. Ainda nesta fase, é produzida e/ou validada toda a Informação Descritiva que irá suportar a descoberta e localização do material arquivado. Em paralelo são efectuadas todas as transformações necessárias de

[7] Do inglês *Submission Information Package* (SIP).

modo a tornar o SIP apto para preservação a longo-prazo. Deste processo resulta um Pacote de Informação de Arquivo[8] (AIP), ou seja, uma estrutura de dados que será em última instância mantida e efectivamente preservada pelo repositório.

O componente de ingestão assume, assim, o importante papel de servir de interface entre o repositório OAIS e os vários produtores de informação (B. F. Lavoie, 2004).

A Informação Descritiva, vulgarmente designada por metainformação, pode ser fornecida pelo produtor ou gerada no interior do repositório. Esta informação é posteriormente armazenada e gerida pelo componente Gestão de Dados[9]. Este componente deverá, para além de guardar a informação descritiva, permitir estabelecer relações entre a metainformação descritiva e o material preservado (i.e., AIP), efectuar pesquisas sobre a metainformação e produzir relatórios sobre os conteúdos do repositório.

Por sua vez, o material a preservar (i.e., o AIP) é armazenado no Repositório de Dados[10]. Para além de guardar os objectos digitais, este componente é responsável por gerir a hierarquia de armazenamento, garantir que os objectos não são adulterados pelo suporte físico de armazenamento, efectuar verificações de integridade ao nível lógico e oferecer funcionalidades de salvaguarda e recuperação de dados em situação de desastre, e.g. RAID, cópias de segurança, etc.

O componente Planeamento de Preservação é responsável pela definição de políticas de preservação e de planos de contingência que garantam que o material arquivado permanece acessível e de acordo com os requisitos de qualidade e autenticidade exigidos pela sua comunidade de interesse[11]. Este componente é ainda responsável por monitorizar o ambiente externo ao repositório por forma a detectar modificações no panorama tecnológico vigente ou nos requisitos dos seus utilizadores que possam influenciar a forma como os objectos digitais deverão ser preservados ou disseminados. Mediante a situação, este serviço poderá desencadear eventos de preservação no interior do repositório. É da responsabilidade deste

---

[8] Do inglês *Archival Information Package* (AIP).
[9] Do inglês *Data Management.*
[10] Do inglês *Archival Storage.*
[11] Também conhecido por população potencialmente utilizadora. É de notar que o conceito de comunidade de interesse deverá ser entendido no seu sentido mais lato. Trata-se de um conceito por vezes associado a centros de documentação e bibliotecas especializadas, como é o caso de certas bibliotecas universitárias (e.g. Biblioteca de Física da Universidade do Minho em que a comunidade de interesse são os estudantes e professores de matérias ligadas à Física). Em bibliotecas de carácter geral, como bibliotecas públicas ou nacionais, e na generalidade dos arquivos este conceito não é aplicável ou apenas o será se se considerar que a comunidade de interesse coincide com a totalidade da população.

componente, por exemplo, a elaboração de estratégias de preservação e a definição dos formatos mais adequados para disseminar o material arquivado (Consultative Committee for Space Data Systems, 2002; B. F. Lavoie, 2004). É importante referir que as funções associadas a este componente são vulgarmente desempenhadas por pessoas especializadas em tecnologia e preservação digital.

O componente Acesso estabelece a ponte entre o repositório e a sua comunidade de interesse, i.e., o conjunto de potenciais Consumidores de material custodiado. Este componente é responsável por facilitar a descoberta e localização dos objectos digitais, bem como preparar os mesmos para entrega ao consumidor. A informação é entregue ao consumidor sob a forma de Pacotes de Informação de Disseminação[12], ou DIP. É de realçar que os Pacotes de Informação de Disseminação poderão ser diferentes dos Pacotes de Informação de Arquivo, ou seja, a informação entregue ao consumidor poderá ser um subconjunto da informação arquivada ou uma versão transformada da mesma (ver Migração/conversão na página 26) (Consultative Committee for Space Data Systems, 2002; B. F. Lavoie, 2004).

Por último, o componente Administração é responsável pelas operações de manutenção diárias do repositório. Entre estas encontram-se: a parametrização do sistema, monitorização dos seus processos, a execução de planos de preservação, etc. Este componente interage com todos os restantes de modo a assegurar o correcto funcionamento dos mesmos (B. F. Lavoie, 2004).

2.3 Estratégias de preservação digital

Ao longo da última década têm vindo a ser propostas inúmeras estratégias no sentido de solucionar o problema da obsolescência tecnológica. Segundo Lee et al. estas estratégias podem ser agrupadas em três classes fundamentais: emulação, migração e encapsulamento (Lee et al., 2002).

Thibodeau, por sua vez, organiza as diferentes estratégias num mapa bidimensional posicionando no seu extremo esquerdo as estratégias centradas na preservação do objecto físico e/ou lógico[13] e no extremo oposto as estratégias centradas na preservação do objecto conceptual (Figura 9). No eixo vertical as várias estratégias são dispostas mediante o seu grau

[12] Do inglês *Dissemination Information Package* (DIP).
[13] Também designada na literatura por preservação de tecnologia.

de especificidade, i.e., se são estratégias apenas aplicáveis a uma dada classe de objectos digitais ou se se tratam de estratégias genéricas, passíveis de ser administradas a qualquer classe de objectos digitais (Thibodeau, 2002).

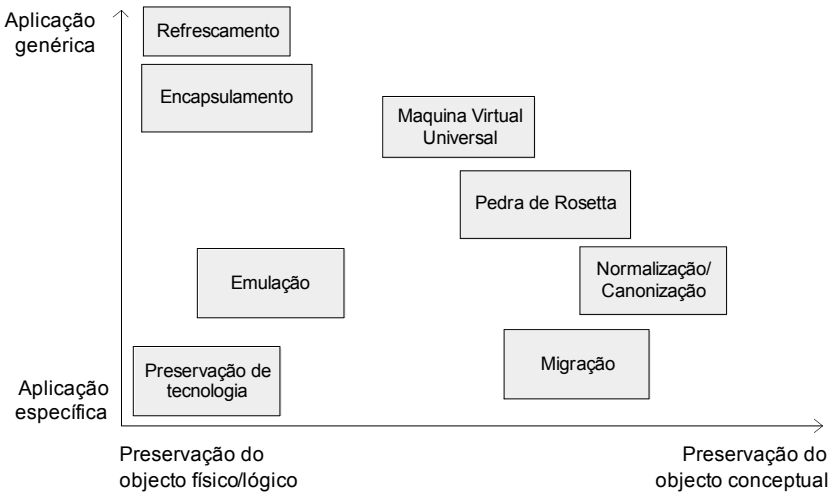


Figura 9 – Classificação das diferentes estratégias de preservação digital.

2.3.1 Preservação de tecnologia

Uma das primeiras estratégias de preservação apresentadas consiste na conservação do contexto tecnológico utilizado originalmente na concepção dos objectos digitais que se pretendem preservar. Esta estratégia consiste, essencialmente, na conservação e manutenção de todo o hardware e software necessários à correcta apresentação dos respectivos objectos digitais (Bearman, 1987; Hendley, 1998; Swade, 1998; Task Force on Archiving of Digital Information et al., 1996). Na prática, esta estratégica consiste na criação de museus de tecnologia.

Nesta estratégia, o foco não se centra na preservação do objecto conceptual, mas sim na preservação do objecto tal como este foi criado, ou seja, na sua forma mais original. Os impulsionadores desta estratégia consideram-na a única suficientemente capaz de assegurar que os objectos digitais são experimentados de forma fidedigna, i.e., que a sua autenticidade não é comprometida (Lee et al., 2002).

Contudo, a história da computação tem vindo a demonstrar que qualquer plataforma tecnológica, mesmo a mais popular, acaba inevitavelmente por se tornar obsoleta, acabando frequentemente por desaparecer sem deixar rasto (Hendley, 1998). Estratégias baseadas na conservação de tecnologia introduzem ainda dificuldades acrescidas ao nível da gestão do

espaço físico, manutenção e custo de operação, tornando-as inadequadas para aplicação a longo-prazo (Lee et al., 2002). Outras desvantagens associadas a este tipo de estratégias têm que ver com o facto de o acesso à informação ficar confinado a apenas alguns locais físicos do globo e com condicionalismos adicionais no que diz respeito à reutilização de informação (Rothenberg et al., 1999).

2.3.2 Refrescamento

Um objecto digital torna-se persistente no momento em que é inscrito num suporte físico de armazenamento (e.g. disquete, disco rígido, CD-ROM). Garantir a integridade do suporte é fundamental para que a informação nele armazenada possa ser correctamente interpretada. Se o suporte físico se deteriorar ou se se tornar obsoleto a ponto de deixarem de existir periféricos capazes de o ler, então a informação nele armazenada perder-se-á de forma irremediável (Hendley, 1998).

O refrescamento consiste na transferência de informação de um suporte físico de armazenamento para outro geralmente mais actual, antes que o primeiro se deteriore ou se torne irremediavelmente obsoleto (Bearman, 1989; Hendley, 1998; Task Force on Archiving of Digital Information et al., 1996; Woodyard, 1998).

O refrescamento não constitui uma estratégia de preservação *per se*. Em vez disso, deverá ser considerado um pré-requisito para o sucesso de qualquer estratégia de preservação (Besser, 2001). A frequente verificação de integridade dos suportes físicos, assim como o seu refrescamento periódico são actividades vitais num contexto de preservação digital.

2.3.3 Emulação

As estratégias de emulação baseiam-se na utilização de um software especial, designado Emulador, capaz de reproduzir o comportamento de uma plataforma de hardware e/ou software, numa outra que de outra forma seria incompatível (Rothenberg et al., 1999). A grande vantagem desta abordagem está na capacidade de preservar, com um elevado grau de fidelidade, as características e as funcionalidades do objecto digital original (Lee et al., 2002).

Tal como acontece em estratégias de preservação baseadas na preservação de tecnologia, as técnicas de emulação centram-se na preservação do objecto lógico não alterando o seu formato original. No entanto, este tipo de estratégias não padece de alguns dos problemas geralmente associados à criação de museus de tecnologia, como por exemplo, o desgaste do hardware e a escassez de peças para substituição.

Existem, fundamentalmente, dois tipos de emuladores: emuladores de sistemas operativos e emuladores de hardware. Os primeiros focam-se na reprodução de um sistema operativo permitindo a execução de diversas aplicações no contexto de um único emulador. Um exemplo deste tipo de emuladores é o Wine[14], um emulador que permite executar aplicações desenvolvidas na plataforma Windows em ambientes Unix. O segundo tipo de emuladores visa mimar o comportamento de uma plataforma de hardware, possibilitando que vários sistemas operativos e correspondentes aplicações possam ser executados no contexto de um único emulador (Granger, 2000; Thibodeau, 2002). Apesar de mais versáteis, este tipo de emuladores obriga à instalação de um sistema operativo completo, assim como todas as aplicações necessárias ao correcto funcionamento ou interpretação do objecto digital. Exemplos deste tipo de emuladores são: VMware Workstation (VMWare, 1998) e o Parallels Desktop (Parallels, 1995), muito utilizados actualmente para virtualizar máquinas, i.e., permitir que várias máquinas virtuais (i.e., baseadas em software e não hardware) possam ser executas concorrentemente sobre um mesmo sistema operativo de base. Existem também vários emuladores de plataformas consideradas obsoletas, e.g. ZX Spectrum (Davidson & Pollard, 2005), Nintendo NES (Krijgsman, 2005), entre outras.

Rothenberg, um dos principais promotores deste tipo de abordagens, defende um modelo teórico capaz de emular plataformas actuais em computadores do futuro. O modelo consiste na conservação do objecto digital original, juntamente com todo o software necessário à sua execução/apresentação (incluindo o sistema operativo), e na criação de uma especificação abstracta da plataforma de hardware que suporta a execução desse software. Essa especificação abstracta deverá ser escrita numa linguagem independente da plataforma e ser suficientemente rica para que um emulador possa ser construído automaticamente num qualquer computador do futuro (Rothenberg et al., 1999).

Hendley considera que a emulação apenas deveria ser utilizada em contextos onde a comunidade de interesse valoriza a preservação do ambiente tecnológico original ou ainda em situações em que os objectos digitais não são passíveis de ser convertidos para formatos mais actuais (Hendley, 1998). Outros autores consideram potencialmente arriscado confiar no software original como forma de preservar objectos digitais, uma vez que este pode ser alvo de vírus ou portador de *bugs* que poderão, no futuro, resultar em perdas substanciais de informação (Thibodeau, 2002; Waugh, Wilkinson, Hills, & Dell'oro, 2000).

[14] http://www.winehq.org/

É importante realçar que a criação de especificações capazes de descrever transversalmente plataformas de hardware não é uma tarefa simples de concretizar. Geralmente, implica recorrer a mão-de-obra altamente especializada, o que por si só poderá constituir um obstáculo considerável para a maioria das organizações (Granger, 2000; Heslop, Davis, & Wilson, 2002; Thibodeau, 2002). Para além do disposto, a criação de especificações imprecisas ou incompletas poderá impossibilitar a construção futura dos respectivos emuladores (Holdsworth & Wheatley, 2001). É também importante salientar que, com o tempo, o próprio emulador irá sofrer de obsolescência, havendo então necessidade de o converter para uma nova plataforma ou desenvolver um novo emulador capaz de emular o primeiro (Thibodeau, 2002).

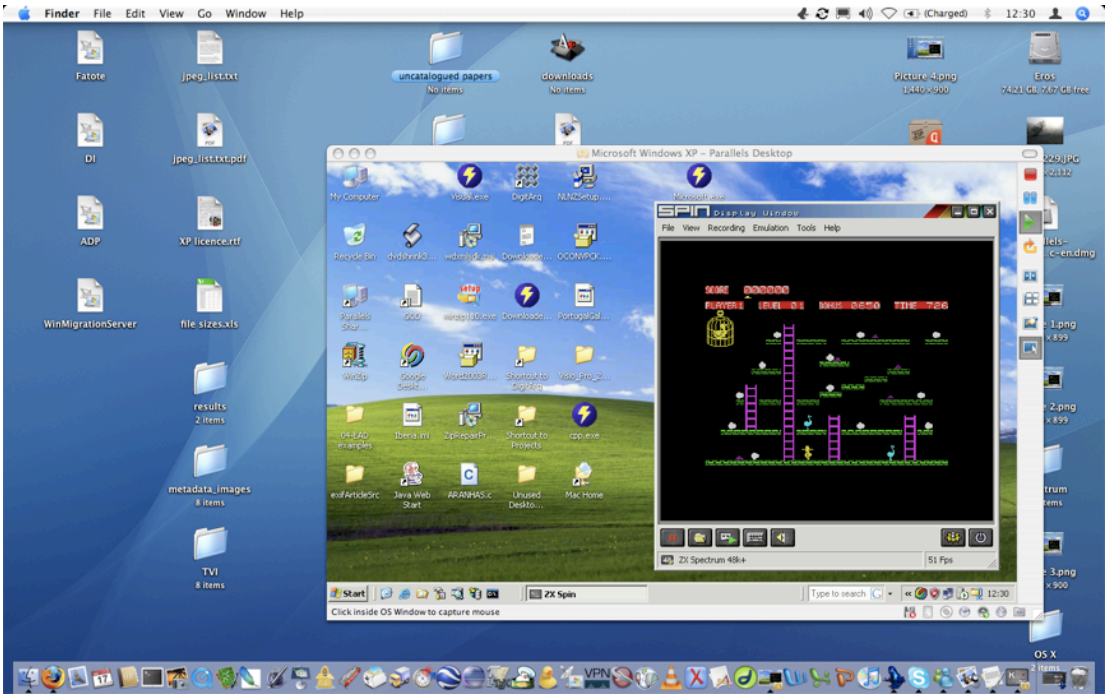


Figura 10 – Exemplo de um cenário de emulação.

O uso de emuladores parte também do pressuposto pouco sustentado de que os utilizadores do futuro serão capazes de operar adequadamente aplicações e sistemas operativos há muito desaparecidos. Por exemplo, não será razoável assumir que num futuro próximo os utilizadores possuam a aptidão necessária para enfrentar as adversidades do sistema operativo MS-DOS (Microsoft Corporation, 1981), nem tão pouco que estes terão a disponibilidade suficiente para ganhar essa capacidade apenas com o objectivo de consumir um objecto digital produzido nesse ambiente tecnológico.

A Figura 10 apresenta um cenário de emulação onde um jogo de computador está a ser executado por um emulador de ZX Spectrum, que por sua vez está a ser executado por um emulador de Windows sobre Mac OS X.

Apesar dos problemas apresentados, as estratégias de emulação continuam a assumir um papel importante na preservação de objectos digitais. Determinados tipos de objectos, especialmente aqueles dotados de características dinâmicas e/ou interactivas, poderão exigir o recurso a emuladores como única forma de garantir uma apresentação fidedigna (Woodyard, 2000). As estratégias de emulação são particularmente relevantes em contextos em que os objectos que se pretendem preservar se tratam de aplicações de software, tal como acontece actualmente com um número crescente de jogos de computador considerados de valor histórico assinalável.

2.3.4 Migração/conversão

A Migração ou Conversão 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 et al., 1996).

Os objectos digitais são constituídos por elementos de estrutura e elementos de informação. O formato de um objecto digital constitui a estrutura pela qual os elementos de informação se encontram organizados. Neste contexto, a migração pode ser vista como o processo responsável pela reorganização dos elementos de informação que constituem um objecto numa nova estrutura (Lawrence, Kehoe, Rieger, Walters, & Kenney, 2000).

Ao contrário das estratégias de preservação já apresentadas, mais focadas na cristalização do objecto digital no seu formato original, as estratégias baseadas em migração centram-se na procura de formatos alternativos para representar o mesmo conteúdo intelectual que constitui o objecto digital. Tratam-se de estratégias orientadas à preservação do objecto conceptual que desvalorizam a preservação do objecto lógico e/ou físico original (Russell, 2000).

A migração tem como principal objectivo garantir que os objectos digitais permanecem compatíveis com tecnologias actuais. Deste modo, um consumidor comum é capaz de interpretar esses objectos sem ter de recorrer a artefactos menos convencionais, como por exemplo, emuladores. No entanto, os processos de migração acarretam algumas desvantagens que deverão ser consideradas. Neste tipo de estratégias existe uma grande probabilidade de algumas das propriedades que constituem os objectos digitais não serem correctamente

transferidas para o formato de destino adoptado (Hedstrom, 2001; Heslop et al., 2002). Isto deve-se, sobretudo, a incompatibilidades estruturais entre os formatos de origem e destino ou à utilização de conversores com pouca capacidade de realizar adequadamente as tarefas a que se propõem (Ferreira, Baptista, & Ramalho, 2006a; Lawrence et al., 2000; Rauber & Aschenbrenner, 2001).

Adicionalmente, não é espectável que uma estratégia de migração possa resolver permanentemente os problemas de preservação. O formato de destino encontra-se, também este, sob constante ameaça de obsolescência, o que significa que será apenas uma questão de tempo até que uma nova migração tenha de ser ministrada. Não obstante, a migração é de longe a estratégia de preservação mais aplicada, tanto em contextos institucionais como no domínio doméstico (Lee et al., 2002).

Existem diversas variantes de migração que poderão ser consideradas: migração para suportes analógicos, actualização de versões, conversão para formatos concorrentes, normalização, migração a-pedido e migração distribuída.

Migração para suportes analógicos

A migração para suportes analógicos consiste na conversão de objectos para suportes não digitais com o intuito de aumentar a sua longevidade (Task Force on Archiving of Digital Information et al., 1996). Esta estratégia consiste, essencialmente, na reprodução de um objecto digital em papel, microfilme ou qualquer outro suporte analógico de longa duração e concentrar os esforços de preservação em torno do novo suporte.

Esta estratégia, no entanto, apenas pode ser aplicada a objectos digitais que possuam uma representação aproximada em suportes analógicos, como por exemplo, documentos de texto ou imagens. Objectos interactivos e/ou dinâmicos ficam automaticamente excluídos deste tipo de estratégias.

Actualização de versões

É bastante comum encontrar aplicações de software capazes de abrir ou importar objectos digitais produzidos por versões anteriores da mesma aplicação. Essas aplicações permitem geralmente gravar os objectos importados no formato produzido pela nova aplicação. Esta operação designa-se por actualização da versão do formato.

A actualização de versões é, possivelmente, a estratégia de preservação mais utilizada pela generalidade dos utilizadores. Essencialmente, consiste em actualizar os materiais digitais

produzidos por um determinado software, recorrendo a uma versão mais actual do mesmo (Thibodeau, 2002).

Conversão para formatos concorrentes

O processo e actualização de versões é geralmente assegurado pela organização que desenvolveu uma dada aplicação de software. A qualidade da migração depende, assim, da capacidade dos importadores fornecidos pelo fabricante do software e do grau de retrocompatibilidade oferecido pelo novo formato.

Idealmente, um fabricante asseguraria que todos os atributos presentes numa dada versão de um formato estariam disponíveis na versão que o substitui. No entanto, independentemente do sucesso económico de um fabricante ou produto de software, os formatos encontram-se constantemente sujeitos a descontinuidade (Thibodeau, 2002). Uma forma de garantir que os objectos digitais sobrevivem a este tipo de rupturas tecnológicas consiste em convertê-los para formatos associados a uma linha de produtos concorrente.

Certos formatos não são dependentes de qualquer aplicação de software. Isso permite que aplicações distintas sejam capazes de abrir e manipular objectos codificados nesses formatos, tal como acontece com grande parte dos formatos de imagem, como por exemplo, o JPEG, o TIFF ou o PNG.

Normalização

A normalização tem como objectivo simplificar o processo de preservação através da redução do número de formatos distintos que se encontram num repositório de objectos digitais (Lee et al., 2002; Thibodeau, 2002). Um número controlado de formatos permite que uma estratégia de preservação seja aplicada de forma transversal a um grande número de objectos digitais. A aplicação deste tipo de políticas de ingestão introduz uma redução generalizada dos custos de preservação, facilitando a gestão e a aplicação de eventos de preservação (Hofman, 2001).

Considere-se um exemplo. Existe um leque alargado de opções no que diz respeito a formatos para representação de imagens bidimensionais (e.g. BMP, GIF, JPEG, PNG, TARGA). Se durante o processo de ingestão todas as imagens digitais forem convertidas para um único formato, futuras intervenções ao nível da sua preservação poderão ser realizadas de forma mais simples e, consequentemente, mais económica.

A escolha do formato de normalização é um factor determinante para o sucesso desta estratégia. Sempre que possível, deverão ser escolhidos formatos reconhecidos pela comunidade de interesse e baseados em normas internacionais abertas (Heslop et al., 2002). Isto poderá evitar futuras complicações ao nível dos direitos de autor e a necessidade de pagamento de royalties (Ayre & Muir, 2004). Paralelamente, o formato de normalização deverá ser suficientemente rico para que as características fundamentais dos vários formatos possam ser devidamente incorporadas.

A normalização promove, também, a interoperabilidade entre sistemas distintos. Ao serem utilizados formatos abertos e independentes da plataforma, diferentes configurações de hardware e software serão capazes de os interpretar (Howel, 2004; Thibodeau, 2002).

A normalização de formatos pode ser implementada de diversas formas. Determinados repositórios procedem à conversão automática dos objectos depositados para formatos únicos de preservação. Outros, definem políticas de arquivo que limitam os formatos em que aceitam informação, o que significa que cabe aos produtores de informação a tarefa de converter os seus objectos digitais para os formatos estipulados (Hedstrom, 1998; Hodge & Frangakis, 2004). O argumento que suporta a segunda abordagem assenta no pressuposto de que os produtores de informação serão as entidades mais adequadas para avaliar a qualidade da conversão efectuada.

Migração a-pedido

O sucesso de uma migração depende, fundamentalmente, da qualidade dos conversores utilizados e da capacidade apresentada pelo formato de destino em acomodar o conjunto de propriedades do formato de partida. Poder-se-á assumir que sempre que é efectuada uma migração, os objectos digitais resultantes são de alguma forma diferentes dos objectos de partida. Ao fim de algumas conversões sucessivas, os objectos preservados poderão ser substancialmente diferentes dos objectos originais (Figura 11). Para combater este fenómeno surgiu uma estratégia designada por migração a-pedido (Mellor, Wheatley, & Sergeant, 2002).

Neste tipo de migração, ao invés das conversões serem aplicadas ao objecto mais actual, estas são sempre aplicadas ao objecto original (Figura 12). Deste modo, se de uma dada conversão resultar um objecto substancialmente diferente do original, numa futura conversão o problema poderá ser resolvido recorrendo a um conversor de melhor qualidade ou a um formato de destino mais adequado.

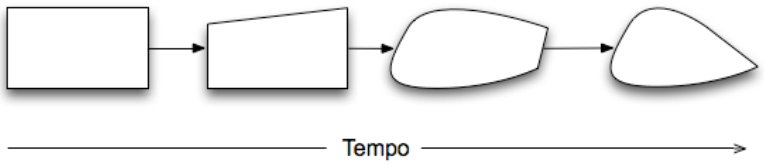


Figura 11 – Degradação do objecto digital ao longo de sucessivas migrações.

Esta abordagem possui como principal vantagem o facto de, uma vez construído o módulo de descodificação do conversor (i.e., o módulo capaz de ler as propriedades do formato de origem), apenas ser necessário desenvolver o codificador específico para cada formato de saída. Não obstante, será necessário suportar ao longo do tempo um conjunto alargado de conversores de modo a garantir a capacidade de transformar os objectos armazenados nos seus formatos originais para formatos que sirvam adequadamente as necessidades dos seus consumidores.

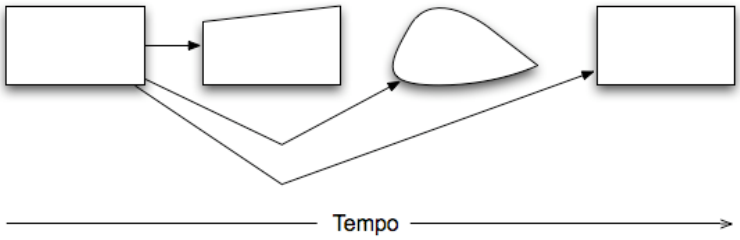


Figura 12 – Migração a-pedido.

Migração distribuída

Os mais recentes desenvolvimentos no contexto da migração introduzem arquitecturas distribuídas de conversores (Figura 13). Neste tipo de migração, existe um conjunto de serviços de conversão que se encontram acessíveis através da rede ou da Internet e que poderão ser invocados remotamente recorrendo a um pequeno módulo de software ou aplicação-cliente.

Existem actualmente várias iniciativas que visam o desenvolvimento deste tipo de conversores. O Typed Objects Model (TOM) implementa 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).

No Lister Hill National Center for Biomedical Communications (LHNCBC) foi desenvolvido um Web service que converte cinquenta formatos distintos para PDF. Para além do serviço disponibilizado, o LHNCBC oferece uma aplicação designada MyMorph que permite a qualquer utilizador tirar partido do serviço publicado (Walker & Thoma, 2003, 2004, 2005).

Hunter e Choudhury dão um passo em frente no seu projecto PANIC propondo uma rede de serviços de conversão suportada por uma descrição semântica que possibilita a sua descoberta e invocação automática por agentes de software (Hunter & Choudhury, 2004, 2005, 2006).

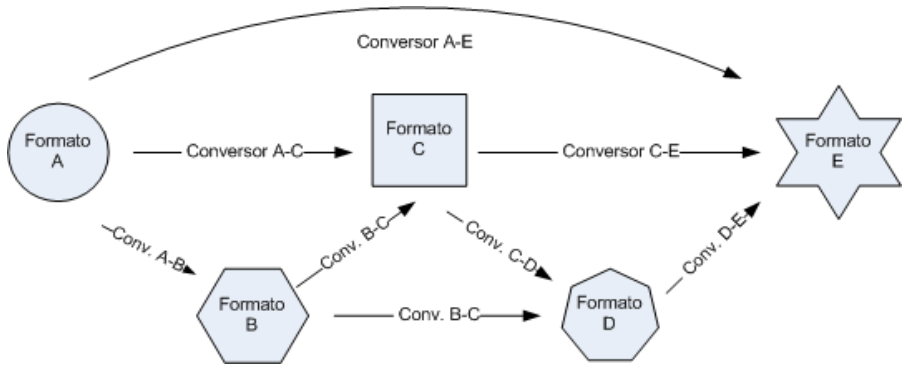


Figura 13 – Migração distribuída baseada em Serviços Web.

Este tipo de migração apresenta algumas vantagens face às estratégias de migração mais convencionais, nomeadamente:

- A utilização de serviços de conversão 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;
- A existência de múltiplos caminhos de migração permite à solução resistir ao desaparecimento gradual de parte dos conversores;
- Este tipo de abordagem é compatível com uma série de variantes de migração, como por exemplo, normalização e migração a-pedido;
- A criação de uma rede global de conversores poderá conduzir a uma redução generalizada dos custos de preservação. Pequenas e grandes organizações poderão amortizar o seu investimento no desenvolvimento de conversores, publicando-os na rede de serviços e cobrando uma pequena taxa pela sua utilização.

Apesar das vantagens apresentadas, a migração distribuída poderá não ser uma solução adequada a todos os contextos de utilização. Um repositório de informação digital pode facilmente conter milhares de itens, atingindo níveis de armazenamento na ordem dos Terabytes. Transferir através da Internet um volume de informação desta natureza acarreta custos que poderão ser impeditivos para muitas organizações. Para além disso, requisitos em termos de largura de banda, segurança dos dados e tempo de transferência poderão ser factores determinantes para o insucesso de estratégias desta natureza.

2.3.5 Encapsulamento

Por vezes não é fácil determinar o valor intrínseco de determinados objectos digitais. Poderão passar-se muitos anos até que a comunidade de consumidores revele um particular interesse por uma determinada colecção de objectos (Heminger & Robertson, 2004). Durante esse tempo, o material custodiado poderá nunca ser consultado. Neste tipo de cenários, estratégias de preservação que carecem de uma diligência contínua (e.g. migração) poderão revelar-se demasiado onerosas. As soluções baseadas em encapsulamento procuram resolver este problema, mantendo os objectos digitais inalterados até ao momento em que se tornam efectivamente necessários.

A estratégia de encapsulamento consiste em preservar, juntamente com o objecto digital, toda a informação necessária e suficiente para permitir o futuro desenvolvimento de conversores, visualizadores ou emuladores. Esta informação poderá consistir, por exemplo, numa descrição formal e detalhada do formato do objecto preservado (Digital Preservation Testbed, 2001).

O Formato Universal de Preservação[15] (UPF) trata-se de uma iniciativa que visa criar um formato normalizado e auto-descritivo para armazenar informação digital. Este formato é independente da aplicação, do sistema operativo e do suporte físico utilizados na criação do objecto digital (T. Shepard & MacCarn, 1998, 1999).

Raymond Lorie propõe uma alternativa a esta estratégia substituindo a especificação formal por uma aplicação de software compilada para uma máquina virtual universal, por exemplo, para a Java Virtual Machine (Raymond A. Lorie, 2001; Raimond A. Lorie, 2002). Esta aplicação é na realidade um descodificador[16] e tem como finalidade apresentar uma visão lógica do objecto digital permitindo, deste modo, uma navegação simples através das suas

[15] Do inglês *Universal Preservation Format*.
[16] Do inglês *decoder*.

propriedades. Lorie argumenta que a máquina virtual universal é suficientemente simples para que possa ser implementada em qualquer arquitectura de hardware futura.

2.3.6 Pedra de Roseta digital

O povo egípcio deixou uma infindável quantidade de vestígios da sua presença na Terra. Entre estes encontram-se as famosas pirâmides de Gizé[17] e inúmeras peças de arte. Muitos destes artefactos eram adornados com hieróglifos. Apesar de estes existirem há mais de 5000 anos, só a partir do século XIX é que foi possível decifrar o seu significado. Tudo aconteceu em 1799 quando um grupo de soldados franceses descobriu no delta do Nilo um bloco de granito que ficou conhecido como a Pedra de Roseta (Figura 14). Nela encontrava-se escrito em três línguas distintas, egípcio hieroglífico, egípcio cursivo e grego clássico, um decreto emitido em 196 a.C. por Ptolomeu V Epifânio. Em 1822 o paleógrafo francês Jean-François Champollion descodificou a versão egípcia do texto recorrendo aos seus conhecimentos de grego clássico, um idioma bem conhecido dos historiadores da época (Wikipedia contributors, 2005). Esta descoberta conduziu à descodificação de inúmeros outros textos egípcios encontrados nos mais variados locais e suportes (e.g. monumentos, rochas, papiros).

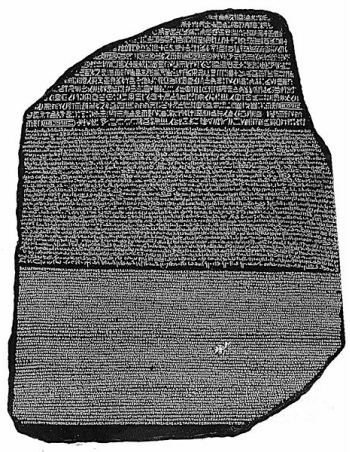


Figura 14 – Pedra de Roseta.

Heminger e Robertson propõem a utilização de uma estratégia semelhante para recuperar objectos digitais para os quais não existe informação suficiente sobre o seu formato (Heminger & Robertson, 2004). Nesta estratégia, em vez de se preservar as regras que permitem descodificar o objecto digital, são reunidas amostras de objectos que sejam representativas do

[17] Gizé ou Guiza, nome mais próximo do original.

formato que se pretende recuperar. Estas amostras deverão existir num formato que possa ser directamente interpretado pelo ser humano. Trata-se do conjunto de referência, i.e., a versão grega do decreto inscrito na Pedra de Roseta. Com esta informação seria possível inferir as regras necessárias para traduzir/converter o objecto original para um qualquer formato contemporâneo (Heminger & Robertson, 1998; Thibodeau, 2002).

Um exemplo de aplicação desta estratégia consiste em imprimir em papel um conjunto representativo de documentos de texto juntamente com a sua representação binária. No futuro, as regras necessárias para interpretar e migrar os objectos para um novo formato poderiam ser inferidas, comparando os documentos impressos com a sua representação binária (Thibodeau, 2002).

Trata-se sobretudo de uma ferramenta de arqueologia digital e não propriamente de uma estratégia sólida para preservação de objectos digitais (Heminger & Robertson, 2004). Esta apenas deverá ser considerada em situações em que todos os outros esforços de preservação fracassaram.

2.4 Directórios de formatos

Uma das formas de minimizar a ansiedade de todos os profissionais envolvidos na preservação de objectos digitais consiste na criação de directórios centralizados de informação técnica sobre formatos digitais. Esta informação inclui, por exemplo, a identificação dos produtores de um dado formato, a sua data de criação, informação sobre as aplicações que o suportam, especificações técnicas, grau de obsolescência, entre outros.

Para além de disponibilizar este tipo informação, os directórios de formatos poderão prestar serviços avançados de apoio à preservação digital. Por exemplo, um directório de formatos poderá disponibilizar serviços ou ferramentas para detecção e identificação de formatos e promover o uso de vocabulários controlados para os seus descritores. Poderá ainda fornecer especificações técnicas sobre formatos que permitam a qualquer instituição desenvolver descodificadores, bem como disponibilizar um conjunto de informações relevantes de apoio às actividades de preservação digital, como por exemplo, informação sobre a cota de mercado de um dado formato, tendências de utilização ou produzir recomendações quanto aos formatos mais apropriados para preservação a longo-prazo.

Existem actualmente diversas iniciativas que visam a implementação de directórios deste tipo. Alguns exemplos são: os Mime Media Types (Freed & Borenstein, 1996), o PRONOM (UK

National Archives, 2002), o Global Digital Format Registry (Abrams & Seaman, 2003) e o projecto Typed Object Model (Ockerbloom, 1998).

Actualmente, o sistema de identificação de formatos mais utilizado é o MIME Media Types (Freed & Borenstein, 1996). Este sistema é amplamente utilizado na Internet para especificar as regras de codificação/descodificação de documentos anexados a mensagens de correio electrónico e para identificar os formatos de dados trocados entre servidores Web e browsers. Não obstante, este sistema não possui a granularidade necessária para identificar de forma unívoca todos os formatos existentes. Por exemplo, as várias versões da família PDF, desde a versão 1.2 à 1.7, PDF/X da versão 1 à 3 e PDF/A, são todas elas identificadas através do mesmo descritor: `application/pdf`.

O PRONOM Technical Registry[18] é uma iniciativa dos Arquivos Nacionais do Reino Unido que visa a concentração de informação técnica sobre software e formatos associados (Darlington, 2003; UK National Archives, 2002). O modelo de dados que suporta o PRONOM incorpora vários elementos de informação, tais como: descritores de formatos, identificadores únicos de formato, esquemas de codificação de caracteres[19], algoritmos de compressão, sistemas operativos de suporte, hardware específico e ligações para outras fontes de informação. O PRONOM disponibiliza ainda uma ferramenta de identificação de formatos de nível local – o Droid (UK National Archives, 2005). O Droid é uma pequena aplicação multiplataforma que permite identificar o formato de um objecto digital recorrendo à base de dados de informação disponibilizada pelo PRONOM.

O Global Digital Format Registry[20] (GDFR) apresenta-se como uma alternativa aos actuais MIME Media Types, introduzindo um mecanismo de identificação de formatos mais preciso e rigoroso. O GDFR possui, ainda, como objectivo a reunião de informação sobre a sintaxe e semântica dos diversos formatos digitais por ele reconhecidos. A sua criação está a cargo de um grupo de trabalho internacional, constituído por membros de diversas bibliotecas e arquivos nacionais, assim como bibliotecas académicas, num total de 18 instituições (Abrams & Seaman, 2003).

O projecto Typed Object Model[21] (TOM) assenta no pressuposto de que todos os formatos digitais podem ser vistos como objectos (i.e., possuidores de propriedades e métodos) e, como

---

[18] http://www.nationalarchives.gov.uk/PRONOM/
[19] Do inglês *encoding*.
[20] http://hul.harvard.edu/gdfr/
[21] http://tom.library.upenn.edu/

tal, será possível construir uma arquitectura baseada em herança, capaz de descrever a estrutura de cada formato, as suas instâncias e as relações existentes entre os mesmos (Ockerbloom, 1998). Este projecto introduz uma taxionomia classificativa de formatos e um sistema distribuído de conversores baseado em agentes mediadores. Apesar da sua complexidade e riqueza, não se antevê que o TOM possa vir a tornar-se uma norma *de facto* no contexto dos directórios de formatos, uma vez que a sua utilização é meramente residual.

Para além das iniciativas anteriormente descritas existem outras que também merecem ser mencionadas. A Biblioteca do Congresso disponibiliza um conjunto de páginas Web com informação sobre formatos e seus variantes[22] (Brown, 2008). Apesar de apenas reunir informação sobre um conjunto reduzido de formatos, a informação disponibilizada é extremamente rica, incluindo informação descritiva sobre o formato, características técnicas, relações com outros formatos, documentação produzida pelo fabricante e informação especifica sobre a sua preservação.

Na Universidade de Maryland foi desenvolvido um projecto designado FOCUS[23] (Format Curation Service) que tem como objectivo servir de prova de conceito de um directório de formatos global baseado em tecnologias Web, tais como o LDAP e Web services. O directório foi desenhado de modo a suportar uma vasta gama de serviços, incluindo: identificação de formatos, verificação de integridade, disponibilização de aplicações de visualização, migração, caracterização de formatos, entre outros (Brown, 2008; Geremew, Song, & J. JaJa, 2006). Infelizmente, a partir de Janeiro de 2008 o serviço de demonstração deste projecto deixou de estar disponível.

O Digital Curation Centre (DCC) também tem vindo a desenvolver o seu próprio directório de formatos – o Representation Information Registry Repository[24] (RIRR). Este directório tem como principal objectivo implementar e estender o modelo de dados de informação de representação definido pela norma OAIS (Brown, 2008). Futuros desenvolvimentos em torno desta iniciativa são esperados no âmbito do projecto CASPAR[25].

[22] http://www.digitalpreservation.gov/formats/
[23] http://www.umiacs.umd.edu/research/adapt/focus/
[24] http://registry.dcc.ac.uk/omar/
[25] http://www.casparpreserves.eu/

2.5 Autenticidade

O conceito de autenticidade está longe de ser consensual entre os profissionais da preservação. Este poderá assumir significados consideravelmente diferentes consoante a comunidade que o manipula. Para um historiador um objecto é autêntico se a sua identidade e integridade não forem comprometidas (i.e., se o objecto for original) e, não menos importante, se o objecto for verdadeiro (Cullen, 2000). Na perspectiva de um arquivista, a autenticidade de um objecto não pressupõe que este seja verdadeiro. Um arquivista preocupa-se, sobretudo, com a prova que um documento poderá constituir. Este poderá conter incorrecções, erros ou até falsidades, mas isso não invalida a sua importância como testemunho de que algo aconteceu (Hirtle, 2000). Um documento falsificado, por exemplo, pode ser considerado autêntico uma vez que constitui prova de que alguém falsificou um documento (Hofman, 2002b).

Definições mais abrangentes de autenticidade giram em torno de conceitos como autenticação, integridade, completude, veracidade, validade, conformidade com o original, significância e adequabilidade ao fim a que se destina (Rothenberg, 2000).

Em termos genéricos, o conceito de autenticidade traduz-se na capacidade de descrever os elementos diplomáticos que permitem evidenciar que um dado objecto é autêntico. Trata-se da identificação do "porquê", do "quando", do "onde" e do "por quem" de um objecto digital (Hofman, 2002a). A autenticidade num contexto digital tem que ver com a capacidade de se conseguir demonstrar que um objecto digital é aquilo que se propõe ser (Authenticity Task Force, 2002; Hofman, 2001; B. Lavoie & Gartner, 2005; C. Lynch, 2000; MacNeil et al., 2001; Millar, 2004). Para atingir esse objectivo é fundamental documentar convenientemente a proveniência do objecto, contextualizar a sua criação e existência, descrever a sua história custodial e atestar que a sua integridade não foi comprometida, i.e., provar que o conjunto de propriedades que se consideram essenciais à interpretação do objecto não foram adulteradas ao longo do tempo (Diessen & Werf-Davelaar, 2002; B. Lavoie & Gartner, 2005). Num contexto digital, autenticidade não tem tanto que ver com o demonstrar que um objecto é original, mas sim, que está conforme o original.

Os problemas associados à determinação da autenticidade de um objecto não estão limitados à documentação digital. Na idade média, por exemplo, a reprodução de livros era realizada manualmente. Cada cópia de um livro apresentava, frequentemente, um conjunto de diferenças face ao original. A maior parte dessas diferenças resultavam de infelizes erros de transcrição. No entanto, não seriam raras as vezes em que escrivães mais perspicazes introduziam deliberadamente "melhorias" durante o processo de transcrição do documento

(Akester, 2004). No contexto digital, os problemas relacionados com a autenticidade são em tudo semelhantes aos do mundo analógico. Contudo, a simplicidade com que alterações podem ser introduzidas, a rapidez com que estas podem ser disseminadas e a dificuldade inerente à sua detecção tornam este problema sensivelmente mais complexo.

No contexto analógico, o conteúdo e o suporte são geralmente duas entidades inseparáveis. As propriedades físicas que caracterizam o suporte fornecem, geralmente, pistas suficientes para que a autenticidade do seu conteúdo possa ser aferida (Hofman, 2002b). No mundo digital este tipo de pistas não existe. O ambiente tecnológico é propício à introdução de modificações, provocando um clima generalizado de desconfiança em relação à autenticidade deste tipo de material (Akester, 2004; P. Graham, 2000; C. Lynch, 2000; MacNeil et al., 2001).

As estratégias de preservação de informação no domínio analógico manifestam-se, sobretudo, pela tentativa de conservar o suporte inalterado durante o máximo de tempo possível. No contexto digital, a preservação do suporte ou da sequência de *bits* que constitui o objecto, não é condição suficiente para garantir que a informação permanece acessível, reutilizável e autêntica ao longo do tempo (The Cedars Project Team, 2001). Preservar informação digital consiste, por vezes, em modificar ou transformar deliberadamente o objecto físico ou lógico que veicula a mensagem (ver Migração/conversão na página 26). Para que essas modificações não perturbem a mensagem, é fundamental definir quais as propriedades da mensagem que deverão ser conservadas durante o processo de transformação.

Paralelamente, a informação não é um conceito ou substância concreta. Esta materializa-se através de um processo de interpretação que transforma um conjunto de símbolos em algo com significado (Diessen, 1997). A interpretação desse significante está sujeita a influências adicionais. O hardware e o software que servem de mediadores nesse processo podem diferir substancialmente de consumidor para consumidor (Diessen & Werf-Davelaar, 2002). Neste contexto, a definição de essência[26] de um objecto digital é de extrema importância, pois caracteriza o conjunto de propriedades que deverão ser mantidas e preservadas de forma intacta para que o objecto possa ser considerado autêntico, ou seja, de acordo com o original (Hofman, 2002b).

O conjunto de propriedades significativas, i.e., aquelas que definem a essência do objecto, não é universal nem tão pouco absoluto. A sua definição deverá ter em conta a natureza da organização responsável pela preservação, as características da colecção e, acima

[26] Propriedades significativas e essência de um objecto digital são duas expressões vulgarmente utilizadas para representar o mesmo conceito.

de tudo, os requisitos e exigências da sua comunidade de interesse (Beagrie et al., 2002; Hofman, 2002b). A definição das propriedades significativas de um objecto digital influencia directamente a forma como este deverá ser preservado. Quanto maior for o número de propriedades significativas, maiores serão os requisitos relativamente à infra-estrutura tecnológica necessária para suportar a sua preservação (Rusbridge, 2003; The Cedars Project Team, 2001).

Embora desejável, a definição de um conjunto de propriedades significativas para cada objecto digital existente num repositório não é economicamente viável. Torna-se, assim, necessária a criação de políticas de preservação que exprimam, para cada classe de objectos, o conjunto das propriedades significativas que serão asseguradas pelo repositório (Rusbridge, 2003).

A título de exemplo, considere-se uma biblioteca responsável por preservar artigos científicos (o repositório institucional da Universidade do Minho, por exemplo). Se a sua política de preservação apenas especificar a propriedade significativa: preservação do conteúdo textual dos artigos científicos depositados; então, estes estarão a ser adequadamente preservados, se se mantiverem apenas os caracteres ASCII[27] que os constituem. Se por outro lado a política de preservação especificar propriedades significativas adicionais como a disposição do texto na página ou a sua formatação em termos de parágrafos e tipos de letra, então a preservação dos caracteres ASCII deixa de ser suficiente, passando a ser necessário recorrer a formatos mais complexos, como por exemplo o PDF.

Ainda neste contexto surge o conceito de canonização. Lynch apresenta-o como uma forma de avaliar o sucesso de uma migração (C. Lynch, 1999). O formato canónico tem como objectivo representar de forma unívoca as características essenciais de uma classe de objectos digitais. O formato canónico especifica a ordem e a estrutura das propriedades que constituem os objectos digitais e assume valores por omissão para as propriedades que não possuem valores associados. O método funciona comparando os objectos canónicos obtidos a partir dos objectos originais e convertidos (Figura 15). Por exemplo, se se considerar dois objectos digitais, o original e o convertido, poder-se-á afirmar que o objecto convertido preserva as características essenciais do objecto original se os objectos canónicos obtidos a partir destes forem iguais (C. Lynch, 1999).

[27] American Standard Code for Information Interchange. Trata-se de um conjunto de códigos capaz de representar letras, dígitos e outros símbolos, amplamente utilizado por computadores para troca de informação textual.

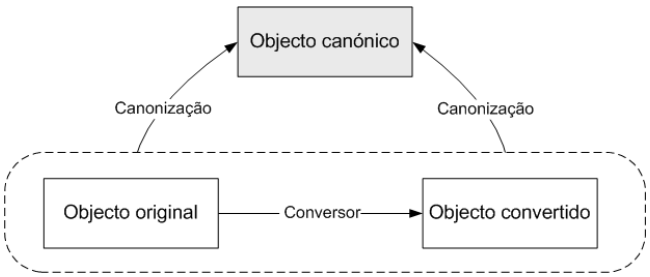


Figura 15 – Verificação da qualidade de uma migração através de canonização.

2.6 Metainformação de preservação

A metainformação de preservação tem como objectivo descrever e documentar os processos e actividades relacionados com a preservação dos materiais digitais. A metainformação de preservação é responsável por reunir, junto do material custodiado, informação detalhada sobre a sua proveniência, autenticidade, actividades de preservação, ambiente tecnológico e condicionantes legais (B. Lavoie & Gartner, 2005).

No que diz respeito à proveniência, a metainformação de preservação procura descrever a história custodial dos materiais, i.e., o caminho percorrido por estes desde a sua criação até à sua incorporação no repositório (B. Lavoie & Gartner, 2005). Esta assume também a responsabilidade de garantir a autenticidade dos mesmos. Para tal, agrega um conjunto de metainformação que descreve detalhadamente as actividades desenvolvidas no interior do repositório, especialmente aquelas que interagem directamente com os objectos digitais custodiados (B. Lavoie & Gartner, 2005).

A metainformação de preservação serve também para descrever o ambiente tecnológico necessário à correcta execução e apresentação dos objectos digitais (i.e., hardware, sistemas operativos e software) (B. Lavoie & Gartner, 2005).

2.6.1 PREMIS

O modelo de referência OAIS constituiu um ponto de partida para a discussão em torno da necessidade de criar um conjunto de elementos de metainformação capazes de dar suporte às actividades relacionadas com a preservação digital (Consultative Committee for Space Data Systems, 2002; B. Lavoie & Gartner, 2005). Desde o seu aparecimento que diversas instituições têm vindo a propor dicionários de metainformação que reflectem as necessidades individuais dos projectos em que estão ou estiveram envolvidas (Lupovici & Masanès, 2000;

National Library of Australia, 1999; The Cedars Project Team, 2002). Em 2002, o consórcio Online Computer Library Center e Research Libraries Group (OCLC/RLG) compilou o conhecimento resultante desses projectos num único documento onde se destacam as diversas classes de informação que deverão estar presentes num esquema de metainformação de preservação (OCLC/RLG Preservation Metadata Working Group, 2002).

Em 2003, a OCLC/RLG constituiu um segundo grupo de trabalho designado PREMIS (PREservation Metadata: Implementation Strategies) com o objectivo de continuar o desenvolvimento deste esquema de metainformação. O grupo de trabalho foi constituído por um comité internacional com mais de trinta especialistas em preservação digital. Deste trabalho resultou o Dicionário de Dados PREMIS[28], um documento que identifica e descreve um conjunto básico de elementos de metainformação de suporte à preservação digital, bem como um conjunto de recomendações quanto à forma como estes deverão ser utilizados no contexto de um arquivo digital (PREMIS Working Group, 2005).

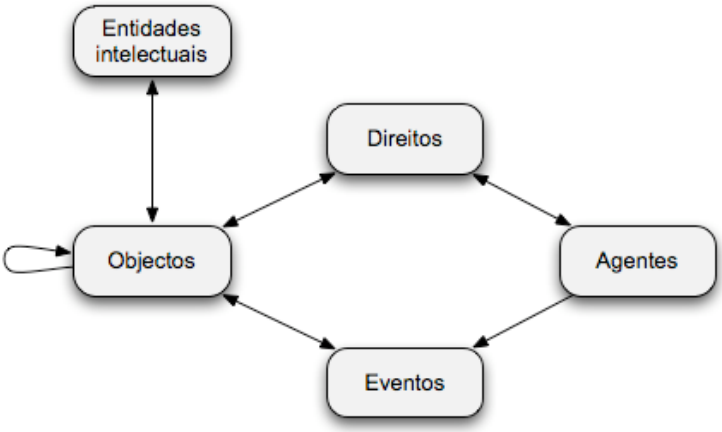


Figura 16 – Entidades presentes no Dicionário de Dados PREMIS.

O Dicionário de Dados PREMIS identifica cinco componentes fundamentais: entidades intelectuais, agentes, eventos, objectos e direitos[29] (Figura 16).

Uma entidade intelectual é um conjunto coerente de informação que pode ser identificado e descrito como uma unidade. Um livro, uma fotografia ou uma base de dados são exemplos do que pode ser considerado uma entidade intelectual.

[28] Do inglês *PREMIS Data Dictionary*.
[29] Do inglês *Intellectual Entities, Agents, Events, Objects* e *Rights*.

Uma entidade intelectual pode conter outras entidades intelectuais no seu interior. Um sítio *Web*, por exemplo, pode ser constituído por várias páginas *Web* e cada uma destas ser composta por um conjunto de imagens. Cada uma dessas páginas pode ser vista como uma entidade intelectual. De modo análogo, cada uma das suas imagens pode ser considerada uma entidade intelectual por si só. Tudo depende da granularidade a que se pretende estabelecer o conceito.

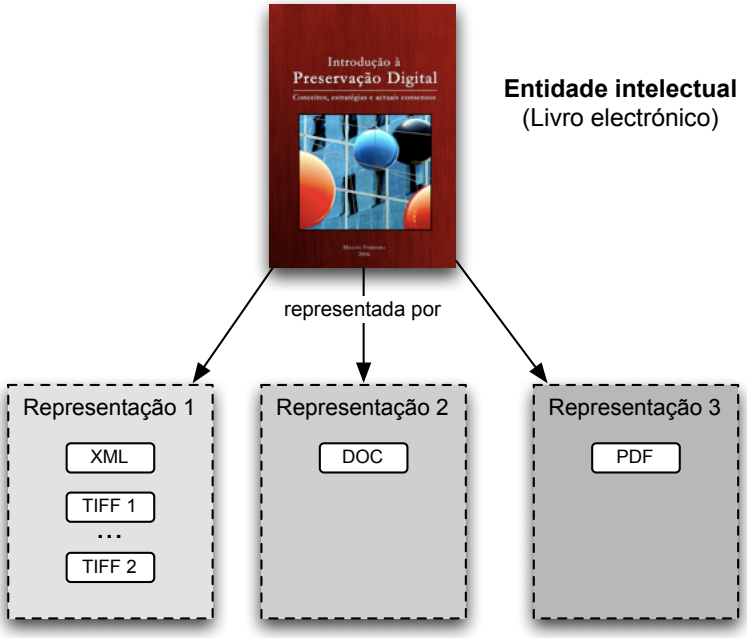


Figura 17 – Diferentes representações para a mesma entidade intelectual.

É importante realçar que uma entidade intelectual pode estar associada a mais do que uma representação. Um livro electrónico, por exemplo, pode ser representado de várias formas, como por exemplo: através de conjunto de imagens em formato TIFF e um ficheiro XML que descreve a sequência correcta de apresentação das mesmas, através de um documento Word ou em formato PDF. A entidade intelectual[30] que descreve cada uma das representações é sempre a mesma, apesar da sua manifestação física variar significativamente (Figura 17).

A entidade agente descreve qualquer pessoa, organização ou aplicação de software envolvida num evento de preservação. Por sua vez, um evento agrega informação sobre as acções de preservação realizadas em torno de um objecto (e.g. verificação de integridade, migração,

[30] De notar o paralelismo entre o conceito de entidade intelectual e objecto conceptual.

rastreio de vírus, etc.). O registo das acções de preservação, especialmente aquelas que têm como finalidade modificar o objecto digital, é considerado uma actividade fundamental para a manutenção e conservação da autenticidade dos materiais arquivados.

A entidade objecto é responsável por descrever o conjunto de representações, ou manifestações físicas, de uma entidade intelectual. Um objecto pode ainda ser dividido em três subtipos: representação, ficheiro ou sequência de bits[31]. Uma representação é um conjunto de ficheiros com metainformação estrutural associada. Este conjunto de ficheiros permite a apresentação de um objecto conceptual de forma completa.

Um ficheiro é um conjunto ordenado de bits reconhecido por um sistema operativo. Um ficheiro pode assumir propriedades como: tamanho, data de criação/modificação, permissões de acesso, etc. Por último, uma sequência de bits é um conjunto de dados coeso e com particular interesse para efeitos de preservação e que pode ser identificado e extraído do interior de um ficheiro, e.g. a faixa de áudio num ficheiro de vídeo.

A entidade direitos reúne informação sobre os direitos de propriedade intelectual e permissões associadas ao objecto e/ou agente.

O dicionário de dados é acompanhado de um conjunto de esquemas XML que auxiliam a utilização e promovem a implementação do PREMIS. Apesar de os esquemas publicados não terem sido criados para ser utilizados de forma escrupulosa, a sua existência facilita a implementação do dicionário de dados por quem desenvolve repositórios digitais.

Em Março de 2008, o comité responsável pela manutenção do dicionário de dados PREMIS (i.e., PREMIS Maintenance Activity) publicou uma nova versão do documento com as seguintes revisões (Guenther et al., 2008; B. F. Lavoie, 2008):

- A secção de direitos foi inteiramente revista e expandida de forma a suportar mais detalhe e oferecer maior aplicabilidade;
- Foram adicionados mais atributos relacionados com propriedades significativas e informação de preservação;
- Foram incluídos mecanismos que permitem expandir e personalizar o dicionário de dados.

[31] Do inglês *Representation*, *File* e *Bitstream*.

Para além das alterações anteriormente descritas foram ainda efectuadas pequenas revisões do documento no que toca à qualidade da documentação e exemplos fornecidos, estrutura do documento e especificação dos formatos a utilizar, nomeadamente, para designar datas (B. F. Lavoie, 2008).

2.7 Considerações finais

Este capítulo teve como objectivo descrever e contextualizar as principais actividades que têm vindo a ser realizadas internacionalmente no domínio da preservação digital. O capítulo começa por definir preservação digital e introduzir o conceito de objecto digital. Este é apresentado sob uma perspectiva semiótica, sendo feita uma análise dos diferentes níveis de abstracção a que pode ser conisderado: físico, lógico e conceptual. Esta visão multidimensional do objecto digital promove uma melhor compreensão e enquadramento das diferentes estratégias de preservação apresentadas ao longo do capítulo.

De seguida foi apresentado o modelo de referência OAIS (Open Archival Information System), uma norma internacional que visa a identificação dos principais componentes funcionais e objectos de informação presentes num sistema de arquivo com aspirações de preservação a longo-prazo. O modelo de referência serviu sobretudo para introduzir alguma da terminologia utilizada ao longo desta tese. Foram também descritas e contextualizadas as principais estratégias de preservação que têm vindo a ser propostas pela comunidade científica, bem como as mais relevantes iniciativas no que toca a directórios de formatos.

O capítulo termina com uma breve discussão sobre autenticidade, salientando-se a importância do conceito de propriedade significativa na elaboração de políticas de preservação. Paralelamente, procurou-se realçar a necessidade da utilização de metainformação como meio para assegurar a autenticidade dos materiais custodiados, dando especial ênfase ao Dicionário de Dados PREMIS.

O amadurecimento do domínio científico da preservação digital levou a que duas estratégias de preservação ganhassem maior destaque: a migração e a emulação. Não será inadequado afirmar que durante algum tempo se assistiu a uma batalha ideológica entre aqueles que defendiam a utilização das estratégias de migração e os que eram a favor de estratégias baseadas em emulação. Esta discussão teve origem em questões relacionadas com a autenticidade dos materiais no domínio digital.

Em estratégias derivadas da migração, assume-se que os objectos digitais irão ser alvo de modificações sucessivas ao longo do tempo. Determinadas migrações poderão mesmo originar

perdas substanciais de informação. Para os defensores da emulação, assumir de antemão que a informação que se procura preservar será sistematicamente adulterada ao longo do tempo viola os pressupostos mais elementares da preservação (Rothenberg et al., 1999).

Esta questão, no entanto, não está confinada ao domínio digital. Na arquivística tradicional, há quem defenda que a preservação do material no seu estado original deverá ser considerada como a única medida de sucesso. Há, no entanto, quem opte por transferir os seus materiais para suportes menos volumosos, como por exemplo o microfilme, tomando uma decisão explícita pela poupança de espaço em detrimento da originalidade (B. F. Lavoie & Dempsey, 2004). O material digital, no entanto, possui características que fazem com que estas questões acabem por ser, em boa medida, amplificadas. O material digital é estruturalmente mais complexo que o seu equivalente analógico. Diferentes tipos de informação podem ser combinados num único objecto (e.g. texto, vídeo, som) e este, pode ainda, exibir características dinâmicas e/ou interactivas. Para além disso, pode facilmente ser modificado, desconstruído e recombinado de formas inovadoras usando o software adequado (B. F. Lavoie & Dempsey, 2004).

Não obstante, a preocupação obstinada pela originalidade tem vindo a diminuir à medida que aumenta a compreensão generalizada sobre os processos de preservação. Começa-se a difundir a ideia de que o foco da preservação não deverá estar na retenção do objecto físico original, mas na conservação da experiência sensorial que é produzida por esse objecto (Heslop et al., 2002).

Neste contexto, Burkel questiona-se sobre o papel da tecnologia no processo de interpretação de informação digital – "(…) as entradas e saídas de qualquer sistema digital são na forma de linguagens humanas. A tecnologia e as suas linguagens próprias apenas asseguram um processamento mais eficiente dessa informação no interior do computador" (Burkel, 2003).

Reforçando esta ideia, Thibodeau argumenta que no futuro, tal como hoje, os consumidores desejarão servir-se das tecnologias mais modernas ou daquelas que melhor conhecem para manipular mais eficientemente a informação que necessitam. A opção por uma estratégia de emulação poderá conduzir ao incumprimento desta necessidade básica (B. F. Lavoie & Dempsey, 2004; Thibodeau, 2002).

A batalha ideológica – migração versus emulação – tem, assim, tendência a esgotar-se. Instala-se o reconhecimento generalizado de que diferentes estratégias de preservação deverão ser implementadas dependendo do contexto específico da organização preservadora e do tipo de objectos a preservar (Waters, 2002). A selecção de estratégias de preservação deve ter em

conta factores diversos, como: as características intrínsecas dos objectos, o custo de implementação e manutenção, os interesses do arquivo ou da sua comunidade de interesse. Para diversos autores este último ponto é de extrema importância. A informação terá pouca utilidade se não for preservada e disseminada de acordo com as necessidades da sua comunidade de interesse (Bennett, 1997; Hedstrom, 1998; B. F. Lavoie & Dempsey, 2004).

A tendência actual vai no sentido de combinar um conjunto de técnicas como o refrescamento automático de suportes, a normalização para formatos de preservação durante o processo de ingestão, a conservação do objecto original (como salvaguarda e para fins arqueológicos) e migração a-pedido para adaptar os formatos de preservação a formatos mais adequados à sua disseminação (Thibodeau, 2002). Os objectos digitais predominantemente dinâmicos ou interactivos são geralmente preservados nos seus formatos originais e apresentados recorrendo a técnicas de emulação (Hodge & Frangakis, 2004).

Apesar do aparecimento de ferramentas de software que auxiliam o processo de arquivo e preservação (e.g. OCLC Digital Archive[32], DSpace[33], LOCKSS[34], Fedora[35], Eprints[36], PANDAS[37], DIAS[38], JHove[39], Droid[40], Xena[41], etc) existe ainda uma escassez assinalável no que toca a produtos comerciais com capacidades de preservação (Hodge & Frangakis, 2004). Isto faz com que cada organização se sinta de certa forma responsável pelo desenvolvimento do seu próprio sistema de preservação, bem como pela definição e implementação de políticas de preservação adequadas.

A definição de uma política de preservação envolve, geralmente, todas as facetas de um arquivo. Implica a criação de políticas de avaliação e selecção de materiais, a identificação de esquemas de metainformação apropriados (e.g. descritiva, técnica, estrutural e de preservação), a definição de estratégias de preservação adequadas a cada classe de objectos digitais, a criação de planos de sucessão (para a eventualidade da organização detentora cessar a sua actividade), a utilização de modelos sustentáveis de financiamento, entre outros.

---

[32] http://www.oclc.org/digitalarchive/
[33] http://www.dspace.org
[34] http://www.lockss.org
[35] http://www.fedora.info
[36] http://www.eprints.org/software/
[37] http://pandora.nla.gov.au/pandas.html
[38] http://www-5.ibm.com/nl/dias/
[39] http://hul.harvard.edu/jhove/
[40] http://www.nationalarchives.gov.uk/aboutapps/pronom/droid.htm
[41] http://sourceforge.net/projects/xena

Uma política de preservação deverá descrever claramente as estratégias adoptadas para assegurar a preservação dos materiais em cada um dos níveis de abstracção a que estes podem ser considerados, i.e., físico, lógico e conceptual, mas também a níveis superiores, como o social, o económico e o organizacional (Ambacher et al., 2007; Beagrie et al., 2002; Digital Curation Centre & DigitalPreservationEurope, 2007).

A Tabela 1 enumera algumas das possíveis estratégias que poderão ser utilizadas para preservar cada um dos níveis de abstracção anteriormente descritos.

Nível de abstracção	Estratégias a aplicar
Físico	Acondicionamento adequado dos suportes físicos, utilização de suportes de longa duração, salas de prevenção contra desastres naturais, etc.
Lógico	Refrescamento, *backup*, replicação local e/ou remota, etc.
Conceptual	Migração, emulação, encapsulamento, etc.
Social	O sistema de preservação deverá ser capaz de impedir ou de corrigir a ocorrência de erros provocados por operadores ou atacantes externos, e.g. implementação de mecanismos de *undo,* registo de actividades, autenticação e gestão de permissões, etc.
Económico	Definição de modelos de financiamento sustentáveis. As despesas com a preservação deverão fazer parte dos orçamentos de base das organizações.
Organizacional	Definição de planos de sucessão que garantam a sobrevivência dos materiais face à eventual cessação de actividade por parte da organização detentora.

Tabela 1 – Possíveis estratégias de preservação por nível de abstracção.

Não obstante, e para além da definição de uma política de preservação e do estabelecimento de estratégias de preservação adequadas, é fundamental adoptar um sistema de arquivo digital, i.e., um repositório capaz de albergar os objectos, bem como facilitar a implementação dessas políticas e respectivas estratégias de preservação. O recurso a um repositório digital facilita a gestão dos objectos, bem como a sua localização, ambas operações fundamentais em qualquer sistema de arquivo.

Até à data, nenhum dos principais repositórios digitais (e.g. DSpace, Fedora, Eprints) oferece funcionalidades que permitam a implementação de políticas de preservação de forma transversal, nem tão pouco suportam esquemas de metainformação de preservação, essenciais para garantir a autenticidade dos materiais custodiados. No entanto, oferecem já a capacidade de armazenar, organizar, descrever e disseminar esses materiais. Será portanto espectável que a curto prazo estas plataformas comecem a incorporar funcionalidades de preservação que permitam garantir o acesso a longo-prazo aos materiais digitais custodiados.

Capítulo 3

Automatização de processos de migração

Apesar dos progressos ocorridos ao longo dos últimos anos no domínio da preservação digital, continua a existir um vazio assinalável no que diz respeito à automatização dos processos que lhe são inerentes (Ross & Hedstrom, 2005). Mais ainda, vários problemas permanecem por resolver, como por exemplo: como garantir que os materiais digitais permanecem autênticos após sucessivas intervenções de preservação; como validar formalmente o sucesso de uma intervenção; ou, como melhorar os processos de preservação no sentido de se conseguir uma redução generalizada dos custos de preservação?

Todas as intervenções de preservação envolvem escolhas. Os recursos disponíveis nas organizações são finitos, muitas das vezes escassos, pelo que qualquer intervenção de preservação carece de uma fase de análise e planeamento. É fundamental assegurar que os requisitos da organização, da colecção de objectos digitais e da comunidade de interesse são satisfeitos, mesmo na presença de condicionantes estruturais, por vezes, difíceis de contornar. Estas condicionantes poderão manifestar-se de diversas formas: falta de capacidade técnica, orçamentos limitados, imposições legais, equipamento insuficiente, restrições de tempo, espaço, formação, etc. (Rauch & Rauber, 2004).

Neste contexto, as estratégias de preservação baseadas em migração não são diferentes das restantes. Uma análise detalhada dos objectivos, meios para os alcançar e resultados obtidos é fundamental para que uma estratégia de migração possa ser considerada bem sucedida.

Neste capítulo pretende-se descrever o conjunto de actividades que geralmente está associado à implementação de uma estratégia de migração, nomeadamente: a selecção de uma alternativa de migração, a sua execução e controlo de qualidade dos resultados obtidos.

A secção 3.1 começa por descrever detalhadamente cada uma dessas actividades. A secção 3.2 apresenta um conjunto de argumentos que realçam as vantagens inerentes à utilização de sistemas distribuídos na implementação deste tipo de estratégias. A secção 3.3 apresenta um cenário onde se evidenciam o tipo de problemas de preservação que geralmente emergem num contexto organizacional. O mesmo cenário é utilizado na secção 3.4 para ilustrar de que forma uma arquitectura de serviços de preservação poderá facilitar a implementação automática de uma estratégia de migração. Ainda nesta secção, para cada um dos serviços identificados é apresentada uma lista de ferramentas, produtos e/ou serviços desenvolvidos por terceiros que poderão ser utilizados para suportar o seu funcionamento. A secção 3.5 reproduz o cenário apresentado na secção 3.3, salientando a forma como os serviços de preservação previamente identificados facilitariam o desenvolvimento e implementação de estratégias de preservação. O capítulo termina, na secção 3.6, com um sumário e uma reflexão sobre os conceitos e temáticas abordadas ao longo do capítulo.

3.1 Actividades inerentes a um processo de migração

A implementação de uma estratégia de migração pressupõe a realização de um conjunto mínimo de actividades. Entre estas encontram-se a selecção de uma alternativa de migração adequada ao problema de preservação em questão, execução da respectiva alternativa e a análise e avaliação dos resultados obtidos de modo a aferir a qualidade da selecção efectuada, i.e., controlo de qualidade (Ferreira, 2005; Ferreira et al., 2006a).

3.1.1 Selecção de uma alternativa de migração

A selecção de uma alternativa de migração consiste sobretudo na obtenção de uma resposta para duas questões fundamentais, nomeadamente:

- Qual o formato de destino que deverá ser utilizado para acomodar as propriedades essenciais do objecto original?

- Que conversor, ou cadeia de conversores, apresenta maior aptidão para realizar essa transformação?

É do interesse da entidade responsável pela preservação que a melhor combinação entre formato de destino e conversores a utilizar seja seleccionada, i.e., aquela que garante a preservação do maior número de propriedades significativas do objecto original, ao menor custo possível.

O custo deverá ser entendido sob uma perspectiva multidimensional, i.e., factores como a velocidade de conversão, preço do software, complexidade da implementação, abertura dos formatos envolvidos, o seu nível de adopção e todos os restantes custos de operação deverão ser considerados de forma concertada durante esta fase de preparação.

A actividade de selecção de uma alternativa de migração é particularmente complexa em contextos onde poderá existir um elevado número de opções no que toca a formatos e aplicações de conversão. Este é, aliás, o caso num ambiente de migração distribuída como aquele que é descrito na secção 2.3.4 na página 30 (Ferreira, Baptista, & Ramalho, 2007).

3.1.2 Conversão de materiais

A conversão de materiais tem que ver com a reestruturação dos elementos de informação que os constituem segundo as regras de um novo formato (Lawrence et al., 2000). No contexto de uma organização, a tarefa conversão pode ser realizada de duas formas distintas: adquirindo software capaz de realizar a conversão pretendida ou desenvolvendo conversores específicos adequados ao problema de migração em questão.

Em ambos os casos, o processo de conversão implica custos para a organização e a sua implementação requer, geralmente, a presença de intervenientes humanos, em especial durante a fase de preparação da actividade e, posteriormente, durante a fase de controlo de qualidade (Becker, Ferreira et al., 2008; Ferreira, Baptista, & Ramalho, 2006b).

A velocidade de conversão é também um factor determinante. Repositórios detentores de um elevado número de objectos digitais poderão requerer um tempo de conversão suficientemente longo para que haja preocupação com a durabilidade dos suportes físicos que os sustêm (Halem et al., 1999; Hedstrom, 2001). Por exemplo, se se considerar a migração de um conjunto de objectos na ordem dos 100 Terabytes e assumindo que cada Megabyte demoraria 2 segundos a converter, a migração da totalidade dos objectos iria estender-se ao

longo de aproximadamente 7 anos, tempo suficiente para que os suportes físicos de armazenamento se tornassem obsoletos.

3.1.3 Controlo de qualidade

Após uma conversão é fundamental avaliar os resultados obtidos, i.e., verificar em que medida os objectos que resultaram da conversão satisfazem os requisitos definidos *a priori* pela entidade preservadora. Este processo consiste, usualmente, na análise e comparação dos objectos que resultaram da migração com os objectos originais, tendo por base o conjunto de propriedades significativas definido previamente pela entidade preservadora (Hofman, 2002b; Rusbridge, 2003). Este conjunto de propriedades significativas constitui o nível de compromisso assumido pela organização no que toca à preservação dos materiais digitais.

Após a conversão, uma avaliação abaixo das expectativas poderá implicar a selecção de uma nova alternativa de migração e a repetição de todo o processo de conversão (Ferreira et al., 2006a). Esta actividade de controlo de qualidade, devido às suas características e ao facto de ser frequentemente realizada por profissionais qualificados, é considerada morosa e extremamente dispendiosa (Rauch, Pavuza et al., 2005). Ao longo desta tese procurar-se-á mitigar estes dois problemas implementando mecanismos automáticos de controlo de qualidade em processos de migração.

3.2 Migração em ambientes distribuídos

Numa secção anterior, foi possível constatar como redes distribuídas de conversores poderão contribuir para um aumento da flexibilidade na implementação de estratégias de migração (ver Migração distribuída, na secção 2.3.4, na página 30).

Qualquer agente de software capaz de invocar serviços remotos, como por exemplo Web services, estará automaticamente habilitado a realizar conversões entre formatos sem que haja necessidade de adquirir ou implementar localmente soluções específicas de conversão. Paralelamente, a utilização de serviços remotos dotados de redundância assegura a fiabilidade do sistema perante situações de ruptura parcial da rede e a existência de múltiplos caminhos de conversão confere a este tipo de soluções uma longevidade superior comparativamente a estratégias de migração mais convencionais.

Uma tendência recente no domínio da preservação digital dirige-se no sentido da criação de arquitecturas de serviços que facilitem a implementação de estratégias de preservação

(Hitchcock, Brody, Hey, & Carr, 2007). Este tipo de arquitecturas designam-se genericamente por arquitecturas orientadas ao serviço.

Uma arquitectura orientada ao serviço[42] ou SOA é um sistema baseado em software cujas funções se encontram distribuídas através de diferentes componentes de acordo com os processos de negócio que implementam. Estas funções podem ser acedidas a partir da rede e utilizadas na construção de sistemas cada vez mais complexos (Erl, 2005).

Num ambiente SOA não há limitações de interoperabilidade, nomeadamente ao nível dos sistemas operativos, linguagens de programação e/ou outras tecnologias de suporte (Newcomer & Lomow, 2005). Os serviços comunicam entre si trocando mensagens em formatos neutros que poderão servir, tanto para transportar dados, como para coordenar os diferentes serviços cooperantes (SOA Reference Model TC, 2008). Alguns dos princípios fundamentais que governam este tipo de arquitecturas são (Balzer, 2004):

- **Granularidade, modularidade e capacidade de reutilização** – a lógica de negócio encontra-se dividida em módulos simples e atómicos de forma a promover a sua reutilização em contextos distribuídos;

- **Possibilidade de composição** – os serviços são desenvolvidos de forma a possibilitar a sua composição (i.e., invocação e execução em sequência);

- **Interoperabilidade** – os serviços são baseados em normas de forma a promover a sua interoperabilidade funcional e informacional;

- **Autonomia** – cada serviço é responsável pela sua própria lógica de negócio;

- **Auto-descrição** – um serviço não carece de documentação extra para além daquela que lhe é intrínseca para que possa ser utilizado eficazmente.

Os Web services, como tecnologia de suporte à implementação de SOA, ganharam aceitação generalizada por parte da indústria, sobretudo devido ao facto de se basearem em normas internacionais abertas promovidas por entidades independentes como a W3C (World Wide

[42] Do inglês *Service Oriented Architecture* (SOA).

Web Consortium)[43]. No entanto, outras tecnologias concorrentes poderão ser utilizadas para implementar o mesmo conceito, tais como: Jini[44], CORBA[45] ou REST[46] (Fielding, 2000).

3.3 Cenário de preservação

Num ambiente organizacional existe uma série de problemas que são comuns ocorrer e que requerem o uso de ferramentas e conhecimentos específicos da área da preservação para que possam ser eficazmente solucionados. O cenário que se segue pretende ser ilustrativo quanto a este tipo de ocorrências. Este cenário apresenta um gestor de informação que é deparado com a necessidade de preservar uma colecção de objectos digitais e expõe o tipo de dificuldades que geralmente emergem deste tipo de contextos (Ferreira et al., 2006b):

> Num dado momento, uma empresa de dimensão média decide que todos os relatórios técnicos produzidos no decurso da sua actividade deverão estar acessíveis a todos os seus colaboradores à distância de um clique. Para tal, foi contratado um gestor de informação cuja função seria implementar e administrar um repositório digital com o objectivo de preservar e dar acesso aos mesmos através da Intranet da empresa.
>
> Todos os relatórios existentes até à data de implementação do repositório haviam sido elaborados com recurso ao Microsoft Word 95. À medida que o tempo foi passando, novas versões da aplicação Word começaram a ser exploradas no interior da organização. Paralelamente, alguns colaboradores mais adeptos do movimento *open-source* começaram a utilizar a ferramenta OpenOffice para produzir os seus relatórios técnicos. Consequentemente, o número de formatos existentes no repositório aumentou a ponto de se tornar impossível consultar transversalmente o conjunto de relatórios de um dado projecto sem que houvesse necessidade de instalar software adicional para os poder visualizar.
>
> Para agravar ainda mais a situação, o responsável pelo repositório verificou que existiam vários relatórios cuja extensão não lhe era familiar, dificultando assim a identificação da aplicação adequada à sua visualização. Paralelamente, a Microsoft anuncia que a nova versão do seu pacote de aplicações Office não irá suportar o formato Word 95.

---

[43] http://www.w3.org
[44] http://www.jini.org
[45] http://www.corba.org/
[46] http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm

O responsável pela administração do repositório digital conclui de que algo teria de ser feito no sentido de homogeneizar os conteúdos do repositório. Primeiro, decide realizar um levantamento de todos os formatos existentes no repositório. Olhando para a extensão dos ficheiros foi possível determinar qual a aplicação que fora utilizada na sua criação. No entanto, para determinar a versão da mesma teria que ser encontrada uma estratégia mais eficaz.

O gestor do repositório decidiu então investigar quais os mecanismos que poderiam ser utilizados para identificar o formato dos objectos custodiados. Após alguma pesquisa, encontra uma série de pequenas aplicações que proclamavam ser capazes de identificar o formato de um ficheiro confrontando-o com uma base de dados de *magic numbers*[47] e cabeçalhos predefinidos.

Após produzir uma listagem com os formatos existentes no repositório, o responsável procedeu à identificação daqueles que estavam em risco de se tornarem obsoletos mais rapidamente. Após alguma pesquisa descobriu um guia que definia critérios que permitiam apurar o risco incorrido ao conservar objectos em determinados formatos. Esse guia chamava-se *Risk Management of Digital Information: a file format investigation* (Lawrence et al., 2000).

No entanto, a realização de uma análise de risco para todos os formatos existentes no repositório revelou-se uma tarefa demasiado morosa, pelo que o funcionário optou por confiar no seu instinto e optou por converter todos os relatórios para a última versão do Microsoft Word, baseado no pressuposto de que o documento Word era, efectivamente, o formato mais abundante no repositório.

Para realizar essa tarefa, o gestor do repositório necessitou de adquirir um conjunto de aplicações de conversão. Algumas das conversões necessárias não puderam ser realizadas directamente, i.e., foi necessário converter para um formato intermédio e depois utilizar outra aplicação para realizar a migração para o formato designado.

Após o processo de migração, o funcionário inspeccionou alguns dos relatórios convertidos e constatou que a aparência dos mesmos não era exactamente igual à dos originais. Em alguns casos a paginação havia sido alterada fazendo com que os índices

[47] Tratam-se de pequenas sequências de bytes geralmente encontradas no início de um ficheiro que permitem determinar o seu formato.

incluídos nos documentos ficassem desactualizados. Noutros casos, certas imagens haviam perdido detalhe, o que dificultava consideravelmente a sua compreensão.

Foi necessário informar os utilizadores do repositório que aqueles documentos haviam sido convertidos e que, devido a esse facto, as suas propriedades significativas haviam sido adulteradas. No entanto, especificar quais propriedades e o nível de degradação que cada relatório havia sofrido revelou-se uma tarefa demasiado penosa para uma pessoa só. Tornou-se evidente que seria necessário encontrar algo que permitisse automatizar e simplificar todo esse processo.

3.4 Serviços de preservação

Uma análise atenta ao cenário apresentado permite identificar um conjunto de funcionalidades que deverão fazer parte de um sistema capaz de prestar serviços de preservação. O desenvolvimento do conjunto de serviços identificados permite automatizar os processos de preservação que garantem o acesso continuado à informação custodiada num repositório digital. Entre estes, encontram-se os seguintes serviços:

- **Serviço de identificação de formatos** – responsável por determinar o formato de um dado objecto digital e também por verificar a integridade lógica dos mesmos (i.e., verificar se a codificação de um objecto respeita a sintaxe do formato identificado);
- **Serviço de selecção de estratégias de migração** – responsável por determinar e sugerir estratégias de migração adequadas às necessidades da instituição preservadora e sua comunidade de interesse (i.e., formato de destino e a aplicação de conversão);
- **Serviço de conversão** – serviço responsável pela migração de formatos;
- **Serviço de controlo de qualidade** - serviço que determina quais os atributos do objecto original que não foram devidamente preservados durante o processo de migração;
- **Serviço de notificação de obsolescência** – serviço que verifica e disponibiliza informação sobre os formatos que estão em risco de se tornar obsoletos no interior de um repositório.

A Figura 18 apresenta uma visão geral de uma arquitectura que disponibiliza o conjunto de serviços previamente identificados. A figura encontra-se dividida em duas partes fundamentais: o cliente (em cima) e o provedor de serviços (em baixo).

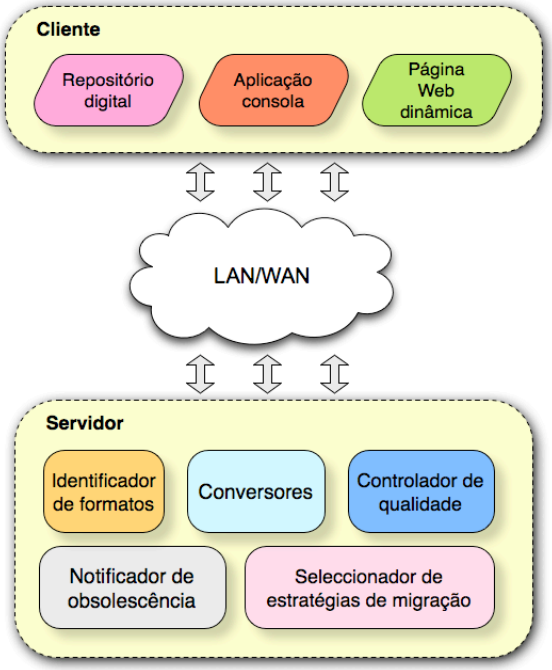


Figura 18 – Arquitectura de um sistema de preservação.

A parte da figura referente ao cliente fornece exemplos de aplicações que poderão tirar partido dos serviços disponibilizados por esta plataforma. Entre estes encontram-se repositórios digitais (e.g. DSpace[48], Eprints[49], Fedora[50]), aplicações-cliente baseadas na consola ou aplicações desenvolvidas para a Web. Os exemplos apresentados pretendem ser ilustrativos e não prescritivos, i.e., qualquer aplicação capaz de invocar um serviço remoto poderá tirar partido das funcionalidades disponibilizadas por uma plataforma com estas características.

Na camada inferior da figura encontra-se a plataforma de serviços, assim como todos os componentes que a constituem. Cada um destes é responsável por uma tarefa específica e funciona de forma independente dos restantes. Esta abordagem permite que cada componente possa ser administrado por uma entidade distinta, facilitando ainda a distribuição da carga por vários servidores. Não obstante, os vários componentes poderão colaborar entre si em torno de um objectivo comum. Fazem-no trocando mensagens e invocando mutuamente os serviços disponibilizados por cada um dos componentes.

[48] http://www.dspace.org
[49] http://www.eprints.org
[50] http://www.fedora-commons.org/

As secções que se seguem descrevem detalhadamente cada um dos serviços identificados na Figura 18 e apresentam algumas ferramentas capazes de implementar os conceitos que lhes são subjacentes.

3.4.1 Identificador de formatos

O componente Identificador de formatos, tal como o nome indica, é um serviço que permite determinar o formato de um objecto digital.

Organizações que tenham como missão preservar objectos digitais terão de ser capazes de verificar e monitorizar a integridade lógica dos seus objectos sem necessidade de intervenção humana, i.e., deverão ser capazes de reconhecer o formato de um objecto e verificar se este está de acordo com o formato identificado. Um serviço de identificação de formatos é fundamental no cumprimento deste requisito.

Uma vantagem que advém da utilização de um serviço com estas características tem que ver com o facto de os objectos digitais serem identificados de acordo com um único vocabulário. O uso transversal de um vocabulário controlado para designar formatos garante a interoperabilidade lexical entre todos os componentes da plataforma e torna a orquestração de tarefas um processo simples e harmonioso.

Existem várias soluções capazes de suportar a construção de um serviço com estas características. Entre estas destacam-se as seguintes:

- JHOVE[51] – software desenvolvido conjuntamente pela JSTOR[52] e pela Biblioteca da Universidade de Harvard[53], especialmente desenhado para identificar e caracterizar formatos digitais. Na prática, o JHOVE é mais do que um identificador de formatos. Esta aplicação é capaz de extrair metainformação técnica a partir de diversos formatos digitais. A principal desvantagem desta aplicação é que apenas suporta onze formatos distintos, nomeadamente: AIFF, WAVE, ASCII, HTML, PDF, XML, UTF-8, GIF, JPEG, JPEG 2000 e TIFF.

- Droid[54] (Digital Record Object Identification) – software desenvolvido pelos Arquivos Nacionais do Reino Unido[55], os criadores do directório de formatos PRONOM (ver

[51] http://hul.harvard.edu/jhove/
[52] http://www.jstor.org/
[53] http://hul.harvard.edu/
[54] http://droid.sourceforge.net

Directórios de formatos na página 34), foi desenhado especificamente para identificar formatos digitais. Esta ferramenta permite processar sequencialmente vários objectos e produz designações de formato que congregam o nome e a versão do mesmo. As principais vantagens desta ferramenta advêm do facto de esta suportar centenas de formatos distintos e da sua base de dados de formatos estar em constante crescimento. As actualizações desta aplicação são realizadas automaticamente durante o arranque da mesma.

- Unix file[56] – comando que acompanha as distribuições de sistemas operativos Unix/Linux que permite identificar o formato de ficheiros através da linha de comandos. O comando file, apesar de não ser multiplataforma como as duas aplicações anteriores, oferece uma velocidade de processamento inigualável e apresenta suporte para uma elevada quantidade de formatos.

- FILExt[57] (The File Extension Source) – trata-se de um sítio Web que reúne informação sobre formatos tendo por base a extensão que geralmente é associada ao formato em causa. O portal disponibiliza um serviço de pesquisa por extensão e fornece informações como: nome da aplicação de leitura/produção do formato identificado e o seu fabricante, Mime Types associados ao formato, *magic numbers* e hiperligações para descarregar aplicações de leitura.

3.4.2 Conversores

O componente designado por Conversores representa os serviços que permitem efectuar transformações entre formatos (Figura 18). Os conversores poderão ser utilizados para construir conversores mais complexos, recorrendo à composição de serviços.

Vários exemplos de serviços de conversão foram já apresentados na secção Migração distribuída na página 30, nomeadamente, o TOM (Ockerbloom, 1998, 2003), o MyMorph (Walker & Thoma, 2003, 2004, 2005) e o PANIC (Hunter & Choudhury, 2003). Para além destes, existem outros exemplos que, apesar de não terem sido idealizados como serviços com fins de preservação, nem tão pouco implementarem os requisitos necessários para que possam ser considerados SOA, poderiam ser utilizados eficazmente para suportar uma rede de serviços de conversão. Entre estes, destacam-se os seguintes:

---

[55] http://www.nationalarchives.gov.uk/
[56] http://darwinsys.com/file/
[57] http://www.filext.com

- Media-convert[58] – trata-se de um sítio Web que oferece aos seus utilizadores a capacidade de efectuar conversões entre dezenas de formatos: vídeo, documentos de texto, folhas de cálculo, áudio, imagem matricial, imagem vectorial e apresentações multimédia. Os objectos a converter são enviados para o sítio Web através de um HTTP-POST e os resultados da conversão são descarregados pelo browser acedendo a um URL. O sítio Web é suportado financeiramente por publicidade.

- Zamzar[59] – trata-se de um sítio Web em tudo semelhante ao anterior, diferindo apenas no método de retorno dos objectos convertidos. Neste caso, o URL onde se encontram os objectos convertidos não se encontra imediatamente disponível. Ao invés disso, o URL é enviado para o cliente por correio-electrónico. O modelo de financiamento que suporta este sítio é baseado numa subscrição mensal que quanto mais elevada, melhor a qualidade de serviço fornecido, tanto em termos de velocidade de processamento, como em volume de dados suportado.

3.4.3 Controlo de qualidade

O serviço designado por Controlador de qualidade tem como missão detectar perdas de informação nos objectos digitais que resultam das migrações efectuadas. Este componente deverá ser capaz de comparar os objectos digitais submetidos a migração com as suas versões convertidas e produzir um relatório evidenciando detalhadamente as diferenças detectadas. Esse relatório permite documentar a intervenção de preservação e determinar o nível de qualidade associado à intervenção realizada. Com base nesta informação é possível determinar quais os conversores que prestam o melhor serviço de conversão, i.e., determinar aqueles que garantem a conservação do maior número de propriedades significativas do objecto digital original.

Neste contexto há sobretudo uma iniciativa que merece ser destacada:

- XCEL/XCDL – o XCEL (*eXtensible Characterisation Extraction Language*) é um dialecto XML que permite definir regras para extracção de propriedades de objectos digitais codificados num dado formato digital. Uma vez criado o documento XCEL de um formato, este é processado, conjuntamente com um objecto digital, por uma aplicação designada *Extractor*. O *Extractor* interpreta as regras definidas pelo documento XCEL e produz um documento XCDL (*eXtensible Characterisation Definition Language*) que

[58] http://media-convert.com/
[59] http://www.zamzar.com/

comporta, numa linguagem abstracta e uniformizada, as propriedades extraídas do objecto digital. Os documentos XCDL produzidos a partir de dois objectos em formatos distintos podem ser comparados e as suas diferenças facilmente detectadas (Becker, Rauber, Heydegger, Schnasse, & Thalle, 2008). O principal obstáculo encontrado nesta abordagem encontra-se no processo de criação de documentos XCEL. Para determinados formatos, estes documentos são extremamente complexos e a elaboração dos mesmos requer geralmente a colaboração do produtor do formato. Até ao momento este projecto apenas produziu especificações XCEL para os formatos TIFF e PNG.

3.4.4 Notificador de obsolescência

O serviço de Notificação de obsolescência é responsável por informar as entidades-cliente que determinados formatos se encontram em risco de se tornar obsoletos. Este serviço deve ser consultado regularmente pela entidade cliente de modo a determinar quais os objectos presentes no seu repositório que poderão vir a tornar-se inacessíveis devido a alterações significativas no panorama tecnológico vigente ou devido à existência de determinadas características consideradas inadequadas num contexto de preservação (Ferreira et al., 2006a). Várias iniciativas poderão servir de base à construção de um serviço com estas características:

- O relatório *Risk Management of Digital Information: A File Format Investigation* apresenta os resultados de um estudo que procura medir o impacto que a migração pode ter na integridade dos objectos digitais e quais os riscos incorridos ao manter objectos em determinados formatos (Lawrence et al., 2000);

- A metodologia INFORM procura prever a durabilidade de formatos digitais identificando um conjunto de características que poderão inviabilizar o acesso à informação, como por exemplo, DRM[60], algoritmos de compressão, encriptação, assinaturas digitais, dependência de hardware e software específico, etc. (Stanescu, 2004);

- A DigiCULT[61] e a Digital Preservation Coalition[62] publicam periodicamente relatórios que procuram identificar as principais tendências no uso de tecnologias. Apesar destes

[60] Digital Rights Management
[61] http://www.digicult.info/pages/techwatch.php
[62] http://www.dpconline.org/graphics/reports/

relatórios não terem como principal objectivo alertar a comunidade para os formatos que se estão a tornar obsoletos, estes poderão, em boa medida, servir de base para precaver este tipo de ocorrências.

- Existem também vários serviços na Web especializados em monitorizar o lançamento de novas versões de software. Apesar deste tipo de serviços não estar especialmente vocacionado para detectar novos formatos, na maioria dos casos o lançamento de uma nova versão de um software é motivo suficiente para que haja preocupação com a obsolescência dos formatos associados às suas versões precedentes. Exemplos deste tipo de serviços são: o VersionTracker[63] e o SUMo[64];

- Uma iniciativa liderada pelos Arquivos Nacionais da Austrália e pela Australian Partnership for Sustainable Repositories[65] elaborou um sistema chamado AONS[66] (Automatic Obsolescence Notification Service) que cumpre escrupulosamente os objectivos identificados para este componente, i.e., providenciar um serviço que notifica entidades detentoras de objectos digitais de que determinados formatos estão em vias de se tornar obsoletos e que portanto devem desenvolver-se diligências no sentido de se preservar os objectos codificados nesses formatos. O AONS recolhe informação sobre formatos digitais a partir de vários parceiros, nomeadamente o PRONOM (Darlington, 2003; UK National Archives, 2002, 2005) e a iniciativa LCSDF da Biblioteca do Congresso (Library of Congress, 2004a), e constrói a sua própria base de dados de formatos. Posteriormente, monitoriza essa base de dados em busca de formatos para os quais existam novas versões, formatos com pouco suporte aplicacional, formatos proprietários ou formatos que apenas são suportados por software obsoleto. Repositórios (por agora, apenas DSpace e Fedora) poderão registar as suas colecções de objectos e esperar notificações quando algum dos seus formatos se encontrar em risco de obsolescência (Curtis, Koerbin, Raftos, Berriman, & Hunter, 2007; Pearson, 2008).

3.4.5 Seleccionador de estratégias de migração

O componente Seleccionador de estratégias de migração tem como principal objectivo identificar os serviços de conversão mais adequados para resolver um problema de preservação específico.

---

[63] http://www.versiontracker.com
[64] http://www.kcsoftwares.com/?sumo
[65] http://www.apsr.edu.au
[66] http://www.apsr.edu.au/aons2 e http://sourceforge.net/projects/aons/

Os conversores disponíveis na rede poderão ser mais ou menos aptos dependendo dos requisitos da entidade-cliente. Para identificar o serviço de conversão mais adequado, é necessário confrontar os requisitos enumerados pelo cliente com as características do conversor e encontrar a melhor combinação possível.

Rauch e Rauber desenvolveram um método capaz de comparar e seleccionar alternativas de preservação tendo em conta as necessidades individuais de cada entidade preservadora (Rauch, Pavuza et al., 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 decisão em projectos complexos no domínio da engenharia civil e economia.

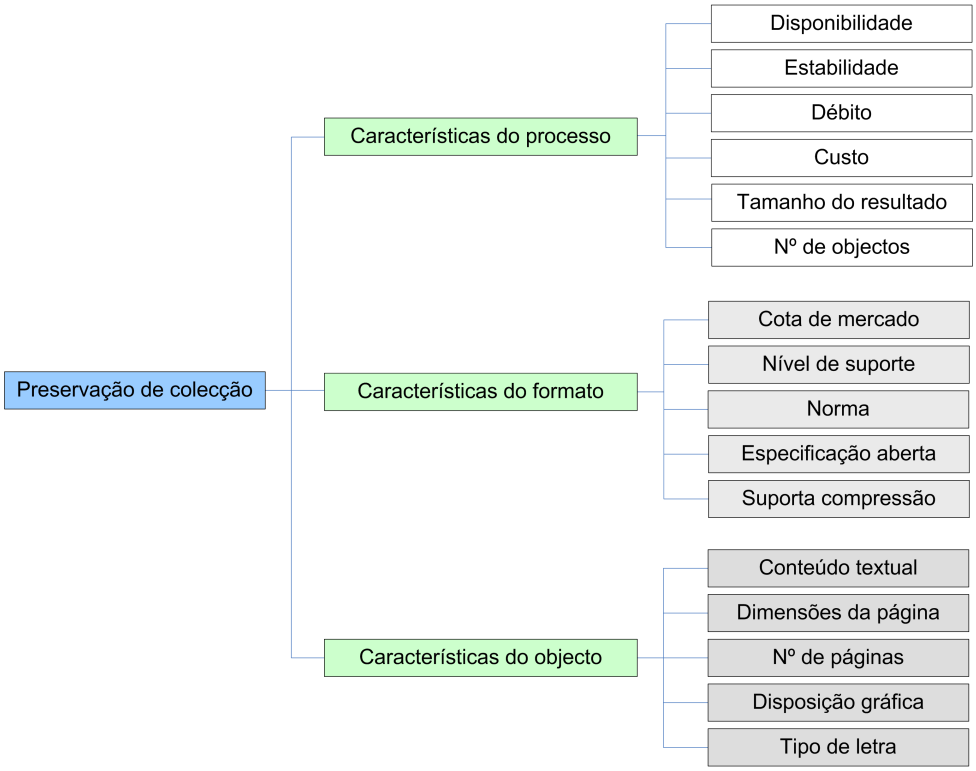


Figura 19 – Exemplo de árvore-objectivo.

O método desenvolvido segue o seguinte protocolo:

1) Inicialmente é construída uma árvore-objectivo onde são reunidos e organizados hierarquicamente os vários critérios que serão utilizados para aferir o grau de adequabilidade de uma estratégia de preservação (Figura 19);

2) Numa segunda fase são associadas unidades de medida a cada um desses critérios, e.g. milímetro, segundo, Mb/s, Euro, etc.;

3) Num terceiro passo é reunido um conjunto representativo de objectos digitais que será utilizado para testar cada uma das alternativas de preservação;

4) A quarta fase consiste na selecção de um conjunto de alternativas que poderão ser utilizadas para preservar a colecção de objectos de teste. Estas alternativas serão comparadas e ordenadas de acordo com a sua capacidade de satisfazer os critérios de avaliação reunidos;

5) No quinto passo cada uma das alternativas é executada face ao conjunto de objectos de teste. O resultado de cada intervenção é então avaliado à luz dos vários critérios que constituem a árvore-objectivo (Figura 20 – 1);

6) No sexto passo os resultados das avaliações são normalizados, i.e., transformados em unidades numéricas comparáveis (Figura 20 – 2);

7) No sétimo são atribuídos pesos a cada um dos critérios que constituem a árvore-objectivo. Os pesos atribuídos representam as preferências de preservação de quem está a avaliar as alternativas e irão determinar a estratégia mais adequada (Figura 20 – 3);

8) O passo oito consiste na agregação de valores parciais e totais obtidos a partir das experiências realizadas (Figura 20 – 4);

9) Finalmente, todas as alternativas são ordenadas mediante o grau de adequação que apresentam face aos requisitos manifestados pela entidade-cliente.

É importante realçar que a construção da árvore-objectivo é, por si só, uma tarefa complexa, morosa e que geralmente requer o envolvimento de profissionais da área tecnológica, arquivística, produtores de informação e respectivos consumidores.

Rauch e Rauber têm promovido a construção de árvores-objectivo para diversas classes de objectos digitais através da realização de *workshops* no seio de organizações detentoras de informação digital. Durante esses *workshops,* um conjunto de pessoas é convidado a sugerir critérios de avaliação que consideram importantes no sentido de garantir a preservação de um

dado conjunto de objectos digitais. Estes critérios são então organizados em classes e subclasses de forma constituir uma árvore-objectivo semelhante à apresentada na Figura 19.

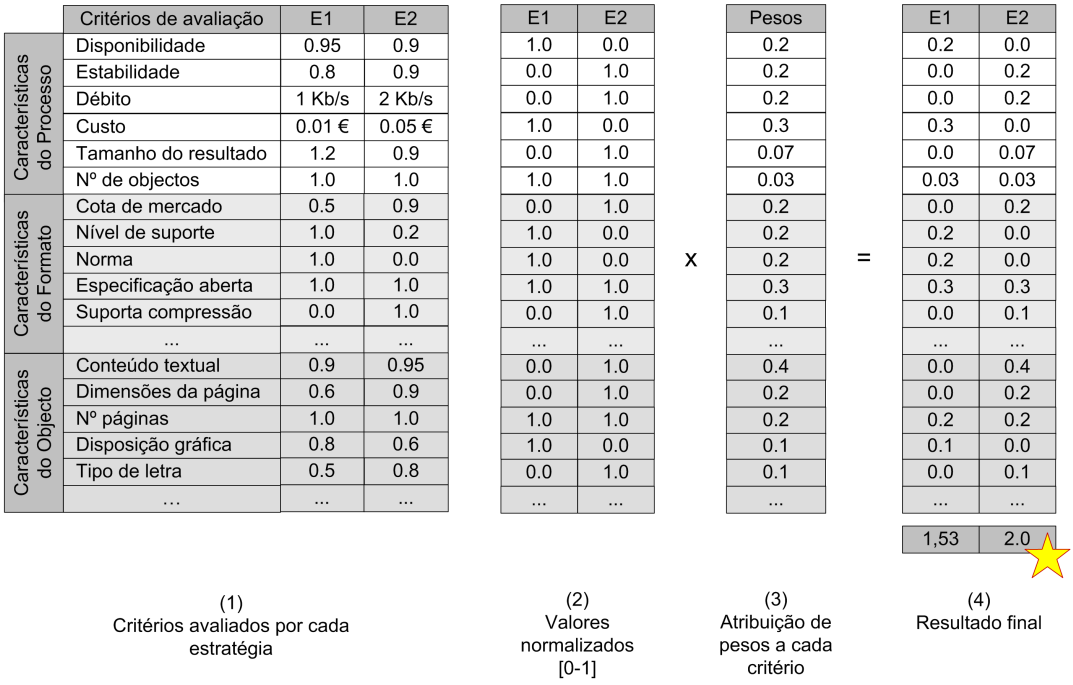


Figura 20 – Processo de selecção de estratégias de preservação.

A árvore-objectivo da Figura 19 descreve um conjunto de critérios para a avaliação de estratégias para a preservação de documentos de texto. Nela podemos encontrar critérios relativos ao processo de preservação (e.g. disponibilidade, estabilidade, débito, custo, etc.), aos formatos envolvidos na preservação (e.g. cota de mercado, nível de suporte, se se trata de um formato normalizado, etc.) e critérios relacionados com os objectos propriamente ditos (e.g. conteúdo textual, dimensões da página, nº de páginas, etc.).

Este último conjunto de critérios pode ser entendido como o conjunto das propriedades significativas associadas a uma respectiva classe de objectos digitais, neste caso documentos de texto (Ferreira et al., 2006a).

3.5 Cenário revisto

O cenário apresentado anteriormente permitiu identificar os vários serviços de preservação necessários para implementar uma estratégia de migração de forma transversal e automática. Assumindo a existência de uma infra-estrutura capaz de disponibilizar os serviços descritos, o mesmo cenário poderia ser reescrito da seguinte forma:

Num dado momento, uma empresa de dimensão média decide que todos os relatórios técnicos produzidos no decurso da sua actividade deverão estar acessíveis a todos os seus colaboradores à distância de um clique. Para tal, foi contratado um gestor de informação cuja função seria implementar e administrar um repositório digital de informação com o objectivo de preservar e dar acesso aos mesmos através da Intranet da empresa.

Todos os relatórios existentes até à data de implementação do repositório haviam sido elaborados com recurso ao Microsoft Word 95. À medida que o tempo foi passando, novas versões da aplicação Word começaram a ser exploradas no interior da organização. Paralelamente, alguns colaboradores mais adeptos do movimento *open-source* começaram a utilizar a ferramenta OpenOffice para produzir os seus relatórios técnicos. Consequentemente, o número de formatos existentes no repositório aumentou a ponto de se tornar impossível consultar transversalmente o conjunto de relatórios de um dado projecto sem que houvesse necessidade de instalar software específico para os poder visualizar.

Para agravar ainda mais a situação, o responsável pelo repositório verificou que existiam vários relatórios cuja extensão não lhe era familiar, dificultando assim a identificação da aplicação adequada à sua visualização. Paralelamente, a Microsoft anuncia que a nova versão do seu pacote de aplicações Office não irá suportar o formato Word 95.

O responsável pela gestão do repositório digital conclui que algo teria de ser feito no sentido de homogeneizar os conteúdos do repositório. Assim, desenvolve uma pequena aplicação capaz de interagir com os serviços fornecidos pela plataforma de preservação que se encontra acessível através da Internet. A aplicação começa por enviar os objectos cuja extensão é desconhecida para o serviço designado Identificador de formatos. De seguida, a aplicação-cliente consulta o serviço de Notificação de obsolescência com a finalidade de determinar quais os formatos que se encontram em risco de se tornar obsoletos. O serviço invocado determina que existe um formato no repositório que se encontra em vias de se tornar obsoleto e que deverão ser desencadeadas medidas preventivas ao nível da sua preservação.

A fim de determinar qual a estratégia de migração mais adequada para preservar esses documentos, a aplicação desenvolvida pelo gestor invoca um serviço disponibilizado

pela plataforma que devolve uma lista de critérios relevantes para efeitos de avaliação e controlo de qualidade (i.e., o Seleccionador de estratégias de migração). O colaborador passa a atribuir pesos a cada um desses critérios tendo em consideração os requisitos de preservação definidos pela sua organização. Entre estes encontram-se itens como: preservação do conteúdo textual, preservação da apresentação gráfica do documento, custo de migração (€/conversão), velocidade de conversão (Kb/s), etc.

O gestor do repositório decide que o conteúdo textual e a apresentação gráfica dos documentos são propriedades importantes e que portanto deverão ser preservados a todo o custo. Os restantes critérios foram menos valorizados, pelo que, o peso que lhes foi atribuído foi expressivamente inferior.

O serviço remoto, após receber as preferências manifestadas pelo gestor, responde com uma listagem de formatos para os quais os documentos Word 95 poderão ser convertidos. Estes formatos maximizam os requisitos de preservação manifestados pelo funcionário. Entre estes formatos encontram-se o PDF, Word 2003 e OpenOffice 2. O formato PDF foi apontado pelo sistema como sendo o mais favorável.

O gestor decide adoptar a sugestão fornecida pelo sistema e requisita uma lista de possíveis serviços de conversão capazes de realizar a respectiva conversão. Baseado nos pesos previamente atribuídos pelo gestor, o sistema remoto sugere um serviço de conversão, que embora não seja gratuito, garante resultados de elevada qualidade. O gestor passa então a enviar os seus documentos para o sistema remoto, invocando o serviço de Conversão disponível, e dá início ao processo de migração dos seus objectos digitais.

Após cada migração, a aplicação-cliente recebe uma versão PDF do documento técnico submetido a conversão e um registo de metainformação produzido pelo serviço de Controlo de qualidade que poderá ser utilizado para documentar a intervenção de preservação. Nesse registo encontra-se informação variada como uma descrição do serviço de migração utilizado, a data e a hora da conversão e o nível de degradação incorrido em cada uma das propriedades significativas do documento original.

Depois de realizar os mesmos passos para os restantes formatos existentes no repositório, o gestor do repositório constata que o PDF é quase sempre sugerido como o formato mais adequado para preservar os relatórios técnicos armazenados no

repositório. O funcionário decide, então, elaborar uma política de ingestão onde é recomendado que todos os relatórios técnicos sejam convertidos para PDF antes de serem submetidos ao repositório.

O gestor do repositório desenvolveu também os mecanismos necessários para que o repositório pudesse consultar regularmente o serviço de notificação de obsolescência de formatos. Assim, saberia de imediato se algum dos formatos que mantém no seu repositório se encontra em risco de se tornar obsoleto e passa a poder agir em conformidade e de forma antecipada.

3.6 Considerações finais

Este capítulo teve como principal objectivo apresentar o conjunto mínimo de serviços considerados essenciais para a implementação transversal de estratégias de preservação baseadas em migração num contexto organizacional.

O capítulo começa por descrever as três actividades fundamentais que geralmente acompanham um processo de migração, nomeadamente: a selecção de uma alternativa de migração, a execução da respectiva conversão e a análise dos resultados obtidos (i.e., controlo de qualidade).

O capítulo continua com uma definição de arquitectura orientada ao serviço (SOA) e com a apresentação deste tipo de plataformas como sendo adequadas a contextos de preservação, evidenciando as vantagens que advêm da sua utilização.

Ainda neste capítulo, é apresentado um problema de preservação que foi solucionado de duas formas distintas. No primeiro caso, a inexistência de uma plataforma de serviços de auxilio à preservação obrigou a que a generalidade das actividades de preservação fossem realizadas manualmente pelo gestor de um repositório digital; no segundo, a presença de uma plataforma de serviços de preservação viabiliza a automatização de processos e simplifica todo o processo administrativo.

O primeiro cenário serve também de ponto de partida para a apresentação de um conjunto de serviços considerados fundamentais no que diz respeito à automatização de processos de migração. Entre estes encontram-se os seguintes: um serviço de identificação de formatos, um serviço que permite identificar as alternativas de migração mais adequadas para solucionar o problema de preservação específico de um cliente, um serviço capaz de realizar conversões de formatos, um serviço de controlo de qualidade e um serviço de notificação de obsolescência.

Para cada um dos serviços apresentados procurou-se seleccionar um conjunto de ferramentas e/ou tecnologias capazes de dar suporte à sua implementação. Estas ferramentas serviram de base para o desenvolvimento do CRiB, uma arquitectura orientada ao serviço que disponibiliza um conjunto de funcionalidades que permitem implementar de forma transversal e automática estratégias de preservação baseadas em migração. Esta plataforma é descrita, em detalhe, no capítulo que se segue.

Capítulo 4
CRiB – Plataforma de serviços de preservação

O capítulo anterior procurou evidenciar de que forma uma arquitectura baseada em serviços poderia facilitar a implementação de estratégias de preservação, especialmente aquelas baseadas na migração de formatos. Na presença de uma arquitectura deste tipo, qualquer indivíduo ou instituição com capacidade para invocar serviços remotos passa a poder implementar os seus próprios processos de preservação, construídos a partir dos serviços disponibilizados.

No capítulo anterior foram identificados vários serviços de preservação, bem como possíveis formas de os implementar. Entre estes, encontram-se um notificador de obsolescência, um identificador de formatos, um conjunto de conversores de formatos, um módulo de controlo de qualidade e um componente capaz de auxiliar o cliente na escolha da alternativa de migração mais adequada à resolução do seu problema de preservação.

Este capítulo introduz a plataforma CRiB[67], uma arquitectura orientada ao serviço que procura implementar os conceitos e serviços anteriormente descritos. As secções que se seguem

[67] CRiB é um acrónimo que deriva da expressão Conversion and Recommendation of Digital Object Formats.

descrevem de forma detalhada como cada um destes serviços foi desenvolvido e quais as suas dependências funcionais.

O capítulo encontra-se organizado da seguinte forma: a secção 4.1 apresenta uma visão geral da arquitectura desenvolvida, descrevendo sucintamente os componentes e serviços por ela implementados; a secção 4.2 descreve em detalhe o componente CRiB Core Preservation Services que tem como missão servir de interface entre a plataforma de serviços e os seus utilizadores. Esta secção apresenta, ainda, as mensagens trocadas entre ambos os intervenientes e um conjunto de diagramas de sequência que facilita a compreensão de todo o processo de interacção. As secções 4.3, 4.4, 4.5, 4.6, 4.7 e 4.8 descrevem em detalhe os restantes componentes do sistema, nomeadamente o Service Registry, o Format Identifier, o Migration Broker, o Object Evaluator, o Format Evaluator e o Migration Advisor; o capítulo termina, na secção 4.9, com um sumário e algumas considerações finais relativamente ao trabalho realizado.

4.1 Visão geral

O CRiB trata-se de uma arquitectura orientada ao serviço que tem como objectivo auxiliar tanto instituições, como utilizadores individuais, na implementação de estratégias de preservação baseadas em migração. O conjunto de serviços disponibilizados por esta plataforma permite a qualquer entidade cliente implementar de forma transversal e automática todas as tarefas subjacentes à preservação de objectos digitais.

A Figura 21 apresenta a arquitectura geral da plataforma de serviços CRiB. Esta, encontra-se divida em três camadas: a camada de aplicação (*application layer*), a lógica de negócio (*business layer*) e a camada de dados e/ou fontes de informação (*data layer*).

Na camada de aplicação podem ver-se exemplos de aplicações-cliente semelhantes às já apresentadas no modelo abstracto da Figura 18. Entre estes encontram-se repositórios digitais de âmbito geral como o DSpace, o Fedora Commons ou o Eprints, e algumas aplicações específicas como o repositório RODA (Barbedo et al., 2007; Faria et al., 2007; Portuguese National Archives & University of Minho, 2006; Ramalho, Ferreira, Castro et al., 2007; Ramalho, Ferreira, Faria, & Castro, 2007) ou a aplicação Plato do projecto Planets (Becker, Ferreira et al., 2008; Becker, Kulovits, Rauber, & Hofman, 2008).

A camada de negócio (*business layer*) identifica os principais componentes responsáveis por realizar todos os serviços disponibilizadas pelo CRiB. O componente CRiB Core Preservation Services serve de mediador entre as aplicações-cliente e o resto dos

componentes do sistema. Para além disso, é também responsável por orquestrar todas as mensagens trocadas no interior do sistema garantindo, deste modo, o correcto funcionamento da arquitectura (Ferreira et al., 2006b).

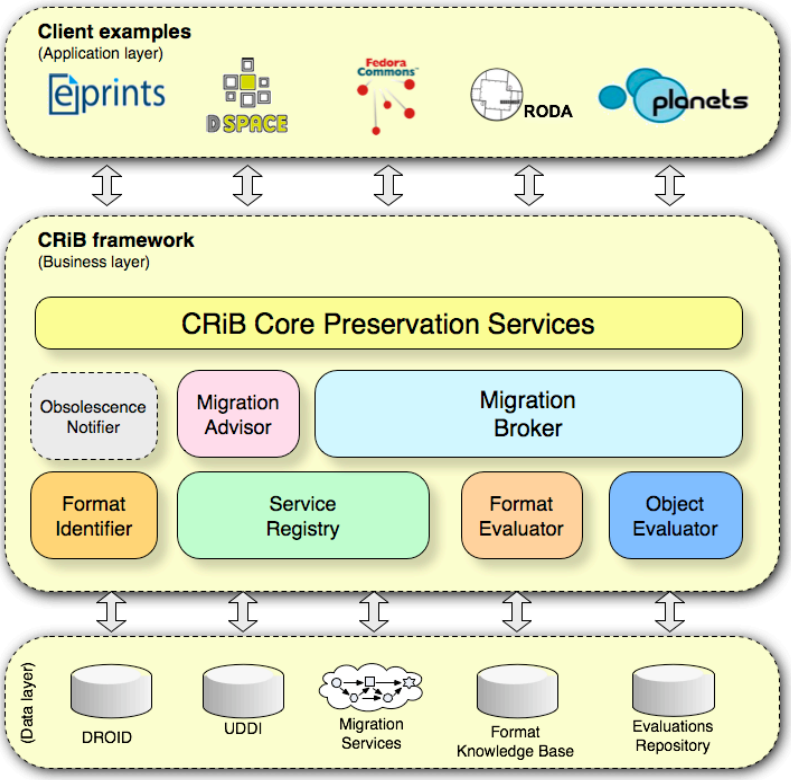


Figura 21 – Arquitectura geral da plataforma CRiB.

O componente Format Identifier disponibiliza um serviço de identificação de formatos que poderá ser invocado por uma aplicação-cliente.

O componente Service Registry oferece um conjunto de métodos que permite registar e localizar serviços de migração disponibilizados através da plataforma. Este serviço é ainda capaz de descobrir conversões compostas calculando no momento do pedido o fecho transitivo entre os diferentes serviços de conversão registados.

O acesso aos conversores registados na plataforma é realizado através do componente Migration Broker. Este componente é responsável por efectuar duas operações fundamentais: executar os processos de conversão (quer estes sejam individuais ou compostos) e medir a performance computacional dos mesmos (Becker, Ferreira et al., 2008; Ferreira, 2006a; Ferreira et al., 2005, 2006b, 2007).

O componente Object Evaluator é responsável pelo controlo de qualidade, ou seja, cabe a este componente a tarefa de detectar possíveis perdas de informação incorridas durante o processo de migração (Becker, Ferreira et al., 2008; Ferreira, 2006a; Ferreira et al., 2005, 2006b, 2007).

O componente Format Evaluator fornece informação técnica sobre os formatos suportados pela plataforma e permite determinar quais os formatos que possuem o conjunto de características mais propício para preservar informação durante longos períodos de tempo.

A informação disponibilizada pelo Format Evaluator, combinada com a informação produzida pelo Object Evaluator e pelo Migration Broker, permite ao componente Migration Advisor determinar qual a estratégia de migração mais adequada para resolver um determinado problema de preservação. Este componente produz uma lista de serviços de migração que garantem à entidade-cliente a melhor solução em termos de performance, conservação das propriedades significativas dos objectos digitais e aptidão dos formatos para reter informação por longos períodos de tempo (Becker, Ferreira et al., 2008; Ferreira, 2006a; Ferreira et al., 2005, 2006b, 2007).

Para dar suporte aos componentes anteriormente descritos, o CRiB recorre a alguns serviços de informação. Estes encontram-se representados na camada inferior da Figura 21 designada por *data layer*. O Droid[68], utilizado pelo Format Identifier, fornece o motor de identificação de formatos; o jUDDI[69] implementa funcionalidades de registo e descoberta de serviços e é utilizado pelo componente Service Registry; o Format Knowledge Base materializa uma base de dados com informação relevante sobre os formatos suportados e alimenta o componente Format Evaluator; e o Evaluations Repository armazena todos os relatórios produzidos pelos componentes responsáveis pelo de controlo de qualidade, nomeadamente o Format Evaluator, Object Evaluator e Migration Broker, e dá suporte ao motor de recomendação implementado pelo Migration Advisor (Becker, Ferreira et al., 2008; Ferreira, 2006a; Ferreira et al., 2005, 2006b, 2007).

É de realçar que os componentes Format Knowledge Base e Evaluations Repository foram totalmente desenvolvidos no âmbito deste projecto. Os restantes componentes associados à *data layer,* Droid e UDDI, foram desenvolvidos por terceiros e encontram-se descritos nas secções 3.4.1 e 4.3, respectivamente.

---

[68] Ver secção 3.4.1 na página 58.
[69] Ver secção 4.3 na página 83.

É importante referir que a tecnologia que suporta os serviços descritos neste capítulo é baseada em Web services, ou seja, toda a comunicação realizada entre as aplicações-cliente e a plataforma CRiB, assim como todas as mensagens trocadas no seu interior são asseguradas por protocolos abertos baseados em XML/SOAP (S. Graham et al., 2002; Newcomer & Lomow, 2005; W3C, 2002).

4.2 Core preservation services

O CRiB disponibiliza um conjunto de serviços de preservação úteis a qualquer instituição, ou indivíduo, com um problema específico de preservação. Para melhor compreender de que forma uma instituição poderá utilizar os serviços disponibilizados, é apresentado um conjunto de diagramas que descrevem as sequências de interacção que modelam a comunicação entre o cliente e o sistema, as suas interfaces aplicacionais e os objectos trocados.

É importante referir que cada um dos subcomponentes que constituem o sistema pode ser acedido directamente de forma independente dos restantes. No entanto, o componente `CRiB Core Preservation Services` (i.e., a interface aplicacional do sistema[70]) simplifica o *workflow* no interior do CRiB, desdobrando cada pedido do cliente num conjunto de mensagens que serão resolvidas ordenadamente pelos restantes subcomponentes do sistema.

A interface aplicacional disponibilizada pelo CRiB encontra-se ilustrada na Figura 22. Aqui encontram-se representadas as várias funções que permitem ao utilizador realizar tarefas como: identificação de formatos (i.e., `identifyFormat`), selecção de estratégias de migração (i.e., `getEvaluationCriteria` e `getRecommendation`), migração de formatos com controlo de qualidade associado (i.e., `convert`) e alguns serviços adicionais que facilitam a exploração do sistema e a obtenção de informações relevantes para o cliente (i.e., `getSupportedSourceFormats`, `getSupportedTargetFormats`, `getMigrationPaths` e `getConverterMetadata`).

[70] Do inglês *Application Programming Interface* (API).

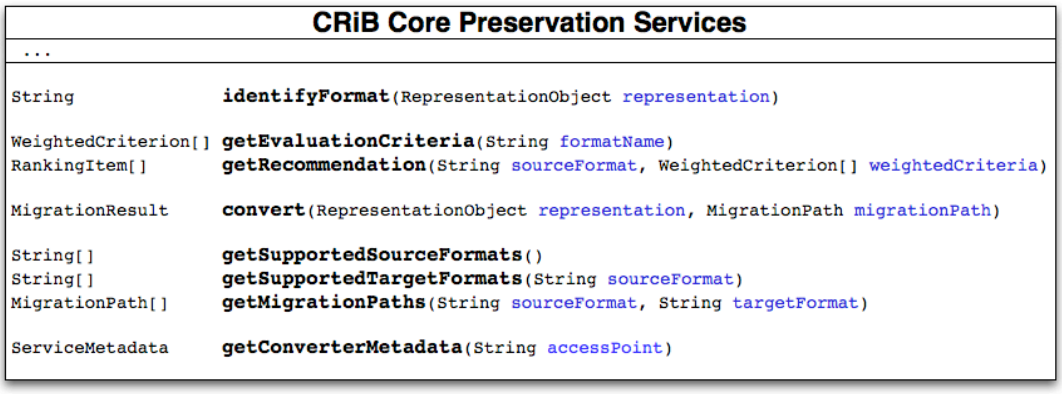


Figura 22 – Interface do componente Core Preservation Services.

As estruturas de dados trocadas durante a invocação dos serviços enumerados encontram-se representadas no diagrama de classes da Figura 23.

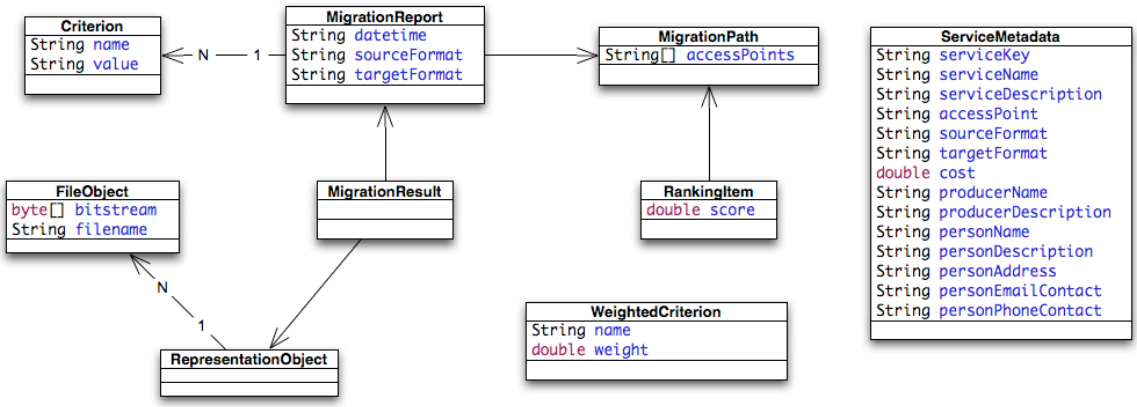


Figura 23 – Diagrama de classes das mensagens trocadas pelo CRiB.

Cada um dos serviços anteriormente apresentados é descrito com maior detalhe nas secções subsequentes.

4.2.1 Identificação de formatos

Para identificar o formato de uma representação digital, um cliente apenas necessita de invocar o método remoto designado `identifyFormat`, enviando a respectiva representação como argumento do mesmo. Após analisar a representação submetida, o sistema responde com o nome e versão do formato detectado ou com o termo `Unknown Format`, caso este não seja reconhecido pelo sistema (Figura 24).

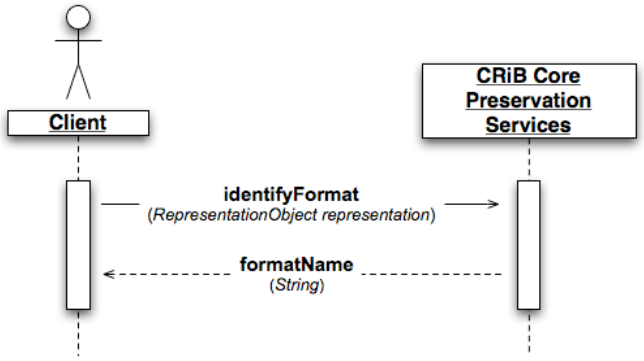


Figura 24 – Diagrama de sequência da identificação de formatos.

A representação enviada como argumento segue a estrutura da mensagem `RepresentationObject` descrita no diagrama de classes da Figura 25. Esta estrutura segue o modelo definido no Dicionário de dados PREMIS para descrever uma representação digital. Segundo este modelo uma `representação digital` é composta por um ou mais `ficheiros`, e um ficheiro é composto por um ou mais `bitstreams`, i.e., sequências de bits (Guenther et al., 2008; PREMIS Working Group, 2005).

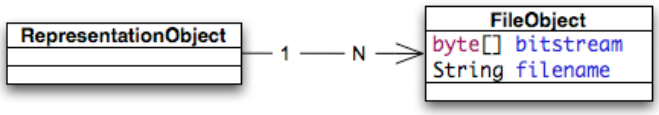


Figura 25 – Diagrama de classes de uma representação.

A estrutura `FileObject` possui, ainda, um atributo adicional designado `filename`, que tem como objectivo identificar o ficheiro veiculado pela estrutura. Este atributo facilita a identificação de formatos e permite preservar os nomes originais dos ficheiros que constituem a representação.

4.2.2 Selecção de estratégias de migração

Após identificar o formato da representação que se pretende preservar, é possível proceder à selecção de um conjunto de alternativas de migração consideradas aptas para resolver o problema de preservação da instituição-cliente. Todas as alternativas conhecidas pelo sistema são avaliadas e ordenadas mediante o nível de aptidão demonstrado em resolver o problema específico do cliente.

Para que o serviço seja capaz de ordenar as alternativas de migração de acordo com a sua aptidão, este necessita de conhecer os requisitos específicos do cliente. Assim, numa primeira iteração, o cliente invoca o método `getEvaluationCriteria` de modo a obter a lista de critérios de avaliação que são suportados para uma dada classe de objectos. Este processo encontra-se ilustrado na Figura 26. A Figura 27 apresenta os objectos trocados entre o sistema e o cliente.

Os critérios de avaliação suportados pelo CRiB não dependem do formato, mas sim da classe de objectos a que o formato pertence. Por exemplo, um objecto pertencente à classe `documentos de texto` é avaliado à luz de um conjunto específico de critérios, enquanto que um objecto pertencente à classe `imagens matriciais` é avaliado por um conjunto de critérios completamente distinto. O CRiB trata de identificar automaticamente a classe de objectos a partir do formato indicado.

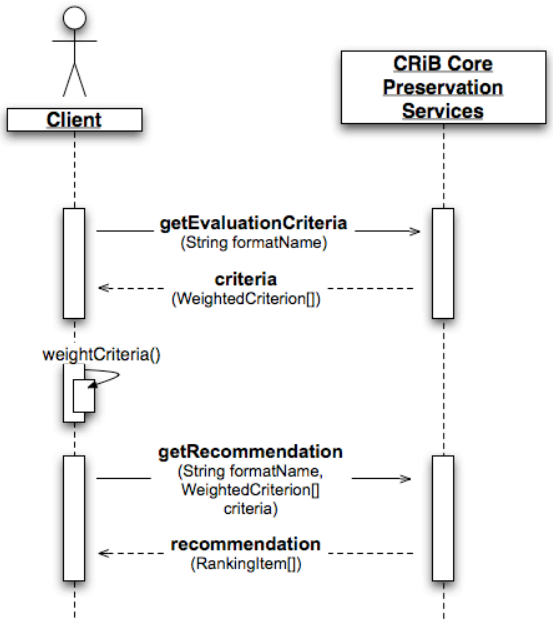


Figura 26 – Diagrama de sequência relativo à selecção de uma alternativa de migração.

Ao receber a lista de critérios de avaliação fornecida pelo CRiB (i.e., `WeightedCriterion[]`), o cliente deverá associar um peso ou importância a cada um dos critérios constituintes. Os pesos atribuídos pelo cliente são, efectivamente, a sua manifestação de preferências ou, por outras palavras, a formalização do seu problema específico de preservação. Por exemplo, nesta fase o cliente poderá informar o sistema que considera a

velocidade de conversão um critério da máxima importância, enquanto que o custo da mesma não deverá ser considerado decisivo.

Os pesos atribuídos a cada um dos critérios de avaliação deverão pertencer ao conjunto [0, 1], com 0 a representar um critério considerado pouco relevante e 1 a representar um critério com elevada influência na decisão final. O sistema é capaz de analisar três categorias distintas de critérios[71]:

- Critérios associados ao processo de migração (e.g. disponibilidade, custo, débito, etc.);
- Critérios relacionados com aspectos técnicos dos formatos envolvidos na migração (e.g. abertura do formato, quota de mercado, facilidade de descodificação, etc.);
- Critérios associados ao próprio objecto digital (e.g. conteúdo, apresentação gráfica, framerate, nº de cores, nº de páginas, etc.).

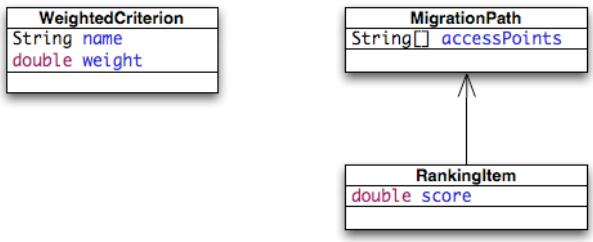


Figura 27 – Mensagens envolvidas na selecção de uma alternativa de migração.

Uma vez atribuídos os pesos por parte do cliente, este deverá invocar o método `getRecommendation`, enviando a lista de critérios previamente pesados e indicando o formato dos objectos que pretende preservar (Figura 26).

O serviço passa então a ser capaz de identificar as alternativas de migração mais aptas para satisfazer as suas necessidades de preservação e devolve ao cliente uma lista de caminhos de migração à qual este poderá recorrer, bem como a pontuação atribuída a cada um destes. A lista devolvida é, efectivamente, a recomendação produzida pelo sistema. O cliente é livre de seleccionar qualquer uma das opções sugeridas.

[71] O conjunto global de critérios de avaliação suportados pelo CRiB encontra-se descrito em detalhe na secção 4.6.2, Taxionomias de avaliação, na página 105.

4.2.3 Migração de formatos e controlo de qualidade

Após obter uma recomendação, o cliente poderá invocar qualquer um dos caminhos de migração sugeridos pelo sistema de forma a migrar os seus objectos para o formato de destino recomendado. O CRiB disponibiliza um método, convenientemente designado `convert`, que permite realizar esta operação.

Para que possa ser utilizado, este método necessita de saber qual o caminho de migração a executar, i.e., `MigrationPath`, e a representação que se pretende converter, i.e., `RepresentationObject` (Figura 28). O método remoto trata de compor todos os serviços de conversão e executar a respectiva migração, avaliando, em simultâneo, a performance da mesma.

É importante referir que a lista de pontos de acesso incluídos numa mensagem do tipo `MigrationPath` funciona como um identificador único para um dado conversor composto. O CRiB irá procurar na sua lista de serviços se existe alguma conversão composta pelos pontos de acesso fornecidos. Se não existir, este irá devolver ao cliente uma excepção, identificando claramente o problema detectado. Caso contrário, o CRiB trata de invocar todos os serviços de migração necessários para satisfazer o pedido do cliente.

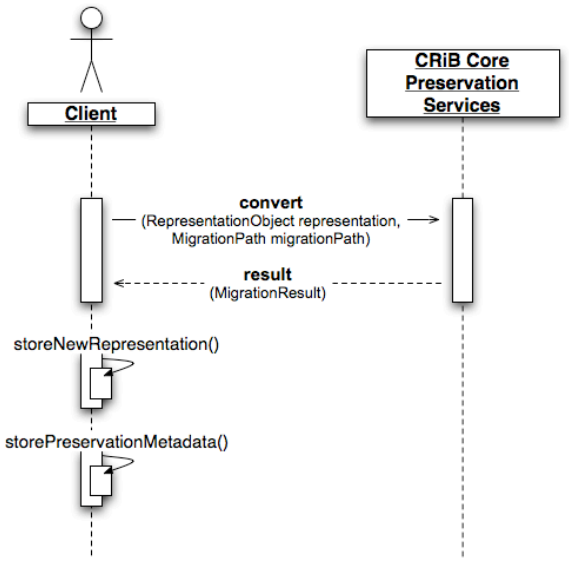


Figura 28 – Diagrama de sequência do processo de conversão.

Após terminar a conversão requisitada pelo cliente, o CRiB desencadeia um conjunto de acções de controlo de qualidade ao nível da performance dos conversores utilizados, aptidão

dos formatos envolvidos na conversão e capacidade apresentada pelos conversores em preservar as propriedades significativas dos objectos submetidos a conversão. Estas acções são realizadas pelos componentes Migration Broker[72], Format Evaluator[73] e Object Evaluator[74], respectivamente.

Os resultados destas acções de controlo de qualidade são reunidos numa estrutura de dados designada MigrationReport (Figura 29). Esta estrutura contém informação suficiente para documentar a intervenção de preservação. Esta inclui detalhes sobre os conversores utilizados durante a migração (e.g. nome, descrição, produtor, etc.), a data e hora da intervenção, os formatos envolvidos, a lista de critérios que foram avaliados pelo sistema, bem como os resultados dessa avaliação.

Estes relatórios permitem informar os futuros consumidores da informação que modificações foram introduzidas nas suas propriedades significativas. Ao consultar esta informação, o consumidor será capaz de aferir o grau de fidelidade apresentado pelo objecto preservado em relação à sua representação original. Este relatório serve assim para verificar a autenticidade das representações intervencionadas.

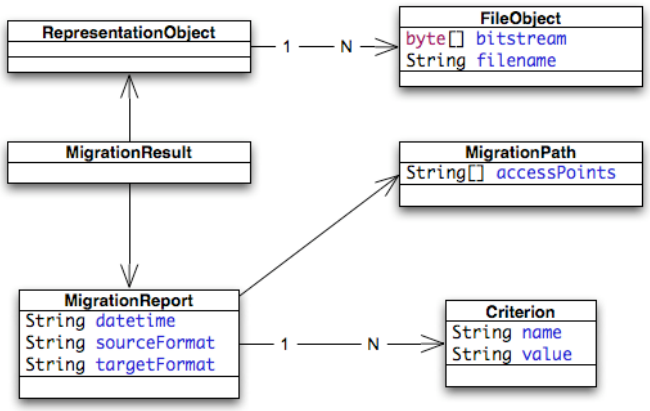


Figura 29 – Diagrama de classes associadas ao processo de conversão.

Para além do disposto, o relatório produzido permite à entidade preservadora aferir a qualidade da intervenção, expondo de forma objectiva o nível de degradação introduzido pelo

[72] Este componente encontra-se descrito em detalhe na secção 4.5 na página 89.
[73] Este componente encontra-se descrito em detalhe na secção 4.6 na página 96.
[74] Este componente encontra-se descrito em detalhe na secção 4.7 na página 110.

processo de migração e permitindo à mesma determinar se a intervenção realizada satisfaz os seus requisitos mínimos de qualidade.

Após a conversão, é também devolvida ao cliente uma estrutura de dados contendo a nova representação digital (i.e., o objecto convertido). Ambas as estruturas descritas, i.e., o relatório de qualidade e a nova representação, são encapsuladas numa mensagem designada `MigrationResult` (Figura 29).

Após receber o resultado da migração (i.e., `MigrationResult`), o cliente deverá desenvolver localmente duas acções fundamentais: gravar a nova representação no seu sistema de armazenamento e, no caso de pretender reter metainformação de preservação, anexar o relatório de migração à metainformação de preservação que acompanha os seus objectos digitais.

4.2.4 Serviços adicionais

Para além dos serviços de preservação anteriormente descritos, o CRiB disponibiliza um conjunto de métodos remotos que facilitam a descoberta de serviços de conversão (Figura 30). Entre estes, encontra-se um método que permite descobrir quais os formatos de origem suportados pelo CRiB, i.e., a partir de que formatos existem conversores registados no CRiB – este método designa-se por `getSupportedSourceFormats`.

Ainda neste contexto, é possível conhecer, para um dado formato, quais os formatos de destino disponíveis na plataforma – `getSupportedTargetFormats`.

Para conhecer os conversores disponíveis entre dois formatos, o cliente poderá invocar o método `getMigrationPaths`. O sistema, irá devolver todos os caminhos de migração disponíveis entre os dois formatos desejados. Este pedido poderá resultar numa lista relativamente extensa de caminhos de migração. Para determinar qual o caminho de migração mais adequado, o cliente deverá invocar o método `getRecommendation` descrito anteriormente na secção 4.2.2.

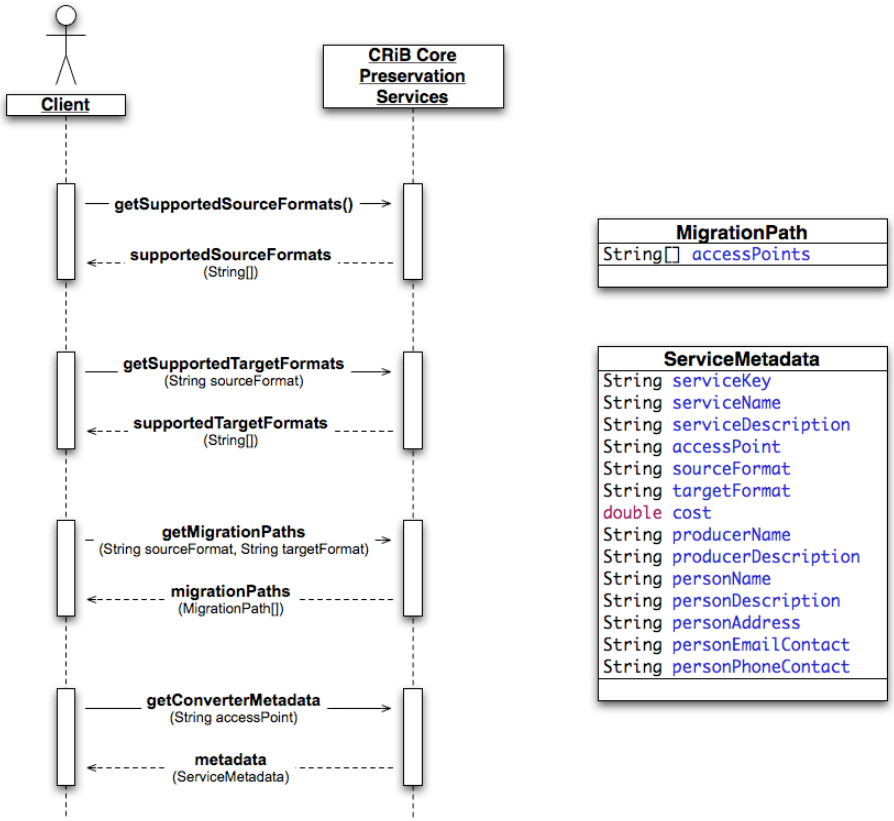


Figura 30 – Outros métodos disponibilizados pelo CRiB.

Adicionalmente, o componente Core Preservation Services disponibiliza um método que permite ao cliente conhecer melhor um dado serviço de migração – getConverterMetadata. Este método permite obter informação variada, como: o nome e descrição do serviço, formatos suportados, custo de utilização e dados relativos à entidade produtora do serviço, entre outros (Figura 30).

As secções que se seguem descrevem detalhadamente cada um dos subcomponentes do sistema que permitem, em conjunto, realizar as tarefas anteriormente descritas, disponibilizadas pelo CRIB Core Preservation Services.

4.3 Service Registry

O componente Service Registry tem como principal objectivo reunir informação sobre os diversos serviços de conversão existentes na rede. Esta informação dá suporte à localização e invocação desses mesmos serviços e permite aos clientes obter informação detalhada sobre os agentes de software que intervieram em processos de migração.

O Service Registry tem como base a norma Universal Description, Discovery and Integration, vulgarmente designada por UDDI (OASIS, 2005). A norma UDDI resulta de uma iniciativa aberta conduzida por um grupo de representantes da indústria (e.g. Ariba, IBM e Microsoft) e é actualmente suportada pela Organization for the Advancement of Structured Information Standards (OASIS)[75]. Nascida em 2000, esta norma assegura o registo, publicação e pesquisa de informação sobre serviços disponibilizados na Web, seus produtores e a forma como podem ser invocados por uma qualquer aplicação-cliente. Estas informações encontram-se organizadas em três unidades semânticas designadas Business Entity, Service Entity e Binding Entity, respectivamente (S. Graham et al., 2002). As relações existentes entre cada uma destas entidades encontram-se ilustradas na Figura 31.

O componente Service Registry implementado no CRiB é, na prática, suportado por um servidor de UDDI designado Apache jUDDI[76]. O jUDDI trata-se de um servidor *open-source* desenvolvido pela Apache Software Foundation[77] que implementa a norma UDDI versão 2.0 (Bryan et al., 2002).

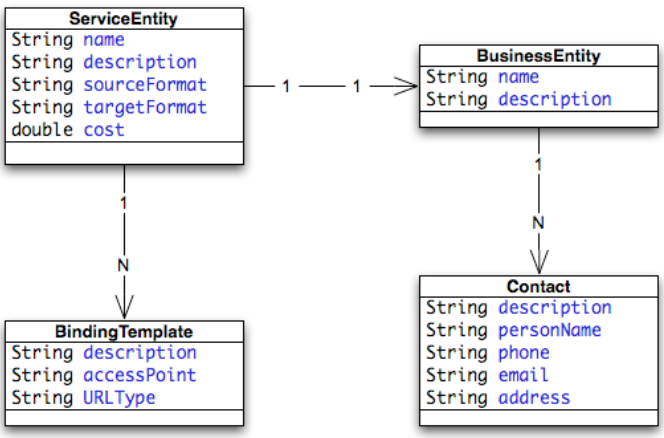


Figura 31 – Relações entre entidades que descrevem um serviço de conversão.

O UDDI foi estendido no âmbito do CRiB no sentido de suportar alguns atributos de informação complementares, considerados fundamentais para descrever serviços de migração. Entre estes, encontram-se: o sourceFormat, targetFormat e cost (Tabela 3). A

[75] http://www.oasis-open.org
[76] http://ws.apache.org/juddi/
[77] http://www.apache.org/

inclusão dos dois primeiros elementos permite a identificação e localização imediata de serviços de conversão tomando por base os formatos que suportam. Para além disso, possibilita a detecção de migrações compostas através da combinação de formatos de destino e formatos de partida. O último elemento permite associar um custo de utilização, em unidades monetárias, a cada serviço de conversão.

Neste contexto, é importante referir que, para que seja possível identificar e executar conversões compostas, é fundamental que aos atributos sourceFormat e targetFormat sejam associados valores obtidos a partir de um vocabulário controlado. No caso do CRiB, os valores utilizados são baseados nos descritores de formato produzidos pelo Droid (ver Format Identifier na página 88).

Cada serviço de conversão adicionado ao Service Registry é descrito pelos atributos apresentados na Tabela 2, Tabela 3, Tabela 4 e Tabela 5.

Business Entity		
Elemento descritivo	**Obrigatoriedade**	**Descrição**
name	Obrigatório	Nome da organização que desenvolveu o serviço de conversão.
description	Opcional	Descrição da organização.
contacts	Opcional	Contacto dos responsáveis pela criação e manutenção do serviço (ver Tabela 5).

Tabela 2 – Elementos de metainformação sobre a organização que desenvolveu o serviço de conversão.

Service Entity		
Elemento descritivo	**Obrigatoriedade**	**Descrição**
name	Obrigatório	Nome do serviço de conversão.
description	Opcional	Descrição do serviço de conversão.
sourceFormat	Obrigatório	Formato de origem da conversão (baseado num vocabulário controlado).
targetFormat	Obrigatório	Formato de destino da conversão (baseado num vocabulário controlado).
cost	Obrigatório	O custo de execução do conversor em unidades monetárias.
bindingTemplates	Obrigatório	Informação sobre a localização do serviço (ver Binding Templates).
businessEntity	Obrigatório	Informação sobre a organização que desenvolveu o serviço (ver Business Entity).

Tabela 3 – Elementos de metainformação que descrevem serviços de conversão.

Binding Templates		
Elemento descritivo	**Obrigatoriedade**	**Descrição**
description	Opcional	Descrição do localizador de serviço.
accessPoint	Obrigatório	Endereço onde reside o serviço.
URLType	Obrigatório	Protocolo de acesso ao serviço (e.g. mailto, http, https, ftp, fax, phone, other).

Tabela 4 – Elementos de metainformação que descrevem a localização do serviço.

Contacts		
Elemento descritivo	**Obrigatoriedade**	**Descrição**
description	Opcional	Descrição do contacto.
personName	Obrigatório	Nome da pessoa responsável.
phone	Opcional	Telefone do responsável.
email	Opcional	Endereço de correio-electrónico do responsável.
address	Opcional	Morada do responsável.

Tabela 5 – Elementos de metainformação que descrevem os contactos de uma organização.

O jUDDI é utilizado pelo Service Registry para armazenar a informação que descreve os serviços de migração. A comunicação entre estes dois componentes é realizada através de mensagens XML/SOAP. Para facilitar a comunicação foi utilizada uma biblioteca designada UDDI4J[78] que facilita a construção e envio dessas mensagens (Figura 32).

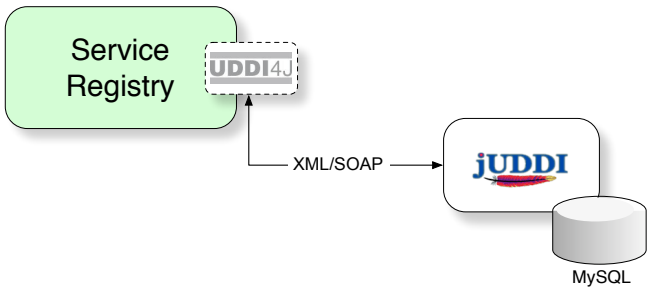


Figura 32 – Arquitectura detalhada do Service Registry.

O componente Service Registry disponibiliza todos os métodos definidos pela norma UDDI e complementa-os com métodos especificamente desenvolvidos para manipular serviços de conversão. A Figura 33 ilustra os principais métodos disponibilizados por este componente.

[78] http://uddi4j.sourceforge.net/

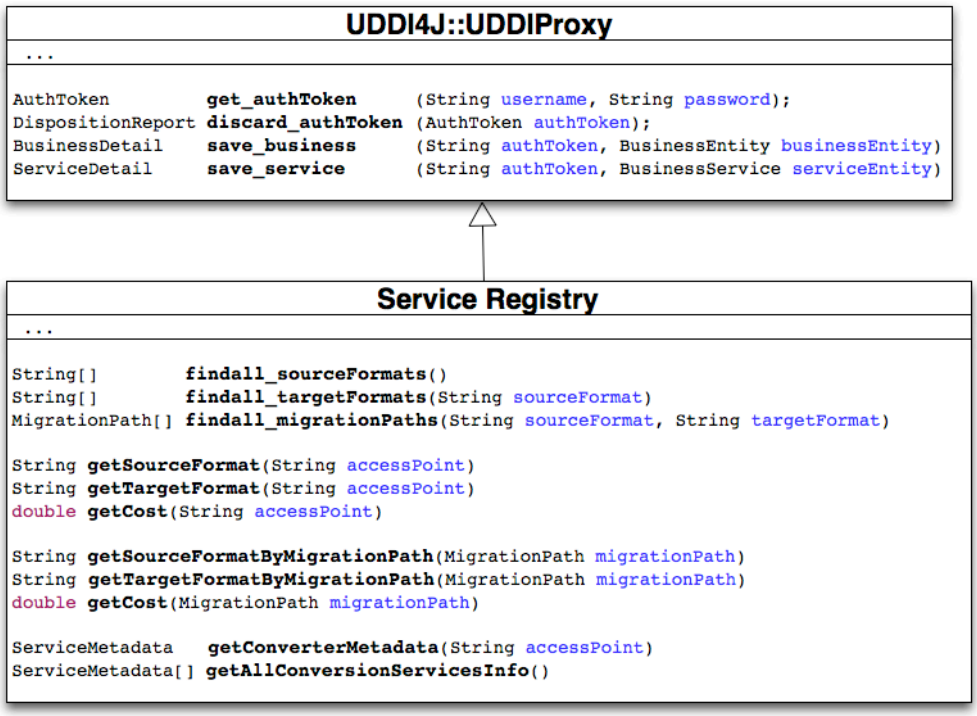


Figura 33 – Métodos disponibilizados pelo Service Registry.

Os métodos get_authToken e discard_authToken são herdados da classe UDDIProxy que acompanha a biblioteca UDD4J e permitem que um cliente se autentique e, subsequentemente, termine uma sessão e trabalho num servidor de UDDI. Os métodos save_business e save_service oferecem a capacidade de registar novos produtores e serviços de migração no directório de serviços, respectivamente.

Os restantes métodos oferecem funcionalidades básicas de consulta de serviços de conversão, como por exemplo: identificar todos os formatos de partida suportados (i.e., findall_sourceFormats), todos os formatos de destino para os quais existem conversores registados (i.e., findall_targetFormats), todos os caminhos de migração entre dois formatos (i.e., findall_migrationPaths), consultar o custo de invocação de um dado serviço (i.e., getCost), recolher toda a metainformação descritiva de um dado serviço (i.e., getConverterMetadata), determinar o formato de partida e de chegada de um dado serviço ou caminho de migração (i.e., getSourceFormat, getTargetFormat, getSourceFormatByMigrationPath e getTargetFormatByMigrationPath, respectivamente) e, ainda, um método que permite obter toda a metainformação armazenada no Service Registry (i.e., getAllConvertionServicesInfo).

4.4 Format Identifier

O CRiB incorpora também um serviço de identificação de formatos. Este serviço é assegurado pelo componente Format Identifier e é baseado no Droid[79], um software desenvolvido pelos Arquivos Nacionais do Reino Unido, os responsáveis pelo directório de formatos PRONOM[80].

A interface do Format Identifier disponibiliza dois métodos remotos que se distinguem apenas pelos argumentos que recebem (Figura 34). Um, recebe um RepresentationObject, i.e., um conjunto de ficheiros compostos por sequências binárias que definem uma representação digital (e.g. uma página Web constituída por um ficheiro HTML e várias imagens em formato JPEG); o outro recebe apenas uma sequência de bits, facilitando assim a transmissão de representações constituídas apenas por um ficheiro.

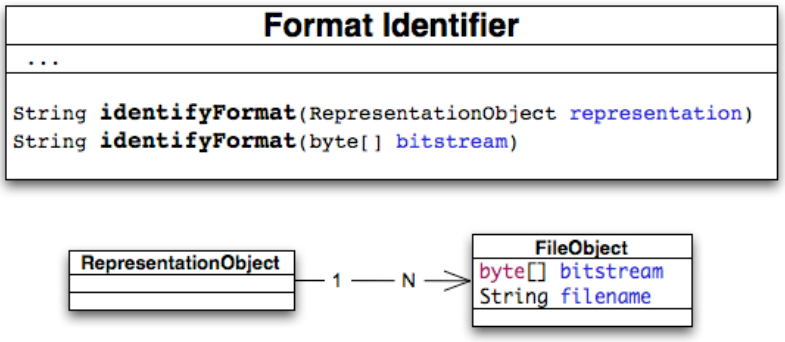


Figura 34 – Métodos disponibilizados pelo Format Identifier.

A designação de formato devolvida pelo método identifyFormat segue a seguinte estrutura:

designação de formato [, version **versão do formato**]

A parte referente à versão do formato é opcional, sendo apenas incluída quando a versão do mesmo é positivamente identificada. Seguem-se alguns exemplos de designações de formato produzidas por este componente:

- Tagged Image File Format, version 3

[79] Ver secção 3.4.1, na página 58.
[80] Ver secção 2.4, na página 49.

- JPEG File Interchange Format, version 1.02
- Microsoft Word for Windows Document, version 97-2003
- Graphics Interchange Format, version 1989a
- JPEG 2000

É importante referir que estas designações de formato são utilizadas no preenchimento dos atributos `sourceFormat` e `targetFormat` dos descritores de serviços de migração armazenados no `Service Registry`[81], garantindo deste modo a interoperabilidade terminológica entre os diversos componentes do sistema.

4.5 Migration Broker

Sempre que é solicitada uma migração ao CRiB, cabe ao componente `Migration Broker` invocar os serviços de conversão necessários para realizar a respectiva migração. Este componente tem como responsabilidade compor os serviços de conversão requeridos e coordenar todo o processo de forma a torná-lo transparente para o utilizador. Na prática, o `Migration Broker` garante que todo o processo de migração é executado de forma atómica do ponto de vista dos restantes componentes do sistema, independentemente do número de serviços que forem necessários para a concretizar (Figura 35).

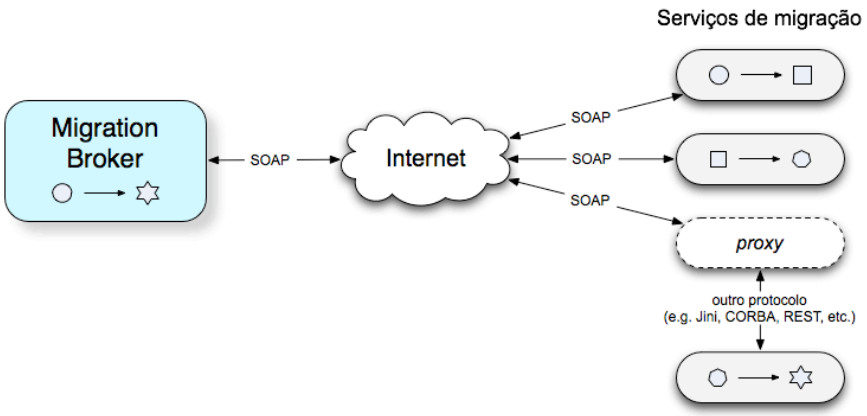


Figura 35 – Arquitectura detalhada do Migration Broker.

A interface apresentada pelo `Migration Broker` encontra-se ilustrada na Figura 36. Este componente disponibiliza dois métodos fundamentais: o método `convert`, que transforma uma representação num novo formato, recorrendo, se necessário, a uma sequência de serviços

[81] Para mais informação sobre o componente Service Registry, consulte secção 4.3 na página 83.

de migração (i.e., MigrationPath); e, tratando-se este de um componente capaz de realizar avaliações quanto ao desempenho de um caminho de migração, um método designado getEvaluationCriteria que permite ao cliente conhecer os critérios de avaliação implementados por este componente.

É importante referir que todos os componentes do CRiB dotados de capacidades de avaliação implementam a interface Evaluator (Figura 36). Esta interface garante aos actores externos a capacidade de conhecer os critérios de avaliação suportados pelo componente para um dado formato ou classe de objectos digitais.

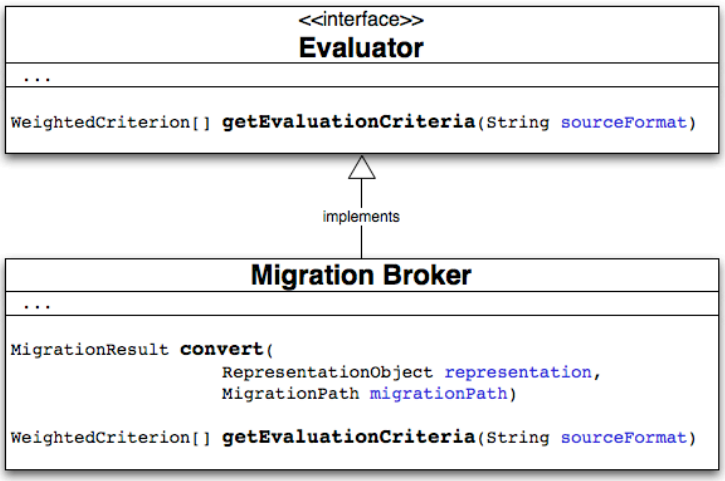


Figura 36 – Métodos disponibilizados pelo Migration Broker.

As mensagens trocadas entre uma aplicação-cliente e o Migration Broker encontram-se representadas na Figura 37.

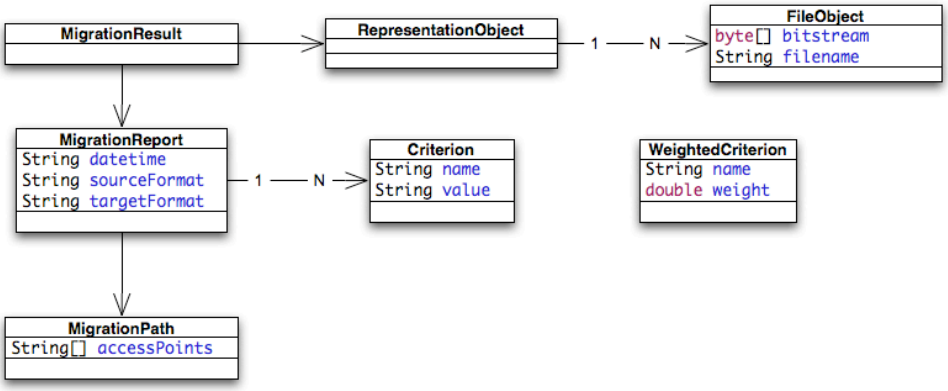


Figura 37 – Mensagens trocadas pelo Migration Broker.

Para que um serviço de migração possa ser utilizado pelo Migration Broker e, consequentemente, pelos clientes do CRiB, este deverá respeitar uma interface predefinida. Esta interface define um método que todos os serviços de migração deverão implementar – o método convert. Este método recebe como parâmetro a representação que se pretende converter e tem como objectivo devolver uma representação desse objecto num novo formato. A interface genérica de um serviço de conversão e alguns exemplos de conversores encontram-se ilustrados na Figura 38.

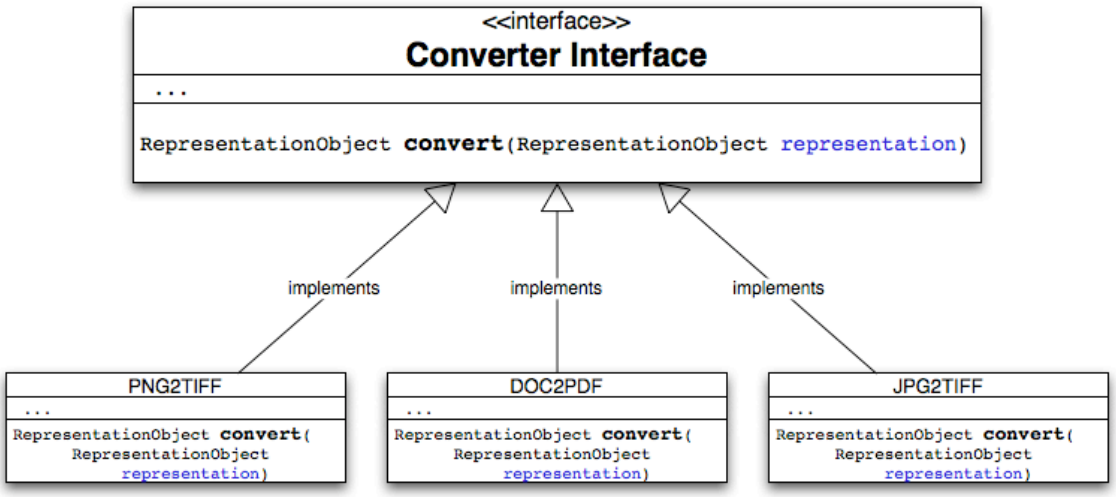


Figura 38 – Interface comum a todos os serviços de conversão.

É importante referir que um dado serviço de migração poderá encontrar-se fisicamente localizado em qualquer parte do globo. Para que possa ser integrado na rede de serviços de migração disponibilizada pelo CRiB, este apenas necessita de estar acessível através da Internet via XML/SOAP e de respeitar a interface definida. Serviços que não respeitem estas duas condições, quer porque se baseiam em protocolos diferentes, quer porque simplesmente respeitam interfaces incompatíveis, podem ser integrados recorrendo a *proxies*.

Um *proxy* é um componente de software que permite traduzir um protocolo num outro inicialmente incompatível. A Figura 35 apresenta um exemplo de um serviço de conversão que opera sobre um protocolo diferente do XML/SOAP e que foi integrado no CRiB por meio de um *proxy*. Durante o desenvolvimento do CRiB foram testados *proxies* que permitiam integrar os serviços de conversão criados no âmbito do projecto Typed Object Model[82] (TOM) com os restantes serviços de migração integrados no CRiB. O TOM utiliza um protocolo

[82] O projecto TOM encontra-se descrito na secção Directórios de formatos na página 34.

próprio para realizar as suas conversões compostas, no entanto, disponibiliza um conjunto de bibliotecas que permitem a qualquer programador tirar partido dos serviços que disponibiliza.

Para além de efectuar conversões compostas, o Migration Broker é responsável por avaliar a performance de cada caminho de conversão. As avaliações realizadas por este componente permitem determinar quais os caminhos de migração que oferecem melhor qualidade de serviço segundo múltiplos critérios, nomeadamente: disponibilidade, estabilidade, débito, custo de utilização, taxa de crescimento em bytes e taxa de crescimento em número de ficheiros (Ferreira et al., 2007). Cada um destes critérios é descrito em detalhe nas secções subsequentes.

4.5.1 Disponibilidade

A disponibilidade[83] é definida como a probabilidade de um serviço se encontrar acessível e operacional no momento em que é requisitado (Jiang & Schulzrinne, 2003).

A disponibilidade de um serviço de conversão é calculada dividindo o número de vezes que este foi invocado com sucesso, pelo número total de vezes que foi invocado (independentemente do sucesso da sua invocação) (Jiang & Schulzrinne, 2003; Zeng, Benatallah, Dumas, Kalagnanam, & Sheng, 2003) – Fórmula 1.

$$availability = \frac{\#\ successful\ invocations}{\#\ invocations}$$

Fórmula 1 – Disponibilidade.

Um serviço de conversão com baixa disponibilidade é um serviço que nem sempre está acessível no momento em que é necessário. Em processos de migração que envolvam várias centenas de representações, a indisponibilidade momentânea de um serviço de conversão poderá atrasar ou até mesmo inviabilizar todo o processo de migração.

4.5.2 Estabilidade

A estabilidade[84] é definida como a probabilidade de um serviço de conversão ser capaz de concluir com sucesso as tarefas a que se propõe. Por outras palavras, a estabilidade representa a capacidade de um serviço não falhar durante a sua execução (i.e., o seu nível de tolerância a

[83] Do inglês *availability*.
[84] Do Inglês *stability*. Zeng *et al.* designam este conceito por Confiabilidade (do inglês *Reliability*).

falhas). Esta, é calculada dividindo o número de conversões bem sucedidas pelo número total de conversões requisitadas (Zeng et al., 2003).

$$stability = \frac{\#\ successful\ conversions}{\#conversion\ requests}$$

Fórmula 2 – Estabilidade.

Este critério de avaliação é particularmente importante quando se efectua composição de serviços. O primeiro serviço da composição pode operar em perfeitas condições, mas um dos serviços intermédios poderá falhar recorrentemente. Isto significa que esse caminho de migração tem elevada disponibilidade mas uma estabilidade reduzida (Figura 39).

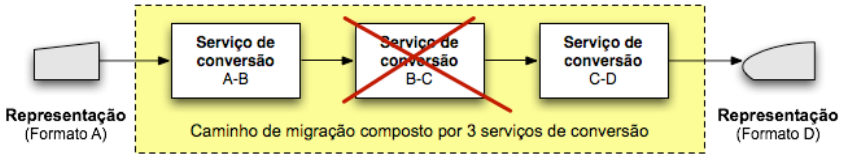


Figura 39 – Caminho de migração com baixa estabilidade.

4.5.3 Débito

O débito[85] mede a quantidade de trabalho que um serviço de conversão é capaz de realizar por unidade de tempo (Menascé, 2002) – Fórmula 3. A carga imposta a um conversor, i.e., o trabalho a realizar, é determinado pelo tamanho em bytes do objecto digital submetido a conversão. Trata-se obviamente de uma simplificação, uma vez que o tempo de conversão de um objecto digital não depende exclusivamente do seu comprimento em bytes. A complexidade do próprio objecto influencia significativamente o tempo necessário para a sua conversão. Não obstante, a simplificação introduzida constitui um ponto de partida considerado razoável.

$$throughput = \frac{object\ length}{migration\ time}$$

Fórmula 3 – Débito de conversão

[85] Do Inglês *throughput*.

Foi também efectuada uma simplificação no que diz respeito à medição do tempo de migração. A arquitectura proposta pelo CRiB impossibilita a medição individual do tempo de transmissão e do tempo efectivamente gasto em conversão. Esta limitação deve-se ao facto de este critério ser avaliado por um agente externo aos conversores utilizados, i.e., o Migration Broker (Zeng et al., 2003). Devido a esse facto, estes dois elementos temporais são considerados conjuntamente, ou seja, o tempo de migração é medido a partir do momento em que a representação é enviada para o primeiro conversor, até ao momento em que a nova representação é recebida por este componente (Figura 40).

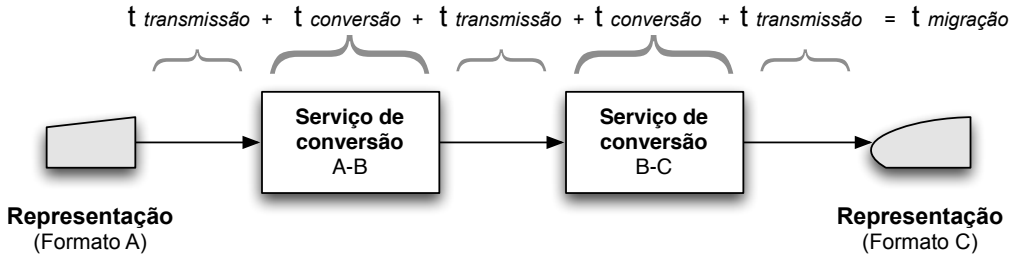


Figura 40 – Cálculo do tempo de migração.

4.5.4 Custo de utilização

O custo de utilização diz respeito ao valor, em unidades económicas, que uma organização terá que despender para tirar partido de um determinado serviço de conversão (Zeng et al., 2003). O custo é definido por um valor constante a cobrar por cada invocação de serviço.

O custo de uma conversão composta é calculado através do somatório dos vários custos individuais associados a cada serviço de conversão que compõe o caminho de migração (Figura 41).

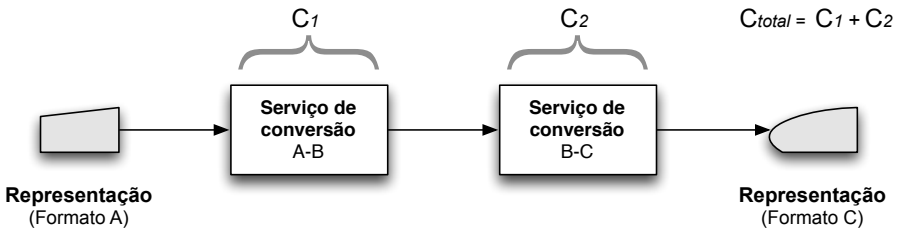


Figura 41 – Cálculo do custo de utilização de uma migração composta.

A introdução deste critério de avaliação tem como objectivo estimular o desenvolvimento de conversores, uma vez que estes poderão ser publicados e "vendidos" através da plataforma de serviços de preservação. Apesar do modelo económico apresentado ser demasiado simplista para que possa ser posto em prática de forma efectiva, este poderá constituir uma ferramenta importante no sentido de determinar em que medida o custo pode influenciar decisões tomadas a favor ou contra determinadas alternativas de migração.

4.5.5 Taxa de crescimento em bytes

Em cenários de preservação onde se manipulam grandes volumes de informação, o custo de armazenamento é uma variável de extrema importância. O custo de um sistema de armazenamento com algumas dezenas de Terabytes poderá facilmente ultrapassar a fasquia de um milhão de euros e este valor não inclui os custos de manutenção, electricidade, refrigeração, administração, etc.

A dimensão das representações digitais influencia directamente as necessidades de armazenamento de uma organização e, indirectamente, os custos de preservação. O Migration Broker faz uma análise contínua da razão existente entre a dimensão em bytes das representações submetidas a migração e a dimensão das representações que resultam dessa actividade.

Para um dado caminho de migração é possível determinar se as representações resultantes irão exigir mais ou menos espaço de armazenamento que as representações originais. Isto permite planear com maior detalhe uma intervenção de preservação e garantir que existe espaço de armazenamento suficiente para acomodar as novas representações.

A taxa de crescimento em bytes de uma dada representação submetida a conversão é calculada pela Fórmula 4.

$$outcome_length_ratio = \frac{source\ representation\ length}{target\ representation\ length}$$

Fórmula 4 – Taxa de crescimento em bytes de representações convertidas.

A fórmula apresentada valoriza a redução do tamanho das representações após a sua conversão, i.e., representações finais de menores dimensões produzem valores mais elevados deste critério de avaliação. Isto significa que valores superiores a 1 representam efectivamente uma redução do tamanho das representações após conversão.

4.5.6 Taxa de crescimento em número de ficheiros

Do mesmo modo que a dimensão das representações influencia o custo de armazenamento e preservação, o número de ficheiros que as constituem influencia directamente a capacidade da sua gestão. A decomposição de objectos digitais complexos nas suas partes constituintes é uma abordagem de preservação amplamente utilizada (Hunter & Choudhury, 2006), no entanto, quanto maior for o número de ficheiros associados a uma representação, maior será a dificuldade ao nível da gestão do armazenamento, descrição técnica dos seus constituintes e gestão dos relacionamentos existentes entre os diversos ficheiros.

O Migration Broker foi dotado de funcionalidades que permitem medir a taxa de crescimento de uma representação no que diz respeito ao número de ficheiros. A Fórmula 5 permite calcular essa taxa de crescimento.

$$outcome_number_files_ratio = \frac{source\ representation\ number\ of\ files}{target\ representation\ number\ of\ files}$$

Fórmula 5 – Taxa de crescimento em número de ficheiros.

Tal como acontece com a fórmula de cálculo da taxa de crescimento em bytes, este critério valoriza a redução dos seus valores. O critério foi invertido de modo a que taxas de crescimento inferiores a 1 pudessem ser consideradas mais benéficas para efeitos de preservação.

4.6 Object Evaluator

O componente Object Evaluator tem como missão determinar o nível de degradação infligido a uma representação digital durante um processo de migração. Este objectivo é alcançado, calculando as diferenças entre a representação submetida a migração e a representação que resulta da mesma.

Após uma migração, a representação resultante é comparada com a representação original à luz de um conjunto preestabelecido, mas extensível, de critérios de avaliação. Estes critérios, designados, neste contexto, por propriedades significativas, identificam o subconjunto de atributos que constituem a representação digital e para os quais existe um compromisso institucional no sentido da sua preservação. Estes atributos caracterizam a essência da representação digital e qualificam-na como uma entidade intelectual inteligível (ver secção sobre Autenticidade na página 37).

O processo de avaliação levado a cabo pelo Object Evaluator tem, sobretudo, dois objectivos:

- Documentar a intervenção de preservação através da produção de relatórios com informação detalhada sobre as propriedades significativas que não foram devidamente preservadas durante o processo de migração (i.e., controlo de qualidade);
- Alimentar o Evaluations Repository para suportar o auxílio à tomada de decisão quanto aos formatos e serviços de migração mais adequados para preservar uma dada colecção de objectos digitais (i.e., selecção de uma estratégia de migração).

Os relatórios de qualidade produzidos pelo Object Evaluator baseiam-se na entidade Eventos incluída no dicionário de dados PREMIS (Caplan et al., 2005; Guenther et al., 2008; PREMIS Working Group, 2005). Esta entidade semântica regista todas as acções desenvolvidas em torno de um objecto digital e é, em grande parte, responsável por assegurar a autenticidade dos materiais preservados. Nela são reunidos elementos descritivos como: tipo de evento (e.g. migração), data e hora de ocorrência, descrição detalhada da acção, informação sobre o sucesso da intervenção e informação sobre o agente de software responsável pela sua realização.

O segundo objectivo é alcançado, avaliando o desempenho médio de cada serviço de migração, particularmente no que diz respeito a perdas de informação. Os resultados dessas avaliações são enviados ao utilizador e, simultaneamente, armazenados num repositório de dados designado Evaluations Repository. Isto permite ao componente Migration Advisor identificar as alternativas de migração que melhor se adequam às necessidades da entidade-cliente (ver secção sobre o componente Migration Advisor na página 118).

É importante referir que o Object Evaluator não é o responsável pelo registo dos resultados no repositório de avaliações, mas sim o componente que o invoca. Na arquitectura apresentada, esse componente é o Core Preservation Services, o serviço que coordena todos os processos de preservação que ocorrem no interior do CRiB.

A arquitectura interna do Object Evaluator encontra-se caracterizada em detalhe na Figura 42. Após obter duas representações digitais (i.e., a representação original e uma versão convertida), o componente começa por extrair os valores das propriedades significativas incluídas em ambas as representações. Para tal, pressupõe-se a existência de um descodificador para cada formato suportado, i.e., um componente de software que auxilia no processo de extracção destes valores. Este componente permite transformar as sequências de bits que

constituem uma representação numa estrutura lógica onde os valores das suas propriedades podem ser facilmente inspeccionados de forma automática.

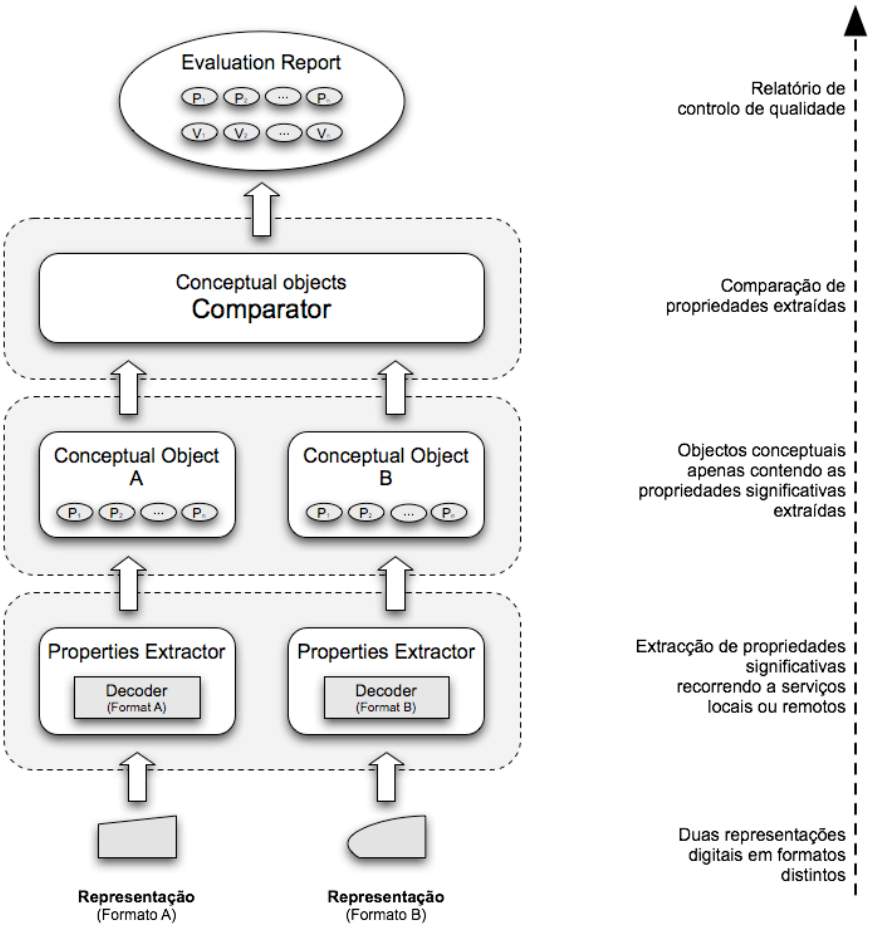


Figura 42 – Arquitectura detalhada do Object Evaluator.

É importante referir que o extractor de valores de propriedades significativas é, ele próprio, um serviço que poderá ser construído à custa de outros serviços acessíveis, local ou remotamente. Isto é particularmente importante quando existem formatos que são dependentes de uma plataforma tecnológica e apenas podem ser descodificados no contexto tecnológico correspondente. Por exemplo, o descodificador mais apto para extrair propriedades de documentos Word é o software Microsoft Word. Uma vez que esta aplicação não existe em ambientes Linux (i.e., o sistema operativo que suporta a generalidade dos serviços do CRiB), o extractor de propriedades para o formato Word teve de ser desenvolvido num ambiente Windows e, posteriormente, invocado pelo Object Evalutator.

Após a extracção dos valores das propriedades significativas, estes são guardados numa estrutura de dados neutra que facilita a sua manipulação. Essa estrutura trata-se, efectivamente, do objecto conceptual, i.e., aquele que carrega a semântica da representação digital mas que é desprovido de características técnicas específicas de um dado formato (ver secção A anatomia de um objecto digital na página 14).

Uma vez obtidos os objectos conceptuais, é possível calcular as diferenças entre duas instâncias e assim determinar o nível de degradação incorrido durante a migração. A comparação de objectos conceptuais é assegurada por subcomponentes comparadores específicos para cada classe de objectos digitais, i.e., `Comparator` (Figura 43).

O subcomponente responsável por comparar as propriedades que constituem os objectos conceptuais implementa uma função de similaridade para cada tipo de propriedade (Figura 43). Por exemplo, para determinar o nível de similaridade entre dois documentos de texto em termos do seu conteúdo textual é necessário recorrer a uma função de similaridade capaz de comparar cadeias de caracteres. Por outro lado, para determinar a similaridade gráfica entre duas imagens será necessária uma função capaz de comparar matrizes de cor.

Os resultados obtidos após a aplicação das funções de similaridade pertencem ao domínio [0, 1], com o valor 1 a representar o valor máximo de similaridade (i.e., igualdade), e 0 a máxima distância entre dois valores possíveis (i.e., a desigualdade máxima). O conjunto de valores produzidos pelas várias funções de similaridade irão fazer parte do relatório de controlo de qualidade (i.e., `Evaluation Report`). Este relatório determina o nível de similaridade existente entre duas representações digitais e identifica objectivamente os critérios analisados e os níveis de similaridade obtidos para cada um deles.

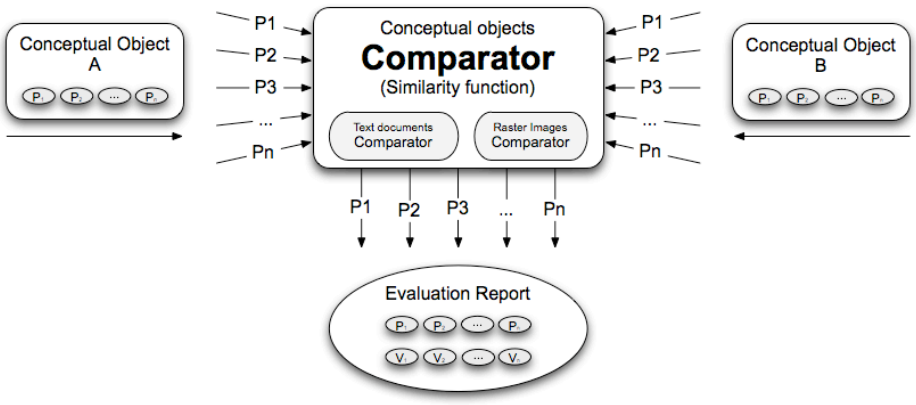


Figura 43 – Arquitectura detalhada do comparador de objectos conceptuais.

As propriedades significativas avaliadas durante o processo de controlo de qualidade dependem da classe de objectos digitais em análise. Por exemplo, objectos pertencentes à classe documentos de texto são avaliados à luz de propriedades como: número de páginas, apresentação gráfica do documento, dimensões de página, etc. (Tabela 6).

Critério	Descrição
appearance::static_page::pages::size	Dimensões da página
appearance::static_page::pages::layout	Organização vários elementos gráficos na página
appearance::static_page::pages::numbering	Número de páginas do documento
appearance::static_page::pages::headline	Cabeçalho das páginas
appearance::static_page::pages::footline	Rodapé das páginas
appearance::static_page::pages::break	A página quebra junto do mesmo texto
appearance::static_page::pages::margins	Tamanho das margens da página em milímetros
appearance::static_page::letters::size	Tamanho de letra
appearance::static_page::letters::special_characters	Apresentação e validade dos caracteres

Tabela 6 – Exemplo de uma taxionomia de avaliação de documentos de texto.

Contudo, objectos pertencentes à classe de objectos áudio são avaliados segundo um conjunto de critérios completamente distinto, tais como: resolução, volume médio, nível de ruído, duração, etc. (Tabela 7).

Critério	Descrição
appearance::audio::quality::resolution	Largura de banda em bits/amostra
appearance::audio::quality::drop_out	Pequenos momentos de silêncio no som
appearance::audio::quality::level	Volume do som
appearance::audio::quality::sample_rate	Frequência de amostragem
appearance::audio::quality::compression_rate	Grau de compressão do ficheiro de som
appearance::audio::funcionality::stereo	Se o som é mono ou estéreo
appearance::audio::funcionality::dolby_surround	Se o ficheiro suporta a tecnologia dolby surround
appearance::audio::funcionality::speed_variance	Descreve se há variações na velocidade reprodução do som

Tabela 7 – Exemplo de uma taxionomia de avaliação de objectos áudio.

No âmbito deste trabalho foram definidas taxionomias de avaliação para as classes documentos de texto e imagens matriciais. Para efeitos de prova de conceito apenas foram realizadas experiências em torno da classe imagens matriciais.

Os métodos disponibilizados pelo Object Evaluator encontram-se caracterizados na Figura 44. O componente disponibiliza dois métodos fundamentais, o método compare, que dadas duas representações e respectivos formatos é capaz de determinar o nível de

similaridade entre ambas as representações, e o método `getEvaluationCriteria`[86], que devolve os critérios que o componente é capaz de analisar para uma dada classe de objectos.

O método `compare` utiliza a informação sobre os formatos das representações para que se possa invocar os respectivos extractores de valores de propriedades.

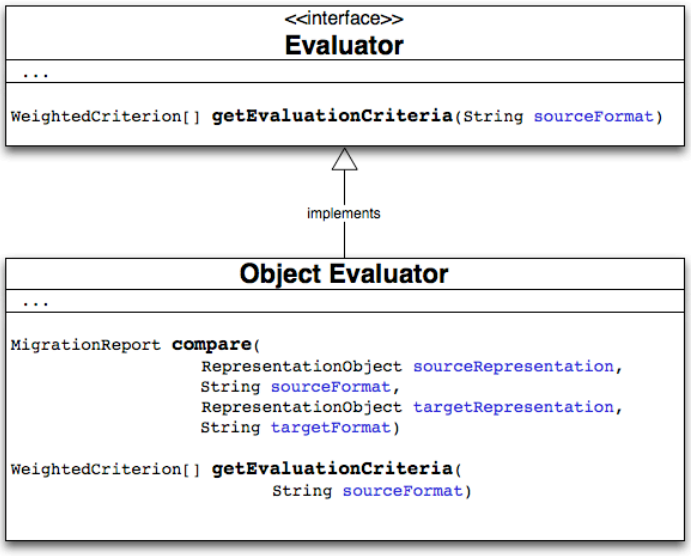


Figura 44 – Métodos disponibilizados pelo Object Evaluator.

O conjunto de mensagens suportadas por este componente encontra-se ilustrado na Figura 45. É de notar que o relatório de avaliação (i.e., `Migration Report`) produzido pelo `Object Evaluator` inclui a propriedade `MigrationPath`. No entanto, esta não é preenchida pelo mesmo, uma vez que o único componente que tem conhecimento do caminho de migração previamente executado é o `Migration Broker`. A propriedade existe porque todos os componentes avaliadores produzem relatórios com a mesma estrutura.

As secções que se seguem descrevem as classes de objectos suportadas pelo CRiB, os critérios de avaliação que lhes são subjacentes, os extractores de valores de propriedades e as funções de similaridade associadas.

[86] Tratando-se de um serviço que realiza avaliações no contexto da plataforma CRiB, este implementa a interface Evaluator.

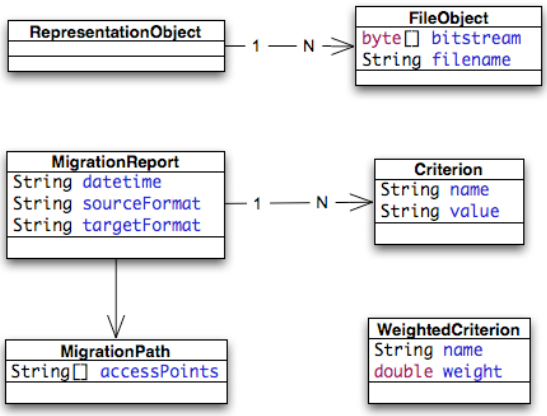


Figura 45 – Mensagens trocadas pelo Object Evaluator.

4.6.1 Classes de objectos

O CRiB oferece suporte para duas classes de objectos distintas: documentos de texto e imagens matriciais. A Tabela 8 enumera os formatos suportados pelo CRiB para cada uma destas classes.

Classe de objectos	Formatos associados
Imagens matriciais	JPEG File Interchange Format, version 1.00 JPEG File Interchange Format, version 1.01 JPEG File Interchange Format, version 1.02 Tagged Image File Format, version 3 Portable Network Graphics, version 1.0 Portable Network Graphics, version 1.1 Graphics Interchange Format, version 1989a Graphics Interchange Format, version 1987a Raw JPEG Stream Windows Bitmap, version 3.0 Exchangeable Image File Format (Compressed), version 2.1 Exchangeable Image File Format (Compressed), version 2.2 Exchangeable Image File Format (Uncompressed), version 2.1 Exchangeable Image File Format (Uncompressed), version 2.2 JPEG 2000
Documentos de texto	Microsoft Word for Windows Document, version 97-2003 Portable Document Format, version 1.4 Rich Text Format, version 1.0 Rich Text Format, version 1.4 Rich Text Format, version 1.6 Rich Text Format, version 1.7 OpenDocument Text Format, version 1.0

Tabela 8 – Formatos suportados pelo CRiB.

Em termos arquitecturais, o CRiB está preparado para suportar um número arbitrário de classes e formatos. No entanto, as restrições temporais a que um projecto de doutoramento está sujeito levaram a que apenas fosse possível integrar um número limitado de classes e

formatos. As razões que levaram à escolha das classes imagens matriciais e documentos de texto encontram-se resumidas de seguida.

Imagens matriciais

Várias instituições, especialmente as de cariz cultural, como os arquivos e as bibliotecas, recorrem frequentemente à transferência de suporte como forma de preservar os seus materiais analógicos. A preservação destes materiais é assegurada, limitando o seu manuseamento pelo público em geral, fornecendo como alternativa uma representação do mesmo num outro formato ou suporte.

Um dos suportes mais utilizados neste tipo de contextos é o microfilme. Porém, a digitalização tem vindo a afirmar-se como uma tecnologia com vantagens acrescidas ao nível da facilidade de reprodução e disseminação. Neste contexto, os esforços de preservação deixam de estar centrados unicamente no material analógico, passando também a estar focados na preservação dos seus equivalentes digitais.

O Arquivo Distrital do Porto[87], por exemplo, disponibiliza aos seus utentes um serviço de digitalização a-pedido de todos os itens incluídos no seu acervo (Ferreira, 2006b; Ferreira & Ramalho, 2004a, 2004b, 2004c; Ramalho, Ferreira, Ferros, Lima, & Sousa, 2006). As reproduções digitais requisitadas são descritas e arquivadas, recorrendo a um sistema de Gestão de Objectos Digitais desenvolvido especificamente com essa finalidade (Ramalho et al., 2006). Este sistema é também responsável por colocar em linha versões de baixa resolução dessas reproduções, permitindo ao utente pré-visualizar e, posteriormente, adquirir as mesmas através de um balcão electrónico também disponível através do portal do Arquivo (Sousa, Ferros, Ramalho, & Lima, 2007). A preservação dessas reproduções é, para o Arquivo Distrital do Porto, uma actividade crítica no suporte ao seu negócio.

Outro fenómeno relevante é o aparecimento de câmaras fotográficas digitais. Estas foram colocadas no mercado em 1990 pela Logitech e desde então a sua adopção não tem parado de aumentar (Wikipedia contributors, 2007). Gigantes como a Nikon já anunciaram o abandono progressivo da produção de câmaras fotográficas analógicas e o alinhamento dos seus planos de marketing na promoção dos seus produtos digitais (Musgrove, 2006).

Os utilizadores deste tipo de equipamentos, quer sejam amadores ou profissionais, são responsáveis pela produção e armazenamento de uma grande quantidade de imagens em

[87] http://www.adporto.pt

formatos digitais[88]. Exemplo disso são os vários acervos de imagens existentes na Web, como por exemplo, o Flickr[89], o Picasa Web Albums[90] ou o Kodak Gallery[91], dedicados fundamentalmente à publicação de fotografias por parte de um público amador. Não obstante, existe também um grande número de sítios Web dedicados à publicação e venda de imagens de cariz profissional. Exemplos disso são os serviços de venda de imagens Shutterstock[92], Dreamstime[93], Stockxpert[94], 123RF[95] e iStockPhoto[96].

Para reforçar um pouco mais a importância deste tipo de objectos digitais, uma consulta aos perfis de preservação publicados pelo Registry of Open Access Repositories[97] (ROAR) permite concluir que, logo após aos documentos de texto, as imagens (em formato JPEG e TIFF) são as classes de objectos mais prevalecentes nos repositórios institucionais[98] actualmente implementados (University of Southampton, 2007).

Houve, portanto, duas razões fundamentais que conduziram à escolha desta classe de objectos para integração na plataforma CRiB. A primeira, teve que ver com a elevada ubiquidade deste tipo de material. Uma estratégia de preservação deve preocupar-se em primeiro lugar com os materiais mais prevalecentes (isto, à falta de métrica mais eficaz na identificação de prioridades relativamente a que objectos preservar). A segunda razão, teve que ver com a simplicidade do ponto de vista técnico inerente ao processamento deste tipo de objectos. Optou-se por encetar o desenvolvimento da plataforma de serviços com uma classe de objectos sobre a qual houvesse documentação suficiente e ferramentas disponíveis capazes de os processar eficazmente.

Documentos de texto

Uma análise aos perfis de preservação publicados pelo projecto ROAR permitiu concluir que os documentos de texto são claramente a classe de objectos digitais mais prevalecente

---

[88] Apesar de, na sua grande maioria, as câmaras digitais guardarem fotografias em formato JPEG, é muito comum, especialmente em contextos profissionais, a gravação de imagens em formatos RAW que são diferentes consoante o fabricante.

[89] http://www.flickr.com

[90] http://picasaweb.google.com

[91] http://www.kodakgallery.com

[92] http://www.shutterstock.com

[93] http://www.dreamstime.com

[94] http://www.stockxpert.com

[95] http://www.123rf.com

[96] http://www.istockphoto.com

[97] http://roar.eprints.org/

[98] Em Janeiro de 2008 haviam sido incluídos nesta estatística 968 repositórios.

nos repositórios digitais actualmente existentes (University of Southampton, 2007). Este tipo de repositórios é responsável por arquivar e preservar todo o tipo de material que seja produto intelectual de uma dada organização (Sarmento, Baptista, & Ramos, 2005). A grande maioria destes repositórios é mantida por organizações de carácter académico, como universidades ou centros de investigação e neles podemos encontrar documentos diversos como artigos, monografias, relatórios técnicos, teses, dissertações, entre outros (Ferreira et al., 2008).

Retornando ao argumento da prevalência e ubiquidade, os documentos de texto foram eleitos como a segunda classe de objectos a considerar durante o desenvolvimento do CRiB. Os documentos de texto acarretam complexidade adicional na medida em que são compostos por texto, imagens, formatação, disposição gráfica, entre outros; tornando-os consideravelmente mais difíceis de preservar.

4.6.2 Taxionomias de avaliação

A criação de taxionomias de avaliação de objectos digitais, i.e., conjuntos de propriedades significativas que têm como objectivo avaliar a qualidade de uma migração ou, por outras palavras, determinar o nível de degradação incorrido durante uma migração, não é uma tarefa simples de concretizar (Ferreira & Baptista, 2005). Rauch e Rauber têm vindo a desenvolver esforços no sentido de reunir conjuntos de critérios de avaliação em torno de várias classes de objectos digitais. Este processo compreende a organização de pequenos eventos, semelhantes a *workshops,* onde especialistas de diversas áreas analisam conjuntos representativos de objectos digitais, codificados em diferentes formatos, com o propósito de identificar um conjunto comum de propriedades com relevância num contexto de preservação digital (Rauch, 2004; Rauch, Krottmaier, & Tochtermann, 2007; Rauch, Pavuza et al., 2005; Rauch & Rauber, 2004; Rauch, Rauber et al., 2005). O Arts and Humanities Data Service (AHDS) e a Biblioteca do Congresso têm vindo a publicar relatórios com informação técnica sobre diferentes classes de objectos digitais e onde se pode encontrar um número assinalável de propriedades consideradas relevantes num contexto de avaliação de estratégias de preservação (Arts and Humanities Data Service, 2006; Library of Congress, 2004b).

As taxionomias de avaliação suportadas pelo Object Evaluator foram construídas tomando como base a bibliografia existente. De todos os critérios identificados foram seleccionados apenas aqueles que seriam passíveis de ser extraídos e avaliados automaticamente por componentes de software. Todos os critérios que careciam de intervenção humana para que pudessem ser avaliados foram excluídos durante o desenvolvimento deste componente.

Outras fontes de informação relevantes para a construção destas taxionomias de avaliação foram: o documento "Assessing the Durability of Formats in a Digital Preservation Environment" (Stanescu, 2004) e a Wikipedia[99] onde se pode encontrar uma quantidade assinalável de informação técnica sobre formatos e as várias aplicações de software que os suportam.

Imagens matriciais

A taxionomia de avaliação utilizada pelo Object Evaluator para determinar o nível de degradação incorrido durante a migração de um objecto pertencente à classe imagens matriciais encontra-se ilustrada na Figura 46.

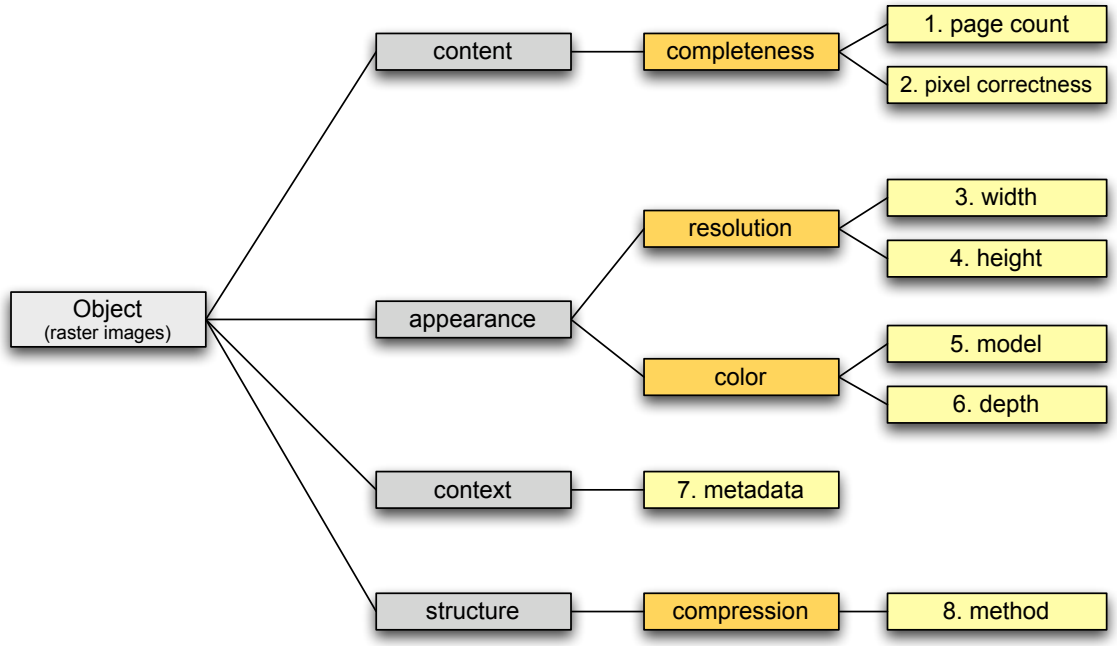


Figura 46 – Taxionomia de avaliação de imagens matriciais.

[99] http://www.wikipedia.org

As propriedades presentes na taxionomia previamente apresentada encontram-se descritas em detalhe na Tabela 9.

ID	Critério de avaliação	Descrição
1	Número de páginas	Avalia se o número de páginas que constituem uma imagem não foi alterado durante a migração.
2	Conformidade gráfica	Nível de similaridade existente entre os *píxeis* que compõem uma dada imagem e os *píxeis* de uma outra, vulgarmente designada por original. No caso de imagens compostas por múltiplas páginas, a comparação é realizada página-a-página e um valor de similaridade global é calculado.
3	Largura	Critério que determina se a largura da imagem em *píxeis* foi preservada.
4	Altura	Critério que determina se a altura da imagem em *píxeis* foi preservada.
5	Modelo de cor	O modelo de cor (*color model*) trata-se de um modelo matemático que descreve a forma como as cores são codificadas num dado formato de imagem (e.g. RGB, sRGB, HSL, HSV, YUV, CMYK). Este critério procura determinar se o modelo de cor foi preservado durante a migração.
6	Profundidade de cor	A profundidade de bits (*color depth*) determina o número de bits utilizado para representar a cor de um *pixel* (Wikipedia contributors, 2006a). Este critério determina se a profundidade de bits de uma imagem foi modificada pelo processo de migração.
7	Metainformação embebida	Vários formatos de imagem suportam metainformação embebida. Este critério avalia se numa migração a metainformação de uma imagem foi preservada.
8	Método de compressão	Certos formatos de imagem suportam compressão. Esta pode introduzir perdas de informação (*lossy compression*) ou preservarem todos os pormenores da imagem original (*lossless compression*). Alguns exemplos de compressão sem perdas são: Run-length encoding e LZW. Exemplos de algoritmos de compressão que introduzem perdas são: redução de cores; Chroma subsampling e Fractal compression (Wikipedia contributors, 2006b). Este critério procura determinar se o método de compressão se manteve inalterado durante o processo de migração.

Tabela 9 – Propriedades associadas a imagens matriciais.

Documentos de texto

A Figura 47 apresenta a taxionomia de avaliação utilizada pelo Object Evaluator para determinar o nível de degradação incorrido durante a conversão de documentos de texto.

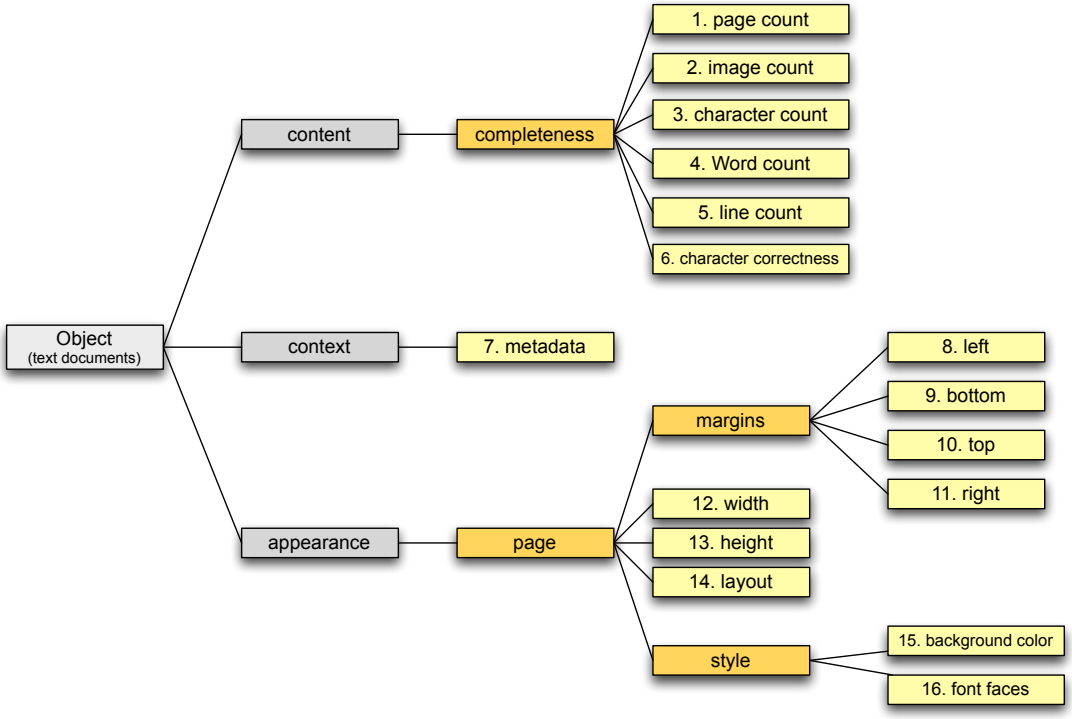


Figura 47 – Taxionomia de avaliação de documentos de texto.

As propriedades apresentadas na taxionomia previamente apresentada encontram-se descritas em detalhe na Tabela 10.

ID	Critério de avaliação	Descrição
1	Número de páginas	Este critério avalia se o número de páginas do documento foi preservado.
2	Número de imagens	Critério que determina se o número de imagens contidas numa imagem foi preservado durante a migração.
3	Número de caracteres	Este critério avalia se o número de caracteres que constitui o documento foi preservado.
4	Número de palavras	Este critério avalia se o número de palavras que constitui o documento foi preservado.
5	Número de linhas	Este critério avalia se o número de linhas e parágrafos que constitui o documento foi preservado.
6	Conformidade dos caracteres	Determina o nível de conformidade entre os caracteres existentes no documento convertido em relação aos caracteres do documento original. Este critério é responsável pode determinar se ocorreu degradação do texto que constitui o documento.
7	Metainformação embebida	Certos documentos de texto carregam consigo metainformação. Este critério procura determinar se essa metainformação foi devidamente preservada durante uma migração de formatos.
8	Margem esquerda	Determina se as dimensões da margem esquerda do documento foram preservadas. Este critério é calculado página-a-página e um valor de similaridade global é obtido através da média dos valores parciais.
9	Margem inferior	Determina se as dimensões da margem inferior do documento foram preservadas. Este critério é calculado página-a-página e um valor de similaridade global é obtido através da média dos valores parciais.

10	Margem superior	Determina se as dimensões da margem superior do documento foram preservadas. Este critério é calculado página-a-página e um valor de similaridade global é obtido através da média dos valores parciais.
11	Margem direita	Determina se as dimensões da margem direita do documento foram preservadas. Este critério é calculado página-a-página e um valor de similaridade global é obtido através da média dos valores parciais.
12	Largura de página	Determina se a largura do documento em milímetros foi preservada. Este critério é calculado página-a-página e um valor de similaridade global é obtido calculando a média dos valores parciais.
13	Altura de página	Determina se a altura do documento em milímetros foi preservada. Este critério é calculado página-a-página e um valor de similaridade global é obtido calculando a média dos valores parciais.
14	Conformidade gráfica	Determina se a disposição gráfica dos elementos em cada página foi devidamente preservada durante a conversão.
15	Cor de fundo	Determina se a cor de fundo do documento foi preservada.
16	Tipos de letra	Verifica se a colecção de tipos de letra utilizada no documento convertido é igual à colecção usada no documento original.

Tabela 10 – Propriedades associadas a documentos de texto.

4.6.3 Extractores de valores de propriedades

O Object Evaluator é acompanhado de um conjunto de subcomponentes capazes de extrair os valores de propriedades significativas de objectos digitais. Estes subcomponentes podem facilmente ser estendidos para suportar novas propriedades e/ou formatos.

Cada um dos valores das propriedades anteriormente descritas é extraído de uma representação digital, recorrendo a um subcomponente deste tipo. Estes designam-se genericamente por Property Extractors.

Os Property Extractors fazem uso de bibliotecas e ferramentas externas que permitem descodificar os formatos suportados e obter os valores das propriedades que os constituem. Para mais informações sobre os extractores de valores de propriedades implementados pelo CRiB consulte-se o Apêndice 8.1 na página 197.

4.6.4 Funções de similaridade

Os valores associados a uma dada propriedade significativa são caracterizados por um tipo de dados e por uma interpretação do que significa preservar essa propriedade. Para determinar se uma propriedade se manteve inalterada ao longo do tempo é necessário comparar os valores extraídos da representação original com os valores da representação convertida. Para comparar estes valores recorre-se a funções de similaridade.

Cada propriedade significativa deve ser comparada através de uma função de similaridade específica. Por exemplo, para determinar se o comprimento em bytes de uma representação se

manteve inalterado durante um processo de migração não é suficiente verificar se o seu comprimento é igual ao comprimento da representação convertida. É fundamental utilizar uma métrica que respeite as relações de proporcionalidade entre ambos os valores e que, ao mesmo tempo, tenha em consideração a dimensão das suas grandezas. Uma representação que tenha passado de 100 Kilobytes para 150 Kilobytes foi alvo de um aumento de 50%. No entanto, uma representação que tenha crescido de 100 Megabytes para 120 Megabytes sofreu apenas um aumento de 20%. Não obstante, no primeiro caso a diferença absoluta foi de 50 Kilobytes, enquanto que no segundo foi de 20 Megabytes, um valor cerca de 410 vezes superior ao do primeiro exemplo.

Todas as funções de similaridade utilizadas durante o desenvolvimento do CRiB encontram-se descritas em detalhe no Apêndice 8.3 na página 203.

4.7 Format Evaluator

O componente Format Evaluator tem como missão fornecer informação técnica sobre os formatos digitais suportados pela plataforma CRiB. Esta informação permite ao componente Migration Advisor determinar quais os formatos que apresentam o conjunto de características mais favorável para preservar uma dada classe de objectos digitais.

Por exemplo, considere-se uma representação codificada num formato que requer o pagamento *royalties* aquando da sua produção e/ou utilização. Agora, imagine-se um formato para o qual esta representação poderia ser convertida, livre deste tipo de encargos. A realização dessa migração traria benefícios significativos no que diz respeito aos custos de preservação desta representação. Formatos que requerem o pagamento de *royalties* são geralmente maus candidatos a formatos de preservação devido aos custos inerentes à sua utilização. Estes custos poderão tornar-se incomportáveis a longo-prazo.

Se por outro lado se equacionar uma migração de um formato não-comprimido para um formato comprimido, sendo que o novo formato é baseado em algoritmos de compressão com perdas (e.g. JPEG), então poder-se-ia assumir que se estaria a diminuir a capacidade de preservar adequadamente a respectiva representação. O uso de algoritmos de compressão com perdas são contra-indicados em contextos de preservação, pois degradam irremediavelmente os objectos digitais e torna-os mais vulneráveis a corrupção involuntária (i.e., a modificação de um único bit implica geralmente danificação da totalidade do objecto). A assumpção de que a realização desta migração iria diminuir as capacidades de preservação do objecto digital pode ser efectuada sem que haja necessidade de consumar a respectiva conversão. As características inerentes aos formatos envolvidos na migração são suficientes para se aferir se a realização da

respectiva conversão iria trazer benefícios ou prejuízos no que toca à capacidade de preservar objectos digitais a longo-prazo.

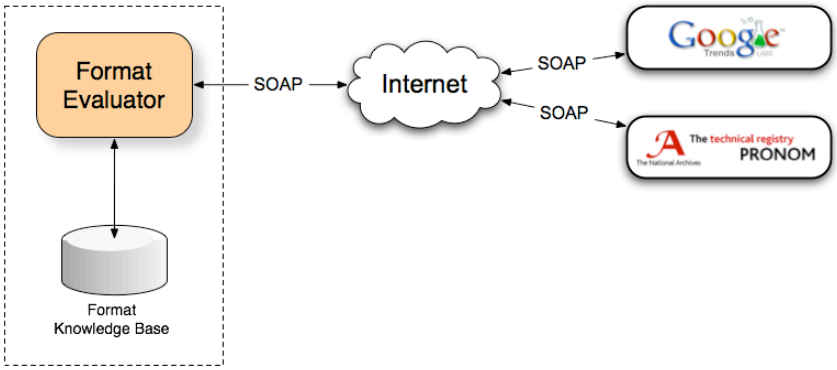


Figura 48 – Arquitectura do Format Evaluator.

O componente Format Evaluator é suportado por uma base de dados designada Format Knowledge Base (Figura 48). Esta base de dados é baseada em XML e pode ser livremente consultada no sítio Web do projecto[100].

De futuro, outras fontes de informação poderão ser integradas, permitindo ao Format Evaluator obter informação continuamente actualizada sobre o estado de cada formato digital. Por exemplo, o serviço Trends da Google (Google, 2006) permite determinar a popularidade de um termo de pesquisa ao longo do tempo. Recorrendo a este serviço é possível calcular a popularidade instantânea de um dado formato e qual a sua tendência ao longo do tempo. Formatos com uma tendência negativa de popularidade são menos desejáveis num contexto de preservação, pois teme-se que num futuro próximo haja necessidade de migrar os objectos codificados nesse formato para um outro mais reconhecido pela sua comunidade de interesse.

Outro serviço externo que poderá vir a tornar-se compatível com o Format Evaluator é o PRONOM Technical Registry (ver Directórios de formatos na página 34). Este serviço reúne um conjunto alargado de informação técnica sobre formatos digitais. No entanto, para poder ser utilizado pelo Format Evaluator é fundamental que essa informação possa ser consultada através de uma interface remota, por exemplo, via XML/SOAP, algo que não acontece actualmente.

[100] http://crib.dsi.uminho.pt

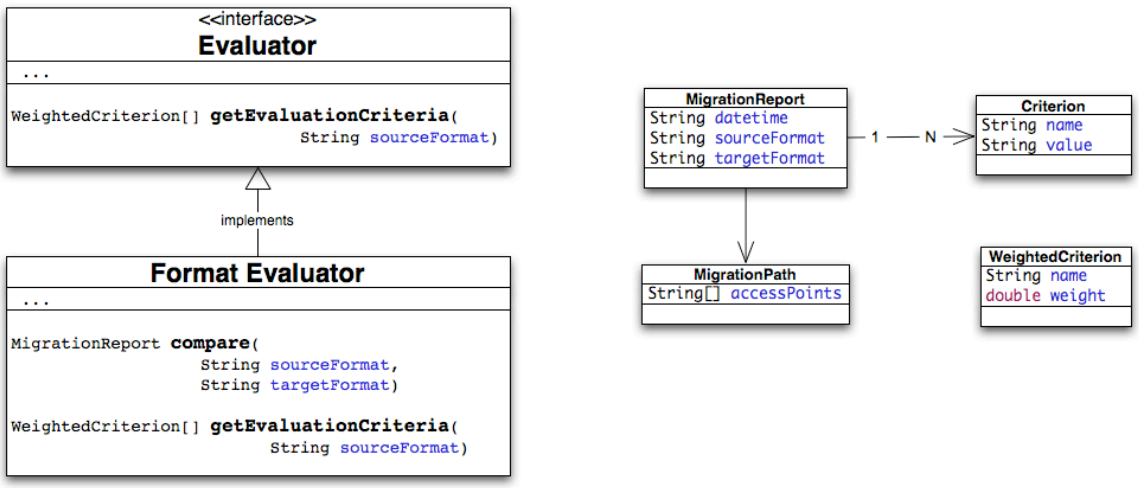


Figura 49 – Diagrama de classes associadas ao Format Evaluator.

A Figura 49 apresenta o diagrama de classes associadas ao Format Evaluator, bem como todas as mensagens trocadas por este componente. Tal como acontecia com os restantes componentes avaliadores, existe um método, designado getEvaluationCriteria, que permite ao cliente conhecer os vários critérios de avaliação suportados pelo serviço.

O componente disponibiliza ainda um método designado compare que realiza efectivamente o trabalho de avaliação de formatos. Ao contrário do método com o mesmo nome incluído no Object Evaluator, este apenas recebe como parâmetro as designações dos dois formatos que se pretendem comparar. O resultado é uma lista de características técnicas com o nível de benefício que se obteria se se realizasse a respectiva conversão entre os dois formatos.

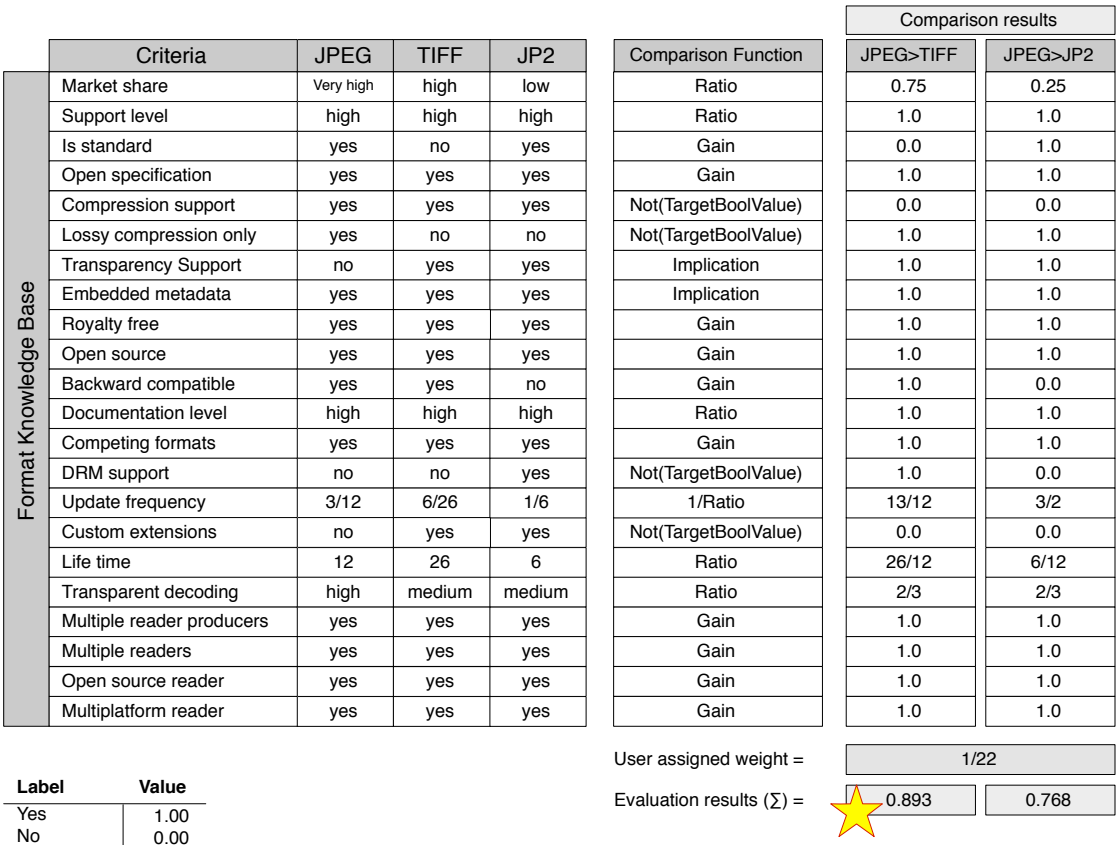
O benefício é determinado, aplicando funções de cálculo de benefício aos valores apresentados pelas características de cada formato. Todas as características técnicas suportadas pelo Format Evaluator encontram-se descritas na Tabela 11.

Característica técnica	Descrição
Quota de mercado	Se o formato é amplamente aceite ou é simplesmente um formato de nicho. A quota de mercado também é conhecida como grau de "adopção". A "adopção" refere-se ao grau de utilização do formato por parte dos criadores primários, disseminadores e/ou utilizadores dos recursos de informação. Um elevado nível de adopção é considerado favorável para fins de preservação.
Nível de suporte técnico	O nível de suporte técnico dado pelo criador oficial do formato. Um elevado nível de suporte é preferível num contexto de preservação.
É uma norma	Se o formato foi publicado por uma organização oficial de normalização. Formatos normalizados são preferíveis aos não normalizados.
Especificação aberta	Se a especificação do formato pode ser inspeccionada/verificada de forma independente. O uso de formatos abertos é fortemente recomendado em

	contextos de preservação.
Suporta compressão	Se o formato suporta qualquer tipo de compressão. Formatos não comprimidos são geralmente preferidos pela comunidade dedicada à preservação digital.
Apenas suporta compressão com perdas de informação	Se o formato suporta exclusivamente um tipo de compressão que provoca perda de informação ou deterioração do objecto original. Os esquemas de compressão com perda são grandemente desaconselhados.
Suporta transparência	Se o formato oferece funcionalidades de transparência. Este critério é específico de determinado tipo de formatos (p. ex. imagens de mapa de bits). Se o formato de origem contém funcionalidades de transparência, o formato de destino deve ter também suporte para essa propriedade.
Metainformação embebida	Se o formato contem metainformação embebida. O formato de destino deve ter capacidade de incluir/acomodar a metainformação embebida do formato de partida.
Royalties (taxas de utilização)	Se a utilização ou produção do formato requer o pagamento de *royalties* ou taxas de utilização. Existe preferência por formatos livres de *royalties*.
Código-aberto	Se existem aplicações cujo código pode ser inspeccionado/verificado de forma independente. A existência de aplicações de código aberto é amplamente recomendada.
Retro-compatível	Se as revisões aos formatos incluem suporte para as versões anteriores. A retrocompatibilidade é uma característica desejável.
Nível de documentação	Se as especificações do formato estão bem documentadas. O sistema favorece a existência de formatos bem documentados.
Existem formatos concorrentes	Se existem formatos concorrentes ou similares. A existência de formatos concorrentes torna um formato mais atractivo para preservação, uma vez que a informação poderá ser mais facilmente convertida.
Implementa DRM	Se é possível a utilização de Gestão de Direitos Digitais (DRM), encriptação ou assinaturas digitais. Desaconselha-se a existência de qualquer tipo de funcionalidade que possa constituir obstáculo no acesso à informação.
Frequência de actualização	Qual a frequência de revisão de um formato desde a sua publicação inicial. Este critério é definido de acordo com a seguinte fórmula: número de revisões / (ano actual – ano de disponibilização). Os formatos estáveis são preferenciais. Se a frequência de revisões é muito grande, o arquivo poderá ter dificuldade em acompanhar o ritmo das mesmas.
Permite extensões marginais	Se o formato permite a inclusão de extensões, tais como secções executáveis ou características marginalmente suportadas. Desaconselha-se a utilização de formatos que suportam tais funcionalidades.
Idade	Quantos anos passaram desde que o formato foi disponibilizado oficialmente. Os formatos de longa duração têm geralmente preferência sobre formatos novos e pouco estabelecidos.
Interpretação/descodificação transparente	Complexidade inerente à codificação: legibilidade por parte de um ser humano recorrendo a um editor do texto simples. Têm preferência os formatos que podem ser facilmente inspeccionados e/ou interpretados.
Vários produtores de aplicações de leitura	Se existem várias entidades que produzem leitores/visualizadores. Para finalidades de preservação não se deve apostar em leitores produzidos somente por uma única entidade.
Várias aplicações de leitura	Se o formato pode ser lido/interpretado por diversas aplicações informáticas. Para finalidades da preservação não se deve apostar em formatos que apenas podem ser lidos/visualizados por uma aplicação específica.
Aplicações de leitura em código-aberto	Se o código fonte da aplicação de leitura pode ser inspeccionada/verificada de forma independente. A existência de leitores/visualizadores em código aberto é uma característica altamente desejável.
Existem leitores/visualizadores para várias plataformas	Se a aplicação de leitura/visualização pode ser executada ou tem versões para várias outras plataformas (por exemplo, sistemas operativos ou hardware). A existência de aplicações executáveis em plataformas concorrentes é uma característica altamente desejável num contexto de preservação.

Tabela 11 – Características técnicas avaliadas pelo Format Evaluator.

Considere-se, ainda, o seguinte exemplo. Uma instituição pretende preservar uma colecção de imagens codificadas em formato JPEG 1.02 que resultaram de um recente projecto de digitalização. A instituição deseja saber qual o formato mais adequado para garantir o acesso continuado a esses objectos. Ao mesmo tempo, pretende que o formato escolhido minimize o número de intervenções de preservação necessárias no futuro. Por outras palavras, a instituição pretende conhecer o formato de preservação mais adequado para suster as representações que perfazem a sua colecção. O componente Format Evaluator pode ser consultado para obter esta informação.



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 (∑) =	0.893	0.768

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

Figura 50 – Cálculo do benefício de migração.

O componente compara sempre dois formatos, o formato de partida e um potencial formato de destino e determina o benefício em termos de preservação que se obteria se se realizasse essa conversão. A Figura 50 ilustra como esse cálculo é realizado. Nela, pode encontrar-se uma

análise do benefício que se obteria ao converter as digitalizações do formato JPEG 1.02 para TIFF 6 e JPEG 2000, respectivamente. A figura apresenta, ainda, as características apresentadas por cada um destes formatos, obtidas a partir da Format Knowledge Base, as funções de cálculo de benefício utilizadas e o resultado final dessa avaliação. É importante referir que no exemplo apresentado foi atribuído o mesmo nível de importância a todas as características avaliadas.

Observando a figura é possível concluir que o formato TIFF 6 foi considerado mais benéfico do que o JPEG 2000 para preservar a colecção de objectos originalmente em formato JPEG 1.02. Isto deve-se, sobretudo, ao facto de o formato TIFF se apresentar como um formato mais prevalecente e maduro que o JPEG 2000, ou seja, apresentou maiores níveis de quota de mercado[101] e uma idade[102] substancialmente superior.

O Format Evaluator recorre a quatro funções de cálculo de benefício para comparar as características técnicas de dois formatos distintos: ganho de preservação[103], implicação[104], negação[105] e razão[106]. Cada uma destas funções encontra-se descrita ao longo das secções que se seguem.

4.7.1 Ganho de preservação

A função ganho de preservação, ou gain, procura quantificar o benefício em termos de capacidade de preservação que se obtém ao converter uma representação para um novo formato. O ganho de preservação é calculado de acordo com a Tabela 12.

$\text{Criterion}_{\text{source}}$	$\text{Criterion}_{\text{target}}$	Gain
0	0	**0.5**
0	1	**1**
1	0	**0**
1	1	**1**

Tabela 12 – Cálculo da função Gain.

Esta função valoriza a adopção de formatos que introduzam características favoráveis à preservação. Por exemplo, se um formato for dotado de uma especificação aberta (i.e., $\text{Open Specification}_{\text{source}} = 1$), algo considerado positivo num contexto de preservação, mas, no

[101] *Market share.*
[102] *Life time.*
[103] *Gain.*
[104] *Implication.*
[105] *Not.*
[106] *Ratio.*

entanto, o formato de destino não possuir esta característica (i.e., $Open\ Specification_{target} = 0$), então o resultado produzido por esta função será um valor pejorativo de 0. Se, por outro lado, ambos os formatos possuírem essa característica, uma potencial conversão entre estes não iria piorar a sua aptidão para preservar objectos digitais, ou seja, o ganho de preservação seria de 1. Todavia, por contraposição com o exemplo anterior, se um formato não possuir uma dada característica favorável à preservação e o formato de destino também não a possuir, então o valor de *ganho de preservação* será de 0.5, reforçando a ideia de que apesar de não se estar a perder uma característica técnica favorável, se esta tivesse sido introduzida pelo novo formato estar-se-ia a beneficiar mais em termos de preservação do objecto digital.

4.7.2 Implicação

A função *implicação*, ou *implication*, é bastante semelhante à anterior. No entanto, apenas desvaloriza conversões onde uma dada característica existente no formato de partida não é suportada no formato de destino. A tabela de verdade associada a esta função encontra-se definida na Tabela 13.

$Criterion_{source}$	$Criterion_{target}$	Implication
0	0	**1**
0	1	**1**
1	0	**0**
1	1	**1**

Tabela 13 – Cálculo da função Implication.

Exemplos de propriedades avaliadas através desta função são: *suporte para transparência* ou *metainformação embebida*. Neste contexto, o facto de um dado formato de destino não suportar uma dada característica não diminui nem aumenta a sua aptidão para preservar objectos digitais. O novo formato apenas seria desfavorável se o formato de partida fosse dotado dessa característica. Nesse caso, haveria propriedades da representação original que não seriam suportadas pelo formato de preservação. Todos os restantes casos não são considerados prejudiciais.

4.7.3 Negação

A *negação*, ou *not*, é uma função que apenas tem em consideração as características do formato de destino, i.e., é indiferente às propriedades apresentadas pelo formato de partida. O facto de um formato de destino possuir ou não determinada propriedade é suficiente para tirar

ilações quanto ao benefício introduzido pela sua utilização. A tabela de verdade associada a esta função encontra-se definida na Tabela 14.

$Criterion_{source}$	$Criterion_{target}$	Not
0	0	**1**
0	1	**0**
1	0	**1**
1	1	**0**

Tabela 14 – Cálculo da função Not.

Se um formato de destino, por exemplo, suportar exclusivamente compressão com perdas, implementar DRM ou permitir extensões não normalizadas, não é necessário analisar as características do formato de partida para concluir que o formato de destino possui características que são consideradas desfavoráveis para retenção a longo-prazo.

4.7.4 Razão

A função `razão`, ou `ratio`, distingue-se das anteriores na medida em que não é baseada numa tabela de verdade. Na realidade, esta função calcula a razão existente entre o valor de uma característica existente no formato de partida e o mesmo valor no formato de destino correspondente.

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

Fórmula 6 – Ratio.

A título de exemplo, partindo do pressuposto que um formato com um elevado nível de prevalência é preferido face a um formato de nicho pouco utilizado, a função `ratio` permite determinar o benefício obtido ao converter uma representação de um destes formatos para o outro.

O exemplo apresentado na Fórmula 7 demonstra o benefício obtido ao converter uma representação em formato JPEG 2000 para TIFF 6 e vice-versa.

$$MarketShare_{JPEG2000} = low = 0.25$$
$$MarketShare_{TIFF6} = high = 0.75$$

$$Ratio(MarketShare_{JPEG2000}, MarketShare_{TIFF6}) = \frac{0.75}{0.25} \approx 3.0$$
$$Ratio(MarketShare_{TIFF6}, MarketShare_{JPEG2000}) = \frac{0.25}{0.75} \approx 0.33$$

Fórmula 7 – Exemplo de aplicação da função Ratio.

Nota: os valores utilizados neste exemplo são meramente ilustrativos e foram obtidos da tabela apresentada na Figura 50.

4.8 Migration Advisor

O Migration Advisor é um serviço capaz de processar as avaliações produzidas pelos componentes anteriormente descritos e, a partir destas, sugerir alternativas de migração adequadas à resolução de um problema específico de preservação. Uma entidade-cliente pode manifestar os seus requisitos de preservação atribuindo pesos ou importâncias aos critérios de avaliação suportados pelos componentes avaliadores. Os critérios pesados são comunicados à plataforma no momento em que a recomendação é requisitada ao sistema.

Sempre que, no contexto do CRiB, é efectuada uma conversão, são realizadas três avaliações distintas por parte dos componentes Migration Broker, Object Evaluator e Format Evaluator. Cada um destes componentes é responsável por aferir o desempenho, susceptibilidade a perdas de informação e aptidão técnica para a preservação dos serviços de migração utilizados. O Migration Broker, por exemplo, foca-se na avaliação do processo de migração. Durante a sua avaliação, considera critérios como o débito do serviço de migração, a sua disponibilidade, estabilidade, taxa de crescimento em bytes das representações submetidas a conversão, entre outros (ver secção 4.5).

Por sua vez, o Object Evaluator mede o nível de degradação infligido às representações durante o processo de migração. Foca-se, sobretudo, nos objectos e nas suas propriedades intrínsecas e não apenas no processo de migração. Este componente verifica se determinadas propriedades consideradas significativas (e.g. número de páginas, largura e altura de página, tipos de letra, etc.) se mantiveram intactas durante o processo de conversão (ver secção 4.6).

O Format Evaluator, tal como o nome indica, faz uma análise dos formatos envolvidos na conversão, comparando as suas características técnicas e calculando o benefício que se

obteria em termos de capacidade de preservação se se realizasse uma dada conversão entre dois formatos. Exemplos de características técnicas consideradas por este componente são a quota de mercado de um dado formato, o seu nível de suporte e abertura, existência de software multiplataforma, etc. (ver secção 4.7).

O conjunto integral de critérios suportados pela plataforma designa-se por taxionomia geral de avaliação (ver apêndice 8.2 na página 202). Esta taxionomia é composta por critérios relacionados com o desempenho do processo de migração (Migration Broker), propriedades significativas dos objectos digitais (Object Evaluator) e características técnicas dos formatos envolvidos (Fomat Evaluator). O Migration Advisor combina os relatórios de avaliação produzidos por cada um destes componentes com os pesos atribuídos pela entidade-cliente e ordena todas as alternativas de migração de acordo com a sua capacidade de satisfazer as preferências manifestadas. A Figura 52 apresenta a arquitectura geral do Migration Advisor.

A Figura 51 apresenta o diagrama de sequência que descreve todo o processo de recomendação. Este processo é conduzido da seguinte forma: o cliente começa por informar o sistema sobre qual o formato que pretende preservar e o Migration Advisor responde com a taxionomia geral de avaliação associada ao formato respectivo (método getEvaluationCriteria). Uma vez na posse da taxionomia de avaliação, a entidade-cliente deve atribuir pesos a cada um dos critérios que a constituem, manifestando desta forma as suas preferências e requisitos de preservação. Nesta fase, o cliente poderá especificar, por exemplo, que considera o débito um factor importante a ter em conta, pois gostaria que a migração da sua colecção de objectos fosse realizada da forma mais expedita possível. Em contrapartida, pode definir que o custo de conversão é um factor pouco relevante, pois deseja obter o melhor nível de performance e qualidade possível independentemente dos custos envolvidos. Adicionalmente, o utilizador poderá estipular que os objectos submetidos a migração não deverão sofrer qualquer tipo de degradação introduzida pelo processo de migração.

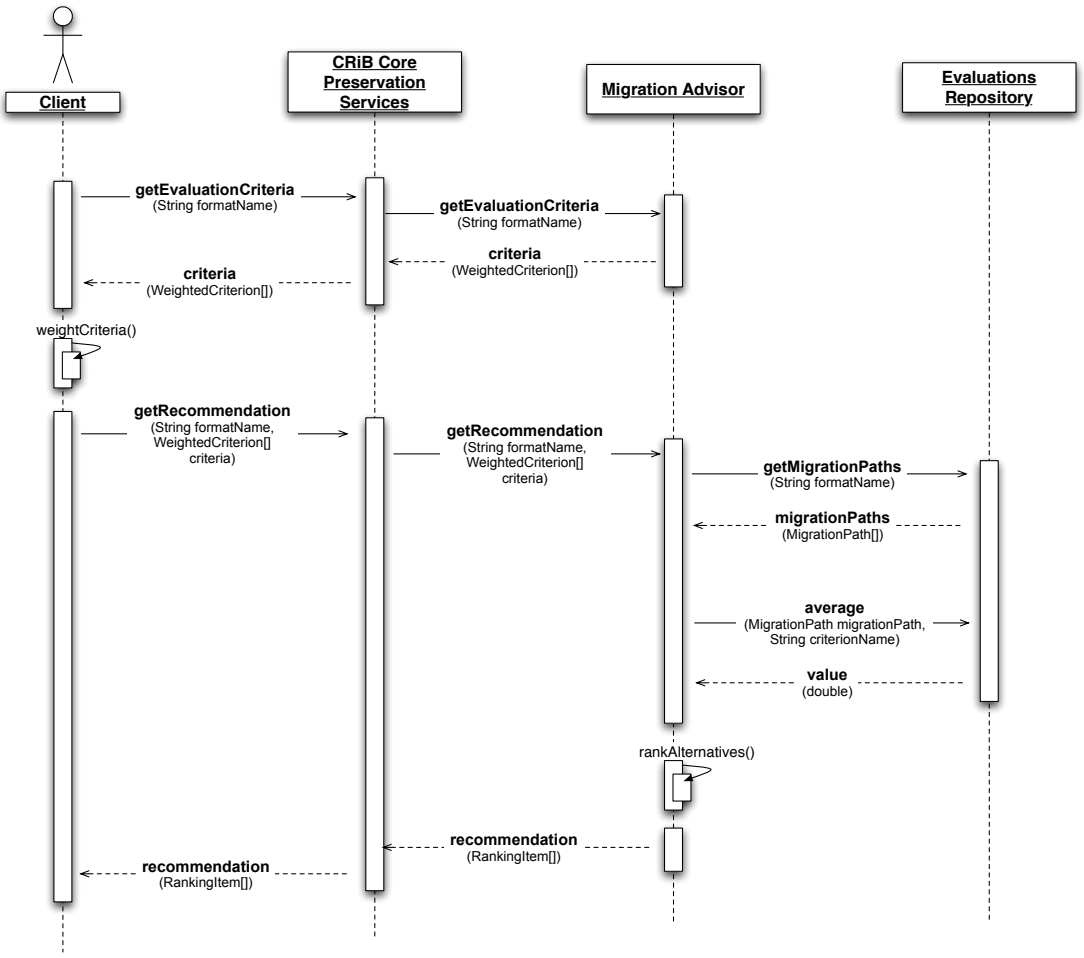


Figura 51 - Diagrama de sequência do processo de recomendação.

Após processar as preferências manifestadas pelo cliente, o sistema é capaz de determinar qual a alternativa de migração mais adequada ao seu contexto específico de preservação. Para tal, o Migration Advisor determina, para cada critério, o comportamento esperado que cada um dos vários caminhos de migração poderá oferecer. Este cálculo é efectuado consultando a informação armazenada no Evaluations Repository, uma base de dados que acumula todos os relatórios de avaliação produzidos pelos vários componentes avaliadores ao longo do tempo. Para um dado critério, o comportamento esperado de um caminho de migração é determinado, analisando a conduta e desempenho de um subconjunto de todas as migrações passadas.

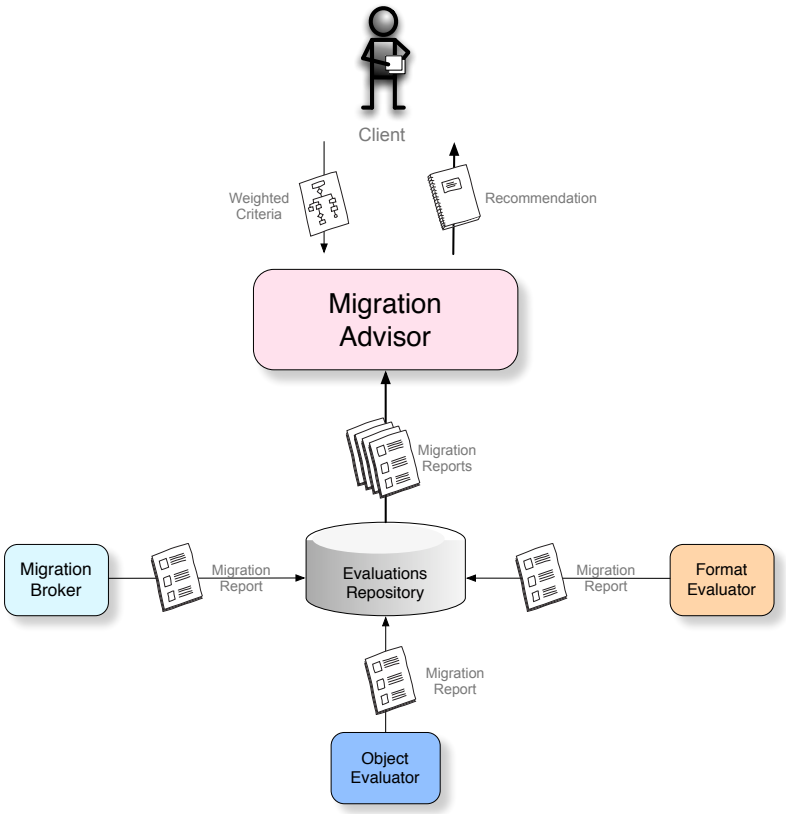


Figura 52 – Arquitectura do Migration Advisor.

Para além da lista ordenada de caminhos de migração, o Migration Advisor informa o cliente sobre a pontuação atribuída a cada um dos caminhos de migração (i.e., *score*). Isto garante ao cliente um nível superior de controlo e segurança no momento da decisão sobre que caminho de migração tomar.

A Figura 53 apresenta o diagrama de classes e mensagens associadas ao componente Migration Advisor. O método getEvaluationCriteria permite ao cliente conhecer a taxionomia geral de avaliação. O cliente deverá pesar cada um dos critérios incluídos na taxionomia, definindo o valor do atributo weight, e devolvê-la ao Migration Advisor, invocando o método getRecommendation. Como resposta, o utilizador irá receber uma lista de caminhos de migração, ordenados pelo seu grau de adequabilidade aos requisitos manifestados.

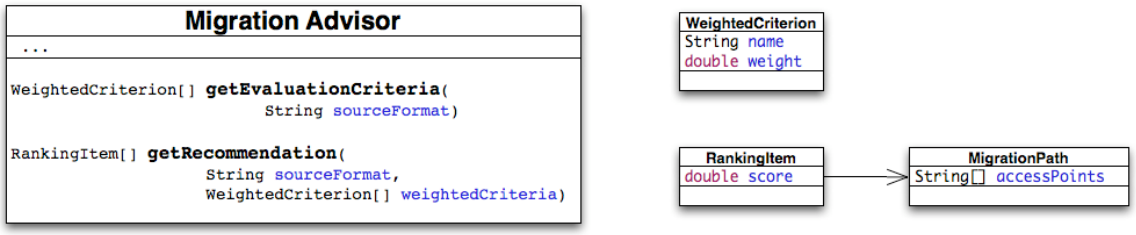


Figura 53 – Diagrama de classes e mensagens trocadas pelo Migration Advisor.

4.8.1 Algoritmo de recomendação

O algoritmo de recomendação que suporta o Migration Advisor é baseado no método de Análise de Utilidade descrito na secção 3.4.5, página 62. Este algoritmo recebe como parâmetros duas estruturas necessárias ao cálculo da recomendação: a taxionomia geral de avaliação, previamente pesada pelo utilizador, e o conjunto de todos os relatórios de migração associados a cada um dos caminhos de migração conhecidos pelo CRiB, i.e., caminhos para os quais existem relatórios de migração registados no Evaluations Repository (Figura 54).

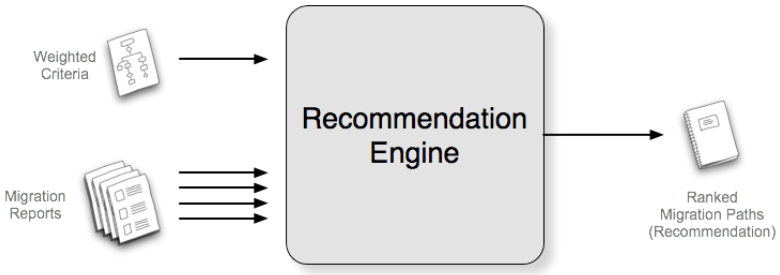


Figura 54 – Arquitectura geral do motor de recomendação.

O processo de recomendação passa essencialmente por quatro fases distintas: 1) normalização dos pesos da taxionomia geral de avaliação, 2) cálculo do desempenho médio de cada caminho de migração, 3) normalização dos desempenhos médios e 4) agregação de resultados e atribuição de pontuação final a cada alternativa de migração (Figura 55). Cada uma destas fases encontra-se descrita nas secções que se seguem.

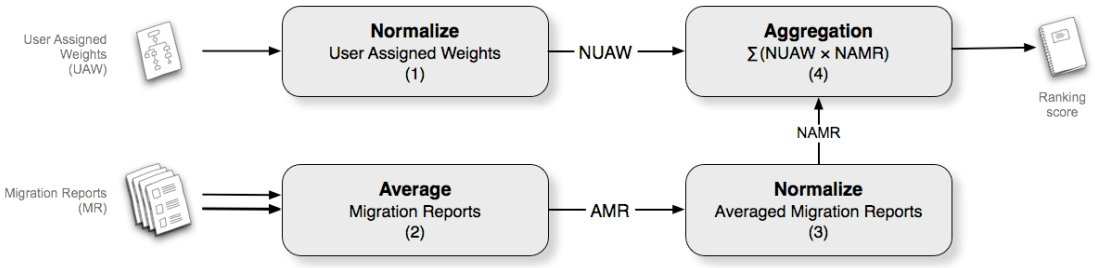


Figura 55 – Cálculo de pontuação de um caminho de migração.

Normalização de pesos

O algoritmo de recomendação exige que o somatório dos pesos atribuídos a cada nível da taxionomia de avaliação seja igual a 1. No entanto, nada na estrutura de dados fornecida ao cliente impõe esse invariante. O estabelecimento dessa restrição na estrutura de dados colocaria dificuldades ao nível da atribuição dos pesos, sendo necessário a construção de uma interface gráfica de auxílio ao utilizador que verificasse esse invariante e o ajudasse na definição dos mesmos.

Na abordagem seguida, o utilizador é livre de atribuir os pesos que achar mais convenientes, não estando limitado a uma escala predefinida. Por exemplo, para cada critério o utilizador poderá atribuir pesos de acordo com uma escala Likert de 1 a 5 (Figura 56).

O processo de normalização responsabiliza-se por reajustar os pesos atribuídos pelo utilizador, preservando a sua importância relativa e garantindo o invariante imposto pelo algoritmo de recomendação.

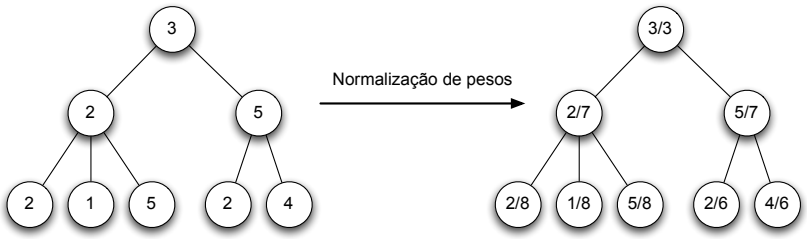


Figura 56 – Exemplo de normalização de taxionomia pesada segundo uma escala Likert de 1 a 5.

O algoritmo de normalização de pesos encontra-se definido na Fórmula 8 onde w_i representa o peso atribuído pelo utilizador ao *i*-nésimo critério de um dado nível da taxionomia geral de avaliação e w_i', o seu valor normalizado. w_i' é calculado dividindo o peso

atribuído pelo utilizador pelo somatório de todos os pesos existentes num dado nível da taxionomia.

$$w_i' = \frac{w_i}{\sum_{i=1}^{n} w_i}$$

Fórmula 8 – Normalização de pesos.

É importante realçar que em taxionomias cujos pesos já respeitem o invariante, o processo de normalização não produz alterações nos pesos atribuídos pelo utilizador.

Cálculo de desempenho médio de um caminho de migração

O Migration Advisor é capaz de determinar o desempenho médio exibido por cada um dos caminhos de migração. O cálculo do desempenho médio baseia-se na análise dos relatórios de migração acumulados ao longo do tempo no Evaluations Repository.

Para determinar o desempenho médio de um caminho de migração é calculada a média dos valores aferidos para cada um dos critérios de avaliação. A Tabela 15 apresenta os valores reais de 5 avaliações ($V_1 \ldots V_5$) efectuadas a um caminho de migração que faz parte da rede de conversores incluída no CRiB. Para facilitar a interpretação, apenas foram incluídos três critérios de cada tipologia de avaliadores.

O vector M resultante passa a representar o desempenho médio do respectivo caminho de migração.

		Avaliações					**Média**
Tipo	**Critério**	V_1	V_2	V_3	V_4	V_5	$M = \frac{\sum_{i=1}^{n} V_i}{n}$
Processo	Débito	6.86818	2.33179	7.18863	9.56329	12.21235	**7.632848**
	Estabilidade	1	1	1	1	1	**1**
	Taxa de crescimento em bytes	1.40623	1.40623	1.42928	1.40623	1.42928	**1.41545**
Objecto	Conformidade gráfica	0.99192	0.99192	0.99118	0.9919	0.99118	**0.99162**
	Largura	1	1	1	1	1	**1**
	Altura	1	1	1	1	1	**1**
Formato	Idade	0.57142	0.57142	0.57142	0.57142	0.57142	**0.57142**
	Quota de mercado	0.00196	0.00196	0.00196	0.00196	0.00196	**0.00196**
	Especificação aberta	1	1	1	1	1	**1**

Tabela 15 – Cálculo de desempenho médio de um caminho de migração.

Normalização do desempenho médio

Uma vez obtidos os vectores de desempenho médio para os vários caminhos de migração registados no sistema, procede-se à normalização dos valores associados a cada critério. A normalização tem como objectivo tornar os diversos valores médios comparáveis, fazendo-os pertencer a uma escala comum. Este processo de normalização é fundamental, pois há critérios que não são balizados superiormente, e.g. débito de conversão, taxa de crescimento em bytes, idade de um formato, etc. Este processo faz com que todos os valores recolhidos pelos componentes avaliadores se situem numa escala compreendida entre 0 e 1.

Os vectores de desempenho médio são normalizados segundo a Fórmula 9. A aplicação desta fórmula faz com que os valores máximos registados assumam o valor 1 e os valores mínimos, o valor 0. Todos os valores situados entre ambos os extremos são distribuídos linearmente ao longo do intervalo.

$$N_{ij} = \frac{M_{ij} - \min(M_j)}{\max(M_j) - \min(M_j)}$$

Fórmula 9 – Normalização de vectores de desempenho.

A Tabela 16 apresenta o resultado da aplicação de um procedimento de normalização sobre cinco caminhos de migração distintos $M_1,...,M_5$.

		Vectores de desempenho médio					**Vectores normalizados**				
Tipo	**Critério**	M_1	M_2	M_3	M_4	M_5	N_1	N_2	N_3	N_4	N_5
Processo	Débito	7.630	6.344	5.333	8.423	4.544	0.796	0.464	0.203	1.000	0.000
	Estabilidade	1.000	0.700	1.000	1.000	0.860	1.000	0.000	1.000	1.000	0.533
	Taxa de crescimento em bytes	1.410	7.570	0.230	0.802	3.63	0.161	1.000	0.000	0.078	0.463
Objecto	Conformidade gráfica	0.990	1.000	0.732	1.000	1.000	0.963	1.000	0.000	1.000	1.000
	Largura	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Altura	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Formato	Idade	0.570	1.571	2.000	0.444	0.857	0.081	0.724	1.000	0.000	0.265
	Quota de mercado	0.002	0.0103	0.588	0.009	0.300	0.000	0.014	1.000	0.012	0.509
	Especificação aberta	1.000	0.000	1.000	1.000	0.000	1.000	0.000	1.000	1.000	0.000

Tabela 16 – Normalização de desempenho médio de um caminho de migração.

Agregação de resultados e cálculo de pontuação final (*score*)

Após a normalização dos vectores de desempenho, procede-se à agregação de resultados e ao cálculo da pontuação final associada a cada caminho de migração. A pontuação define a ordem pela qual os vários caminhos de migração serão apresentados na recomendação. Pontuações

elevadas representam caminhos de migração com maior aptidão para satisfazer os requisitos de preservação manifestados pelo utilizador.

O processo de agregação de resultados começa pela hierarquização dos vectores de desempenho, transformando-os em árvores compatíveis com a taxionomia geral de avaliação pesada pelo utilizador. De seguida, os pesos normalizados atribuídos pelo utilizador são multiplicados pelas folhas da árvore de desempenho normalizada, associada a cada caminho de migração. Após a multiplicação, os resultados são adicionados e agregados no elemento ascendente na taxionomia de avaliação. O processo é recursivamente aplicado até se obter uma pontuação final para o respectivo caminho de migração (Figura 57). Este processo é repetido para cada caminho de migração.

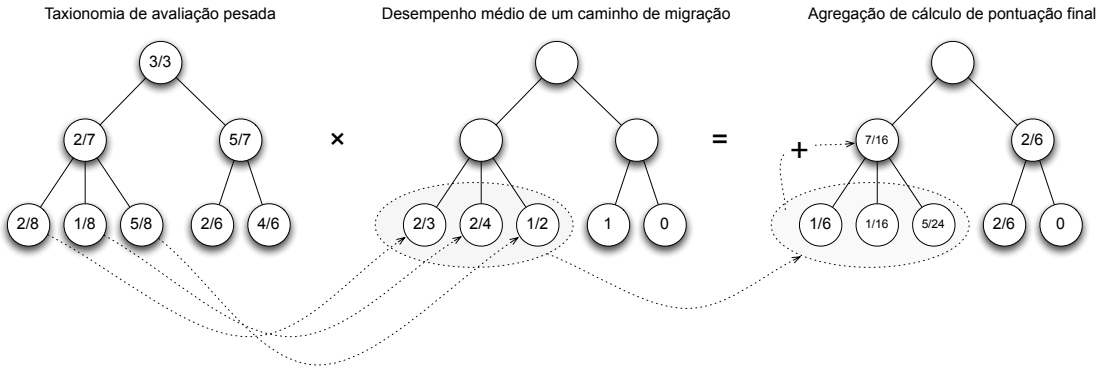


Figura 57 – Agregação de resultados e cálculo de pontuação.

Após obter uma pontuação para cada um dos caminhos de migração, é possível construir um *ranking* com as alternativas mais adequadas para preservar uma dada colecção de objectos. O *ranking* constitui a recomendação produzida pelo Migration Advisor e que é efectivamente enviada ao cliente. Este poderá optar pelo caminho de migração com a pontuação mais elevada ou qualquer um dos outros caminhos apresentados, conhecendo de antemão as vantagens e desvantagens inerentes a essa tomada de decisão.

4.9 Considerações finais

Este capítulo teve como principal objectivo descrever detalhadamente a plataforma CRiB, uma arquitectura orientada ao serviço que disponibiliza um conjunto de serviços de preservação capaz de auxiliar instituições e indivíduos na implementação de estratégias de preservação baseadas em migração.

O capítulo começa por fornecer uma visão geral sobre a arquitectura desenvolvida, expondo exemplos de aplicações-cliente que poderão tomar partido dos serviços disponibilizados e identificando objectivamente os componentes que a constituem. Nesta fase, são ainda identificadas todas as fontes de informação que suportam esses mesmos componentes.

Tratando-se de uma arquitectura orientada ao serviço, qualquer entidade-cliente é livre de aceder e utilizar directamente os serviços que melhor poderão satisfazer as suas necessidades. Não obstante, o CRiB é acompanhado de um componente designado Core Preservation Services que serve de interface entre as aplicações-cliente e todos os componentes que constituem o sistema. Este componente introduz um conjunto de 8 métodos que facilitam a realização de tarefas complexas. Estes métodos são:

- identifyFormat – um método que permite identificar o formato de uma dada representação;
- getEvaluationCriteria – um método que, dado um formato, permite conhecer quais os critérios de controlo de qualidade suportados pela plataforma (i.e., a taxionomia geral de avaliação);
- getRecommendation – um método que permite ao cliente conhecer as alternativas de migração mais adequadas para resolver o seu problema específico de preservação;
- convert – um método que realiza migrações de formatos, recorrendo se necessário à composição de serviços de conversão;
- getSupportedSourceFormats – um método que devolve uma lista de formatos, para os quais existem conversores registados na plataforma;
- getSupportedTargetFormats – um método que devolve os formatos para os quais uma dada representação poderá ser convertida;
- getMigrationPaths – um método que disponibiliza o conjunto de caminhos de migração entre dois formatos;
- getConverterMetadata – um método que permite obter informação detalhada sobre um dado conversor registado na plataforma.

Para além da descrição dos métodos disponibilizados pelo Core Preservation Services, foi também incluída neste capítulo uma descrição das estruturas de dados por estes manipuladas e os diagramas de sequência que caracterizam os procedimentos de interacção entre os clientes e a plataforma.

O capítulo continua com uma descrição detalhada de todos os subcomponentes que constituem a plataforma, expondo todas as suas estruturas de dados, interfaces aplicacionais e ferramentas de suporte. Estes componentes são, designadamente:

- Service Registry – um componente que tem como missão armazenar informação sobre os vários serviços de migração registados na plataforma. Este componente baseia-se na norma UDDI e estende-a de modo a suportar metainformação específica para este tipo de serviços.
- Format Identifier – um componente que permite determinar o formato de uma representação digital. Este componente baseia-se na ferramenta Droid desenvolvida pelos Arquivos Nacionais do Reino Unido;
- Migration Broker – um componente responsável por encapsular a composição de serviços de migração e avaliar o seu desempenho de execução;
- Object Evaluator – um componente responsável pode determinar o nível de degradação incorrido ao nível das propriedades significativas que constituem um objecto digital durante um processo de migração. Para tal, recorre a um conjunto de extractores de propriedades e funções de similaridade que permitem quantificar as diferenças existente entre o objecto que foi submetido a migração e a nova representação que resultou deste processo;
- Format Evaluator – um componente que analisa características técnicas de dois formatos e determina o benefício em termos da capacidade para preservação a longo-prazo que se obteria ao realizar uma migração entre ambos os formatos;
- Migration Advisor – um componente que processa o histórico de avaliações produzidas pelos três componentes anteriormente descritos e produz um *ranking* de alternativas de migração que melhor poderão resolver um determinado problema de preservação. Este componente baseia-se no método de Análise de Utilidade.

4.9.1 Limitações

Apesar dos desenvolvimentos realizados, existe um sem-número de melhorias que poderiam ser introduzidos na plataforma e que fariam com que esta se tornasse mais segura, versátil, eficiente e precisa. Os parágrafos que se seguem procuram descrever e apontar o caminho para alguns desses desenvolvimentos.

O componente Service Registry é utilizado tanto para armazenar metainformação descritiva sobre os vários serviços de migração acessíveis a partir do CRiB, como para calcular

os caminhos de migração existentes entre quaisquer dois formatos. O método utilizado para calcular os caminhos de migração é baseado num algoritmo exaustivo que explora todas as rotas existentes no grafo de migração até encontrar o conjunto de caminhos possíveis entre os dois vértices desejados. Isto faz com que o cálculo dos caminhos de migração demore, por vezes, vários segundos, dependendo da dimensão da rede de migração.

Uma forma de optimizar este processo seria estender o componente Service Registry de forma a calcular o fecho transitivo da rede de migração durante o registo de um novo serviço de migração, guardando para cada par de formatos um conjunto pré-calculado de caminhos de migração prontos a ser utilizados. Outra forma de acelerar este processo seria utilizar um motor de orquestração de serviços[107] baseado numa linguagem de orquestração como a Web services Business Process Execution Language[108] (WS-BPEL). Este tipo de tecnologia permite definir fluxos de execução de Web services, ou seja, permite criar novos serviços baseados na composição de serviços pré-existentes e gerir todo o processo de execução dos mesmos de forma transparente para o utilizador. Esta tecnologia poderia também ser utilizada para substituir o componente Migration Broker na sua função de compositor de serviços, no entanto, esta teria de ser estendida para suportar métricas de avaliação de desempenho.

Ainda em relação aos serviços de migração, o modelo de negócio introduzido na plataforma baseia-se numa arena onde programadores poderão registar e vender os seus serviços de migração e onde clientes poderão tomar partido daqueles que lhes oferecem melhor qualidade de serviço ao menor preço. A plataforma CRiB age apenas como intermediário, prestando serviços de localização e controlo de qualidade. As avaliações de controlo de qualidade realizadas pela plataforma são utilizadas para informar os clientes sobre que migrações oferecem melhor qualidade de serviço.

O modelo de negócio suportado actualmente pela plataforma baseia-se na atribuição de um valor fixo a cada serviço de migração que será cobrado ao cliente durante a invocação do mesmo. Este modelo de negócio, porém, é demasiado simplista para que possa ser posto em prática de forma eficaz. Futuramente serão estudados novos modelos de negócio mais elaborados, baseados, por exemplo, na dimensão das representações digitais a converter, na complexidade dos objectos (e.g. número de páginas, número de tabelas, número de imagens, número de cores, resolução), descontos de quantidade, etc.

---

[107] Um exemplo deste tipo de motores é o Apache ODE disponível em http://ode.apache.org/
[108] http://docs.oasis-open.org/wsbpel/2.0/

No modelo apresentado, os provedores de serviços de migração são responsáveis pela implementação distribuída dos mesmos. Isto incorpora uma falha fundamental que poderá inviabilizar a utilização de uma arquitectura com estas características em determinados contextos de aplicação. Esta falha tem que ver com a confidencialidade dos dados. Os serviços de migração podem operar sobre protocolos seguros como o Hypertext Transfer Protocol sobre Secure Socket Layer (HTTPS), assegurando deste modo que os dados trocados entre o cliente e a plataforma intermédia, e entre esta e os provedores de serviço não são susceptíveis de inspecção por terceiros. No entanto, é difícil garantir a idoneidade dos provedores de serviço que obterão, necessariamente, acesso aos dados a converter. Uma forma de combater este problema seria estabelecer contratos de prestação de serviço que garantissem a confidencialidade e segurança dos dados por parte dos provedores de serviços de migração. Esses contratos seriam estabelecidos no momento do registo de um serviço de migração na plataforma. Não obstante, estes contratos seriam apenas baseados na confiança mútua, pois é complexo implementar mecanismos de monitorização e certificação dos processos desenvolvidos do lado do provedor de serviço. Formas mais criativas na área do direito e da segurança de dados terão que ser investigadas no sentido de mitigar este problema.

Ainda neste contexto, é importante salientar que a transferência de grandes quantidades de dados através da rede, nomeadamente, através da Internet é ainda uma operação excessivamente pesada. O tempo de trânsito dos dados pode facilmente exceder o seu tempo de migração, fazendo com que uma solução centralizada ofereça vantagens consideráveis ao nível da performance de conversão. Não obstante, o CRiB tanto pode ser implementado de forma distribuída, através da Internet, como de forma centralizada na rede do próprio cliente. Na presença de redes locais na ordem dos Gigabits esta alternativa torna-se mais apelativa. Refira-se ainda que a implementação local da plataforma CRiB acarreta a vantagem adicional de resolver o problema da segurança dos dados.

Outro aspecto que poderia ser melhorado é a forma como o débito de uma migração é calculado. Neste momento este parâmetro é determinado, dividindo o comprimento em bytes da representação a converter pelo tempo de migração. No entanto, o tempo de conversão não está directamente relacionado com o comprimento da representação. Representações com conteúdos marcadamente complexos (e.g. um documento com muitas tabelas e imagens) poderão demorar mais tempo a converter do que representações bastante maiores em termos de tamanho, mas de complexidade inferior. Trabalho futuro poderá centrar-se na identificação dos factores que influenciam directamente o tempo de conversão através da análise detalhada das propriedades geralmente associadas a uma dada classe de objectos ou formatos.

No que toca ao Format Identifier, é necessário referir que se poderia enriquecer os resultados que produz se se utilizasse uma combinação de várias ferramentas de identificação de formatos como o Unix file ou o JHove. No entanto, seria necessário criar mapeamentos entre os descritores de formatos utilizados por cada uma destas ferramentas de modo a assegurar a sua coerência e o controlo das designações utilizadas.

Ainda neste contexto, é de referir que o projecto Registry of Open Access Repositories[109] (ROAR) utiliza descritores de formatos semelhantes aos produzidos pelo CRiB, diferindo apenas no facto de a versão do formato ser apresentada entre parênteses após a designação do mesmo, e.g. *Portable Document Format (1.3)*, ao invés de separada pela expressão ", version" como acontece no CRiB, e.g. *Portable Document Format, version 1.3*.

Em relação ao Object Evaluator, é importante referir que o cálculo de similaridade entre dois objectos digitais pode ser afectado pela qualidade dos extractores de propriedades que acompanham este componente. O objectivo do cálculo de similaridade é determinar se houve perdas de informação durante o processo de migração de um objecto digital. Ao extrair valores de propriedades de dois objectos em formatos distintos recorrendo a extractores manifestamente diferentes, poderá incorrer-se precisamente no problema que procura evitar, ou seja, os extractores de propriedades poderão comportar-se de forma errónea e introduzir anomalias nas propriedades extraídas, o que iria influenciar a avaliação realizada. Não obstante, os erros introduzidos pelos extractores de propriedades poderão ser considerados constantes ao longo de todas as avaliações realizadas por este componente, ao passo que os erros introduzidos pelo processo de migração variam consoante o caminho de migração tomado. Isso faz com que o componente permaneça imparcial no que toca às avaliações realizadas, mantendo a relação de ordem entre os vários caminhos de migração utilizados.

Ainda neste contexto, é importante referir que os critérios de avaliação suportados pelos vários componentes avaliadores (i.e., Migration Broker, Object Evaluator e Format Evaluator) foram desenvolvidos como *add-ons* à plataforma, o que significa que o desenvolvimento e a instalação de novos critérios de avaliação podem ser realizados de forma simples, sem que haja necessidade de reprogramação da plataforma.

No que diz respeito ao Migration Advisor alguns dos possíveis melhoramentos futuros passam pela optimização do processo de cálculo de desempenho médio dos vários caminhos de migração recorrendo a técnicas de *Data warehousing,* por exemplo, armazenando valores

[109] http://roar.eprints.org/

acumulados ao invés de os calcular sempre que são requeridos (Caldeira, 2008; Kimball & Ross, 2002). Neste momento, o desempenho médio é calculado realizando um conjunto de questões à Migration Knowledge Base, o que, dependendo do número de avaliações armazenadas e do número de caminhos de migração registados, poderá ser uma tarefa bastante complexa e demorada.

Para além da optimização de processos, seria profícuo a realização de um estudo recorrendo a técnicas de análise de sensibilidade (Saltelli, 2004; Stanley & Stewart, 2002). Este estudo teria como objectivo verificar em que medida pequenas perturbações nos pesos atribuídos por parte das entidades-cliente à taxionomia geral de avaliação poderiam influenciar as recomendações produzidas pelo Migration Advisor.

Para concluir, a plataforma proposta beneficiaria com um aumento dos critérios de avaliação suportados, bem como de formatos reconhecidos. Para além do disposto, a adição de novas classes de objectos digitais tornaria a plataforma mais apta para recomendar estratégias de migração. Adicionalmente, esta deveria ser melhorada para implementar mecanismos de controlo de qualidade que suportassem migrações de formato entre classes de objectos distintas.

Capítulo 5
Metodologia e avaliação

Ao longo do capítulo anterior foram apresentados, em detalhe, todos os componentes que compõem a plataforma CRiB, um sistema capaz de assistir organizações e indivíduos na selecção e execução de intervenções de preservação baseadas em migração .

Este trabalho teve como principal objectivo aferir se seria possível automatizar os processos inerentes à preservação de objectos digitais recorrendo a estratégias de migração.

A implementação de uma estratégia de migração pressupõe o desenvolvimento de três actividades fundamentais, nomeadamente: a selecção de uma alternativa de migração adequada aos objectivos da entidade preservadora e aos objectos digitais que se pretendem preservar, a conversão dos materiais propriamente dita, e a avaliação e controlo de qualidade da respectiva intervenção. O CRiB disponibiliza um conjunto de serviços suportados por componentes de software que têm como objectivo implementar cada uma destas actividades. Esses componentes são, respectivamente, o Migration Advisor, o Migration Broker e o Object Evaluator.

Este capítulo tem como missão descrever a metodologia utilizada durante a validação destes componentes, bem como as conclusões que daí resultaram. É importante referir que o componente Migration Broker não foi validado. Este componente apresenta apenas dois

estados possíveis de execução: sucesso ou insucesso. Os casos de insucesso ocorrem quando os serviços de migração foram incapazes de completar uma dada tarefa de conversão. Os restantes dois componentes, dada a sua complexidade, exigiram um maior rigor e esforço de validação.

Este capítulo está organizado da seguinte forma: a secção 5.1 descreve detalhadamente as experiências realizadas em torno do componente Object Evaluator e a secção 5.2 apresenta a metodologia e os processos de avaliação desenvolvidos em torno do Migration Advisor. Em ambas as secções são ainda apresentados os protocolos experimentais adoptados, a caracterização das colecções de objectos de teste utilizados ao longo da experiência, os detalhes dos estudos comparativos realizados e uma secção de resultados e conclusões.

5.1 Avaliação do Object Evaluator

O componente Object Evaluator tem como missão identificar e quantificar o nível de degradação introduzido nos objectos digitais durante um processo de migração. Este tipo de ocorrências deve-se sobretudo ao facto de as aplicações de conversão não incorporarem todas as funcionalidades necessárias à correcta transformação dos valores das propriedades que constituem o objecto de partida ou simplesmente porque existem incompatibilidades entre os formatos de partida e os formatos de destino.

O modo de funcionamento do Object Evaluator baseia-se na extracção dos valores das propriedades significativas pertencentes aos objectos digitais submetidos a migração e aos seus equivalentes convertidos, e no consequente cálculo da similaridade entre estes. Uma conversão pode ser considerada bem sucedida se os valores das propriedades associadas a um objecto digital não sofreram alterações durante a conversão. Minimizar a degradação destes valores tem como consequência fundamental a melhoraria da qualidade da intervenção de preservação e a garantia da integridade dos objectos a longo-prazo. O propósito do Object Evaluator é, precisamente, controlar a qualidade da migração efectuada, determinando o nível de degradação incorrido nessas propriedades e registando os resultados dessa avaliação para efeitos de documentação do processo de migração.

As propriedades significativas analisadas durante o processo de controlo de qualidade dependem sobretudo da classe dos objectos digitais. Por exemplo, a comparação de dois objectos pertencentes à classe documentos de texto poderá envolver propriedades como o tamanho da página, apresentação gráfica do documento, número de páginas, margens, tipos de letra, cores, etc. No entanto, se os objectos pertencerem à classe áudio, as

propriedades significativas a analisar seriam consideravelmente diferentes, e.g. resolução, volume, nível médio de ruído, duração, etc[110].

Ainda neste contexto, é importante referir que existem dois tipos de propriedades significativas: propriedades de carácter objectivo e propriedades de carácter subjectivo. Considere-se o seguinte exemplo. A largura e altura (em *pixel*) de uma imagem são propriedades marcadamente objectivas, i.e., tratam-se de propriedades que poderão facilmente ser extraídas e comparadas por um qualquer processo automático baseado em software. O mais rudimentar dos visualizadores de imagens, por exemplo, é capaz de ler e apresentar a largura e altura de uma imagem desde que o seu formato seja reconhecido pela aplicação. Calcular a similaridade entre duas propriedades com estas características é um processo, geralmente, simples, não levantando grandes dúvidas relativamente aos resultados obtidos.

No entanto, há um conjunto de propriedades que devido às suas características se tornam difíceis de comparar automaticamente. Por exemplo, se se pedir a dois intervenientes humanos para quantificar o nível de similaridade percepcionado entre duas imagens parecidas (porém não iguais), é possível constatar que a taxa de concordância entre ambos os avaliadores, apesar de elevada, não é inteiramente consensual. Isto significa que certas propriedades são caracterizadas por uma certa subjectividade, o que torna consideravelmente mais complexo o cálculo de similaridade recorrendo a processos automáticos.

A avaliação do componente Object Evaluator teve como principal objectivo aferir o nível de concordância existente entre os valores de similaridade produzidos por este componente e os valores de similaridade produzidos por intervenientes humanos. A avaliação deste componente contemplou a realização de um conjunto de experiências com foco nas propriedades significativas do domínio das imagens matriciais consideradas subjectivas, i.e., conformidade gráfica e metainformação embebida.

É importante realçar que as experiências realizadas em torno do Object Evaluator apenas incluíram formatos pertencentes à mesma classe de objectos. Apesar de ser possível, em teoria, realizar conversões entre formatos pertencentes a classes de objectos distintas, este tipo de cenários foi deliberadamente remetido para trabalho futuro. Entre classes distintas, o número de propriedades significativas comparáveis é mais reduzido (trata-se da intersecção dos conjuntos de propriedades significativas de cada classe de objectos), o que implica o estudo de

[110] Para uma listagem completa das propriedades significativas suportadas pelo Object Evaluator, consulte a secção 4.6.2 na página 105.

formas eficazes de lidar com informação incompleta. Ou seja, de estratégias capazes de determinar que valores de similaridade deverão ser considerados quando uma determinada propriedade apenas está associada a um dos objectos em comparação. No limite, os conjuntos de propriedades significativas associados a cada classe de objectos poderão ser disjuntos, impossibilitando o cálculo adequado de similaridade.

Para ilustrar este ponto, considere-se o seguinte exemplo. É possível admitir que a versão sonorizada de um livro[111] possa ser interpretada como uma representação, ou manifestação, alternativa da sua versão textual, mais convencional. É possível também imaginar um processo capaz de converter uma instância textual dessa obra na sua versão sonorizada, destinada por exemplo ao consumo por invisuais[112]. Este cenário é representativo de uma conversão entre formatos pertencentes a classes de objectos distintas: documentos de texto e documentos áudio. Como já havia sido referido anteriormente, o conjunto de propriedades significativas associadas a cada uma destas classes difere consideravelmente, o que torna o cálculo automático de similaridade bastante complexo ou até mesmo impraticável.

As secções que se seguem descrevem detalhadamente as experiências realizadas em torno do componente Object Evaluator de modo a aferir a sua capacidade em avaliar o nível de degradação introduzido em propriedades significativas de carácter subjectivo.

5.1.1 Protocolo experimental

No seu conjunto, as experiências realizadas em torno do Object Evaluator tiveram como objectivo aferir com que precisão este componente seria capaz de calcular a similaridade entre valores de propriedades significativas consideradas subjectivas extraídos a partir de objectos digitais em formatos distintos. A exactidão e precisão do Object Evaluator foram determinadas comparando os valores por ele produzidos com valores produzidos por um conjunto de avaliadores humanos, sendo estes considerados os valores de referência.

Assim, as experiências realizadas em torno deste componente seguiram o seguinte protocolo:

1. **Construção de uma colecção de teste** – cada experiência realizada obrigou à construção de uma colecção de teste constituída por um conjunto alargado de objectos digitais. Cada colecção de teste foi preparada de modo a incluir as propriedades que se

[111] Hoje em dia é bastante comum encontrar no mercado livros sonorizados em áudio, algo que vulgarmente se designa por "audio book".

[112] As versões mais actuais do software Acrobat Reader já são capazes de sonorizar um documento de texto ao activar uma opção chamada "Read Out Loud".

pretendiam avaliar e de forma a conter objectos em diversos formatos, mas pertencentes à mesma classe de objectos digitais.

2. **Avaliação manual da colecção de teste** – todos os objectos pertencentes às colecções de teste foram avaliados manualmente por um conjunto de intervenientes humanos. As métricas utilizadas e o número de pessoas envolvidas na avaliação dependeram da propriedade significativa em causa. As avaliações realizadas por humanos designam-se vulgarmente por *avaliações subjectivas* (Biström, 2005; Telecommunication Standardization Sector of ITU, 2004).

3. **Avaliação automática da colecção de teste** – esta actividade é, em tudo, semelhante à actividade descrita no ponto anterior. No entanto, realiza-se através de processos automáticos, neste caso recorrendo ao componente Object Evaluator. Este tipo de avaliações designa-se vulgarmente por *avaliações objectivas*, isto apesar das propriedades sob avaliação serem marcadamente subjectivas (Biström, 2005; Telecommunication Standardization Sector of ITU, 2004).

4. **Estudo comparativo dos resultados** – realização de um estudo comparativo entre as avaliações produzidas pelo Object Evaluator e as avaliações realizadas pelo conjunto de avaliadores humanos. Este estudo permitiu quantificar a capacidade do Object Evaluator em determinar correctamente a similaridade entre propriedades subjectivas (por comparação com a mesma avaliação realizada por humanos).

De modo a avaliar os resultados produzidos pelo componente Object Evaluator foi constituída uma colecção de teste composta exclusivamente por imagens matriciais. Após uma análise das propriedades significativas associadas a esta classe de objectos digitais (ver secção 4.6.2 na página 106), foram consideradas para efeitos de avaliação as seguintes propriedades: conformidade gráfica e metainformação embebida, as únicas dotadas de características marcadamente subjectivas.

5.1.2 Propriedade significativa: conformidade gráfica

Num contexto de migração poderá ocorrer deterioração do conteúdo gráfico de um objecto digital. A propriedade conformidade gráfica diz respeito à determinação do grau de semelhança, real ou percepcionada, entre o conteúdo gráfico de um objecto convertido e de um outro considerado original.

As secções que se seguem procuram determinar experimentalmente a capacidade do componente Object Evaluator em aferir o nível de conformidade gráfica existente entre duas imagens matriciais.

Ao longo desta experiência foram avaliados quatro algoritmos de similaridade de imagem, nomeadamente, o Normalized Root Mean Squared Error (Shrestha, O'Hara, & Younan, 2005), o Universal Image Quality Index (Z. Wang & Bovik, 2002), o Structural Similarity Index Metric (Z. Wang, Bovik, Sheikh, & Simoncelli, 2004) e o Content-Based Image Quality Metric (Gao, Wang, & Li, 2005).

Caracterização da colecção de teste

Nos métodos clássicos da estatística, quanto maior for a dimensão da amostra mais fiáveis serão as conclusões que dela se podem extrair. Usualmente a dimensão é determinada a partir de critérios pré-estabelecidos tais como a minimização dos custos de amostragem, a minimização da variância do estimador de um certo parâmetro de interesse, entre outros. No caso concreto desta experiência, a dimensão da amostra foi definida de modo a que o tempo necessário à sua avaliação por parte de um interveniente humano não excedesse os 30 minutos, temendo que o desgaste físico do avaliador pudesse comprometer a qualidade dos resultados. Após algumas interacções experimentais, o número a que se chegou foi o de uma amostra de tamanho 30, i.e., cada avaliação teria uma duração média de 1 minuto.

A colecção de teste utilizada para determinar a capacidade do Object Evaluator em calcular a conformidade gráfica entre duas imagens digitais foi constituída por 10 imagens base a partir das quais foram criadas 3 derivadas com diferentes níveis de deformação (totalizando 40 imagens). Entre as deformações introduzidas encontravam-se artefactos de compressão, a alteração de cores e a alteração do número de bits de cor que constituíam a imagem.

As derivadas incluídas na colecção de teste foram geradas segundo parâmetros de conversão aleatórios. Esta opção teve como objectivo maximizar a diversidade das deformações existentes na colecção de teste.

É importante referir que deformações como redimensionamento (i.e., alteração da largura ou altura), o corte de imagens e deformações que vão para além da mera conversão de formatos (e.g. aplicação de filtros ou introdução manual de artefactos) não fizeram parte da lista de alterações introduzidas. Este tipo de deformações são capturadas por outras propriedades significativas como a largura e a altura da imagem.

A cada avaliador foi pedido que comparasse cada uma das 10 imagens originais com as 3 derivadas previamente produzidas e atribuir uma classificação de 0 a 10 de acordo com o grau de similaridade percepcionado (variando qualitativamente entre o "Totalmente diferentes" e o "Iguais"). No total, cada interveniente humano seria responsável por avaliar 30 pares de imagens.

As dez imagens utilizadas nesta experiência podem ser agrupadas em seis categorias distintas: 1) manuscrito digitalizado, 2) página de jornal digitalizada, 3) cartaz colorido digitalizado, 4) fotografia digital, 5) cartoon digital, e 6) desenho digitalizado.

As tabelas que se seguem descrevem detalhadamente o conjunto completo de imagens que constituíam a colecção de teste utilizada nesta experiência.

Imagem	Código	Descrição	Dimensões
Original	01-00	Imagem digitalizada de um livro manuscrito em formato TIFF 256 tons de cinzento (modo indexado).	3481x2448
Derivada 1	01-01	Imagem convertida para formato GIF 256 cores.	
Derivada 2	01-02	Imagem convertida para formato JPEG com um nível de compressão de 5 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 3	01-03	Imagem convertida para formato JPEG 2000 com um nível de compressão de 5 (1 – qualidade mínima, 100 – qualidade máxima).	

Imagem	Código	Descrição	Dimensões
Original	02-00	Página de jornal digitalizada em escala de cinzentos em formato TIFF.	2313x3414
Derivada 1	02-01	Imagem convertida para formato JPEG com um nível de compressão de 2 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 2	02-02	Imagem convertida para formato JPEG com um nível de compressão de 10 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 3	02-03	Imagem convertida para formato JPEG 2000 com um nível de compressão de 90 (1 – qualidade mínima, 100 – qualidade máxima).	

Imagem	Código	Descrição	Dimensões
Original	03-00	Poster digitalizado em formato TIFF com 24 bits de cor.	685x1404
Derivada 1	03-01	Imagem convertida para formato GIF 256 cores.	
Derivada 2	03-02	Imagem convertida para formato JPEG com um nível de compressão de 10 (0 – qualidade mínima, 12 – qualidade máxima)	
Derivada 3	03-03	Imagem convertida para formato JPEG 2000 com um nível de compressão de 1 (1 – qualidade mínima, 100 – qualidade máxima).	

Imagem	Código	Descrição	Dimensões
Original	04-00	Fotografia digital tirada com uma Canon 350D em formato JPEG.	3456x2304
Derivada 1	04-01	Imagem convertida para formato PNG comprimido.	
Derivada 2	04-02	Imagem convertida para formato GIF 256 cores.	
Derivada 3	04-03	Imagem convertida para formato JPEG com um nível de compressão de 2 (0 – qualidade mínima, 12 – qualidade máxima).	

Imagem	Código	Descrição	Dimensões
Original	05-00	Fotografia digital tirada com uma FinePixA101 em formato JPEG.	1280x960
Derivada 1	05-01	Imagem convertida para formato GIF 256 cores.	
Derivada 2	05-02	Imagem convertida para formato JPEG com um nível de compressão de 1 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 3	05-03	Imagem convertida para formato TIFF, 24 bits de cor.	

Imagem	Código	Descrição	Dimensões
Original	06-00	Fotografia digital tirada com uma Olympus C150 em formato JPEG.	1600x1200
Derivada 1	06-01	Imagem convertida para formato JPEG com um nível de compressão de 7 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 2	06-02	Imagem convertida para formato GIF 256 cores.	
Derivada 3	06-03	Imagem convertida para formato JPEG com um nível de compressão de 1 (0 – qualidade mínima, 12 – qualidade máxima).	

Imagem	Código	Descrição	Dimensões
Original	07-00	Fotografia digital tirada com uma Olympus C150 em formato JPEG.	1600x1200
Derivada 1	07-01	Imagem convertida para formato JPEG com um nível de compressão de 10 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 2	07-02	Imagem convertida para formato TIFF, 24 bits de cor.	
Derivada 3	07-03	Imagem convertida para formato GIF 256 cores.	

Imagem	Código	Descrição	Dimensões
Original	08-00	Fotografia digital tirada com uma Olympus FE110 em formato JPEG.	1600x1200
Derivada 1	08-01	Imagem convertida para formato JPEG com um nível de compressão de 10 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 2	08-02	Imagem convertida para formato GIF 256 cores.	
Derivada 3	08-03	Imagem convertida para formato JPEG com um nível de compressão de 1 (0 – qualidade mínima, 12 – qualidade máxima).	

Imagem	Código	Descrição	Dimensões
Original	09-00	Cartoon em formato GIF 256 cores com transparência.	901x1117
Derivada 1	09-01	Imagem convertida para formato JPEG com um nível de compressão de 1 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 2	09-02	Imagem convertida para formato PNG comprimido.	
Derivada 3	09-03	Imagem convertida para formato TIFF, 24 bits de cor.	

Imagem	Código	Descrição	Dimensões
Original	10-00	Desenho digitalizado em formato JPEG.	1034x1455
Derivada 1	10-01	Imagem convertida para formato JPEG com um nível de compressão de 9 (0 – qualidade mínima, 12 – qualidade máxima).	
Derivada 2	10-02	Imagem convertida para formato GIF 256 cores.	
Derivada 3	10-03	Imagem convertida para formato JPEG com um nível de compressão de 5 (0 – qualidade mínima, 12 – qualidade máxima).	

Avaliação manual

O International Telecommunication Union (ITU) com a ajuda do Video Quality Experts Group tem vindo a emitir a normas e recomendações sobre como se deverão processar experiências na área da medição da qualidade de sequências de vídeo (Telecommunication Standardization Sector of ITU, 2004). Estas recomendações são também utilizadas na definição de guiões de procedimentos para a avaliação e comparação de algoritmos de compressão de imagem com perdas (e.g. JPEG, JPEG 2000). Estas mesmas recomendações serviram de base para a construção do guião de procedimentos utilizado ao longo desta experiência.

De acordo com o ITU este tipo de experiências deverá ser realizado com o maior número possível de intervenientes humanos de modo a minimizar a variabilidade das avaliações subjectivas produzidas. Estudos semelhantes foram realizados com grupos de avaliadores na ordem das 25 pessoas (Telecommunication Standardization Sector of ITU, 2004; Z. Wang et al., 2004). No entanto, devido a restrições de tempo e disponibilidade de participantes, esta experiência foi conduzida com apenas 15 pessoas, i.e., o número mínimo de pessoas recomendado pelo ITU (Telecommunication Standardization Sector of ITU, 2004).

Os voluntários que participaram nesta experiência possuíam formação de nível superior, tratando-se sobretudo de alunos de mestrado e de doutoramento, não especialistas no tratamento de imagem, com idades compreendidas entre os 25 e os 50 anos. Nove eram do sexo masculino e seis eram do sexo feminino. Os avaliadores recrutados possuíam formação em áreas variadas como Ciências Sociais, Educação, Engenharia Electrónica e Informática.

Todos os participantes foram sentados confortavelmente em frente a um ecrã a uma distância que variava entre os 25 e os 40 cm. Foi-lhes fornecida uma aplicação[113] que facilitava a visualização simultânea de duas imagens, incorporando ainda funcionalidades como ampliação da área de visualização e navegação sincronizada em ambas as imagens (Figura 58). Todos os participantes utilizaram a mesma aplicação e o mesmo ecrã. Este tipo de avaliação designa-se por *Simultaneous Double Stimulus for Continuous Evaluation* (SDSCE) devido ao facto de as duas imagens serem apresentadas ao avaliador em simultâneo e não em instantes separados.

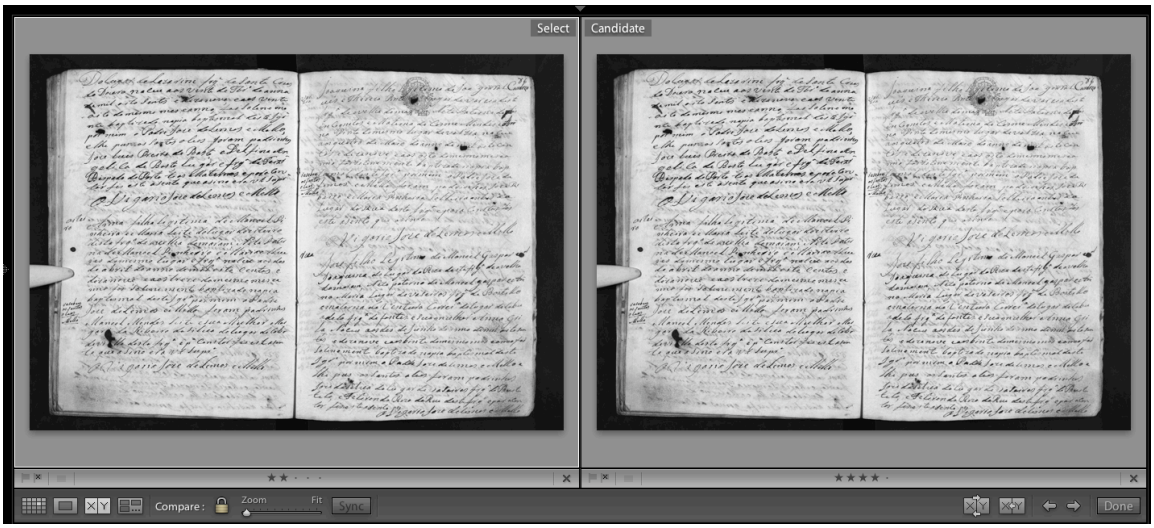


Figura 58 – *Screenshot* da aplicação utilizada para comparar imagens.

Antes de dar início à avaliação da colecção de teste, os participantes receberam instruções detalhadas sobre como a avaliação se iria processar e foi-lhes fornecido um conjunto de imagens de exemplo para que se pudessem ambientar ao tipo de avaliações que se iriam seguir. Esse conjunto de imagens era constituído por 10 pares de imagens com deformações semelhantes às introduzidas na colecção de teste. Esta fase teve como objectivo explicar aos participantes o tipo de avaliações que iriam realizar e, em simultâneo, permitir que estes se familiarizassem com o software de visualização. Esta actividade teve uma duração média de 5 minutos por participante.

Cada participante foi então convidado a observar os vários pares de imagens que compunham a colecção de teste, sem quaisquer restrições de tempo, podendo ainda ampliar e reduzir as respectivas imagens, bem como posicionar a janela de visualização na área da imagem

[113] A aplicação utilizada na experiência chamava-se Adobe Lightroom.

desejada. Após cada observação, foi-lhes pedido que quantificassem o nível de similaridade percepcionado entre ambas as imagens numa escala linear de 0 a 10 (i.e., de "Totalmente diferentes" a "Iguais").

Após reunir as classificações dos 15 intervenientes aos 30 pares de imagens, estas foram agrupadas num único valor designado por Mean Opinion Score ou, simplesmente, MOS (Tabela 17). O valor de MOS representa a média das classificações atribuídas por todos os avaliadores a cada par de imagens (Petrov, Vatolin, Parshin, & Titarenko, 2006; Telecommunication Standardization Sector of ITU, 2004; Z. Wang et al., 2004). A tabela inclui também o desvio-padrão verificado.

	Avaliações subjectivas																	
Par *k*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**MOS**	**σ**	**Valor-P Kolmogorov-Smirnov**
1	10	10	10	9	10	10	10	10	10	10	10	10	10	9	10	9.87	0.35	**0.001**
2	9	6	8	9	10	8	9	8	9	8	9	10	8	8	10	8.60	1.06	0.473
3	6	7	6	5	8	7	7	8	7	6	6	9	7	7	8	6.93	1.03	0.538
4	7	6	6	8	8	7	8	8	5	6	7	8	7	6	8	7.00	1.00	0.347
5	8	9	10	10	10	8	9	9	8	9	10	9	9	8	10	9.07	0.80	0.510
6	9	9	10	10	10	10	9	10	9	10	9	10	10	9	10	9.60	0.51	0.023
7	6	8	10	6	10	7	7	7	7	5	8	10	8	5	10	7.60	1.76	0.718
8	8	9	10	5	9	9	9	8	8	7	9	10	9	6	10	8.40	1.45	0.262
9	8	10	10	9	10	10	10	9	10	8	9	10	10	5	10	9.20	1.37	0.093
10	10	10	10	10	10	10	9	9	10	10	10	10	10	9	10	9.80	0.41	0.002
11	5	6	8	7	7	7	6	8	6	8	7	8	7	6	8	6.93	0.96	0.587
12	6	7	10	4	6	8	7	7	7	7	5	9	6	7	8	6.93	1.49	0.489
13	6	7	10	8	7	7	8	7	9	8	7	10	9	6	7	7.73	1.28	0.306
14	7	4	7	3	5	5	6	4	7	5	4	7	7	5	6	5.47	1.36	0.558
15	8	9	10	9	9	10	10	6	9	10	9	10	10	8	10	9.13	1.13	0.293
16	8	9	9	8	10	10	8	5	9	8	7	10	9	7	9	8.40	1.35	0.556
17	7	8	8	5	8	8	9	4	8	8	8	9	8	8	6	7.47	1.41	0.026
18	6	5	7	3	5	6	6	3	6	5	4	7	6	9	5	5.53	1.55	0.704
19	7	8	8	8	9	9	8	5	7	8	8	9	9	10	9	8.13	1.19	0.282
20	8	9	9	7	9	9	10	5	10	10	10	10	10	10	10	9.07	1.44	0.185
21	6	4	6	4	5	4	6	4	5	5	4	7	6	7	6	5.27	1.10	0.497
22	10	9	10	10	9	10	10	7	9	10	9	10	10	9	10	9.47	0.83	0.064
23	7	7	7	3	5	6	7	5	7	9	6	8	6	7	6	6.40	1.40	0.578
24	7	5	6	5	5	7	6	5	8	5	5	7	8	5	6	6.00	1.13	0.197
25	8	7	6	4	6	6	6	5	8	7	7	8	8	8	7	6.73	1.22	0.675
26	10	10	10	10	10	10	10	7	10	10	10	10	10	9	10	9.73	0.80	**0.001**
27	10	10	10	9	9	10	10	8	9	10	10	10	10	10	10	9.67	0.62	0.006
28	10	10	10	10	10	10	10	7	9	10	10	10	10	10	10	9.73	0.80	**0.001**
29	10	10	9	9	10	10	10	7	9	10	10	10	10	10	10	9.60	0.83	0.100
30	9	9	8	8	7	8	8	6	9	8	7	8	9	10	8	8.13	0.99	0.322

Tabela 17 – Avaliações produzidas por intervenientes humanos.

O valor MOS encontra-se definido formalmente na Equação 1. N representa o número de participantes envolvidos na avaliação (i.e., $N = 15$), k é o índice do par de imagens avaliado e $classificação_{i,k}$ o valor de similaridade atribuído pelo avaliador i ao par de imagens k.

$$MOS_k = \frac{1}{N}\sum_{i=1}^{N} classificação_{i,k}$$

Equação 1 – *Mean Opinion Score* (MOS).

É importante referir que a generalidade das classificações produzidas pelos vários avaliadores a cada par de imagens segue uma distribuição normal como se pode ver pelos valores-$P > 0.01$ obtidos a partir do teste de normalidade Kolmogorov-Smirnov[114]. As avaliações que não seguem a distribuição normal são aquelas cujo valor médio se situa demasiado perto da pontuação máxima, impedindo, deste modo, a formação da curva em forma de sino característica desta distribuição. Esses valores encontram-se assinalados a negrito na Tabela 17.

Após calcular o valor de MOS para cada par de imagens, foram eliminadas as classificações discrepantes, também conhecidas por *outliers*. *Ouliers* são observações que não obedecem ao padrão do conjunto de dados ao qual pertencem (Silva, 2004). Por outras palavras, quando uma observação, ou neste caso, avaliação, se afasta significativamente das restantes é considerada discrepante ou saliente[115]. A ocorrência de tal observação poderá dever-se a múltiplos factores, no entanto, no contexto desta experiência deduziu-se que o aparecimento destas classificações se deveu a desconcentrações momentâneas por parte dos avaliadores.

Assim, as classificações não pertencentes ao conjunto $\left[MOS - 2\sigma,\ MOS + 2\sigma\right]$[116] foram retiradas da matriz de avaliação, produzindo, deste modo, um novo conjunto de valores de MOS e reduzindo o desvio-padrão médio em cerca de 20% (Telecommunication Standardization Sector of ITU, 2004).

A matriz de avaliação obtida após remoção das classificações discrepantes é apresentada na Tabela 18.

---

114 Em estatística, o teste de normalidade de Kolmogorov-Smirnov é utilizado para determinar se uma variável aleatória, representada por uma amostra de valores, segue uma distribuição normal.

115 Outras traduções possíveis para o termo *outlier* são anormal, suspeito ou discordante.

116 O cálculo do intervalo de valores considerados não discrepantes assume que a amostra segue uma distribuição normal, algo que foi previamente demonstrado pelo teste de Kolmogorov-Smirnov.

Par *k*	Discrepantes removidos	MOS'	σ'
1	2	10.00	0.00
2	1	8.79	0.80
3	0	6.93	1.03
4	0	7.00	1.00
5	0	9.07	0.80
6	0	9.60	0.51
7	0	7.60	1.76
8	1	8.64	1.15
9	1	9.50	0.76
10	0	9.80	0.41
11	0	6.93	0.96
12	1	6.71	1.27
13	0	7.73	1.28
14	0	5.47	1.36
15	1	9.36	0.74
16	1	8.64	1.01
17	1	7.71	1.07
18	1	5.29	1.27
19	1	8.36	0.84
20	1	9.36	0.93
21	0	5.27	1.10
22	1	9.64	0.50
23	1	6.64	1.08
24	0	6.00	1.13
25	1	6.93	1.00
26	1	9.93	0.27
27	1	9.79	0.43
28	1	9.93	0.27
29	1	9.79	0.43
30	1	8.29	0.83

Tabela 18 – MOS e desvio-padrão após remoção de valores discrepantes.

Avaliação automática

Após recolher os valores de similaridade atribuídos pelos avaliadores humanos, o mesmo procedimento foi repetido, mas desta vez recorrendo às capacidades de avaliação do componente `Object Evaluator`. Este foi preparado para suportar quatro algoritmos distintos de cálculo de similaridade entre imagens, nomeadamente:

1. Normalized Root Mean Squared Error – NRMSE (Shrestha et al., 2005);
2. Universal Image Quality Index – UQI (Z. Wang & Bovik, 2002);
3. Structural Similarity Index Metric – SSIM (Z. Wang et al., 2004);
4. Content-Based Image Quality Metric – CBM (Gao et al., 2005).

Os quatro algoritmos implementados encontram-se descritos em detalhe no Apêndice 8.3.6 na página 212.

As avaliações produzidas por estes quatro algoritmos pertencem ao conjunto $[0, 1]$, onde 1 significa que duas imagens são iguais e 0 que estas são totalmente diferentes. Uma classificação de 0 apenas acontece quando as imagens comparadas são inversas, i.e., uma é o negativo da outra, situação em que se verifica a distância máxima entre duas componentes de cor.

Uma vez que os valores de MOS previamente recolhidos se encontravam numa escala diferente desta, i.e., $[0, 10]$, estes foram divididos por 10 de modo a torná-los compatíveis com as avaliações produzidas pelos algoritmos.

As avaliações produzidas pelos quatro algoritmos, bem como os valores de MOS normalizados encontram-se resumidos na Tabela 19.

# Par (k)	MOS/10	RMSE	UQI	SSIM	CBM
1	1.000	1.000	1.000	1.000	1.000
2	0.879	0.983	0.866	0.984	0.936
3	0.693	0.975	0.744	0.969	0.877
4	0.700	0.967	0.922	0.992	0.970
5	0.907	0.996	0.922	1.000	0.998
6	0.960	0.998	0.861	1.000	0.999
7	0.760	0.980	0.768	0.973	0.896
8	0.864	0.992	0.928	0.996	0.983
9	0.950	0.995	0.953	0.998	0.993
10	0.980	0.993	0.784	0.996	0.985
11	0.693	0.987	0.632	0.987	0.947
12	0.671	0.987	0.581	0.993	0.974
13	0.773	0.988	0.849	0.986	0.946
14	0.547	0.983	0.670	0.978	0.916
15	0.936	0.997	0.992	0.999	0.998
16	0.864	0.988	0.765	0.987	0.949
17	0.771	0.988	0.867	0.989	0.955
18	0.529	0.980	0.643	0.977	0.909
19	0.836	0.995	0.920	0.997	0.989
20	0.936	0.997	0.986	1.000	0.998
21	0.527	0.980	0.703	0.977	0.907
22	0.964	0.994	0.983	0.999	0.994
23	0.664	0.978	0.867	0.982	0.927
24	0.600	0.967	0.815	0.972	0.891
25	0.693	0.980	0.714	0.993	0.973
26	0.993	1.000	1.000	1.000	1.000
27	0.979	1.000	1.000	1.000	1.000
28	0.993	0.995	0.975	0.999	0.996
29	0.979	0.996	0.983	1.000	0.998
30	0.829	0.989	0.802	0.993	0.970

Tabela 19 – Avaliações produzidas pelos algoritmos RMSE, UQI, SSIM e CBM.

A Figura 59 apresenta o conjunto de projecções que permitem analisar graficamente a correlação existente entre os valores de MOS e os valores produzidos pelos vários algoritmos analisados.

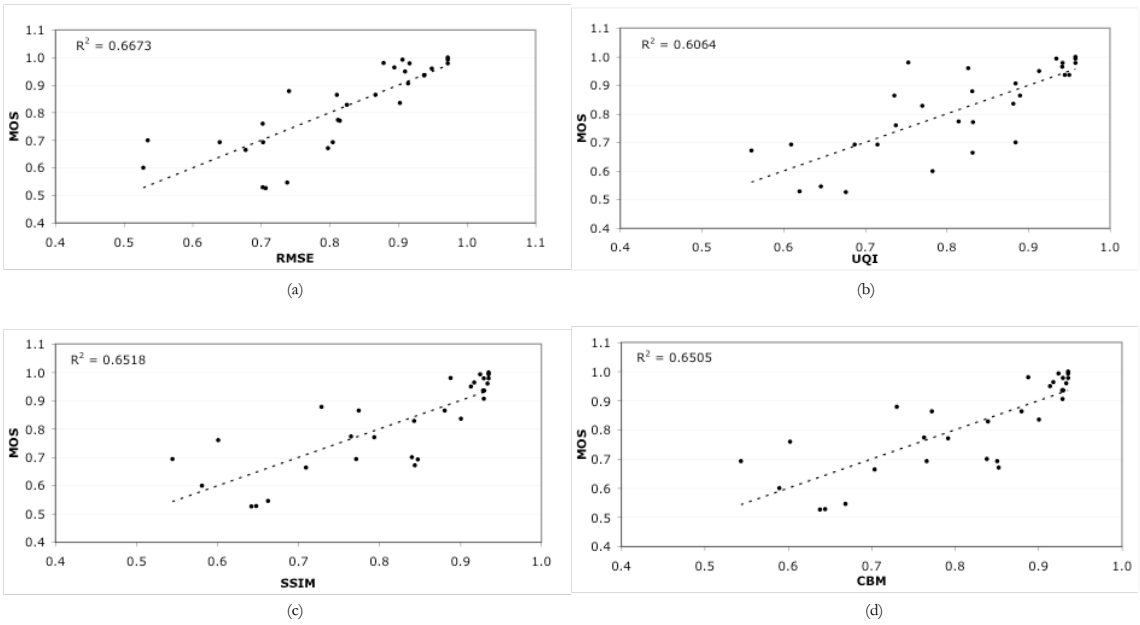


Figura 59 – Projecções de MOS com (a) RMSE, (b) UQI, (c) SSIM e (d) CBM.

Para efeitos de avaliação, o ITU recomenda que os valores produzidos pelos algoritmos de cálculo de similaridade sejam ajustados aos valores de MOS antes de se realizar uma análise comparativa. Este processo permite calibrar o algoritmo de similaridade mediante o tipo de imagem que se está a avaliar, fazendo com que este passe a produzir valores de similaridade mais próximos dos valores de MOS. Após o ajuste, os valores produzidos pelos algoritmos de similaridade tornam-se mais previsíveis e menos erráticos (Telecommunication Standardization Sector of ITU, 2004; Z. Wang & Bovik, 2002).

Recorrendo ao método dos mínimos quadrados, foram determinados os parâmetros necessários para aplicar a regressão linear aos valores produzidos pelos algoritmos (Telecommunication Standardization Sector of ITU, 2004). Os valores ajustados por regressão linear foram obtidos aplicando a fórmula $y = mx + b$, onde y representa o novo valor ajustado, x os valores a ajustar e m e b os parâmetros da função de ajuste.

A Tabela 20 apresenta os valores produzidos pelos algoritmos ajustados aos valores de MOS.

# Par (k)	MOS/10	RMSE m=13.426 b=-12.454	UQI m=0.946 b=0.011	SSIM m=12.541 b=-11.606	CBM m=3.187 b=-2.252
1	1.000	0.972	0.958	0.935	0.935
2	0.879	0.741	0.831	0.728	0.730
3	0.693	0.639	0.715	0.544	0.544
4	0.700	0.535	0.884	0.840	0.838
5	0.907	0.914	0.884	0.929	0.929
6	0.960	0.949	0.826	0.933	0.933
7	0.760	0.702	0.738	0.601	0.602
8	0.864	0.866	0.890	0.881	0.880
9	0.950	0.910	0.913	0.913	0.914
10	0.980	0.879	0.753	0.888	0.887
11	0.693	0.804	0.609	0.771	0.766
12	0.671	0.797	0.561	0.844	0.852
13	0.773	0.812	0.815	0.765	0.763
14	0.547	0.738	0.646	0.662	0.668
15	0.936	0.938	0.950	0.928	0.928
16	0.864	0.810	0.736	0.774	0.772
17	0.771	0.814	0.832	0.793	0.791
18	0.529	0.702	0.620	0.648	0.644
19	0.836	0.902	0.882	0.901	0.900
20	0.936	0.937	0.945	0.929	0.929
21	0.527	0.706	0.677	0.642	0.638
22	0.964	0.894	0.941	0.917	0.918
23	0.664	0.677	0.831	0.709	0.703
24	0.600	0.528	0.783	0.581	0.589
25	0.693	0.703	0.687	0.848	0.850
26	0.993	0.972	0.958	0.935	0.935
27	0.979	0.972	0.958	0.935	0.935
28	0.993	0.906	0.934	0.924	0.924
29	0.979	0.916	0.942	0.929	0.929
30	0.829	0.825	0.770	0.843	0.840

Tabela 20 – Valores de similaridade ajustados aos valores de MOS.

Estudo comparativo

Na sequência do anteriormente exposto, procedeu-se a uma análise comparativa das avaliações produzidas pelos intervenientes humanos (i.e., MOS) e as avaliações produzidas por cada um dos algoritmos propostos: RMSE, UQI, SSIM e CBM. O objectivo deste estudo foi identificar qual dos quatro algoritmos seria capaz de produzir valores de similaridade mais próximos da opinião média de um conjunto de intervenientes humanos (MOS).

O ITU combinou quatro documentos produzidos pelo Video Quality Experts Group (VQEG) e produziu um guia com recomendações para a elaboração de estudos de qualidade na área da compressão de vídeo. O documento produzido pelo ITU estabelece procedimentos e métricas a utilizar na avaliação de diferentes algoritmos de compressão recorrendo a métodos objectivos (algoritmos) e subjectivos (pessoas).

Seguindo as recomendações do ITU, o desempenho de cada um dos algoritmos foi determinado recorrendo a três critérios distintos (Telecommunication Standardization Sector of ITU, 2004):

1. **Precisão** – capacidade do algoritmo em produzir classificações próximas das classificações subjectivas humanas. Este critério é determinado calculando o coeficiente de correlação de Pearson entre as classificações produzidas por cada um dos algoritmos e o MOS (Métrica m_1).
2. **Monotonia** – grau de concordância entre o modelo de previsão automático (i.e., os valores produzidos pelos algoritmos) e as magnitudes relativas das classificações subjectivas atribuídas pelos humanos. Esta métrica é obtida calculando a correlação de Spearman entre os valores produzidos por cada um dos algoritmos e o MOS (Métrica m_2).
3. **Consistência** – capacidade de produzir previsões precisas ao longo de toda a experiência realizada. Esta métrica é obtida calculando a taxa de valores não-discrepantes produzidos pelos algoritmos, i.e., *non-outlier ratio* (Métrica m_3). A taxa de valores não-discrepantes é calculada recorrendo à Equação 2, onde um valor não-discrepante é aquele que pertence ao conjunto $[MOS - 2\sigma, MOS + 2\sigma]$ e N simboliza o número total de avaliações.

$$Taxa\ \ de\ \ não-discrepantes = \frac{\#\ \ não-discrepantes}{N}$$

Equação 2 – Taxa de valores não-discrepantes.

Resultados

A Tabela 21 apresenta os resultados obtidos após a aplicação de cada uma das métricas anteriormente descritas ao conjunto de avaliações produzidas pelos algoritmos considerados.

Algoritmos comparados	Correlação de Pearson (Métrica m_1)	Correlação de Spearman (Métrica m_2)	Percentagem de não-discrepantes (Métrica m_3)	Pontuação $m_1 \times m_2 \times m_3$
MOS-RMSE	**0.817**	**0.854**	**0.900**	**0.6279**
MOS-UQI	0.779	0.779	0.870	0.5280
MOS-SSIM	0.807	0.840	0.870	0.5898
MOS-CBM	0.807	0.840	0.870	0.5898

Tabela 21 – Desempenho dos vários algoritmos de cálculo de similaridade de imagem.

A pontuação final foi determinada multiplicando os resultados produzidos por cada uma das métricas utilizadas. É importante referir que os resultados produzidos pelas três métricas utilizadas pertenciam ao conjunto $[0, 1]$.

Os resultados obtidos nesta experiência são em boa medida surpreendentes. Todos os algoritmos revelaram uma elevada aptidão para determinar o nível de conformidade gráfica existente entre duas imagens. Em boa verdade, qualquer um dos quatro algoritmos poderia ser utilizado pelo `Object Evaluator`, pois serviria adequadamente o seu propósito. Tal já havia sido demonstrado pelo teste preliminar de Wilcoxon apresentado no Apêndice 8.4. Este teste estatístico teve como objectivo verificar se a distribuição das avaliações automáticas era estatisticamente semelhante aos valores de MOS recolhidos.

A pontuação final obtida para cada um dos algoritmos foi muito semelhante, no entanto, o algoritmo que apresentou os melhores resultados, em todas as métricas utilizadas, foi o RMSE. Este resultado contradiz um conjunto de publicações que, recorrendo a estudos semelhantes, concluem que os algoritmos UQI, SSIM e CBM (por ordem crescente) produzem melhores resultados que o simples RMSE (Gao et al., 2005; Y. Wang, 2006; Z. Wang & Bovik, 2002; Z. Wang et al., 2004).

Os algoritmos UQI, SSIM e CBM pertencem à classe de algoritmos baseados no sistema visual humano[117]. Este tipo de algoritmos pondera diferentes componentes da imagem geralmente associados à visão humana durante a sua análise de similaridade. Entre estes encontram-se a luminosidade, o contraste e a estrutura da imagem. Os algoritmos mais simples como o RMSE são agnósticos em relação à generalidade desses parâmetros e limitam-se a calcular matematicamente a distância cromatográfica entre os diferentes *pixel* das imagens.

Os resultados contraditórios que foram obtidos poderão ser justificados pelo tipo de colecção de teste utilizada. Os estudos realizados em torno dos algoritmos UQI, SSIM e CBM foram efectuados sobre colecções de teste que continham deformações profundas provocadas pela aplicação de filtros como ruído, *blur* ou mudanças radicais de cor e não pela introdução de ténues artefactos de compressão. Este tipo de algoritmos é muito eficaz na detecção de erros estruturais acentuados como os que se podem ver na Figura 60. Acontece que as deformações comummente introduzidas por aplicações de conversão não são desta natureza, mas sim caracterizadas pelo aparecimento de subtis artefactos de compressão ou alteração muito ligeira de cores devido a limitações do formato de destino.

---

[117] Do inglês *Human Visual System* (HVS).

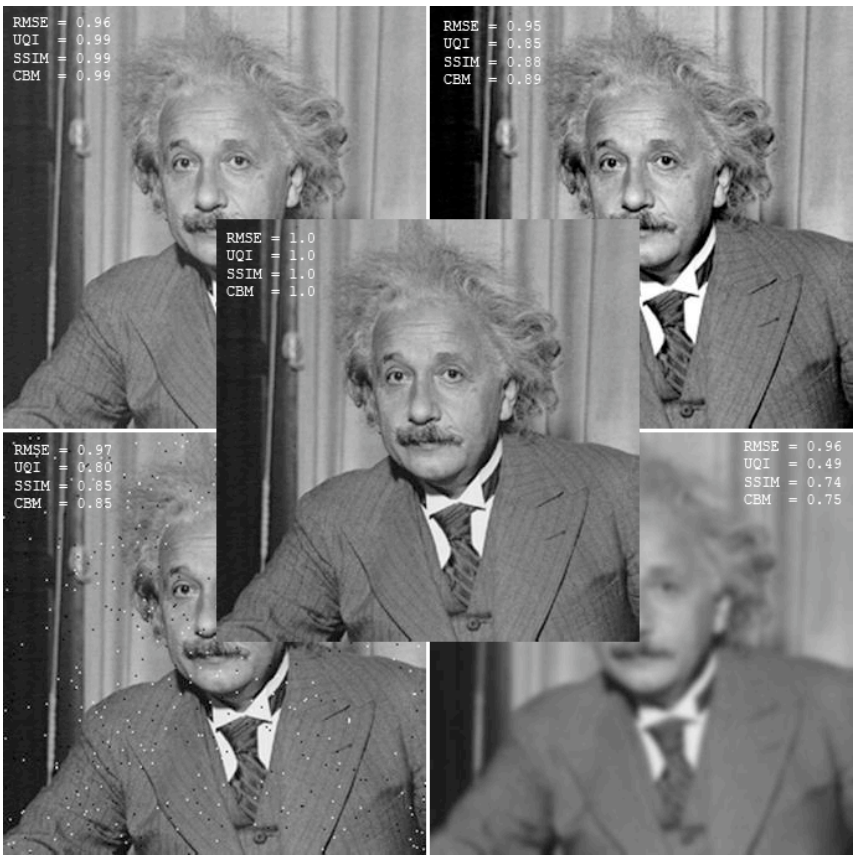


Figura 60 – Conjunto de imagens com RMSE≈0.96 e valores de UQI, SSIM e CBM distintos.

Sugama et al. salientam ainda a influência do tamanho da imagem nas avaliações subjectivas por seres humanos. Sugama et al. argumentam que quanto menores forem as imagens sob avaliação, maiores serão os valores de similaridade atribuídos pelos humanos, resultando em valores de MOS mais elevados.

A colecção de teste utilizada nesta experiência era composta por imagens cuja resolução variava entre o 1 *Megapixel* e os 8,5 *Megapixel*. Em todas as restantes experiências (Gao et al., 2005; Z. Wang & Bovik, 2002; Z. Wang et al., 2004), as colecções de teste eram compostas por imagens na ordem dos 0.3 *Megapixel*. Esse facto, por si só, poderá ter influenciado significativamente os resultados, lançando um novo olhar sobre a eficácia deste tipo de algoritmos perante imagens de grandes dimensões.

Para concluir, o RMSE revelou ser o algoritmo mais preciso na detecção da conformidade gráfica entre duas imagens. As vantagens da utilização deste algoritmo transcendem a sua

precisão. O RMSE é também o algoritmo mais simples de implementar e o que apresenta melhor performance computacional.

5.1.3 Propriedade significativa: metainformação embebida

Um número considerável de formatos de imagem contemplam a possibilidade de se transportar, juntamente com a informação que constitui a imagem, um conjunto de elementos de informação que têm como objectivo descrever e caracterizar a imagem veiculada. Esses elementos de informação e a informação gráfica que constitui a imagem encontram-se armazenados num único objecto digital. Devido a este facto, estes elementos de informação descritiva designam-se por metainformação embebida[118].

O formato TIFF, por exemplo, oferece a possibilidade de adicionar quaisquer elementos de metainformação a uma imagem codificada neste formato. No contexto do formato TIFF, estes elementos são designados por *private tags* (Adobe Developers Association, 1992).

Apesar de não existir propriamente uma norma universal que defina o conjunto de elementos de metainformação passíveis de serem utilizados na descrição de uma imagem, existem algumas especificações, criadas sobretudo por gigantes da indústria do processamento de imagem, que procuram introduzir alguma padronização no sentido de garantir a interoperabilidade entre aplicações de edição de imagem.

A especificação Exif (Exchangeable image file format), apesar de não ser governada por nenhuma organização de normalização, introduziu alguma regulamentação no que diz respeito a metainformação descritiva associada a imagens matriciais, sobretudo nos formatos de imagem JPEG e TIFF (Technical Standardization Committee on AV & IT Storage Systems and Equipment, 2002). A especificação Exif define um conjunto de atributos descritivos que procuram cobrir um largo espectro de casos de utilização. Entre estes encontram-se:

- Atributos que descrevem a data e a hora da aquisição da imagem;
- Atributos relacionados com o dispositivo de captura de imagem, tais como, orientação, nome do dispositivo, fabricante, abertura, velocidade de disparo, distância focal, modo de medição de luz, velocidade do sensor (ISO), etc.;
- Informação de autoria e direitos de autor;
- Atributos de georreferenciação.

---

[118] Do inglês *embedded metadata*.

A Adobe Systems Incorporated introduziu em 2001 um dialecto XML, designado *Extensible Metadata Platform* ou simplesmente XMP, que permite armazenar vários tipos de metainformação no interior de determinados formatos digitais, apenas especificando a sintaxe que deverá ser utilizada e não os elementos descritivos que poderão ser utilizados (Adobe Systems Incorporated, 2004). O XMP segue a sintaxe do dialecto RDF (RDF Core Working Group, 2004) e poderá ser embebido em diversos formatos, tais como: TIFF, JPEG, JPEG 2000, GIF, PNG, HTML, PDF, AI (Adobe Illustrator), SVG/XML, PSD (Adobe Photoshop), PostScript e EPS. O XMP é ainda compatível com o conjunto de atributos descritivos criado pelo International Press Telecommunications Council designado por IPTC Information Interchange Model (IIM), mais conhecido por IPTC headers (International Press Telecommunications Council, 2004), e também com a norma Exif. A norma IPTC IIM define um conjunto de atributos que têm como objectivo descrever objectos produzidos e trocados entre agências noticiosas. Este inclui também imagens, especialmente fotografias (Newspaper Association of America & International Press Telecommunications Council, 1999).

Para além dos elementos descritivos que poderão ser embebidos em imagens digitais, há ainda uma série de atributos de carácter técnico que são exclusivos de determinados formatos. Estes elementos são geralmente armazenados junto dos restantes atributos descritivos.

Tipo de falha	Descrição	Exemplo
Eliminação	Um atributo do objecto de partida não existe no objecto destino	Este tipo de falha é comum quando se efectuam conversões entre um formato que suporta determinados atributos descritivos e um formato que não os suporta, e.g. TIFF para BMP.
Modificação	Um atributo do objecto de partida existe no objecto destino, mas não foi correctamente transferido, i.e., é diferente do original	Certos conversores modificam alguns atributos embebidos nas imagens. Um exemplo típico deste tipo de falha ocorre no atributo "Application" do Exif. O conversor utilizado geralmente introduz a sua marca neste elemento para identificar a aplicação que gerou a nova imagem.
Inserção	Um atributo do objecto destino não existe no objecto de partida	Certas imagens transportam atributos técnicos específicos do formato em que se encontram. Ao converter para certos formatos, estes atributos são automaticamente preenchidos pela aplicação conversora. Nesta tese parte-se do pressuposto de que a introdução de atributos descritivos durante o processo de migração é prejudicial à autenticidade do objecto pois desrespeita o original e poderá introduzir informação que não é verdadeira ou de acordo com o original.

Tabela 22 – Tipos de falhas na metainformação embebida que poderão ocorrer durante uma conversão de formatos.

A experiência descrita nesta secção tem como objectivo determinar a capacidade do componente Object Evaluator em detectar e quantificar a deterioração ao nível da metainformação embebida, ocorrida durante um processo de conversão entre formatos. Neste

contexto, é possível identificar três tipos de falhas possíveis: eliminação, modificação e inserção. Estas encontram-se descritas em detalhe na Tabela 22.

Caracterização da colecção de teste

Para realizar esta experiência foi constituída uma colecção de teste composta por trinta novos pares de imagens contendo metainformação embebida e um número arbitrário de falhas dos três tipos anteriormente descritos.

Para preparar a colecção de teste recorreu-se a uma aplicação profissional de gestão de imagens designada Adobe Bridge. Esta aplicação permite editar os atributos das imagens definidos pelas seguintes normas: IPTC IIM *(Legacy* e *Core)*, Exif, GPS e Camera Raw. As imagens incluídas na colecção de teste encontravam-se codificadas em diversos formatos, nomeadamente, TIFF, JPEG 2000, BMP, JPG, GIF e PNG. A Tabela 23 resume os formatos e a percentagem de modificações introduzidas em cada par de imagens da colecção de teste.

# Par (*k*)	Formato original	Formato destino	# atributos preenchidos	# atributos modificados	% atributos não modificados
1	TIFF	TIFF	18	5	0.722
2	TIFF	JPEG 2000	18	5	0.722
3	TIFF	BMP	18	18	0.000
4	TIFF	JPG	14	3	0.786
5	TIFF	GIF	11	11	0.000
6	TIFF	JPEG 2000	14	4	0.714
7	TIFF	TIFF	38	8	0.789
8	TIFF	JPEG	38	16	0.579
9	TIFF	PNG	38	38	0.000
10	JPEG	TIFF	29	10	0.655
11	JPEG	JPEG	29	11	0.621
12	JPEG	TIFF	21	3	0.857
13	JPEG	TIFF	25	5	0.800
14	JPEG	JPEG 2000	24	3	0.875
15	JPEG	TIFF	25	5	0.800
16	JPEG	JPEG	35	5	0.857
17	JPEG	TIFF	37	8	0.784
18	JPEG	JPEG	35	5	0.857
19	JPEG	JPEG	27	2	0.926
20	JPEG	TIFF	27	1	0.963
21	JPEG	TIFF	27	1	0.963
22	JPEG	JPEG	42	16	0.619
23	JPEG	TIFF	41	15	0.634
24	JPEG	JPEG	53	29	0.453
25	TIFF	JPEG	12	5	0.583
26	TIFF	TIFF	9	6	0.333
27	TIFF	TIFF	22	19	0.136
28	JPEG	JPEG	10	6	0.400
29	JPEG	TIFF	11	6	0.455
30	JPEG	JPEG	10	3	0.700

Tabela 23 – Colecção de teste utilizada na experiência com metainformação embebida.

A percentagem de atributos não modificados apresentada na Tabela 23 representa o nível de similaridade percepcionado entre duas imagens da colecção de teste.

Avaliação automática

Após a preparação da colecção de teste e do cálculo da percentagem de elementos de metainformação que não sofreram alterações, i.e., o nível de similaridade detectado manualmente para a propriedade significativa metainformação embebida, procedeu-se à avaliação automática da colecção de teste.

Foram testados dois métodos diferentes de extrair metainformação das imagens e subsequentemente duas formas de comparar essa mesma informação. O primeiro método extrai o documento XMP embebido no interior das imagens. Uma vez que o XMP é baseado em XML, foi utilizado um método de cálculo de similaridade entre documentos XML designado XML Diff[119] desenvolvido na Universidade de Sannio (Canfora, Cerulo, & Scognamiglio, 2004).

O segundo método de extracção utilizado recorre a uma aplicação designada ExifTool[120] que devolve os atributos encontrados no interior da imagem sob a forma de um conjunto de pares do tipo atributo/valor (Harvey, 2003). Para calcular a similaridade entre os conjuntos de propriedades extraídos recorreu-se ao método designado Coeficiente de Similaridade de Jaccard[121] (Jaccard, 1901; Tan, Steinbach, & Kumar, 2005). Esta métrica é amplamente utilizada para calcular a similaridade entre dois conjuntos de elementos (Xiao, Wang, Lin, & Yu, 2008).

O Coeficiente de Similaridade de Jaccard é obtido dividindo o número de elementos pertencentes à intersecção dos dois conjuntos comparados pelo número de elementos pertencentes à sua reunião (Equação 3).

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Equação 3 – Coeficiente de Similaridade de Jaccard.

[119] Para mais informações sobre este algoritmo de similaridade, consulte o Apêndice 8.3.5 na página 211.
[120] Para mais informações sobre esta ferramenta, consulte o Apêndice 8.1.2 na página 198.
[121] Para mais informações sobre este algoritmo de similaridade, consulte o Apêndice 8.3.4 na página 210.

Os resultados produzidos por ambas as abordagens encontram-se resumidos na Tabela 24. Um valor de 1 significa que dois conjuntos são iguais, enquanto que um valor de 0 significa que os dois conjuntos não possuem qualquer elemento em comum.

# Par (k)	% atributos não alterados	XML Diff	Jaccard
1	0.722	0.789	0.591
2	0.722	0.750	0.522
3	0.000	0.017	0.000
4	0.786	1.000	0.786
5	0.000	0.027	0.000
6	0.714	0.936	0.714
7	0.789	0.818	0.738
8	0.579	0.631	0.585
9	0.000	0.008	0.000
10	0.655	0.905	0.545
11	0.621	0.968	0.625
12	0.857	0.873	0.680
13	0.800	0.807	0.606
14	0.875	0.800	0.724
15	0.800	0.807	0.625
16	0.857	0.858	0.795
17	0.784	0.818	0.636
18	0.857	0.858	0.795
19	0.926	0.848	0.767
20	0.963	0.952	0.867
21	0.963	0.952	0.867
22	0.619	0.767	0.538
23	0.634	0.717	0.532
24	0.453	0.783	0.406
25	0.583	0.844	0.538
26	0.333	0.367	0.300
27	0.136	0.533	0.125
28	0.400	0.613	0.364
29	0.455	0.738	0.385
30	0.700	0.912	0.636

Tabela 24 – Resultados produzidos pelos métodos XML Diff e Jaccard.

Estudo comparativo

Após reunidos os valores de similaridade produzidos pelos dois métodos analisados, XML Diff e Coeficiente de Similaridade de Jaccard, procedeu-se a um estudo comparativo com o objectivo de determinar qual dos dois algoritmos apresentava o melhor desempenho na detecção de falhas na metainformação embebida.

Para efeitos de avaliação, foram utilizadas três métricas distintas:

1. **Precisão** – capacidade demonstrada pelo algoritmo de similaridade para quantificar o nível de falhas introduzidas na colecção de teste. Este critério foi determinado calculando o coeficiente de correlação de Pearson entre os valores produzidos por cada um dos algoritmos considerados e a percentagem de atributos não modificados, i.e., valores de referência (Métrica m_1);
2. **Monotonia** – grau de concordância entre os valores produzidos pelos algoritmos de similaridade e as magnitudes relativas dos valores de referência. Esta métrica foi obtida calculando a correlação de Spearman (Métrica m_2);
3. **Erro médio** – um modelo de previsão é tanto mais eficiente quanto menor for o erro por este apresentado em relação ao valor real utilizado como referência (Fernandes, 1999). Esta métrica é obtida recorrendo à Média do Quadrado do Erro (Métrica m_4).

Resultados

A Tabela 25 apresenta os resultados da aplicação das métricas anteriormente descritas aos valores produzidos pelos dois algoritmos analisados.

Métodos comparados	Correlação de Pearson (Métrica m_1)	Correlação de Spearman (Métrica m_2)	Média do Quadrado do Erro (Métrica m_4)
Referência-xmldiff	0.886	0.713	0.028
Referência-Jaccard	**0.978**	**0.917**	**0.010**

Tabela 25 – Desempenho dos dois métodos de cálculo de similaridade de metainformação embebida.

Ao observar a Tabela 25 conclui-se que, na globalidade das métricas consideradas, o desempenho do método de Jaccard foi superior ao do XML Diff.

Apesar do elevado desempenho demonstrado por ambos os métodos, foi importante explorar as razões que levaram a que estes não apresentassem um comportamento irrepreensível, i.e., que os resultados da sua avaliação não tenham sido mais próximos do valor máximo admitido. À primeira vista nada impedia que ambos os métodos fossem capazes de quantificar exactamente a percentagem de falhas introduzidas na metainformação embebida das imagens analisadas.

No caso do método XML Diff, esta análise foi difícil de realizar. O algoritmo pondera três critérios distintos na sua análise de similaridade: conteúdo, estrutura e posicionamento dos elementos de informação no interior das árvores XML comparadas. Apesar do XML sob

avaliação ser morfologicamente simples, quando foi comparado com o documento XML de referência, apresentou uma correlação inferior a 0.90. Para compreender as razões que levaram este algoritmo a apresentar uma correlação inferior a 1, seria necessário inspeccionar os seus processos internos. No entanto, uma vez que o método de Jaccard apresentava de antemão níveis de desempenho superiores, optou-se por investir mais tempo no aprofundamento de conhecimento sobre este método em detrimento do primeiro.

No caso do método Jaccard, o motivo que levou a que fossem obtidos níveis de correlação inferiores a 1, teve que ver com os dados em si. A similaridade de Jaccard, está preparada para operar sobre conjuntos de elementos. Acontece que os elementos pertencentes aos conjuntos avaliados eram compostos por pares do tipo atributo/valor. A reunião de conjuntos compostos por elementos deste tipo produzia um resultado erróneo na presença de elementos que possuíam o mesmo nome de atributo mas valores distintos.

Considere-se o seguinte exemplo onde se calcula o coeficiente de Jaccard entre dois conjuntos, A e B, constituídos por pares de elementos do tipo (atributo, valor), com a_i a representar o nome do atributo e v_i a representar o seu valor associado (Equação 4).

$$A = \{(a_1,v_1),(a_2,v_2),(a_3,v_3)\}$$
$$B = \{(a_1,v_1),(a_2,v_2),(a_3,v_4)\}$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|\{(a_1,v_1),(a_2,v_2)\}|}{|\{(a_1,v_1),(a_2,v_2),(a_3,v_3),(a_3,v_4)\}|} = \frac{2}{4}$$

Equação 4 – Exemplo da aplicação do coeficiente de Jaccard.

Como se pode constatar, apesar do atributo a_3 existir em ambos os conjuntos, este é contabilizado duas vezes após a reunião de conjuntos. Esta abordagem está matematicamente correcta uma vez que $(a_3,v_3) \neq (a_3,v_4)$, no entanto, não é ajustada ao domínio de aplicação. Assim, propôs-se a seguinte alteração ao `Coeficiente de Jaccard` e a introdução de uma nova função designada *first* que opera sobre conjuntos de pares, transformando-os em conjuntos de elementos singulares constituídos apenas pelo primeiro elemento de cada par. A função *first* e a versão modificada do `Coeficiente de Jaccard` encontram-se definidas na Equação 5.

$$first(S) = \{x\} : \forall (x,y) \in S$$

$$J'(A,B) = \frac{|A \cap B|}{|first(A) \cup first(B)|}$$

Equação 5 – Definição da função *first* e versão modificada do método de Jaccard.

A aplicação da nova métrica ao conjunto de dados conduziu aos resultados apresentados na Tabela 26. Para efeitos de comparação foram novamente representados os valores produzidos pelos anteriores métodos analisados.

# Par (*k*)	% atributos não alterados	XML Diff	Jaccard	Jaccard modificado
1	0.722	0.789	0.591	0.722
2	0.722	0.750	0.522	0.667
3	0.000	0.017	0.000	0.000
4	0.786	1.000	0.786	0.786
5	0.000	0.027	0.000	0.000
6	0.714	0.936	0.714	0.714
7	0.789	0.818	0.738	0.795
8	0.579	0.631	0.585	0.615
9	0.000	0.008	0.000	0.000
10	0.655	0.905	0.545	0.600
11	0.621	0.968	0.625	0.645
12	0.857	0.873	0.680	0.773
13	0.800	0.807	0.606	0.714
14	0.875	0.800	0.724	0.778
15	0.800	0.807	0.625	0.714
16	0.857	0.858	0.795	0.838
17	0.784	0.818	0.636	0.718
18	0.857	0.858	0.795	0.838
19	0.926	0.848	0.767	0.821
20	0.963	0.952	0.867	0.929
21	0.963	0.952	0.867	0.929
22	0.619	0.767	0.538	0.651
23	0.634	0.717	0.532	0.595
24	0.453	0.783	0.406	0.481
25	0.583	0.844	0.538	0.583
26	0.333	0.367	0.300	0.333
27	0.136	0.533	0.125	0.136
28	0.400	0.613	0.364	0.400
29	0.455	0.738	0.385	0.455
30	0.700	0.912	0.636	0.700

Tabela 26 – Resultados produzidos pelo método de Jaccard modificado.

A aplicação do conjunto de métricas de avaliação de desempenho previamente descritas ao novo método de cálculo de similaridade (i.e., Jaccard') veio demonstrar que as modificações efectuadas ao coeficiente de Jaccard vieram melhorar significativamente o seu desempenho

Tabela 27). Apesar de continuar a não apresentar um desempenho perfeito, encontra-se agora muito próximo deste.

Métodos comparados	Correlação de Pearson (Métrica m_1)	Correlação de Spearman (Métrica m_2)	Média do Quadrado do Erro (Métrica m_4)
Referência-XML Diff	0.886	0.713	0.028
Referência-Jaccard	0.978	0.917	0.010
Referência-Jaccard'	**0.991**	**0.960**	**0.002**

Tabela 27 – Desempenho dos dois métodos de cálculo de similaridade de metainformação embebida.

Os casos em que se verificou que os valores produzidos pelo método de Jaccard' não coincidiram com os valores de referência foram analisados manualmente. Estes diferiam, sobretudo, devido a algumas funcionalidades incluídas na aplicação utilizada durante a construção da colecção de teste (i.e., Adobe Bridge) que não se encontravam na ferramenta de extracção de metainformação das imagens (i.e., ExifTool). Por exemplo, um dado par de imagens apresentava os seguintes valores para o atributo Date Time Original: "2003:11:25 12:59:58Z" e "2003:11:25 12:59:58". A primeira, inclui informação sobre fuso horário, representada pela letra "Z" no final do componente representativo da hora. A letra "Z" indica que aquela data/hora pertence ao fuso horário da zona Z, ou seja, ao referencial horário que coincide com o GMT (Tempo Médio de Greenwich). A ferramenta ExifTool assinalou correctamente a diferença existente entre a metainformação embebida em ambas as imagens, no entanto, Adobe Bridge assumiu o fuso horário GMT no caso omisso, fazendo com que ambas as datas/horas fossem consideradas iguais.

Houve, ainda, outros exemplos de assumpção automática de valores por parte da aplicação Adobe Bridge. Por exemplo, quando o valor do atributo *Sharpness* é omisso, esta assume o valor "Normal". O mesmo acontece com o atributo *White Balance* que, quando omisso, é definido como "Auto". A aplicação ExifTool não assume quaisquer valores por omissão. Limita-se a extrair a metainformação encontrada no interior das imagens. Se um dado atributo não existe numa imagem, é retornado o valor nulo e não um outro valor por omissão.

Apesar das falhas detectadas, considerou-se que o desempenho do método de Jaccard' é suficientemente elevado para ser utilizado no contexto do Object Evaluator.

5.2 Avaliação do Migration Advisor

Sempre que é realizada uma conversão no contexto do CRiB, é produzido um relatório que descreve a aptidão para preservação do caminho de migração utilizado. Esse relatório inclui

informação sobre a performance do processo de migração, o grau de degradação incorrido ao nível das propriedades significativas dos objectos digitais e a adequabilidade dos formatos envolvidos para efeitos de preservação digital.

Estes relatórios são armazenados e posteriormente utilizados pelo componente Migration Advisor para determinar o caminho de migração mais apto para preservar uma dada classe de objectos digitais. O componente tem em consideração o desempenho demonstrado por cada caminho de migração e os requisitos particulares de cada utilizador.

Há utilizadores, por exemplo, que valorizam mais a conservação das propriedades significativas e menos a performance de conversão. Essas escolhas influenciam a ordem pela qual os diferentes caminhos de conversão são ordenados e sugeridos ao utilizador.

Cada utilizador pode manifestar as suas preferências atribuindo pesos aos critérios de avaliação suportados pelo CRiB (ver secção 4.6.2 na página 105). Com base nessa informação e no conjunto de relatórios de avaliação produzidos automaticamente pelo sistema, é possível determinar, de entre dezenas de alternativas, qual o formato de destino e o caminho de conversão mais favoráveis para preservar uma dada colecção de objectos digitais.

Para validar a eficácia deste sistema de recomendação foi utilizada uma técnica de validação designada k-fold cross-validation. Esta técnica consiste em particionar um conjunto de dados de teste em *k* partes de igual dimensão. Das *k* partes, *k*-1 são utilizadas para treinar o sistema, enquanto que a partição remanescente é utilizada para o testar. O processo de validação consiste em verificar se o sistema é capaz de recomendar correctamente o conjunto de dados de teste, baseando-se apenas nos dados utilizados para treino. O processo é repetido *k* vezes, alternando a partição de teste. No final, é calculada a média dos resultados obtidos em cada uma das *k* avaliações realizadas (Witten & Frank, 2005). Mais informação sobre esta técnica de validação encontra-se disponível no Apêndice 8.5 na página 219.

A validação do componente Migration Advisor seguiu o seguinte protocolo experimental:

1. **Construção de uma colecção de teste** – para realizar a experiência foi necessário construir uma colecção de objectos de treino/teste suficientemente grande e heterogénea para que o treino do sistema pudesse ser considerado eficaz. Todos os objectos incluídos na colecção de teste teriam obrigatoriamente de ser do mesmo formato.

2. **Selecção de conversores** – antes de se dar início ao processo de avaliação do sistema de recomendação foi identificado um subconjunto relevante de caminhos de conversão a considerar durante a experiência. Este subconjunto de caminhos de conversão teve como invariante o formato de partida, uma vez que este teria necessariamente que coincidir com o formato dos objectos que compunham a colecção de teste.

3. **Particionamento da colecção de teste** – o conjunto de objectos de teste foi dividido em 10 partições construídas aleatoriamente (i.e., *10-fold cross-validation*). Em cada iteração foram utilizados 90% dos objectos para treinar o sistema e 10% para testar as suas recomendações. O processo de avaliação foi composto por 10 iterações (*10 folds*), fazendo-se alternar as partições de treino/teste. A avaliação final do sistema consistiu no cálculo da média dos resultados obtidos em cada uma das 10 iterações.

4. **Treino do sistema** – o treino do sistema consistiu na conversão de todos os objectos pertencentes à partição de treino, recorrendo a todos os caminhos de migração seleccionados no ponto 1, e armazenando no Evaluations Repository os relatórios produzidos pelos vários componentes de avaliação, i.e., Migration Broker, Object Evaluator e Format Evaluator. Os relatórios de avaliação constituem a matéria-prima que permite ao sistema produzir recomendações.

5. **Teste do sistema** – após o treino do sistema, a sua precisão foi determinada comparando as recomendações produzidas pelo Migration Advisor (com base nos relatórios de migração coleccionados durante o treino) com a alternativa de migração que efectivamente apresentou melhor qualidade de serviço. Esta alternativa foi determinada, a-pedido, para cada um dos objectos que constituía a colecção de teste. Isto consistiu em converter cada um dos objectos de teste recorrendo a todas as alternativas de migração disponíveis na plataforma, avaliar cada uma das conversões efectuadas e identificar, de entre estas, qual a melhor opção. Posteriormente, o *ranking ideal* de opções foi comparado com o *ranking* recomendado pelo sistema. Um conjunto de métricas de comparação de *rankings* foi utilizado para aferir a precisão e a exactidão do sistema de recomendação. Nas experiências realizadas foi atribuído o mesmo peso a todos os critérios da taxionomia geral de avaliação.

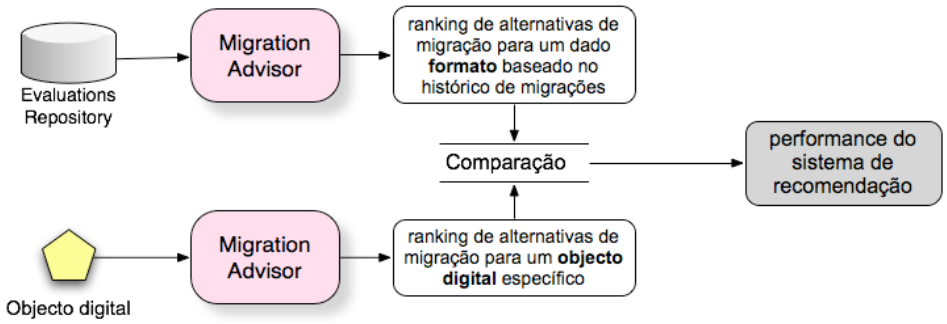


Figura 61 – Teste do sistema de recomendação.

A Figura 61 apresenta os vários passos que compuseram o processo de avaliação do componente Migration Advisor. O processo começa por recorrer ao Evaluations Repository para obter os vários relatórios de migração armazenados durante a fase de treino do sistema. Depois, utilizando essa informação, o Migration Advisor determina para cada formato qual o *ranking* das melhores alternativas de migração, i.e., aquelas que garantem o maior nível de preservação. De seguida, um dado objecto digital é convertido recorrendo a cada uma das alternativas de migração disponíveis na plataforma. Durante estas conversões são também produzidos, porém não armazenados, relatórios de avaliação que relatam o desempenho de cada um dos caminhos de migração utilizados. Esses relatórios são, então, utilizados para produzir um novo *ranking*, o *ranking ideal* das opções de migração. Para avaliar a qualidade das recomendações produzidas pelo Migration Advisor, o *ranking* recomendado é comparado com o *ranking ideal*. O valor que daí resultar determina a performance do sistema de recomendação.

5.2.1 Caracterização da colecção de teste

Um dos formatos de imagem mais utilizados é o JPEG File Interchange Format. O elevado nível de compressão característico deste formato torna-o adequado a transferências de dados através da Internet. Este facto levou os construtores de *Web browsers* a adoptá-lo e a torná-lo num dos formatos mais disseminados do mundo. A recente adesão ao JPEG por parte dos fabricantes de máquinas fotográficas digitais veio fortalecer ainda mais a sua prevalência sobre todos os restantes formatos de imagem.

Um estudo estatístico realizado sobre material publicado em repositórios digitais de Acesso-livre mostra que o JPEG é claramente o formato de imagem mais utilizado para veicular informação de carácter gráfico. Logo de seguida encontra-se o Tagged Image File Format, vulgarmente conhecido por TIFF (University of Southampton, 2007).

Acontece que o formato JPEG possui uma característica que o debilita como formato de preservação. Apesar de ser um formato amplamente disseminado, o que deve ser visto com uma vantagem, o método de compressão que implementa introduz deformações na informação gráfica que constitui a imagem. Neste contexto, procurou-se determinar qual o formato alternativo mais adequado para preservar imagens originalmente produzidas em formato JPEG. Esta experiência serviu também o propósito de avaliar o componente Migration Advisor e a sua capacidade de recomendar opções de migração.

Para dar inicio à experiência, foi reunido um conjunto de imagens em formato JPEG 1.02. Estas imagens foram recolhidas junto de repositórios digitais (sobretudo repositórios institucionais) efectuando pesquisas através das suas interfaces OAI-PMH. O projecto Preserv da University of Southampton foi de extrema importância neste processo, pois permitiu obter ligações directas para todos os objectos digitais de um dado formato partilhados através deste tipo de repositórios (University of Southampton, 2007).

Junto destes repositórios foram recolhidas 8071 imagens em formato JPEG num total de 7.3 GB. A partir destas imagens foram criados subconjuntos de cardinalidade 10, 20, 50 e 100. Cada um destes subconjuntos foi utilizado para treinar e testar o sistema de recomendação recorrendo ao método de validação cruzada descrito anteriormente. As imagens pertencentes a cada um destes conjuntos foram seleccionadas aleatoriamente a partir do conjunto inicial de 8071 imagens. Este plano de treino teve como objectivo verificar se um aumento da cardinalidade da colecção de treino resultaria num aumento da precisão e/ou exactidão do sistema de recomendação.

Cada um dos subconjuntos de objectos digitais utilizados durante o treino/teste do sistema encontra-se descrito na Tabela 28.

#	Descrição da colecção	Dimensão das imagens (Kb)			
		Min	Max	Média	σ
10	Fotografias obtidas a partir de câmaras digitais ilustrando monumentos e peças museológicas.	869	1263	1053	149
20	Fotografias a preto e branco digitalizadas, fotografias digitais de interiores de igrejas, fotografias digitais de exteriores.	15	1274	925	353
50	Fotografias a preto e branco digitalizadas, fotografias digitais de peças museológicas, fotografias digitais de vitrais, monumentos, pinturas e iluminuras.	11	1278	945	378
100	Fotografias a preto e branco digitalizadas, fotografias digitais de peças museológicas, fotografias digitais de vitrais, monumentos, pinturas, iluminuras e cartazes.	16	1278	996	324

Tabela 28 – Descrição das colecções de imagens utilizadas na avaliação do componente Migration Advisor.

5.2.2 Selecção de caminhos de conversão

Para dar início à avaliação do sistema de recomendação foi necessário seleccionar um subconjunto dos vários caminhos de migração passíveis de ser utilizados na respectiva experiência. Uma vez que os objectos incluídos na colecção de teste se tratavam de imagens em formato JPEG File Interchange Format 1.02, os conversores seleccionados teriam obrigatoriamente de suportar este formato.

Entre os caminhos de conversão testados encontrava-se o Conversor-identidade, i.e., um pseudo-conversor que não efectua qualquer tipo de transformação nos objectos digitais, mas que força os mesmos a serem avaliados pelos componentes Migration Broker, Format Evaluator e Object Evaluator. Este conversor especial é utilizado para avaliar o risco incorrido em manter os objectos digitais no seu formato original (i.e., não realizar qualquer migração).

Ao utilizar o Conversor-identidade, os únicos critérios que serão efectivamente avaliados são aqueles que dizem respeito às características técnicas do formato uma vez que a performance de conversão é sempre máxima (os objectos não chegam a ser convertidos, logo, vence a todas alternativas no que toca ao desempenho de conversão) e não há perda de informação (os objectos resultantes da conversão são iguais aos objectos de partida).

A Tabela 29 descreve os caminhos de conversão utilizados na experiência realizada.

Caminho de conversão	Formato de chegada	Descrição
[JPG2TIF]	Tagged Image File Format, version 3	Serviço de conversão suportado pela aplicação ImageMagick.
[JPG2JP2]	JPEG 2000	Serviço de conversão suportado pela aplicação ImageMagick.
[JPG2PNG]	Portable Network Graphics, version 1.0	Serviço de conversão suportado pela aplicação ImageMagick.
[JPG2BMP]	Windows Bitmap, version 3.0	Serviço de conversão suportado pela aplicação ImageMagick.
[JPG2TIF_2]	Tagged Image File Format, version 3	Serviço de conversão suportado pela aplicação "sam2p".
[JPG2BMP] > [BMP2TIF]	Tagged Image File Format, version 3	Serviço de conversão suportado pela composição dos serviços de migração JPG2BMP e BMP2TIF anteriormente descritos.
Conversor-identidade	JPEG File Interchange Format 1.02	A conversão identidade permite ao sistema de recomendação avaliar o potencial em termos de preservação a longo-prazo de não realizar qualquer conversão, i.e., avaliar o risco da não migração.

Tabela 29 – Caminhos de conversão utilizados na avaliação do Migration Advisor.

Após cada conversão, todos serviços de migração, assim como formatos e objectos envolvidos foram avaliados pelos componentes do CRiB responsáveis pelo controlo de qualidade e os resultados dessas avaliações foram armazenados no Evaluations Repository.

É importante referir que estas avaliações foram realizadas por caminho de migração e não por conversão individual, i.e., no caso de migrações compostas por mais do que um conversor (e.g. JPG-BMP seguido de BMP-TIF) as avaliações foram efectuadas entre os objectos/formatos de partida e os objectos/formatos de chegada (i.e., entre JPEG e TIF). Os objectos/formatos intermédios (i.e., BMP) não foram avaliados. Em suma, uma migração composta por vários conversores é vista pelo sistema como uma conversão atómica.

5.2.3 Treino e teste do sistema

O sistema de recomendação foi testado recorrendo ao método de validação cruzada com $K = 10$ (*10 folds*). Foram utilizadas quatro colecções de teste distintas com cardinalidade 10, 20, 50 e 100 com o objectivo de verificar se um aumento da cardinalidade da colecção de treino resultaria num aumento da precisão e/ou exactidão do sistema de recomendação.

A Tabela 30 apresenta o número de conversões efectuadas para treinar/testar o sistema tendo em consideração cada uma das colecções de teste. A tabela apresenta ainda os tempos de conversão e avaliação dos objectos em questão.

#	**Nº de conversões**	**Tempo** (horas)	**Tempo médio** (minutos/objecto)
10	600	06h06	0.55
20	1200	11h03	0.55
50	3000	27h03	0.54
100	6000	58h02	0.58

Tabela 30 – Dados relativos ao treino e teste do componente Migration Advisor.

O número de conversões efectuadas para uma dada colecção de teste é descrito pela Equação 6, onde n representa a cardinalidade da colecção de teste, M o número de conversores utilizados e K o número de partições utilizadas na validação cruzada. Neste caso concreto, $M = 6$ (o conversor identidade não foi contabilizado uma vez que não possui tempo de conversão) e $K = 10$.

$$c(n) = n \times M \times K$$

Equação 6 – Número de conversões mediante o tamanho da colecção de teste.

Quando o Migration Advisor é questionado, este devolve uma lista ordenada (i.e., *ranking*) com as alternativas de migração mais favoráveis para preservar a longo-prazo um dado formato digital tendo em consideração todas as avaliações realizadas no passado pelo sistema. A ordenação das alternativas depende dos pesos atribuídos pelo utilizador aos critérios suportados pelo sistema. Nas experiências realizadas neste contexto foi atribuído o mesmo peso a cada critério que constitui a taxionomia geral de avaliação.

Em termos gerais, um sistema de recomendação pode ser avaliado em termos da sua exactidão e da sua precisão.

As métricas de exactidão têm como função medir empiricamente a proximidade existente entre um *ranking* de itens produzido por um sistema de recomendação e o *ranking* de itens tomado como referência (Herlocker, Konstan, Terveen, & Riedl, 2004). As métricas de exactidão preocupam-se com a ordem dos itens e não com a pontuação que cada um desses itens apresenta no *ranking*.

As métricas de precisão, por sua vez, procuram determinar se um sistema de recomendação é capaz de prever a pontuação que um dado item irá receber no *ranking* de referência, i.e., comparam a pontuação prevista para um item do *ranking* com a pontuação real desse item (Herlocker et al., 2004).

Existem diversas métricas que poderão ser utilizadas para determinar a exactidão de um sistema de recomendação, tais como: a correlação de Pearson, a correlação de Spearman ou a correlação de Kendall Tau (Herlocker et al., 2004). A correlação de Pearson permite determinar se existe uma relação linear entre a pontuação prevista de um dado item no *ranking* (i.e., a pontuação do item no *ranking* recomendado) e a pontuação real que esse item deveria assumir (i.e., a pontuação do item no *ranking* de referência). As correlações de Spearman e Kendall Tau são correlações entre *rankings*, i.e., permitem determinar em que medida dois *rankings* são concordantes independentemente das pontuações assumidas por cada um dos itens que os constituem (i.e., comparam a posição de cada item no *ranking* e não as suas pontuações). A correlação de Kendall Tau é superior à de Spearman no que toca à comparação de *rankings* onde existem empates nas pontuações dos seus itens (Herlocker et al., 2004).

As principais vantagens deste tipo de métricas são: facilitar a comparação de sistemas de recomendação que possuam escalas de pontuação distintas, serem bem conhecidas da comunidade científica e permitirem obter um valor único de avaliação balizado entre dois valores, geralmente -1 e 1.

Para determinar a precisão de um sistema de recomendação poderão ser utilizadas métricas como a Mean Absolute Error (MAE) ou alguma das suas variantes: Mean Squared Error (MSE), Root Mean Squared Error (RMSE) e Normalized Mean Squared Error (NMSE) (Herlocker et al., 2004).

A MSE e a RMSE diferem da MAE por elevarem o erro ao quadrado antes de o agrupar, o que significa que erros superiores irão penalizar mais a avaliação final. A NMSE normaliza os resultados tendo em consideração o domínio dos valores de entrada, permitindo desta forma que os resultados obtidos possam ser comparados com os resultados de outras experiências (Goldberg, Roeder, Gupta, & Perkins, 2001).

O Migration Advisor foi avaliado quanto à sua exactidão segundo as métricas de correlação de Pearson, Spearman e Kendall Tau; e quanto à sua precisão segundo o Normalized Mean Squared Error (NMSE).

5.2.4 Resultados

Os resultados obtidos após o treino do sistema com as colecções de teste anteriormente descritas encontram-se resumidos na Tabela 31.

	Exactidão			Precisão
#	**Pearson**	**Spearman**	**Kendall Tau**	**NMSE**
10	0.869	0.918	0.829	0.197
20	0.828	0.832	0.729	0.223
50	0.682	0.817	0.731	0.276
100	0.757	0.852	0.754	0.254

Tabela 31 – Resultados da validação cruzada efectuada ao Migration Advisor.

Os resultados demonstram que a qualidade geral das recomendações produzidas pelo Migration Advisor é elevada. Todas as colecções de teste utilizadas no treino do sistema resultaram em níveis de correlação superiores a 0.68, sendo na sua maioria superiores a 0.8. O erro de precisão máximo verificado foi de 28%.

A colecção de teste que apresentou melhores resultados foi, curiosamente, a mais pequena, i.e., a colecção de cardinalidade 10. Este facto adveio de um enviesamento imprevisto pelos próprios objectos digitais que constituíram a colecção. É importante relembrar que as colecções de treino/teste foram construídas aleatoriamente a partir de objectos recolhidos junto de repositórios internacionais. Por coincidência, a colecção de teste de cardinalidade 10 era constituída por um conjunto de objectos cujas dimensões possuíam pouca variabilidade,

i.e., o desvio-padrão em relação à média das dimensões era cerca de metade do desvio apresentado pelas restantes colecções de teste. Para determinar a validade desta hipótese, preparou-se manualmente uma colecção de teste constituída por objectos manifestamente diferentes. A nova colecção de teste de cardinalidade 10 encontra-se descrita na Tabela 32.

		Dimensão das imagens (Kb)			
#	Descrição	Min	Max	Média	σ
10_2	Fotografias tiradas com câmaras digitais de paisagens, interiores e peças museológicas, fotografias a cores e preto e branco, posters e páginas de jornal digitalizadas.	7	1264	488	482

Tabela 32 – Características da nova colecção de teste de cardinalidade 10.

Os resultados obtidos após nova experiência encontram-se resumidos na Tabela 33. Como se pode observar, o novo conjunto de resultados corrobora a hipótese levantada. Ao aumentar a variabilidade da colecção de teste, a qualidade das recomendações diminuiu. Na primeira experiência realizada, os objectos utilizados no treino do sistema eram demasiado semelhantes, fazendo com que o sistema de recomendação fosse incapaz de generalizar, ou seja, sofresse de um fenómeno vulgarmente conhecido por *overfitting* (Tetko, Livingstone, & Luik, 1995).

Aumentar o número de objectos de treino e, consequentemente, a variabilidade das suas propriedades fez com que o sistema de recomendação se tornasse mais genérico, i.e., menos preciso, no entanto mais capaz de produzir recomendações adequadas a um maior número de situações distintas.

	Exactidão			Precisão
#	Pearson	Spearman	Kendall Tau	NMSE
10_2	0.553	0.639	0.600	0.349

Tabela 33 – Resultados da validação cruzada efectuada ao Migration Advisor com a nova colecção de teste de cardinalidade 10.

5.3 Considerações finais

Este capítulo teve como principal objectivo descrever os processos de avaliação desenvolvidos em torno do CRiB e dos seus componentes. Estes processos de avaliação incidiram especialmente sobre os componentes Object Evaluator e Migration Advisor, uma vez que são estes que apresentam o maior número de contributos científicos e tecnológicos.

No que diz respeito à avaliação do Object Evaluator, esta teve como principal objectivo aferir em que medida este componente é capaz de determinar o nível de degradação sofrido por um objecto digital durante a sua migração. Uma vez que os resultados produzidos por este componente influenciam directamente as recomendações efectuadas pelo Migration Advisor, tornou-se fundamental garantir que o primeiro produz resultados válidos, de modo a permitir a avaliação eficaz do segundo.

O Object Evaluator é capaz de determinar o nível de degradação sofrido por um objecto digital convertido, comparando-o com o original e calculando o nível de similaridade existente entre ambos. O cálculo de similaridade é efectuado à luz de um conjunto diversificado de critérios. Alguns desses critérios são caracterizados por um elevado nível de subjectividade, i.e., a sua avaliação varia consoante o interveniente que procura determinar a respectiva similaridade.

Entre os vários critérios suportados por este componente, especialmente no que diz respeito a migrações entre formatos de imagem matricial, foram seleccionados dois critérios manifestamente subjectivos: conformidade gráfica e metainformação embebida. As funções de similaridade associadas a estes dois critérios foram avaliadas segundo um protocolo experimental bem definido e um conjunto de métricas de avaliação.

A avaliação do Migration Advisor teve como principal objectivo determinar a capacidade apresentada por este componente em produzir *rankings* de caminhos de migração (i.e., recomendações) adequados à preservação de um dado formato digital. Estes *rankings* foram produzidos tendo em consideração os requisitos manifestados por uma entidade-cliente e os relatórios de migração acumulados ao longo do tempo no Evaluations Repository. Os *rankings* recomendados por este componente foram construídos com base no histórico de migrações e comparados com os *rankings ideais*, calculados a-pedido para um conjunto de objectos digitais de teste.

Para avaliar a qualidade dos *rankings* produzidos por este componente recorreu-se a um método de validação designado *10-fold cross-validation*. Este método consiste em particionar uma colecção de teste em 10 partes iguais, utilizar 9 dessas partes para treinar o sistema e a parte remanescente para testar o mesmo. Este procedimento foi repetido ao longo de 10 iterações, fazendo-se variar a partição de teste ao longo da colecção. Os vários *rankings* foram então comparados segundo um conjunto de métricas distintas, nomeadamente: as correlações de Pearson, Spearman e Kendall Tau e Normalized Mean Squared Error.

As avaliações efectuadas a ambos os componentes revelaram valores elevados de desempenho. No caso do Object Evaluator, este mostrou ser capaz de determinar eficazmente a similaridade gráfica entre duas imagens, apresentando valores de correlação acima dos 0.81 entre as opiniões produzidas por avaliadores humanos e os métodos automáticos de cálculo de similaridade. No que toca à capacidade para quantificar a deterioração ao nível da metainformação embebida, este componente apresentou valores de correlação acima dos 0.96 quando comparada a métrica de Jaccard' com os valores de referência associados à respectiva colecção de teste.

No caso do Migration Advisor, as experiências realizadas revelaram que as recomendações produzidas por este componente possuem um elevado nível de qualidade. As várias colecções de teste utilizadas durante o treino do sistema resultaram em níveis de correlação superiores a 0.68 com erros de precisão inferiores a 28%.

Capítulo 6
Implementações do CRiB

Ao longo do seu desenvolvimento, o projecto CRiB suscitou o interesse de algumas equipas técnicas que lideram projectos na área da preservação digital. O interesse manifestado por estas equipas levou a que, em Novembro de 2007, a plataforma fosse disponibilizada de forma gratuita e em código-aberto para utilização com fins educacionais e/ou de investigação[122]. Desde então, alguns projectos de I&D têm vindo a adoptar a plataforma CRiB, integrando-a com os seus próprios sistemas e aperfeiçoando-a de modo a produzir serviços mais eficientes e adequados aos seus contextos de utilização.

Este capítulo descreve alguns dos projectos que usam actualmente os serviços disponibilizados pelo CRiB ou que construíram serviços inspirados nas funcionalidades oferecidas por esta plataforma.

6.1 Planets

O Planets[123] (Preservation and Long-term Access through Networked Services) trata-se de um projecto de quatro anos co-financiado pela União Europeia no âmbito do 6º Programa

[122] Ver licença de uso e distribuição no Apêndice 8.6, na página 221.
[123] http://www.planets-project.eu/

Quadro que tem como objectivo o desenvolvimento de serviços e ferramentas que facilitem o acesso continuado a informação de âmbito cultural e científico (Farquhar & Hockx-Yu, 2007).

O projecto Planets teve início em Junho de 2006 e é composto pelos seguintes parceiros institucionais: Biblioteca Nacional da Grã-Bretanha, Biblioteca Nacional dos Países Baixos, Biblioteca Nacional Austríaca, Biblioteca Real da Dinamarca, Biblioteca Estatal da Dinamarca, Arquivos Nacionais dos Países Baixos, Arquivo Nacional da Inglaterra, Gales e Reino Unido, Arquivos Federais da Suíça, Universidade de Colónia, Universidade de Freiburg, Universidade de Glasgow, Universidade Técnica de Viena, Centros de Investigação Austríacos, IBM, Microsoft Research Limited e a Tessella Support Services.

O projecto subdivide-se em várias linhas de investigação em torno da preservação digital, tais como: caracterização de objectos digitais, desenvolvimento de serviços de preservação, planeamento de preservação, entre outros (Farquhar & Hockx-Yu, 2007). De uma dessas linhas de investigação resultou uma aplicação designada Plato[124] (Preservation Planning Tool). Esta ferramenta permite a um utilizador planear uma intervenção de preservação, testando um conjunto de acções de preservação pré-definidas contra uma amostra de objectos da colecção que se pretende preservar. A ferramenta executa as várias acções de preservação sobre a amostra fornecida e apresenta os resultados ao utilizador. A ferramenta avalia automaticamente um conjunto de critérios objectivos, definidos para a classe de objectos correspondente e oferece ao utilizador a possibilidade de associar manualmente um nível de satisfação ou qualidade aos restantes critérios considerados subjectivos. Baseada nessa informação, a ferramenta produz uma lista das estratégias adequadas para preservar a colecção de objectos pretendida (Becker, Ferreira et al., 2008; Becker, Kulovits et al., 2008).

O Plato integra na sua lista de acções de preservação o conjunto global de serviços de migração disponibilizados pelo CRiB (Becker, Ferreira et al., 2008). A Figura 62 apresenta um dos ecrãs da ferramenta Plato onde pode ver-se um excerto dos serviços disponibilizados pelo CRiB e a forma como estes podem ser seleccionados pelo utilizador durante o processo de teste das diferentes estratégias de preservação.

Actualmente, o trabalho desenvolvido em torno desta ferramenta centra-se na integração de serviços de caracterização de objectos digitais, serviços de análise e gestão de riscos e desenvolvimento proactivo de relatórios de apoio à tomada de decisão no contexto do planeamento de preservação digital (Becker, Ferreira et al., 2008). Alguns destes serviços

[124] http://www.ifs.tuwien.ac.at/dp/plato/

poderão eventualmente vir a ser construídos a partir de funcionalidades incorporadas no CRiB.

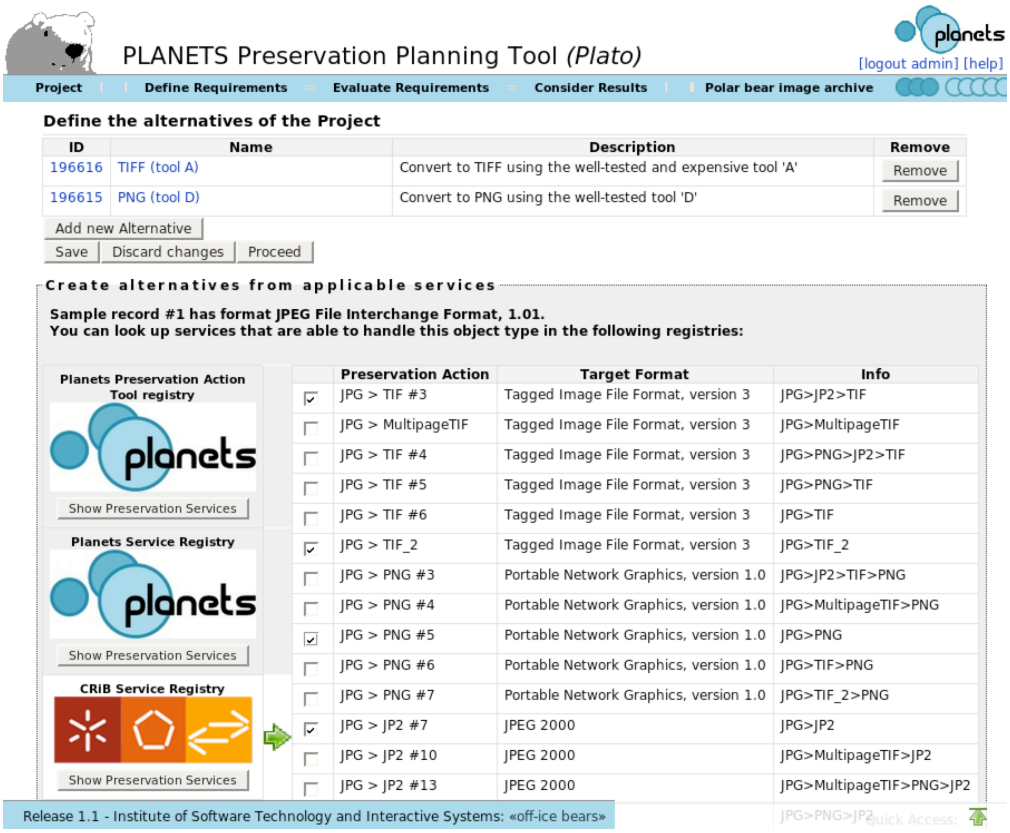


Figura 62 – Plato e os serviços de migração do CRiB.

6.2 RODA

A Direcção-Geral de Arquivos[125] (DGARQ) assume na sua missão institucional a responsabilidade pela identificação e preservação de documentação de valor histórico como meio de garantir e fomentar a memória individual e colectiva nacional. Em paralelo, as iniciativas do Governo Electrónico determinam que a Administração Pública deverá, cada vez mais, basear a sua actividade em processos de negócio electrónicos com o intuito de agilizar e assegurar um serviço mais rápido, completo e transparente para o cidadão. Este cenário evidencia um aumento da produção de informação digital, informação esta que, de acordo com a missão da DGARQ, deverá ver assegurado o seu valor evidencial através da garantia da sua autenticidade (Barbedo et al., 2007).

[125] http://www.dgarq.gov.pt

No sentido de suportar a incorporação e gestão de informação de arquivo produzida em formatos electrónicos a DGARQ empenhou-se ao longo dos últimos anos em desenvolver processos, ferramentas e recursos capazes de dar resposta às necessidades de preservação da informação digital produzida pela Administração Pública, cuja conservação continuada seja considerada importante do ponto de vista patrimonial (Barbedo et al., 2007).

Neste contexto nasce o projecto RODA (Repositório de Objectos Digitais Autênticos), um projecto que visa a promoção da preservação digital a nível nacional através do portal RODA[126] e o desenvolvimento de uma solução tecnológica, ultimada na construção de um repositório digital capaz de incorporar, descrever e dar acesso a todo o tipo de informação digital produzida no contexto da Administração Pública nacional (Ramalho et al., 2008). A Figura 63 apresenta a interface gráfica do RODA.

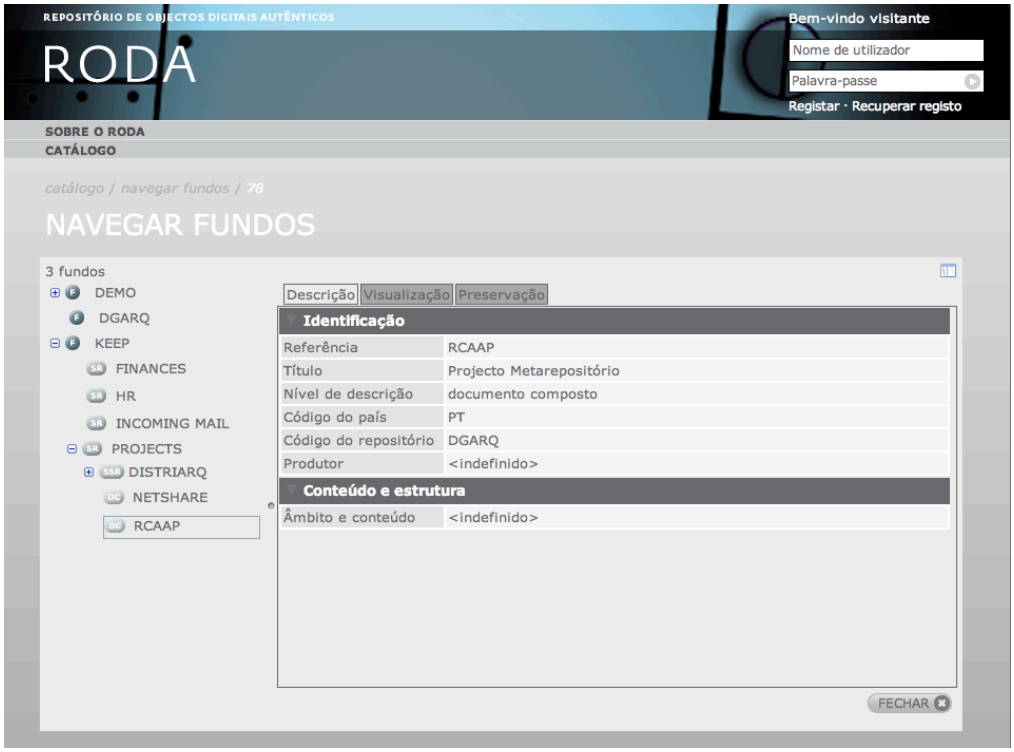


Figura 63 – Interface gráfica do Repositório de Objectos Digitais Autênticos.

Os serviços disponibilizados pelo CRiB são utilizados de forma transversal no âmbito do projecto RODA. O serviço de identificação de formatos é utilizado pelo RODA durante o

[126] http://roda.dgarq.gov.pt

processo de ingestão, de modo a determinar qual a acção de preservação a aplicar no sentido de normalizar os formatos recepcionados. Os serviços de migração são utilizados tanto para normalizar os objectos para formatos de preservação como na transformação destes para formatos mais leves e adequados ao consumo através da Web. Os serviços de avaliação de migração (i.e., Object Evaluator) são também utilizados para descrever o sucesso ou insucesso de uma migração e produzir metainformação de preservação em formato PREMIS (Ramalho et al., 2008).

Ao contrário do que acontece com a ferramenta Plato, cujo acesso aos serviços do CRiB é efectuado de forma remota através da Internet, o RODA implementa o CRiB na sua rede local. Isto garante a segurança e a privacidade dos dados e acelera todo processo de transferência de informação entre os componentes distintos do sistema.

Durante a implementação local do CRiB, a equipa de desenvolvimento do RODA identificou algumas linhas adicionais de desenvolvimento que tornariam os serviços disponibilizados pelo CRiB consideravelmente mais eficientes. Assim, ao invés dos objectos digitais serem transportados no interior de mensagens SOAP, estes passaram a ser referenciados através de um URL, permitindo, deste modo, a utilização de protocolos mais eficientes para transferir as representações (e.g. HTTP ou FTP) e evitar todo um conjunto de operações de codificação e descodificação para Base64 (Josefsson, 2006).

Capítulo 7
Conclusões e trabalho futuro

Este capítulo tem como objectivo apresentar um conjunto de conclusões que resultaram deste trabalho de investigação.

O capítulo começa com uma síntese do trabalho realizado, à qual se segue uma enumeração das principais conclusões que dele foram retiradas. Segue-se uma apresentação dos contributos mais relevantes e um conjunto de linhas de trabalho a realizar no futuro.

7.1 Síntese

A obsolescência tecnológica é um problema que afecta organizações e indivíduos num mundo cada vez mais "digitalizado". Com o aumento da desmaterialização e o crescimento acentuado da "pegada tecnológica" associada a cada individuo, a preservação digital passa a ser relevante, não apenas para quem se preocupa com a salvaguarda de informação de conservação permanente, mas também para todos aqueles que consomem e produzem informação digital no seu dia-a-dia e da qual dependem grande parte dos seus processos de negócio, lazer, comunicação, memória, etc.

Ao longo desta tese abordaram-se várias temáticas relacionadas com a preservação digital. Foram também tocadas diversas áreas científicas na demanda por uma solução tecnológica que permitisse atenuar a ansiedade dos profissionais responsáveis por gerir informação digital.

Deste processo resultou um conjunto de ferramentas que facilita a implementação de estratégias de preservação de informação digital baseadas em migração de formatos.

Ainda neste contexto, procurou-se evidenciar a necessidade de encontrar e implementar mecanismos capazes de auxiliar organizações e indivíduos na realização de tarefas anexas à preservação digital. Foi também argumentado que esses mecanismos deveriam ser, simultaneamente, adaptáveis às necessidades da entidade preservadora, tanto em termos de orçamento como em termos de qualidade de serviço, e reduzir ao máximo a necessidade de intervenção humana sem que houvesse prejuízo da autenticidade dos materiais a preservar.

Este conjunto de objectivos pode ser resumido numa única questão de investigação:

> *Qual o conjunto de serviços que permite implementar, de forma transversal e automática, todos os processos inerentes à migração de objectos digitais num contexto de preservação digital, sem que haja prejuízo da sua autenticidade?*

De forma a dar resposta a esta questão de investigação foi construído um sistema, baseado numa arquitectura orientada ao serviço, composto por um conjunto de serviços independentes que quando invocados de forma orquestrada permitem dar resposta aos objectivos previamente delineados.

Assim, em jeito de resenha, poder-se-á descrever os conteúdos desta tese da seguinte forma: a tese começa com uma introdução à problemática da preservação digital, onde são abordados temas como o conceito de objecto digital, o modelo de referência OAIS, estratégias de preservação digital, directórios de formatos, critérios para a autenticidade, metainformação de preservação e modelos de avaliação de estratégias de preservação.

A tese continua, em espiral, com um enquadramento teórico que facilita a compreensão das diferentes etapas de um processo de migração, estratégia de preservação adoptada ao longo desta tese para efeitos de prova de conceito. É ainda apresentado um cenário de preservação que facilita a identificação das principais dificuldades com as quais um profissional da área da gestão de informação se debate, servindo assim de ponto de partida para a identificação do conjunto mínimo de serviços que garante a automatização de processos de preservação baseados em migração. É ainda descrito, em detalhe, um conjunto de ferramentas que permite implementar serviços de preservação e que serviu de base para a construção do sistema apresentado nesta tese.

Num capítulo subsequente é apresentado o CRiB, um sistema baseado em serviços que procura dar resposta à questão de investigação previamente enunciada. Ainda nesse capítulo, são apresentadas as taxionomias de avaliação utilizadas pelos processos de controlo de qualidade e recomendação implementados pelo CRiB.

Após a descrição do sistema, é apresentada a metodologia de avaliação dos seus componentes e as experiências realizadas em torno da plataforma que permitiram aferir a sua adequabilidade aos objectivos propostos.

Seguiu-se, ainda, uma breve descrição dos projectos RODA e Planets, de relevância nacional e internacional, respectivamente, que adoptaram partes da plataforma CRiB ao longo dos seus desenvolvimentos.

7.2 Conclusões e discussão

Em boa medida as principais conclusões a retirar deste trabalho foram já expostas e discutidas ao longo de capítulos anteriores. Não obstante, esta secção apresenta um compêndio de todas essas notas conclusivas.

Tomando como base a questão de investigação previamente enunciada, pode-se concluir que o seguinte conjunto de serviços é suficiente para implementar procedimentos automáticos de preservação (nesta fase, baseados exclusivamente em migração de formatos) que operem transversalmente sobre colecções de objectos digitais:

- **Serviço de identificação de formatos** – fundamental para a obtenção de um mapa dos formatos que constituem a colecção de objectos a preservar. Mediante esta informação poder-se-á tomar decisões quanto à melhor estratégia de preservação a tomar. No contexto do CRiB, este serviço foi implementado pelo componente `Format Identifier`;

- **Serviço de notificação de obsolescência** – serviço necessário para garantir a automatização dos processos de preservação. Este serviço monitoriza permanentemente o contexto tecnológico vigente e determina o nível de obsolescência de um dado formato e os riscos associados à sua conservação. Numa situação de ruptura tecnológica eminente, informa o sistema de preservação que deverá iniciar uma intervenção de preservação. Este componente não foi desenvolvido no âmbito deste projecto uma vez que já se

encontrava em desenvolvimento pela Biblioteca Nacional da Austrália – AONS (Curtis et al., 2007; Pearson, 2008);

- **Serviço de conversão de formatos** – serviço responsável pela execução de acções de preservação. No contexto desta tese apenas foram exploradas alternativas de preservação baseadas na migração de formatos. O CRiB implementa uma rede de serviços de migração que podem ser invocados de forma individual ou composta e que asseguram a conversão entre dezenas de formatos recorrendo a centenas de caminhos de migração alternativos. O componente responsável pela gestão da rede de serviços de migração designa-se, neste contexto, por Service Registry e o componente responsável por executar as respectivas conversões intitula-se Migration Broker;

- **Serviço de controlo de qualidade** – os conversores não são todos iguais e as conversões não são sempre perfeitas. Este serviço é responsável por aferir a qualidade de uma conversão e subsequentemente do conversor ou conversores utilizados. No contexto do CRiB, este serviço é assegurado por três componentes distintos: o Migration Broker, responsável pelo controlo de qualidade ao nível da performance de migração, o Format Evaluator, responsável por aferir o ganho em termos de capacidade de preservação que se obteria se se realizasse uma conversão entre dois formatos e, finalmente, o Object Evaluator, responsável por determinar o grau de degradação incorrido ao nível das propriedades significativas que constituem o objecto digital que se pretende preservar;

- **Serviço de auxílio à selecção de estratégias de conversão** – um objecto, formato ou classe de objectos pode ser convertido para um grande número de formatos distintos. Uma migração pode ser realizada recorrendo a uma multitude de ferramentas de conversão. Este serviço permite identificar, de entre todas as opções reconhecidas pelo sistema, qual a que garante o maior nível de satisfação da entidade preservadora. O CRiB materializa este conceito através do seu componente Migration Advisor, i.e., um serviço de recomendação de alternativas de migração que tem em consideração factores como performance, adequabilidade dos formatos a preservação a longo-prazo e conservação de propriedades significativas.

Para além da identificação dos serviços necessários à implementação transversal e automática de estratégias de preservação, foi também fundamental assegurar que estes eram capazes de garantir a autenticidade dos materiais. A verificação desta premissa foi alcançada de duas formas distintas. O recurso ao Migration Advisor garante que, num dado instante, uma intervenção de preservação será implementada recorrendo à melhor alternativa de migração conhecida pelo sistema. O Migration Advisor analisa todas as migrações realizadas no passado e determina qual o caminho de migração que maximiza a conservação das propriedades significativas do objecto que se pretende preservar. O caminho recomendado pelo Migration Advisor procura ainda suprir os requisitos da entidade preservadora ao nível do custo, performance e adequabilidade dos formatos envolvidos para preservação a longo-prazo.

O CRiB tem ainda outra medida de salvaguarda no que toca à autenticidade dos materiais. Após uma migração, o objecto digital resultante é comparado com o objecto submetido a migração. Dessa comparação resulta um relatório onde se incluem todas as propriedades significativas do objecto original que foram testadas e informação sobre o nível de degradação detectado. Este relatório constitui, efectivamente, o que geralmente se designa por metainformação de preservação, i.e., metainformação que documenta todas as intervenções de preservação a que um dado objecto foi sujeito e qual o efectivo resultado de cada uma dessas intervenções. A conservação deste relatório junto da metainformação que acompanha o objecto digital é, por si só, condição suficiente para garantir a autenticidade dos materiais preservados. É importante referir que a conservação deste relatório garante a autenticidade dos materiais e não a preservação dos mesmos da forma mais adequada.

No sentido de atestar a viabilidade do CRiB como uma possível materialização dos objectivos delineados para este trabalho, foram implementados processos de validação para os principais componentes que constituem o sistema.

A questão de investigação realça a necessidade de existência de serviços capazes de implementar automática e transversalmente estratégias de preservação baseadas em migração. Como foi visto ao longo desta tese, a implementação de uma estratégia de migração pressupõe o desenvolvimento de três actividades fundamentais: a selecção de uma alternativa de migração, a conversão dos materiais propriamente dita e o controlo de qualidade da respectiva conversão. O CRiB disponibiliza um conjunto de serviços suportados por componentes de software que têm como missão materializar cada uma destas actividades. Esses componentes são, respectivamente, o Migration Advisor, o Migration Broker e o Object Evaluator. Nesta tese foram desenvolvidas experiências no sentido de validar cada um

destes componentes, exceptuando o Migration Broker. Este componente não foi validado uma vez que apenas apresentava dois estados possíveis de execução: sucesso (a conversão resultou num novo objecto digital) ou insucesso (a conversão falhou e não devolveu qualquer objecto). Os casos de insucesso decorrem da submissão de objectos corrompidos ou não compatíveis com os conversores utilizados, ou a falhas na rede que impeçam a comunicação entre os vários componentes do sistema. Os restantes dois componentes, dada a sua complexidade, exigiram um nível superior de rigor ao longo da sua avaliação.

O Object Evaluator tem como missão determinar o nível de degradação incorrido durante um processo de migração ao nível das propriedades significativas de um objecto. Os resultados produzidos por este componente são utilizados pelo Migration Advisor na identificação dos serviços de migração que maximizam a conservação de propriedades significativas de uma dada classe de objectos. Este componente funciona comparando o objecto que resultou de uma migração com o objecto original do qual este foi derivado e determinando o nível de similaridade existente ao nível das suas propriedades significativas.

A avaliação do Object Evaluator foi realizada apenas no domínio das imagens matriciais. Foram seleccionadas duas propriedades significativas consideradas subjectivas: conformidade gráfica e metainformação embebida. Foram ainda analisadas várias funções de similaridade para cada uma destas propriedades, nomeadamente: RMSE, UQI, SSIM e CBM.

Para cada propriedade seleccionada foi construída uma colecção de teste constituída por objectos digitais em diversos formatos pertencentes à classe escolhida (i.e., imagens matriciais). A colecção de teste foi avaliada manualmente por um conjunto de intervenientes humanos e, posteriormente, pelos algoritmos automáticos de cálculo de similaridade. Os resultados produzidos por ambos foram então comparados recorrendo a um conjunto de métricas comummente utilizadas neste tipo de avaliações (para mais detalhes, consultar a Secção 5.1 na página 134).

Este componente revelou ser capaz de determinar eficazmente a similaridade gráfica entre duas imagens, apresentando valores de correlação superiores a 0.81 entre as opiniões produzidas pelos avaliadores humanos e os métodos automáticos de cálculo de similaridade analisados. No que toca à capacidade para quantificar a deterioração ao nível da metainformação embebida, este componente apresentou valores de correlação acima dos 0.96

quando comparada a métrica de Jaccard' com os valores de referência associados à respectiva colecção de teste.

A avaliação do Migration Advisor, por sua vez, teve como principal objectivo determinar a sua capacidade para produzir *rankings* de caminhos de migração (i.e., recomendações de serviços de migração) que maximizassem a qualidade da preservação baseando-se exclusivamente no seu conhecimento de migrações realizadas anteriormente. O conceito de qualidade dependeria, obviamente, dos requisitos manifestados pela entidade-cliente que invocou o serviço de recomendação.

Para avaliar a qualidade dos *rankings* produzidos por este componente recorreu-se a um método de validação designado *10-fold cross-validation*. O sistema de recomendação foi treinado efectuando centenas de conversões entre formatos distintos e acumulando os relatórios de avaliação numa base de conhecimento. Cada conversão realizada permitia ao sistema reconhecer os serviços de migração que exibiam melhor qualidade de serviço em termos de preservação. Os *rankings* produzidos com base em migrações passadas foram então comparados com os *rankings ideais, i.e.,* aqueles que efectivamente maximizavam a qualidade da conversão para um dado objecto digital. Para determinar o *ranking ideal*, todos os objectos pertencentes à colecção de teste foram convertidos recorrendo a todos os serviços de migração conhecidos pelo sistema.

Os *rankings ideais* e os *rankings* baseados em migrações passadas foram comparados recorrendo a um conjunto de métricas frequentemente utilizadas na comparação de *rankings*: correlação de Pearson, Spearman e Kendall Tau e Normalized Mean Squared Error. Na avaliação deste componente foram utilizadas colecções de treino/teste de cardinalidade 10, 20, 50 e 100.

As experiências realizadas em torno do Migration Advisor revelaram que as recomendações baseadas em migrações passadas possuem um elevado nível de correlação com os *rankings ideais* dos objectos digitais incluídos nas colecções de teste. Os testes realizados resultaram em valores de correlação compreendidos entre 0.68 e 0.85 com um erro de precisão máximo de 34.9%.

Com base nos resultados obtidos, conclui-se que o CRiB responde de forma aceitável à questão de investigação definida no início desta investigação. A prova de conceito aqui apresentada permite concluir que é possível materializar um conjunto de serviços capazes de implementar de forma automática processos de migração de objectos digitais sem haja prejuízo da sua autenticidade.

É importante referir que, para que o CRiB possa ser implementado de forma eficaz e prática, é necessário que os objectos digitais se encontrem acessíveis à plataforma de serviços. Isso implica a existência de um agente ou componente de software responsável por desencadear os processos de preservação a partir do ambiente onde os objectos se encontram residentes. Esse agente deverá consultar periodicamente o serviço de notificação de obsolescência e mediante a resposta obtida desencadear medidas reactivas junto dos restantes serviços do CRiB.

Adicionalmente, para garantir a autenticidade dos materiais, é fundamental a existência de um sistema de gestão de informação capaz de associar metainformação de preservação (e também de outros tipos) aos objectos intervencionados. No domínio organizacional, o recurso a repositórios digitais facilita esta tarefa uma vez que o ambiente onde os objectos residem incorpora, de base, as funcionalidades necessárias para suprir esta necessidade.

No domínio doméstico, estes requisitos são mais difíceis de reunir uma vez que os sistemas operativos não os satisfazem de forma natural. Uma solução viável no domínio doméstico consiste no desenvolvimento de uma aplicação que corre em segundo plano (tal como um antivírus) e que tem como objectivo monitorizar o estado de obsolescência dos objectos digitais presentes no sistema. Quando um objecto digital é marcado como estando num formato em vias de se tornar obsoleto são desencadeados mecanismos automáticos de migração e produção de metainformação de preservação que são geridos automaticamente por este agente de software. Esta aplicação poderá também responsabilizar-se por efectuar cópias de segurança de objectos modificados para suportes físicos externos[127].

Há ainda dois aspectos relacionados com o CRiB que não foram devidamente trabalhados e que merecem alguma discussão. Estes são: desempenho e segurança. A comunicação entre os vários componentes do CRiB é efectuada através de Web services. Esta tecnologia apresenta algumas vantagens quando comparada com outras tecnologias de comunicação entre processos. Entre as principais vantagens encontra-se o facto de se basearem em normas suportadas por organismos internacionais, terem um elevado nível de adopção por parte da indústria de software e sobretudo por permitirem a interoperabilidade entre linguagens de programação, sistemas operativos e arquitecturas de hardware[128].

---

[127] A versão 10.5 do sistema operativo Mac OS X é acompanhada de um aplicação chamada Time Machine que efectua automaticamente cópias de segurança dos ficheiros modificados durante a utilização do sistema.

[128] Há relatos de incompatibilidade entre algumas plataformas, mas que poderão ser evitados recorrendo a boas práticas de programação durante o desenvolvimento de Web services.

Apesar das suas consideráveis vantagens, os Web services carecem de muita largura de banda, o que geralmente resulta em tempos de transmissão excessivamente longos. Isto deve-se ao facto de as mensagens trocadas serem codificadas em XML/SOAP que, por ser auto-descritivo, é também demasiado verboso. Uma implementação prática do CRiB necessitaria, portanto, de alguma optimização ao nível da comunicação, como aliás já começou a ser realizada durante a adaptação do mesmo ao projecto RODA (ver Secção 6.2 na página 177). Contudo, é de realçar a vulgarização de redes Gigabit e fibra óptica, assim como o exponencial aumento da largura de banda no acesso à Internet. Com o tempo, o evoluir destas tecnologias irá gradualmente mitigar este problema.

No que diz respeito à segurança, o CRiB abre caminho para um vasto leque de desenvolvimentos adicionais. Num contexto de preservação a segurança dos dados é absolutamente fundamental. O modelo descentralizado defendido nesta tese, em que vários intervenientes competem numa arena comum pela prestação de serviços de migração, propicia ainda mais o problema da segurança dos dados. O modelo apresentado permite que terceiros manipulem os dados que se pretendem preservar, o que poderá constituir um risco à integridade conceptual do objecto preservado. O CRiB incorpora mecanismos de controlo de qualidade que minimizam esse risco. Contudo, no sentido de se construir um ambiente de mútua confiança em torno da plataforma, seria fundamental definirem-se contratos entre os diversos intervenientes, i.e., prestadores de serviço, intermediários e entidades-cliente. Esses contratos deverão incorporar variados aspectos relacionados com serviço prestado, detalhando procedimentos e parâmetros ao nível da segurança e manipulação de dados, confidencialidade, responsabilidades assumidas, garantias e mecanismos de monitorização e/ou fiscalização.

No que toca à segurança dos dados durante a transmissão, deve acrescentar-se que os Web services podem operar sobre protocolos HTTP/SSL (Hypertext Transfer Protocol/Secure Socket Layer) minimizando, assim, o risco de inspecção por terceiros.

Outro ponto que merece ser alvo de discussão é o da obsolescência da própria plataforma de preservação. Como é natural, chegará um momento em que o sistema apresentado deixará de possuir as condições necessárias para poder operar eficazmente. O CRiB, como qualquer outro sistema informático, depende do bom funcionamento de vários elementos, inclusivamente daqueles que constituem a sua infra-estrutura tecnológica (e.g. hardware, sistemas operativos, linguagens de programação, tecnologias de comunicação, etc.). Esses elementos são suportados por diversos fabricantes e fornecedores de serviço que poderão a qualquer momento ser alvo de ruptura institucional (e.g. falência, aquisição por terceiros, cessação de suporte dos seus produtos, entre outros). Esta situação colocaria em risco a

viabilidade da plataforma de serviços aqui apresentada. O CRiB procura mitigar este problema recorrendo a tecnologias abertas amplamente utilizadas pela comunidade de desenvolvimento de tecnologias de informação:

- Hardware – o CRiB foi desenvolvido e testado em arquitecturas de hardware baseadas no x86, a arquitectura de hardware comercialmente mais bem sucedida da história. Adicionalmente, os componentes centrais do CRiB foram desenvolvidos em Java, o que significa que são executados por uma máquina virtual dotada do seu próprio *instruction set*, não estando dependentes de uma arquitectura de hardware específica. O CRiB pode funcionar harmoniosamente em qualquer arquitectura de hardware e/ou sistema operativo para a qual exista uma implementação da Java Virtual Machine. É importante referir que a especificação da Java Virtual Machine é aberta[129] e que existem actualmente implementações para plataformas tão diversas como a Solaris SPARC, Sun Java Desktop System, Linux (todas as distribuições), Windows 98, Windows ME, Windows 2000 (SP4+), Windows XP (SP1 SP2), Vista, Windows 2003, bem como para uma série de dispositivos móveis.

- Sistema operativo – o CRiB é um sistema distribuído onde cada um dos seus componentes pode ser executado num nó de processamento distinto. Cada um desses nós pode possuir o seu próprio sistema operativo desde que possua uma implementação da máquina virtual Java. A maioria dos componentes que constituem o CRiB foram testados em sistemas operativos Linux. No entanto, alguns dos serviços de migração incorporados foram desenvolvidos para plataformas Windows devido ao facto de se basearem em aplicações de software que apenas existiam neste sistema operativo.

- Linguagens de programação – O CRiB foi desenvolvido em Java, uma linguagem de programação bem conhecida da comunidade de desenvolvimento e cuja especificação pode ser publicamente inspeccionada[130]. Esta linguagem foi inicialmente desenvolvida pela Sun Microsystems[131], no entanto, devido ao facto de a sua especificação ser aberta, existem já várias dezenas de implementações paralelas que apresentam optimizações específicas para certas arquitecturas de hardware.

---

[129] http://java.sun.com/docs/books/jvms/
[130] http://java.sun.com/docs/books/jls/
[131] http://www.sun.com/

- Tecnologias de comunicação – A comunicação entre os diversos componentes do sistema foi implementada recorrendo a Web services. Esta tecnologia define um conjunto de protocolos que permite a transferência de informação entre diferentes componentes ou aplicações, independentemente da linguagem de programação ou da infra-estrutura tecnológica que os suporta. Os Web services funcionam sobre o protocolo HTTP (Hypertext Transfer Protocol com ou sem SSL), um protocolo amplamente utilizado e que serve de base à World Wide Web.

O CRiB dever ser visto como um componente externo ao ambiente de preservação onde residem os objectos digitais cujo acesso se deseja continuado. Este sistema tem apenas como objectivo a prestação de serviços de preservação e poderá ser encarado sob uma perspectiva de *outsourcing* aplicacional. O desaparecimento do CRiB não coloca em risco os objectos digitais, apenas os serviços que facilitam a implementação de estratégias de preservação.

Para além do disposto, o sistema apresentado ao longo desta tese deve ser visto como um modelo e não como um produto. Todos os seus componentes poderiam ter sido desenvolvidos recorrendo a tecnologias inteiramente distintas das que foram adoptadas.

As tecnologias adoptadas no âmbito deste projecto foram aquelas que apresentavam um nível de maturidade superior e que facilitavam a rápida prototipagem. Simultaneamente, estas tecnologias permitiram o desenvolvimento de sistemas interoperáveis e multiplataforma. Neste contexto, é importante referir que o problema que a preservação digital se propõe resolver pode ser visto como um problema de interoperabilidade. Um problema de interoperabilidade, não entre sistemas contemporâneos (interoperabilidade no espaço), mas entre sistemas que ainda não foram desenvolvidos (interoperabilidade no tempo). Para que um sistema de preservação possa ser considerado eficaz, este deve ser interoperável pelo menos com os sistemas que lhe são contemporâneos.

7.3 Contributos

Esta tese reúne em si um conjunto de contributos que são considerados relevantes para diferentes contextos de aplicação. Estes foram agrupados de acordo com o público a que se destinam:

Contributos para entidades carentes de preservação digital

- A implementação de mecanismos de controlo de qualidade que permitem aferir de forma automática a quantidade de informação e/ou funcionalidades perdidas durante um processo de migração;

- Capacidade de preservar objectos digitais recorrendo a técnicas de conversão de formatos sem que haja necessidade de implementar localmente complexos sistemas de migração;

- A capacidade de obter relatórios com detalhes técnicos sobre o resultado de uma migração, permitindo assim documentar uma intervenção de preservação e deste modo assegurar a autenticidade dos materiais intervencionados;

- A possibilidade de comparar diferentes alternativas de migração e identificar, de forma objectiva, qual destas é a mais adequada para satisfazer as suas necessidades organizacionais.

Contributos para a indústria de software

- A possibilidade de disponibilizar e/ou vender aplicações de conversão através da infra-estrutura de serviços desenvolvida;

- A capacidade de avaliar de forma objectiva a qualidade geral de aplicações de conversão recorrendo a dezenas de critérios de avaliação;

- A possibilidade de comparar o desempenho de aplicações de conversão com o desempenho de centenas de outras numa arena imparcial que favorece a concorrência;

- Um modelo para a avaliação de migrações que poderá ser implementado em aplicações de software dotadas de capacidade de exportação de dados para vários formatos. Esta funcionalidade permite ao utilizador identificar os formatos de exportação mais adequados para armazenar objectos produzidos no âmbito de uma dada aplicação;

- A agregação de um conjunto de métricas de similaridade de imagens e a sua tradução para a linguagem de programação Java.

Contributos para a investigação em preservação digital

- A publicação de uma revisão de literatura em língua portuguesa que inclui uma introdução aos principais conceitos e estratégias relevantes no domínio da preservação digital. Esta revisão de literatura foi publicada em livro e disponibilizada na Internet em acesso livre[132] - Ferreira, Miguel - "Introdução à preservação digital : conceitos, estratégias e actuais consensos". Guimarães : Escola de Engenharia da Universidade do Minho, 2006. ISBN 978-972-8692-30-8.

- A identificação e caracterização de diferentes serviços e componentes funcionais que possibilitam a implementação de estratégias de preservação baseadas em migração sem que haja prejuízo da autenticidade dos materiais;

- A modelação e desenvolvimento de uma arquitectura orientada ao serviço capaz de avaliar o desempenho de uma migração segundo múltiplos critérios, nomeadamente: performance operacional, aptidão dos formatos envolvidos e quantificação da informação perdida durante a intervenção de preservação;

- A recolha e desenvolvimento de funções de similaridade adequadas a diferentes tipos de propriedades que permitem aferir, de forma objectiva, o nível de degradação incorrido ao nível das propriedades significativas de um objecto digital durante uma migração de formatos.

7.4 Trabalho futuro

Cada desafio conquistado ao longo desta tese abriu portas para novos desenvolvimentos. Tendo consciência do imenso trabalho que ficou por realizar, seguem-se algumas linhas de trabalho futuro.

- O sistema actual pode ser profundamente melhorado se for adicionado suporte para: mais formatos de objectos digitais, mais propriedades significativas e a possibilidade de efectuar migrações entre formatos pertencentes a classes distintas (e.g. migração de documentos de texto para imagens matriciais);

- Ao longo deste trabalho foi possível constatar que, regra geral, os conversores não possuem um comportamento constante, ou seja, conforme as características do

[132] O livro "Introdução à preservação digital – Conceitos, estratégias e actuais consensos" foi até à data descarregado mais de 8000 vezes, maioritariamente por pessoas oriundas do Brasil, Portugal, Argentina, Estados Unidos, Espanha, Peru, Angola e Uruguai.

objecto a processar estes apresentam diferentes níveis de performance computacional. Seria importante realizar um estudo no sentido de se apurar que factores influenciam de forma directa o tempo de conversão de objectos digitais em diferente formatos;

- Investigar métodos e tecnologias que permitam incorporar na arquitectura de serviços outras estratégias de preservação para além da migração, como por exemplo, emulação. Este ponto envolve o desenvolvimento de mecanismos que permitam a execução remota de acções de preservação não baseadas em migração, a reunião de critérios de qualidade adequados à estratégia de preservação adoptada e a implementação de mecanismos capazes de extrair e comparar critérios que garantam a avaliação da sua qualidade;

- Actualmente, o componente Migration Advisor produz recomendações com base nos dados recolhidos a partir de todas as migrações efectuadas no passado. Seria interessante investigar se a implementação de um mecanismo de "esquecimento", onde apenas seriam consideradas as conversões mais recentes, poderia resultar no melhoramento efectivo da qualidade das recomendações. Este mecanismo garantiria que as recomendações eram calculadas com base nas conversões realizadas mais recentemente;

- Estudar formas de garantir a segurança dos dados num ambiente distribuído onde os vários intervenientes têm a capacidade de ler e manipular a informação que se pretende preservar, havendo assim um potencial risco à sua integridade;

- Actualmente, o CRiB recomenda serviços de migração com base no formato do objecto que se pretende preservar. Como trabalho futuro seria fundamental dotar o CRiB de capacidade para analisar os constituintes internos dos objectos digitais de forma a identificar com maior rigor qual o serviço de migração mais adequado à sua conversão. O mesmo se aplica ao componente responsável pelo controlo de qualidade ao nível das propriedades significativas, i.e., o Object Evaluator. Este deveria ser capaz de, por exemplo, comparar uma a uma as imagens contidas num documento de texto com as imagens existentes na sua versão original;

- No sistema actual, a entidade-cliente precisa de especificar os seus requisitos de preservação para que o sistema seja capaz de recomendar um conjunto de migradores adequado às suas preferências. Esta actividade é realizada atribuindo pesos à taxonomia geral de avaliação apresentada pelo sistema no momento que antecede a recomendação de serviços de migração. Seria interessante desenvolver um estudo no

sentido de identificar quais os perfis de preservação mais comuns entre as mais diversas entidades-cliente. Alternativamente, o componente Migration Knowledge Base poderia guardar as taxonomias pesadas pelas várias entidades-cliente e usar essa informação para automaticamente determinar o perfil do utilizador comum;

- Desenvolver um estudo no sentido de determinar quais os formatos de preservação recomendados para as duas classes de objectos digitais suportadas pelo CRiB, i.e., documentos de texto e imagens matriciais;

- Implementar um mecanismo que permitisse ao CRiB obter *feedback* por parte dos seus utilizadores de modo aferir o seu nível de satisfação face às recomendações e migrações realizadas. Este mecanismo poderia ser utilizado para melhorar as recomendações produzidas pelo Migration Advisor;

- Testar o sistema com um motor de orquestração de serviços (e.g. WS-BPEL) de modo a optimizar a selecção e execução de fluxos de Web services (permite optimizar a execução de conversões compostas);

- Actualmente, o CRiB calcula a qualidade de uma migração tendo por base um caminho de migração completo. No entanto, durante um processo de migração um objecto digital pode ser alvo de diversas conversões intermédias. Seria interessante desenvolver um estudo no sentido de aferir se a qualidade associada a uma migração composta é igual ao somatório da qualidade das suas conversões intermédias. Este estudo abriria portas para a criação de uma álgebra capaz de prever o comportamento de redes de conversores;

- Estudar novos modelos de negócio capazes de sustentar a manutenção e o desenvolvimento da plataforma que sejam, simultaneamente, apelativos para produtores de serviços de preservação e seus consumidores.

Capítulo 8
Apêndices

Este capítulo inclui todos os apêndices considerados necessários para garantir a completude desta tese. O capítulo está organizado da seguinte forma: a secção 8.1 descreve as ferramentas e bibliotecas utilizadas pelo componente Object Evaluator na extracção de propriedades significativas de objectos digitais; a secção 8.2 apresenta um exemplo de uma Taxionomia geral de avaliação; a secção 8.3 descreve formalmente e em detalhe as funções de similaridade utilizadas para comparar propriedades extraídas a partir de objectos digitais; a secção 8.4 descreve o teste não-paramétrico de Wilcoxon; a secção 8.5 descreve genericamente o método de validação cruzada; e finalmente, a secção 8.6 apresenta a licença de uso e distribuição da plataforma CRiB.

8.1 Ferramentas de extracção de propriedades

Este apêndice apresenta as bibliotecas e ferramentas utilizadas pelo componente Object Evaluator na extracção de propriedades significativas a partir de objectos digitais.

8.1.1 Image IO

A biblioteca Image I/O[133] que acompanha a linguagem de programação Java desde a sua versão 1.4 constitui uma plataforma extensível que facilita a interpretação e manipulação de imagens matriciais. Esta biblioteca foi utilizada pelo componente extractor de propriedades que acompanha o Object Evaluator para obter o valor de certas propriedades contidas em imagens de diversos formatos.

A Tabela 34 enumera as diferentes propriedades extraídas e formatos suportados por esta biblioteca.

Classe	Propriedade	Formatos suportados
Imagens matriciais	Número de páginas	Tagged Image File Format, version 3 Portable Network Graphics, version 1.0 Portable Network Graphics, version 1.1 Windows Bitmap, version 3.0 JPEG File Interchange Format 1.00 JPEG File Interchange Format 1.01 JPEG File Interchange Format 1.02 Graphics Interchange Format, version 1987a Graphics Interchange Format, version 1989a JPEG 2000
	Conformidade gráfica	
	Largura	
	Altura	
	Modelo de cor	
	Profundidade de cor	
	Método de compressão	

Tabela 34 – Propriedades extraídas e formatos suportados pela biblioteca Java Image I/O.

8.1.2 ExifTool 7.15

O ExifTool[134] trata-se de uma ferramenta independente da plataforma que permite ler e editar metainformação embebida em imagens, ficheiros de áudio e sequências de vídeo. Esta ferramenta suporta várias normas de metainformação como o EXIF, GPS, IPTC, XMP, JFIF, GeoTIFF, ICC Profile, Photoshop IRB, FlashPix, AFCP e ID3 assim como, metainformação específica de alguns fabricantes de câmaras digitais tais como Canon, Casio, FujiFilm, HP, JVC/Victor, Kodak, Leaf, Minolta/Konica-Minolta, Nikon, Olympus/Epson, Panasonic/Leica, Pentax/Asahi, Ricoh, Sanyo, Sigma/Foveon and Sony.

Esta ferramenta foi utilizada pelo Object Evaluator para extrair a metainformação embebida em imagens matriciais. A Tabela 35 enumera os formatos de imagem suportados por esta ferramenta.

---

[133] http://java.sun.com/javase/6/docs/technotes/guides/imageio

[134] http://www.sno.phy.queensu.ca/~phil/exiftool/

Classe	Propriedade	Formatos suportados
Imagens matriciais	Metainformação embebida	Tagged Image File Format, version 3 Portable Network Graphics, version 1.0 Portable Network Graphics, version 1.1 Windows Bitmap, version 3.0 JPEG File Interchange Format 1.00 JPEG File Interchange Format 1.01 JPEG File Interchange Format 1.02 Graphics Interchange Format, version 1987a Graphics Interchange Format, version 1989a JPEG 2000

Tabela 35 – Propriedades extraídas e formatos suportados pela ferramenta ExifTool.

8.1.3 Microsoft Office Word 2003

O Microsoft Word[135] é um processador de texto da Microsoft, criado originalmente em 1983 por Richard Brodie, destinado a computadores IBM PC baseados no sistema operativo DOS. Actualmente, esta aplicação acompanha o pacote de software Microsoft Office[136].

Esta ferramenta utiliza internamente um modelo de dados abstracto que permite manipular programaticamente documentos de texto em formato Word. Este modelo abstracto designa-se por Word Object Model[137]. O Microsoft Office Word, através do Word Object Model, foi utilizado pelo CRiB para extrair as propriedades que acompanham documentos de texto nos formatos Word e RTF. A Tabela 36 enumera as propriedades e os formatos suportados por esta ferramenta.

Classe	Propriedade	Formatos suportados
Documentos de texto	Número de páginas	Microsoft Word for Windows Document, version 97-2003 Rich Text Format, version 1.0 Rich Text Format, version 1.4 Rich Text Format, version 1.6 Rich Text Format, version 1.7
	Número de imagens	
	Conformidade gráfica	
	Margem esquerda	
	Margem inferior	
	Margem superior	
	Margem direita	
	Largura de página	
	Altura de página	
	Cor de fundo	
	Tipos de letra	
	Metainformação embebida	
	Disposição gráfica	

Tabela 36 – Propriedades extraídas pela ferramenta Microsoft Office Word 2003

[135] http://office.microsoft.com/en-us/word/
[136] http://office.microsoft.com
[137] http://msdn.microsoft.com/en-us/library/kw65a0we(VS.80).aspx

8.1.4 OpenOffice.org Writer 2.2

O Writer é um processador de texto multiplataforma, originalmente desenvolvido pela Sun Microsystems[138], que se encontra disponível em código-aberto. Esta aplicação é compatível com um grande número de processadores de texto concorrentes como por exemplo o Microsoft Word e o Corel WordPerfect. Actualmente, a aplicação acompanha o pacote de software OpenOffice.org[139].

O OpenOffice.org Writer disponibiliza uma interface que permite manipular os seus documentos a partir de uma aplicação externa. Essa interface designa-se por Universal Network Object (UNO)[140]. O CRiB tirou partido desta interface para extrair os valores das propriedades significativas existentes em objectos no formato OpenDocument (Tabela 37).

Classe	Propriedade	Formatos suportados
Documentos de texto	Número de páginas	OpenDocument Text Format, version 1.0
	Número de imagens	
	Conformidade de caracteres	
	Margem esquerda	
	Margem inferior	
	Margem superior	
	Margem direita	
	Largura de página	
	Altura de página	
	Cor de fundo	
	Tipos de letra	
	Metainformação embebida	
	Disposição gráfica	

Tabela 37 – Propriedades extraídas pela ferramenta OpenOffice.org Writer 2.2.

8.1.5 PDFBox

A PDFBox[141] trata-se de uma biblioteca Java que permite criar e manipular documentos PDF. Esta biblioteca foi utilizada para extrair as propriedades incluídas na Tabela 38 a partir de documentos PDF.

[138] http://www.sun.com
[139] http://www.openoffice.org/
[140] http://api.openoffice.org/docs/java/ref/overview-summary.html
[141] http://www.pdfbox.org/

Classe	Propriedade	Formatos suportados
Documentos de texto	Número de páginas	Portable Document Format, version 1.4
	Número de imagens	
	Conformidade de caracteres	
	Margem esquerda	
	Margem inferior	
	Margem superior	
	Margem direita	
	Largura de página	
	Altura de página	
	Cor de fundo	
	Tipos de letra	
	Metainformação embebida	
	Disposição gráfica	

Tabela 38 – Propriedades extraídas pela ferramenta PDFBox.

8.2 Taxionomia geral de avaliação

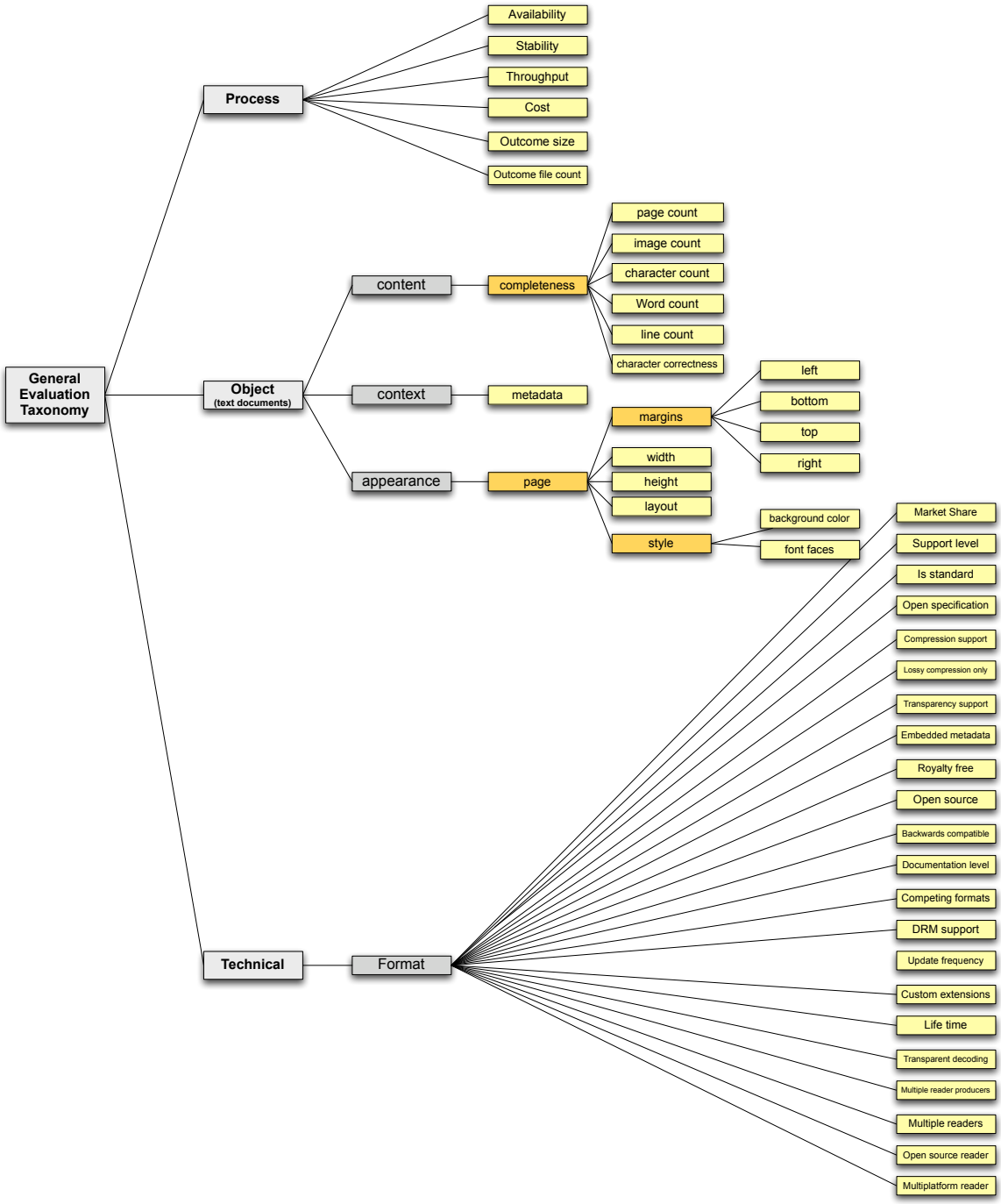


Figura 64 – Taxionomia geral de avaliação.

8.3 Funções de similaridade

O conceito de similaridade diz respeito à proximidade, real ou percepcionada, existente entre dois conceitos ou representações mentais. Estes conceitos são, geralmente, representados por pontos no espaço e a sua similaridade está directamente relacionada com a distância a que estes pontos se encontram nesse espaço (R. N. Shepard, 1962).

Existem diversas métricas que permitem determinar a distância entre dois conceitos. A definição formal de uma função de cálculo de distância é descrita pela Fórmula 10, onde M representa o tipo de dados dos conceitos a analisar.

$$d_M : M \times M \rightarrow \Re^+$$

Fórmula 10 – Definição matemática de distância.

Uma métrica deste tipo deve obedecer ao seguinte conjunto de condições:

- $d(x,y) \geq 0$ – a distância entre dois conceitos é limitada inferiormente;
- $d(x,y) = 0 \text{ sse } x = y$ – a distância entre dois conceitos é zero se e só se os dois conceitos forem iguais;
- $d(x,y) = d(y,x)$ – a distância entre o conceito *x* e o conceito *y* é igual à distância entre o conceito *y* e o conceito *x* (i.e., simetria);
- $d(x,z) \leq d(x,y) + d(y,z)$ – a distância entre dois pontos é sempre a menor distância entre ambos os pontos.

A distância é muitas vezes utilizada para determinar a similaridade entre dois conceitos. A Equação 7 estabelece a relação entre distância e similaridade.

$$similaridade = \frac{1}{1 + distância}$$

Equação 7 – Relação entre similaridade e distância.

Uma função de similaridade pode ser definida formalmente pela Fórmula 11, onde *M* representa o tipo de dados do conceito que se pretende comparar.

$$s_M : M \times M \rightarrow [0,1]$$

Fórmula 11 – Definição matemática de similaridade.

Tal como acontecia com a distância, a similaridade também deve obedecer a um conjunto bem definido de condições, nomeadamente:

- $s(x,y) \leq 1$ – a similaridade entre dois conceitos é limitada superiormente;
- $s(x,y) = 1$ sse $x = 1$ – a similaridade entre dois conceitos é igual a 1 se e só se os dois conceitos forem iguais;
- $s(x,y) = s(y,x)$ – a função de similaridade é simétrica.

A Tabela 39 e a Tabela 40 enumeram as diversas propriedades significativas avaliadas no contexto do CRiB e quais as métricas utilizadas na sua comparação.

Critério de avaliação	**Tipo de dados**	**Métrica de comparação**
Número de páginas	Numérico	Proportional Similarity
Conformidade gráfica	Matriz de cor	NRMSE Similarity UQI Similarity SSIM Similarity CBM Similarity
Largura	Numérico	Proportional Similarity
Altura	Numérico	Proportional Similarity
Modelo de cor	Textual	Relaxed String Equality
Profundidade de cor	Numérico	Proportional Similarity
Metainformação embebida	XML	Property Set Similarity XML Diff
Método de compressão	Textual	Relaxed String Equality

Tabela 39 – Métricas utilizadas para comparar imagens matriciais.

As secções que se seguem descrevem o conjunto de funções de similaridade utilizadas no âmbito do CRiB, em particular pelo componente Object Evaluator. Estas encontram-se organizadas por categorias de acordo com o tipo de dados que manipulam: numérico, vectorial, textual, conjuntos, XML ou informação gráfica do tipo matricial.

Critério de avaliação	Tipo de dados	Métrica de comparação
Número de páginas	Numérico	Proportional Similarity
Número de imagens	Numérico	Proportional Similarity
Conformidade de caracteres	Textual	Jaro Winkler String Similarity
Margem esquerda	Numérico	Proportional Similarity
Margem inferior	Numérico	Proportional Similarity
Margem superior	Numérico	Proportional Similarity
Margem direita	Numérico	Proportional Similarity
Largura de página	Numérico	Proportional Similarity
Altura de página	Numérico	Proportional Similarity
Disposição gráfica	Matriz de cor	NRMSE Similarity UQI Similarity SSIM Similarity CBM Similarity
Cor de fundo	Vectorial	Euclidean distance
Tipos de letra	Textual	Relaxed String Equality
Metainformação embebida	XML	Property Set Similarity XML Diff

Tabela 40 – Métricas utilizadas para comparar documentos de texto.

8.3.1 Similaridade numérica

A similaridade numérica serve para comparar quantidades ou valores absolutos. Este tipo de métricas é amplamente utilizado pela plataforma CRiB para comparar propriedades extraídas de objectos digitais caracterizadas por valores numéricos, como: largura, altura, comprimento em bytes, número de caracteres, etc.

Proportional Similarity

A métrica Proportional Similarity, ou similaridade proporcional, é definida à custa da distância proporcional. Esta distância, tal como o nome indica, procura determinar a diferença entre dois valores numéricos, porém, tem em consideração o nível de grandeza dos mesmos. Por exemplo, a distância entre 3 e 5 é igual a 2; o mesmo acontece com os valores 1003 e 1005. Não obstante, no primeiro exemplo, o valor 5 é 66.6% superior ao valor 3, enquanto que no segundo, o valor 1005 é apenas 0.0019% superior que 1003.

A distância proporcional encontra-se definida na Fórmula 12.

$$\text{ProportionalDistance(a, b)} = \begin{cases} 0 & , a = b \\ \dfrac{|a-b|}{\max(a,b)} & , a \neq b \end{cases}$$

Fórmula 12 – Distância proporcional.

A similaridade proporcional é determinada aplicando a Equação 7 à fórmula de cálculo da distância proporcional.

$$\text{ProportionalSimilarity}(a,b) = \frac{1}{1 + \text{ProportionalDistance}(a,\ b)}$$

Fórmula 13 – Similaridade proporcional.

Esta métrica é utilizada para determinar o nível de degradação sofrido por um objecto digital durante uma conversão em propriedades significativas como: número de páginas, largura e altura, profundidade de cor, dimensões de margens, etc. (ver Tabela 39 e Tabela 40 para uma lista completa das propriedades significativas analisadas por esta métrica).

8.3.2 Similaridade vectorial

A similaridade vectorial é utilizada na comparação de vectores. A noção comum de vector é a de um objecto com tamanho, direcção e sentido, que implementa as operações de adição e multiplicação por números reais . Genericamente, um vector pode ser considerado uma sequência de valores reais sendo representado da seguinte forma: $V = (v_1\ \ v_2,\ \ldots,\ v_n)$.

Similaridade Euclidiana

A similaridade euclidiana permite determinar a semelhança entre dois vectores numéricos. Formalmente, sejam $\mathbf{P} = (p_1, p_2, \ldots, p_n)$ e $\mathbf{Q} = (q_1, q_2, \ldots, q_n)$ dois vectores de comprimento *n,* a distância euclidiana entre ambos é definida pela fórmula $\sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$.

A similaridade euclidiana (Fórmula 14) é determinada à custa da distância, aplicando a transformação definida anteriormente na Equação 7.

$$\text{EuclideanSimilarity}(P,Q) = \frac{1}{1 + \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}}$$

Fórmula 14 – Similaridade euclidiana.

Esta métrica de similaridade é utilizada no âmbito do CRiB para comparar cores. Um cor é definida computacionalmente como um vector de quatro elementos. Três destes definem a intensidade das cores vermelho, verde e azul (i.e., RGB) e o quarto que define a transparência

do ponto de cor (i.e., Alfa), e.g. $C = (r,\ g,\ b,\ a)$. Os valores de cada um dos elementos do vector é um número natural pertencente ao conjunto [0, 254].

8.3.3 Similaridade textual

A similaridade textual[142] tem como missão determinar a proximidade existente entre duas cadeias de caracteres (Navarro, 2001). Por exemplo, as palavras "toca" e "foca" podem ser consideradas sintacticamente semelhantes na medida em que diferem entre si apenas numa letra.

Este tipo de métricas é amplamente utilizado em contextos de recuperação de informação como motores de pesquisa ou sistemas de gestão de bases de dados. São também muito frequentes na detecção de fraude, análise de dados biométricos, sistemas de detecção de plágio, alinhamento de ontologias, análise de ADN, *data mining*, *data cleansing*, etc (Cohen, Ravikumar, & Fienberg, 2003; Navarro, 2001; Soukoreff & MacKenzie, 2001).

Distância de Levenshtein

A distância de Levenshtein é um algoritmo que permite quantificar as diferenças existentes entre duas cadeias de caracteres. Esta medida de distância contabiliza o número de operações de inserção, eliminação e/ou substituição que são necessárias para transformar uma cadeia de caracteres numa segunda (Levenshtein, 1965). Por exemplo, a distância de Levenshtein entre os termos "automovel" e "automóveis" é 4, devido a:

1. Substituição de "o" por "ó"
2. Eliminação de "l"
3. Inserção de "i"
4. Inserção de "s"

O algoritmo da distância de Levenshtein encontra-se definido na Figura 65. A medida de similaridade correspondente pode ser obtida aplicando a fórmula de transformação introduzida na secção 8.2.

[142] Também conhecido por *string matching that allows errors* ou *approximate string matching*.

O CRiB faz uso de uma biblioteca *open-source* designada SimMetrics[143] que implementa um conjunto alargado de algoritmos de similaridade, incluindo a distância de Levenshtein.

```
1       int LevenshteinDistance(char s[1..m], char t[1..n])
2          // d is a table with m+1 rows and n+1 columns
3          declare int d[0..m, 0..n]
4
5          for i from 0 to m
6              d[i, 0] := i
7
8          for j from 0 to n
9              d[0, j] := j
10
11         for i from 1 to m
12             for j from 1 to n
13             {
14                 if s[i] = t[j] then
15                   cost := 0
16                 else
17                   cost := 1
18
19                 d[i, j] := minimum(
20                              d[i-1, j]   + 1,   // deletion
21                              d[i, j-1]   + 1,   // insertion
22                              d[i-1, j-1] + cost // substitution
23                            )
24             }
25
26         return d[m, n]
```

Figura 65 – Algoritmo da distância de Levenshtein.

Relaxed String Equality

A função Relaxed String Equality é utilizada pelo CRiB para determinar se duas cadeias de caracteres podem ser consideradas iguais, apesar de não o serem na sua totalidade.

A função define um nível de similaridade *T* a partir do qual duas cadeias de caracteres são consideradas iguais. A função é definida à custa da similaridade de Levenshtein de acordo com a Fórmula 15. O valor de *T* definido por omissão é 0.7.

Esta métrica é utilizada pelo componente Object Evaluator para determinar se os tipos de letra incluídos em documentos de texto podem ser considerados iguais. A utilização desta métrica torna-se necessária, pois determinados formatos utilizam a designações ligeiramente diferentes para designar o mesmo tipo de letra. Por exemplo, o formato PDF utiliza designações como "TimesNewRomanPSMT" para designar o tipo de letra que o Word interpreta como "Times New Roman".

[143] http://sourceforge.net/projects/simmetrics/

$$\text{RelaxedStringEquality}(s,\ v,\ T) = \begin{cases} 0, & \text{LevenshteinSimilarity}(s,v) < T \\ 1, & \text{LevenshteinSimilarity}(s,v) \geq T \end{cases}$$

Fórmula 15 – Igualdade textual relaxada.

Esta métrica é também utilizada para determinar se os modelos de cor e os métodos de compressão de duas imagens matriciais podem ser considerados iguais.

Jaro Winkler String Similarity

O algoritmo da distância de Levenshtein utiliza uma matriz de caracteres com largura e altura iguais ao comprimento das duas cadeias de caracteres que deverão ser comparadas, i.e., se se pretender determinar a similaridade entre dois documentos de 10 páginas (aproximadamente 45.000 caracteres), isto iria exigir a construção de uma matriz com aproximadamente 45.000×45.000 células. Cada uma destas células seria ocupada por um caracter, algo que necessita de pelo menos 1 byte para que pudesse ser armazenado em memória. Isto resultaria numa matriz de tamanho $45.000 \times 45.000 \times 1$ bytes, ou seja, aproximadamente 1.9 Gigabytes, tornando a utilização deste algoritmo incomportável para a maioria dos documentos e computadores actuais.

Para comparar o conteúdo textual de dois documentos de texto, foi utilizada, em alternativa, a métrica de Jaro Winkler (Winkler, 1999). Esta métrica estende a métrica de Jaro anteriormente publicada (Jaro, 1989, 1995). Dadas duas cadeias de caracteres s_1 e s_2 a similaridade de Jaro é definida pela Fórmula 16 onde m representa o número de caracteres comuns entre s_1 e s_2 (localizados sensivelmente na mesma posição[144]) e t o número de transposições[145] necessárias para que s_1 se transforme em s_2.

$$Jaro(s_1,s_2) = \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right)$$

Fórmula 16 – Métrica de comparação de cadeias de caracteres de Jaro.

[144] Para determinar se dois caracteres estão sensivelmente na mesma posição é utilizada uma janela de tamanho 3.
[145] Uma transposição é uma troca de lugar entre dois membros da mesma sequência de caracteres.

Por exemplo, a similaridade de Jaro entre os termos "toca" e "foca" é determinada da seguinte forma:

- $s_1 = \text{toca}$ e $s_2 = \text{foca}$ (cadeias de caracteres comparadas)
- $|s_1| = 4$ e $|s_2| = 4$ (comprimento das cadeias de caracteres)
- $m = 3$ (número de caracteres em comum)
- $t = 0$ (número de transposições que transformam s_1 em s_2)
- $Jaro(s_1, s_2) = \frac{1}{3}\left(\frac{3}{4} + \frac{3}{4} + \frac{3-0}{3}\right) = 0.833$ (valor de similaridade de Jaro)

A similaridade de Jaro-Winkler difere da métrica de Jaro pelo facto de atribuir valores superiores a cadeias de caracteres que partilham a mesma sequência inicial (Winkler, 1999). Assim, seja P o comprimento do prefixo comum entre s_1 e s_2 e $P' = \max(P, 4)$, a similaridade de Jaro-Wikler é definida pela Fórmula 17.

$$\text{JaroWinkler}(s_1, s_2) = Jaro(s_1, s_2) + \frac{P'}{10}(1 - Jaro(s_1, s_2))$$

Fórmula 17 – Similaridade de Jaro-Winkler.

Esta métrica é utilizada no contexto do CRiB para determinar a similaridade textual entre dois documentos de texto.

8.3.4 Similaridade entre conjuntos

Existem várias métricas que permitem determinar a similaridade entre dois conjuntos. Este tipo de métricas é, tradicionalmente, utilizado em contextos de *data cleansing* para detectar múltiplas representações da mesma entidade (Arasu, Ganti, & Kaushik, 2006; Hadjieleftheriou, Chandel, Koudas, & Srivastava, 2008).

Property Set Similarity

Uma das métricas de comparação de conjuntos mais utilizada designa-se por Coeficiente de Similaridade de Jaccard[146] (Jaccard, 1901; Tan et al., 2005). Esta métrica calcula a similaridade entre dois conjuntos, dividindo o número de elementos que compõem a

[146] Esta métrica é também conhecida por *Jaccard Index* e *Jaccard Similarity*.

intersecção dos dois conjuntos pelo número de elementos de constituem a sua reunião (Fórmula 18). O contradomínio da função é definido pelo intervalo [0, 1].

$$\text{JaccardSimilarity(A, B)} = \frac{|A \cap B|}{|A \cup B|}$$

Fórmula 18 – Coeficiente de similaridade de Jaccard.

Esta métrica foi modificada pelo autor desta tese de modo a adequá-la à comparação de conjuntos de pares ordenados do tipo (atributo, valor). A modificação consistiu na introdução de uma função *first* que, dado um conjunto de pares ordenados, produz um novo conjunto constituído apenas pelo primeiro elemento de cada par (Fórmula 19).

$$X = \{(a_1,\ v_1),\ (a_2,\ v_2),\ ...,\ (a_n,\ v_n)\}$$
$$first(X) = \{a_1,\ a_2,\ ...,\ v_n\}$$

Fórmula 19 – Função *first*.

O Coeficiente de Similaridade de Jaccard foi então enriquecido com a nova função resultando na Fórmula 20. A nova métrica foi designada Property Set Similarity.

$$\text{PropertySetSimilarity(A, B)} = \frac{|A \cap B|}{|first(A) \cup first(B)|}$$

Fórmula 20 – Coeficiente de similaridade de Jaccard modificado.

Esta métrica foi utilizada no contexto do CRiB para determinar a similaridade entre a metainformação embebida em dois objectos digitais distintos.

8.3.5 Similaridade de XML

Os documentos XML assumem actualmente uma grande relevância no contexto da representação de informação e publicação electrónica. Existe uma linha de investigação que se dedica ao desenvolvimento de métricas de similaridade para documentos XML. Este tipo de métricas é utilizado em contextos de recuperação de informação, sistemas de controlo de versões (e.g. CVS, SVN), *data warehousing* (para gestão de índices) e classificação automática de documentos, *clustering,* etc. (Tekli, Chbeir, & Yetongnon, 2006).

XML Diff

A métrica de similaridade XML Diff desenvolvida pela Universidade de Sannio[147] tem como objectivo determinar a proximidade sintáctica entre dois documentos XML (Canfora et al., 2004). Esta métrica combina três características fundamentais durante o processo de comparação de documentos XML, nomeadamente:

- Similaridade estrutural – os documentos comparados deverão apresentar a mesma estrutura;
- Similaridade de conteúdo – os documentos devem possuir o mesmo conteúdo textual;
- Similaridade posicional – o conteúdo textual dos documentos deve encontrar-se nas mesmas posições da árvore documental.

O algoritmo original foi ligeiramente modificado pelo autor de modo a suportar conteúdos armazenados em atributos e não apenas em elementos.

Esta métrica foi utilizada para determinar o nível de similaridade existente entre metainformação extraída a partir de objectos digitais.

8.3.6 Similaridade gráfica

Uma imagem matricial é definida por uma matriz de $M \times N$ pontos coloridos (Figura 66). Cada um destes pontos é constituído por três componentes de cor vermelho, verde e azul, e um quarto componente representando a transparência global do ponto, i.e., $C_{xy} = (r,\ g,\ b,\ a)$.

$$\begin{bmatrix} C_{11} & \cdots & C_{M1} \\ \vdots & \ddots & \vdots \\ C_{1N} & \cdots & C_{MN} \end{bmatrix}$$

Figura 66 – Definição formal de imagem matricial.

Existe um vasto conjunto de métricas que poderão ser utilizadas no cálculo de similaridade entre duas imagens. Estas têm aplicação em variados domínios como: remoção de imagens

[147] http://www.unisannio.it/

duplicadas, recuperação de informação, optimização de algoritmos de compressão, controlo de qualidade, *clustering*, etc.

As métricas de comparação de imagens podem ser divididas em duas categorias: métricas objectivas e métricas subjectivas. As primeiras utilizam cálculos matemáticos para determinar o nível de similaridade entre duas imagens. As segundas recorrem a um conjunto de avaliadores humanos que efectuam a respectiva comparação e atribuem uma classificação ao grau de similaridade percepcionado (Biström, 2005).

Acontece que, o uso de pessoas na realização deste tipo de avaliações impossibilita a automatização destes processos e torna a avaliação demasiado onerosa, tanto em termos de tempo como em termos de dinheiro. Por outro lado, o uso de métricas objectivas nem sempre produz resultados suficientemente precisos e/ou ajustados à realidade. Dependendo da aplicação, as avaliações produzidas por métodos objectivos podem não se correlacionar inteiramente com as percepções dos avaliadores humanos (Biström, 2005; Z. Wang et al., 2004).

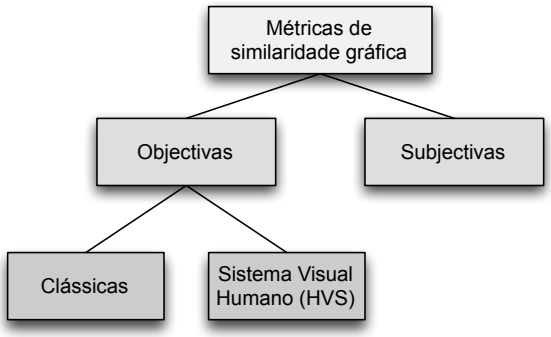


Figura 67 – Classes de métricas de similaridade gráfica.

As métricas objectivas podem ainda ser divididas em duas classes distintas: métricas objectivas clássicas ou métricas baseadas no sistema visual humano[148]. A primeira classe de métricas considera apenas as características matemáticas que são intrínsecas à imagem. A segunda, procura incorporar no seu modelo de avaliação itens que são próprios da percepção humana. Esta abordagem tem como objectivo tornar estes algoritmos mais parecidos com as avaliações subjectivas (Z. Wang & Bovik, 2002).

[148] Do inglês *Human Visual System* (HVS).

Normalized Root Mean Squared Error

Uma das técnicas de comparação de imagens mais utilizadas designa-se por Root Mean Squared Error (RMSE). Este método consiste no cálculo da média das distâncias euclidianas verificadas entre cada ponto de cor que constitui cada uma das imagens comparadas (Shrestha et al., 2005; L. W. Wang, Zhang, & Feng, 2005; Z. Wang et al., 2004).

A Fórmula 21 define formalmente esta métrica, onde *u* e *v* representam duas imagens de tamanho $M \times N$, sendo $u(x,y,i)$ e $v(x,y,i)$ o valor da intensidade da componente de cor *i* na posição *x* e *y* em cada uma das imagens. As funções $\max(u,v,i)$ e $\min(u,v,i)$ determinam os valores de intensidade máximo e mínimo da componente de cor *i* encontrados em ambas as imagens *u* e *v*.

$$RMSE(u,v,i) = \sqrt{\sum_{x=1}^{M}\sum_{y=1}^{N}\left|u(x,y,i) - v(x,y,i)\right|^2}$$

$$NRMSE(u,v) = \frac{1}{4}\sum_{i=1}^{4}\frac{RMSE(u,v,i)}{\max(u,v,i) - \min(u,v,i)}$$

Fórmula 21 – Normalized Root Mean Squared Error (NRMSE).

A métrica NRMSE é uma medida de distância. Para se tornar numa medida de similaridade é necessário aplicar a Equação 7 anteriormente apresentada.

Esta métrica é utilizada no contexto do CRiB para determinar o nível de degradação gráfica sofrido por um objecto digital durante a sua migração, ou seja, corresponde à propriedade conformidade gráfica tanto em imagens matriciais como em documentos de texto.

Universal Quality Index

A métrica Universal Image Quality Index (UQI) pertence à classe de algoritmos que incorpora características do sistema visual humano na sua avaliação de similaridade. Este algoritmo tem em consideração aspectos como luminância, contraste e estrutura das imagens comparadas (Z. Wang & Bovik, 2002).

A métrica UQI encontra-se definida formalmente na Fórmula 22 onde u e v representam duas imagens na sua forma vectorial (ao invés de matricial), i.e., $u = \left(u_i \mid i = 1,2,...,N\right)$ e

$v = \left(v_i \mid i = 1,2,...,N\right)$, com u_i e v_i a representar os pontos de cor que constituem ambas as imagens sob a forma $u_i = \left(r,g,b,a\right)$ e $v_i = \left(r',g',b',a'\right)$.

$$UQI_i(u,v) = \frac{2\bar{u}\bar{v}}{(\bar{u})^2 + (\bar{v})^2} \cdot \frac{2\sigma_u\sigma_v}{\sigma_u^2 + \sigma_v^2} \cdot \frac{\sigma_{uv}}{\sigma_u\sigma_v}$$

Fórmula 22 – Universal Image Quality Index (UQI) de uma componente de cor.

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N} u_i \qquad \bar{v} = \frac{1}{N}\sum_{i=1}^{N} v_i \qquad \text{(Média)}$$

$$\sigma_u^2 = \frac{1}{N-1}\sum_{i=1}^{N}(u_i - \bar{u})^2 \qquad \sigma_v^2 = \frac{1}{N-1}\sum_{i=1}^{N}(v_i - \bar{v})^2 \qquad \text{(Variância)}$$

$$\sigma_{uv} = \frac{1}{N-1}\sum_{i=1}^{N}(u_i - \bar{u})(v_i - \bar{v}) \qquad \text{(Covariância)}$$

Fórmula 23 – Fórmulas auxiliares ao cálculo de UQI.

Por uma questão de clareza, a formulação de UQI apresentada na Fórmula 22 apenas se aplica a uma das quatro componentes de cor que constitui cada uma das imagens comparadas. A mesma fórmula deverá ser aplicada separadamente a cada uma das componentes de cor, sendo o valor global de UQI obtido a partir do valor médio dos UQI_i parciais (Fórmula 24).

$$UQI = \frac{1}{4}\sum_{i=1}^{4} UQI_i(u,v)$$

Fórmula 24 – Valor global de UQI.

Esta métrica é utilizada no contexto do CRiB para determinar o nível de degradação gráfica sofrido por um objecto digital durante a sua migração, ou seja, corresponde à propriedade conformidade gráfica tanto em imagens matriciais como em documentos de texto.

Structural Similarity

A métrica designada por Structured Similarity (SSIM) procura generalizar os conceitos incorporados na métrica UQI tornando esta métrica mais flexível e, ao mesmo tempo, configurável. A nova métrica continua a combinar os conceitos de luminância, contraste e estrutura, mas incorpora constantes na sua formulação, nomeadamente C_1, C_2 e C_3, que

evitam que o algoritmo se comporte de forma instável na presença de imagens com determinadas características, como por exemplo, imagens com grandes superfícies da mesma cor (Z. Wang et al., 2004) – Fórmula 25.

$$l(u,v) = \frac{2\bar{u}\bar{v} + C_1}{(\bar{u})^2 + (\bar{v})^2 + C_1} \qquad \text{(Luminância)}$$

$$c(u,v) = \frac{2\sigma_u\sigma_v + C_2}{\sigma_u^2 + \sigma_v^2 + C_2} \qquad \text{(Contraste)}$$

$$s(u,v) = \frac{\sigma_{uv} + C_3}{\sigma_u\sigma_v + C_3} \qquad \text{(Estrutura)}$$

Fórmula 25 – Fórmulas auxiliares ao cálculo de SSIM.

O novo algoritmo recebe ainda como parâmetros o peso que cada um dos conceitos anteriormente mencionados (i.e., luminância, contraste e estrutura) terá na apreciação global de similaridade, i.e., α, β e γ (Fórmula 26).

$$SSIM_i(u,v) = \left[l(u,v)\right]^{\alpha} \cdot \left[c(u,v)\right]^{\beta} \cdot \left[s(u,v)\right]^{\gamma}$$

Fórmula 26 – Structural Similarity (SSIM) de uma componente de cor.

Tal como acontecia no cálculo de UQI, a fórmula de $SSIM_i$ apenas considera uma das quatro componentes de cor que constituem as imagens. Para obter uma apreciação global de SSIM é necessário, em primeiro lugar, calcular a média dos valores de SSIM obtidos para cada uma das quatro componentes de cor (Fórmula 27).

$$SSIM(u,v) = \frac{1}{4}\sum_{i=1}^{4} SSIM_i(u,v,i)$$

Fórmula 27 – Valor de SSIM que combina as quatro componentes de cor.

Para além do disposto, esta métrica é aplicada não à imagem completa mas apenas a um conjunto aleatório de janelas gaussianas de raio 11 *pixel* (Z. Wang et al., 2004). O valor global

de similaridade é obtido calculando a média dos valores de SSIM resultantes da aplicação do algoritmo a cada uma das *M* janelas previamente recolhidas (Fórmula 28).

$$MSSIM(u,v) = \frac{1}{M}\sum_{j=1}^{M} SSIM(u,v)$$

Fórmula 28 – Valor global de SSIM que combina os valores de SSIM das *M* janelas amostradas.

Esta métrica é utilizada no contexto do CRiB para determinar o nível de degradação gráfica sofrido por um objecto digital durante a sua migração, ou seja, corresponde à propriedade conformidade gráfica tanto em imagens matriciais como em documentos de texto.

Content-Based Image Quality Metric

A métrica designada por Content-Based Image Quality Metric (CBM) estende a métrica SSIM anteriormente descrita na medida em que, para além de considerar propriedades como luminância, contraste e estrutura, considera também os contornos, texturas e regiões planas das imagens comparadas (Gao et al., 2005). O algoritmo começa por particionar as imagens nestas três componentes recorrendo a uma máscara de Sobel, processo descrito detalhadamente em (Duda & Hart, 1973; Li, Chen, Chi, & Lu, 2004; Sobel & Feldman, 1968) – Figura 68.

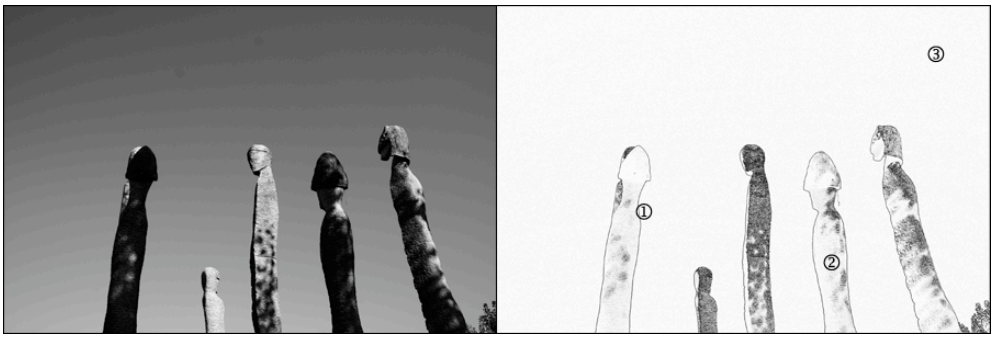


Figura 68 – Detecção de (1) contornos, (2) texturas e (3) regiões planas usando uma máscara de Sobel.

Após o particionamento das imagens, o algoritmo CBM recorre à métrica SSIM para determinar a similaridade em cada uma destas componentes extraídas. Finalmente o valor final de CBM é obtido calculando a média dos valores de SSIM obtidos (Figura 69).

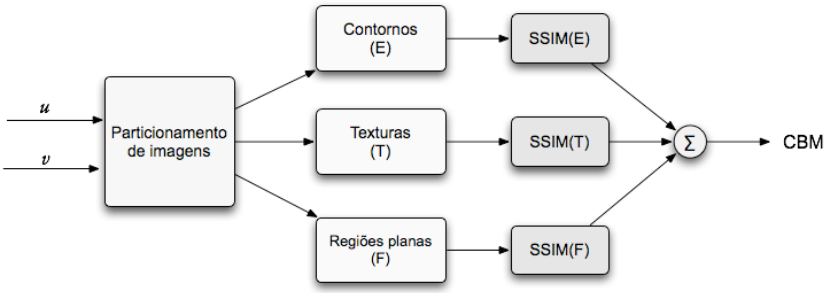


Figura 69 – Diagrama de processamento da métrica CBM.

Esta métrica é utilizada no contexto do CRiB para determinar o nível de degradação gráfica sofrido por um objecto digital durante a sua migração, ou seja, corresponde à propriedade conformidade gráfica tanto em imagens matriciais como em documentos de texto.

8.4 Teste não-paramétrico de Wilcoxon

Com o objectivo de determinar qual dos vários algoritmos de similaridade gráfica estudados apresentava melhores resultados, foi realizado o teste de hipóteses não-paramétrico de Wilcoxon. Com base neste teste foi possível comparar estatisticamente duas amostras independentes e determinar se estas poderiam ser consideradas equivalentes, i.e., se a média da primeira amostra era estatisticamente equivalente à média da segunda amostra – $\mu_1 = \mu_2$.

O teste de Wilcoxon é uma alternativa não-paramétrica ao teste t-Student, geralmente utilizado na comparação de médias. A diferença fundamental entre o teste t-Student e o teste de Wilcoxon reside no facto de o primeiro assumir que as observações seguem uma distribuição Normal enquanto que o segundo não faz qualquer tipo de assumpção relativamente à distribuição subjacente.

Este teste de hipóteses permite determinar se a distribuição das avaliações automáticas é estatisticamente semelhante à das avaliações produzidas pelos avaliadores humanos, e não qual dos algoritmos apresenta o melhor desempenho quando comparado com os valores de MOS. O teste pode ser visto como uma verificação rápida da elegibilidade de uma dada amostra, ou seja, permite rejeitar um algoritmo de similaridade sem ter de realizar uma análise profunda do seu desempenho.

Assim, considerando que d_k representa a diferença entre as medições de similaridade produzidas pelos humanos (X_k) e por cada um dos algoritmos objectivos (Y_k) – Fórmula 29.

$$d_k = X_k - Y_k, \ \ para \ \ k = 1,2,...,30$$

Fórmula 29 – Diferença entre as avaliações subjectivas e os valores objectivos.

Tem-se como hipótese nula verificar se as distribuições de ambos os processos de avaliação se encontram simetricamente distribuídas em torno de uma média comum μ. Por outras palavras, pretende-se descobrir se a diferença entre ambas as médias de ambas as medições é igual a zero.

$$H_0 : \mu = 0 \quad (Hipótese \ \ nula)$$
$$H_1 : \mu \neq 0 \quad (Hipótese \ \ alternativa)$$

Fórmula 30 – Formulação de hipóteses.

Antes de aplicar o teste de Wilcoxon, procedeu-se ao ajuste dos valores objectivos produzidos pelos algoritmos recorrendo à regressão linear. Os valores ajustados utilizados no teste paramétrico apresentam-se na Tabela 20 na página 150.

Amostras comparadas	Wilcoxon Valor-P	Valor-P > 0.05	Conclusão
MOS-RMSE	0.629	Sim	Não há evidência suficiente para rejeitar H_0
MOS-UQI	0.781	Sim	Não há evidência suficiente para rejeitar H_0
MOS-SSIM	0.845	Sim	Não há evidência suficiente para rejeitar H_0
MOS-CBM	0.861	Sim	Não há evidência suficiente para rejeitar H_0

Tabela 41 – Resultados da aplicação do teste de Wilcoxon para comparação de médias.

Os resultados do teste de Wilcoxon encontram-se resumidos na Tabela 41. Considerando um grau de confiança de 95% (α = 0.05), é possível concluir que não há evidências suficientes para se rejeitar a hipótese nula. Assim, uma vez que todos os algoritmos apresentam uma relação suficientemente forte com as avaliações humanas, partiu-se para a aplicação das três métricas previamente descritas de modo a determinar qual dos algoritmos apresentava o melhor desempenho. O conjunto de experiências realizado nesse sentido encontra-se descrito na secção 5.1.2 na página 137.

8.5 Validação cruzada

A validação cruzada ou *cross-validation* foi aplicada pela primeira vez por Seymour Geisser, um profissional de estatística cujos trabalhos incidiram sobre a análise de métodos estatísticos de previsão. A validação cruzada é um método estatístico prático que toma por base uma amostra

de dados subdividida em várias partições: umas são usadas para treinar o sistema e as restantes para o testar. Dentro deste método podemos encontrar várias variantes:

- Camilo Oliveira na sua tese de mestrado descreve o *holdout validation* como um dos métodos mais utilizados, sendo também designado por teste de cálculo simples, em que se divide o conjunto de dados em dois subconjuntos, designados por conjunto de treino e de teste. Este autor considera que é um método de cálculo "pessimista" porque só uma parte dos dados é utilizada para treino (Oliveira, 2001);

- O método *random subsampling* ou validação cruzada com subamostragem aleatória, consiste na separação de um número de elementos de treino de forma aleatória. Numa segunda experiência, separa-se o mesmo número de exemplos, mas desta feita em posições excludentes. Repete-se esta separação em todas as experiências que possam existir (Oliveira, 2001);

- O método *K-fold cross-validation* consiste em dividir um conjunto de amostras de tamanho N em K partições mutuamente excludentes e de igual tamanho (Oliveira, 2001). Das K partes, K-1 serão utilizadas para treinar o sistema, enquanto que a restante será utilizada para o testar. O processo é repetido K vezes. Em cada repetição ensaia-se e valida-se o modelo. No final será calculada a média dos resultados obtidos em cada uma das K validações realizadas (Kohavi, 1995).

- *Leave-one-out cross-validation*, é um método de validação cruzada deixando apenas um individuo de fora para testar o sistema. Neste caso, K=N onde o número de partes é igual ao número de elementos do conjunto original. Por exemplo, para um conjunto de dados com N exemplos, executa-se N experiências. Para cada uma delas, utiliza-se N-1 indivíduos de treino e apenas um é reservado para teste (Oliveira, 2001). Segundo Ron Kohavi, este método é excelente, quase imparcial, mas a sua desvantagem reside na sua alta variância (Kohavi, 1995);

O esquema seguinte apresenta um exemplo *K-fold cross-validation*, utilizando 4 dobras (Oliveira, 2001)

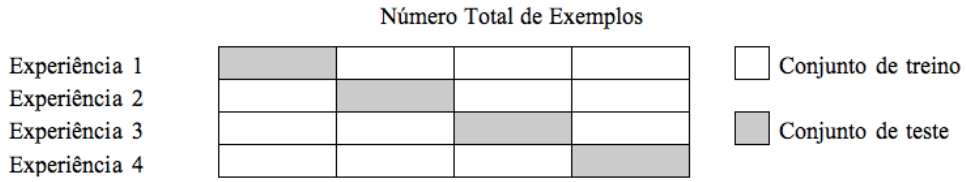


Figura 70 – Exemplo do método de validação cruzada com 4 dobras.

8.6 Licença de uso e distribuição do CRiB

```
CRiB | Conversion and Recommendation of Digital Object Formats
Copyright(c) 2008 Miguel Ferreira <mferreira@dsi.uminho.pt>
All Rights Reserved.

This software was developed with the Department of Information Systems
of the University of Minho, Portugal, under the supervision
of Ana Alice Baptista <analice@dsi.uminho.pt> and José Carlos Ramalho
<jcr@di.uminho.pt>.
------------------------------------------------------------------------
Permission to use, copy, or modify this software and its documentation
for educational and research purposes only and without fee is hereby
granted, provided that this copyright notice and the original authors'
names appear on all copies and supporting documentation. This program
shall not be used, rewritten, or adapted as the basis of a commercial
software or hardware product without first obtaining permission of the
authors. The authors make no representations about the suitability of
this software for any purpose. It is provided "as is" without express
or implied warranty.

THE NAME AND TRADEMARKS OF COPYRIGHT HOLDERS MUST ALWAYS BE INCLUDED
OR ASSOCIATED TO ANY ADVERTISING, PUBLICITY OR DISTRIBUTION OF THIS
SOFTWARE AND ITS DOCUMENTATION. TITLE TO COPYRIGHT THIS SOFTWARE AND
ANY ASSOCIATED DOCUMENTATION WILL AT ALL TIMES REMAIN WITH THE
COPYRIGHT HOLDERS.
------------------------------------------------------------------------
This software is part of the CRiB platform. The CRiB is a Service
Oriented Architecture (SOA) designed to assist cultural heritage
institutions in the implementation of migration-based preservation
interventions. The CRiB works by assessing the quality of distinct
conversion services to produce recommendations of optimal migration
strategies. The recommendations produced by the system take into
account the specific preservation requirements of each client
institution.

For additional information, please refer to the following papers and
Web sites:

- Ferreira, M., Baptista, A. A., & Ramalho, J. C. (2007). An
  intelligent decision support system for digital preservation.
  International Journal on Digital Libraries, 6(4), 295-304.
```

- Ferreira, M., Baptista, A. A. & Ramalho, J. C. (2006). A Foundation for Automatic Digital Preservation. Ariadne(48).

- CRiB homepage: http://crib.dsi.uminho.pt

- Author homepage: http://www.dsi.uminho.pt/~ferreira

Kindly report any suggestions or corrections to mferreira@dsi.uminho.pt

Capítulo 9
Anexos

9.1 Interpretação de valores-P

- Valor-*p* próximo de 0 – Um indicador de que a hipótese nula é falsa.
- Valor-*p* próximo de 1 – Não há evidência suficiente para rejeitar a hipótese nula.
- Normalmente considera-se um valor p de 0,05 como o patamar para avaliar a hipótese nula. Se o valor p for inferior a 0,05 pode-se rejeitar a hipótese nula. Caso contrário, não existe evidência que permita rejeitar a hipótese nula (o que não significa automaticamente que seja verdadeira). Em situações de maior exigência é usado um valor p inferior a 0,05, geralmente 0,01.

REFERÊNCIAS

Abrams, S. L., & Seaman, D. (2003). *Towards a global digital format registry*. Paper presented at the World Library and Information Congress: 69th IFLA General Conference and Council.

Adobe Developers Association. (1992). *TIFF revision 6.0*. Mountain View, USA: Adobe Systems Incorporated.

Adobe Systems Incorporated. (2004). *XMP Specification*. San Jose, USA: Adobe Systems Incorporated.

Akester, P. (2004). Internet law - authenticity of works. Authorship and authenticity in cyberspace. *Computer Law & Security Report, 20*(6).

Ambacher, B., Ashley, K., Berry, J., Brooks, C., Dale, R. L., Flecker, D., et al. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist*: OCLC & CRL.

Arasu, A., Ganti, V., & Kaushik, R. (2006). *Efficient Exact Set-Similarity Joins*. Paper presented at the International Conference on Very Large Data Bases, Seul, Korea.

Arts and Humanities Data Service. (2006). AHDS Repository Policies and Procedures. Retrieved 2006-11-12, from http://ahds.ac.uk/preservation/ahds-preservation-documents.htm

Authenticity Task Force. (2002). *Requirements for Assessing and Maintaining the Authenticity of Electronic Records*. Vancouver, Canada: InterPARES Project.

Ayre, C., & Muir, A. (2004). The Right to Preserve - The Rights Issues of Digital Preservation. *D-Lib Magazine, 10*(3).

Balzer, Y. (2004). Improve your SOA project plans - Strong governance principles ensure a successful outcome. Retrieved 2004-12-12, from http://www-128.ibm.com/developerworks/webservices/library/ws-improvesoa/

Barbedo, F., Corujo, L., Faria, L., Castro, R., Ferreira, M., & Ramalho, J. C. (2007). *RODA: Repositório de Objectos Digitais Autênticos*. Paper presented at the 9º Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, Ponta Delgada, Portugal.

Beagrie, N., Bellinger, M., Dale, R., Doerr, M., Hedstrom, M., Jones, M., et al. (2002). *Trusted Digital Repositories: Attributes and Responsibilities* (Report): Research Libraries Group & Online Computer Library Center.

Bearman, D. (1987). *Collecting Software: A New challenge for Archives & Museums* (No. 1): Archival Informatics.

Bearman, D. (1989). *Archival Methods* (Techical Report No. 1). Pittsburgh: Archives and Museum Informatics.

Becker, C., Ferreira, M., Kraxner, M., Rauber, A., Baptista, A. A., & Ramalho, J. C. (2008). *Distributed Preservation Services: Integrating Planning and Actions*. Paper presented at the European Conference on Research and Advanced Technology for Digital Libraries (ECDL'08), Aarhus, Denmark.

Becker, C., Kulovits, H., Rauber, A., & Hofman, H. (2008). *Plato: A Service Oriented Decision Support System for Preservation Planning*. Paper presented at the Joint Conference on Digital Libraries (JCDL), Pittsburgh, Pennsylvania, USA.

Becker, C., Rauber, A., Heydegger, V., Schnasse, J., & Thalle, M. (2008). *A Generic XML Language for Characterising Objects to Support Digital Preservation*. Paper presented at the Symposium on Applied Computing (SAC), Ceará, Brazil.

Bennett, J. C. (1997). *A Framework of Data Types and Formats, And Issues Affecting the Long Term Preservation of Digital Material* (Report No. 50). West Yorkshire, UK: British Library Research and Innovation Centre.

Besser, H. (2001). Digital Preservation of Moving Image Material? *The Journal of the Association of Moving Image Archivists, 1*(2), 39-55.

Biström, J. (2005). *Comparing Video Codec Evaluation Methods for Handheld Digital TV* (No. 21548C). Helsinki: Helsinki University of Technology.

Brody, T. (2005). Growth of Institutional Archives over Time. Retrieved 2005-12-12, from http://archives.eprints.org/index.php?action=analysis

Brown, A. (2008). *Representation Information Registries* (White Paper No. IST-2006-033789 - PC/3-D7). London, UK: National Archives.

Bryan, D., Draluk, V., Ehnebuske, D., Glover, T., Hately, A., Husband, Y. L., et al. (2002). *UDDI Version 2.04 API Specification*. OASIS.

Burkel, R. (2003). The Role of Microfilm in Information Management. *Information Management Journal, 37*(1), 58-65.

Caldeira, C. P. (2008). *Data Warehousing: Conceitos e Modelos com Exemplos Práticos*: Edições Sílabo.

Canfora, G., Cerulo, L., & Scognamiglio, R. (2004). *Measuring XML document similarity: a case study for evaluating information extraction systems*. Paper presented at the 10th International Symposium on Software Metrics, Chicago, Illinois, USA.

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).

Chen, S.-S. (2001). The Paradox of Digital Preservation. *IEEE Computer, 34*(3), 24-28.

Cohen, W. W., Ravikumar, P., & Fienberg, S. E. (2003). *A Comparison of String Distance Metrics for Name-Matching Tasks*. Paper presented at the Information Integration on the Web (IIWeb), Acapulco, Mexico.

Consultative Committee for Space Data Systems. (2002). *Reference Model for an Open Archival Information System (OAIS) - Blue Book*. Washington: National Aeronautics and Space Administration.

Cullen, C. T. (2000). Authentication of Digital Objects: Lessons from a Historian's Research. In *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources.

Curtis, J., Koerbin, P., Raftos, P., Berriman, D., & Hunter, J. (2007). AONS - An obsolescence detection and notification service for Web archives and digital repositories *New Review of Hypermedia and Multimedia, 13*(1), 39-53.

Darlington, J. (2003). PRONOM - A Practical Online Compendium of File Formats. *RLG DigiNews, 7*(5).

Davidson, A., & Pollard, A. (2005). Jasper - ZX Spectrum Emulator. Retrieved 2005-12-02, from http://www.spectrum.lovely.net/

Diessen, R. J. v. (1997). Model Driven Object-Oriented Development of Systems: A Behavioural-Oriented Approach. Hilversum, The Netherlands.

Diessen, R. J. v., & Werf-Davelaar, T. v. d. (2002). *Authenticity in a digital environment* (Report No. 2). Amesterdam, The Netherlands: Koninklijke Bibliotheek & IBM.

Digital Curation Centre, & DigitalPreservationEurope. (2007). *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*. Glasgow.

Digital Preservation Testbed. (2001). *Migration: Context and Current Status* (White Paper). The Hague.

Duda, R. O., & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*: John Wiley & Sons Inc.

Erl, T. (2005). *Service-oriented Architecture: Concepts, Technology and Design*: Upper Saddle River: Prentice Hall PTR.

Faria, L., Castro, R., Ferreira, M., Ramalho, J. C., Barbedo, F., & Corujo, L. (2007). *RODA - Repository of Authentic Digital Objects*. Paper presented at the International Workshop on Database Preservation, National e-Science Centre, Edinburgh, Scotland.

Farquhar, A., & Hockx-Yu, H. (2007). Planets: Integrated Services for Digital Preservation. *International Journal of Digital Curation, 2*(2).

Fernandes, E. (1999). *Estatística Aplicada*: Serviços de Reprografia e Publicações da Universidade do Minho.

Ferreira, M. (2005). *Automatic Evaluation of Migration Quality in Distributed Networks of Converters*. Paper presented at the Doctoral Consorcium of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL), Vienna, Austria.

Ferreira, M. (2006a). Automatic Evaluation of Migration Quality in Distributed Networks of Converters. *Bulletin of the IEEE Technical Committee on Digital Libraries (TCDL), 2*(2).

Ferreira, M. (2006b). Três anos depois…uma reflexão sobre o projecto DigitArq. In Disciplina de Seminário da Licenciatura em Ciência da Informação da Faculdade de Letras da Universidade do Porto (Ed.). Porto, Portugal.

Ferreira, M., & Baptista, A. A. (2005). *The use of Taxonomies as a way to achieve Interoperability and improved Resource Discovery in DSpace-based Repositories*. Paper presented at the XATA - XML: Aplicações e Tecnologias Associadas, Vila Verde, Braga, Portugal.

Ferreira, M., Baptista, A. A., & Ramalho, J. C. (2005). *Avaliação Automática de Migração em Redes Distribuídas de Conversores*. Paper presented at the Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI), Bragança, Portugal.

Ferreira, M., Baptista, A. A., & Ramalho, J. C. (2006a). A Foundation for Automatic Digital Preservation. *Ariadne*(48).

Ferreira, M., Baptista, A. A., & Ramalho, J. C. (2006b). *CRiB: A service oriented architecture for digital preservation outsourcing*. Paper presented at the XATA - XML: Aplicações e Tecnologias Associadas, Portalegre, Portugal.

Ferreira, M., Baptista, A. A., & Ramalho, J. C. (2007). An intelligent decision support system for digital preservation. *International Journal on Digital Libraries, 6*(4), 295-304.

Ferreira, M., & Ramalho, J. C. (2004a). *Aquisição e Armazenamento de Metainformação no Contexto de um Arquivo*. Paper presented at the XATA - XML: Aplicações e Tecnologias Associadas, Faculdade de Engenharia da Universidade do Porto, Portugal.

Ferreira, M., & Ramalho, J. C. (2004b). *DigitArq - Creating and Managing a Digital Archive*. Paper presented at the ICCC/IFIP International Conference on Electronic Publishing, Brasília, Brazil.

Ferreira, M., & Ramalho, J. C. (2004c). *DigitArq: Creating a Historical Digital Archive*. Paper presented at the 5ª Conferência da Associação Portuguesa de Sistemas de Informação, Lisboa.

Ferreira, M., Saraiva, R., Rodrigues, E., & Baptista, A. A. (2008). Carrots and Sticks - Some ideas on how to create a successful institutional repository. *D-Lib Magazine, 14*(1/2).

Fielding, R. T. (2000). *Architectural Styles and the Design of Network-based Software Architectures*. University of California, Irvine.

Freed, N., & Borenstein, N. (1996). *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types* (RFC No. 2046).

Gantz, J. F., Chute, C., Manfrediz, A., Minton, S., Reinsel, D., Schlichting, W., et al. (2008). *The Diverse and Exploding Digital Universe*. IDC.

Gao, X., Wang, T., & Li, J. (2005). A Content-Based Image Quality Metric. *Springer-Verlag Lecture notes in Computer Science, 3642*(2005), 231-240.

Geremew, M., Song, S., & J. JaJa. (2006). *Using Scalable and Secure Web Technologies to Design a Global Digital Format Registry Prototype: Architecture, Implementation, and Testing*. Paper presented at the IS&T Archiving, Ottawa, Canada.

Goldberg, K., Roeder, T., Gupta, D., & Perkins, C. (2001). Eigentaste: Constant Time Collaborative Filtering Algorithm. *Information Retrieval, 4*(2), 133-151.

Google. (2006). Google Trends. Retrieved 2008-04-21, from http://www.google.com/trends

Graham, P. (2000). Issues in Digital Archiving. In R. Pilette & P. Banks (Eds.), *Preservation: Issues and Planning*. Chicago: IL: American Library Association.

Graham, S., Simeonov, S., Boubez, T., Davis, D., Daniels, G., Nakamura, Y., et al. (2002). *Building Web Services with Java: Making Sense of XML, SOAP, WSDL and UDDI*. Sams Publishing.

Granger, S. (2000). Emulation as a Digital Preservation Strategy. *D-Lib Magazine, 6*(10).

Guenther, R., Caplan, P., Lavoie, B., Bordwell, S., Brandt, O., Clifton, G., et al. (2008). *PREMIS Data Dictionary for Preservation Metadata version 2.0*. Washington DC, USA: Library of Congress.

Hadjieleftheriou, M., Chandel, A., Koudas, N., & Srivastava, D. (2008). *Fast Indexes and Algorithms for Set Similarity Selection Queries*. Paper presented at the International Conference on Data Engineering, Cancun, Mexico.

Halem, M., F., S., Palm, N., Salmon, E., Raghavan, S., & Kempster, L. (1999). *Technology Assessment of High Capacity Data Storage Systems: Can We Avoid A Data Survivability Crisis?* Greenbelt, MD: Earth and Space Data Computing Division, NASA Goddard Space Flight Center.

Harvey, P. (2003). ExifTool by Phil Harvey. Retrieved 2008-01-25, from http://www.sno.phy.queensu.ca/~phil/exiftool/

Hedstrom, M. (1998). Digital Preservation: A time bomb for digital libraries. *Computers and the Humanities, 31*, 189-202.

Hedstrom, M. (2001). Digital Preservation: Problems and Prospects. *Digital Library Network (DLnet)*(20).

Heminger, A. R., & Robertson, S. B. (2004). *A Delphi Assessment of the Digital Rosetta Stone Model.* Paper presented at the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), Big Island, Hawaii.

Hendley, T. (1998). *Comparison of Methods & Costs of Digital Preservation* (No. 106). West Yorkshire: British Library Research and Innovation Center.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems, 22*(1), 5--53.

Heslop, H., Davis, S., & Wilson, A. (2002). An Approach to the Preservation of Digital Records. Camberra, Australia: National Archives of Australia.

Hirtle, P. B. (2000). Archival Authenticity in a Digital Age. In *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources.

Hitchcock, S., Brody, T., Hey, J. M. N., & Carr, L. (2007). Digital Preservation Service Provider Models for Institutional Repositories - Towards Distributed Services. *D-Lib Magazine, 13*(5/6).

Hodge, G., & Frangakis, E. (2004). *Digital Preservation and Permanent Access to Scientific Information: The State of the Practice* (Report No. 2004-3: Rev. 05/04): International Council for Scientific and Technical Information & CENDI.

Hofman, H. (2001). *How to keep digital records understandable and usable through time?* Paper presented at the Long-Term Preservation of Electronic Records, Paris, France.

Hofman, H. (2002a). *A global issue: preservation of digital objects.* Paper presented at the Korean Association of Archives Management, Seoul, Korea.

Hofman, H. (2002b). *Can Bits and Bytes be Authentic? Preserving the Authenticity of Digital Objects.* Paper presented at the International Federation of Library Associations Conference, Glasgow.

Holdsworth, D., & Wheatley, P. (2001). Emulation, Preservation and Abstraction. *DigiNews, Research Library Group, 5*(4).

Howel, A. G. (2004). *Preserving Digital Information: Challenges and Solutions*: Victorian Academic Libraries, Victorian university libraries and State Library of Victoria.

Hunter, J., & Choudhury, S. (2003). *Implementing Preservation Strategies for Complex Multimedia Objects*. Paper presented at the Seventh European Conference on Research and Advanced Technology for Digital Libraries (ECDL'03), Trondheim, Sør-Trøndelag, Norway.

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).

Hunter, J., & Choudhury, S. (2005). Preservation webservices Architecture for Newmedia and Interactive Collections (PANIC). Retrieved 2005-12-12, from http://metadata.net/newmedia/

Hunter, J., & Choudhury, S. (2006). PANIC: an integrated approach to the preservation of composite digital objects using Semantic Web services. *International Journal on Digital Libraries, 6*(2), 174-183.

IEEE History Center. Development of VHS, a World Standard for Home Video Recording, 1976. Retrieved 2008-05-25, from http://www.ieee.org/web/aboutus/history_center/vhs.html

International Press Telecommunications Council. (2004). IPTC Metadata for XMP. Retrieved 2008-01-24, from http://www.iptc.org/IPTC4XMP/

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin del la Société Vaudoise des Sciences Naturelles, 37*, 547-579.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association, 84*, 414-420.

Jaro, M. A. (1995). Probabilistic linkage of large public health data files. *Statistics in Medicine, 14*, 491-498.

Jiang, W., & Schulzrinne, H. (2003). *Assessment of VoIP service availability in the current Internet.* Paper presented at the Passive & Active Measurement Workshop, San Diego, CA.

Josefsson, S. (2006). The Base16, Base32, and Base64 Data Encodings. *RFC 4648* Retrieved 2008-08-17, from http://tools.ietf.org/html/rfc4648

Kenney, A. R., McGovern, N. Y., Entlich, R., Kehoe, W. R., & Olsen, E. (2003). Digital Preservation Management. *Implementing Short-term Strategies for Long-term Problems*, 2009-03-12, from http://www.library.cornell.edu/iris/tutorial/dpm/

Kimball, R., & Ross, M. (2002). *The data warehouse toolkit : the complete guide to dimensional modeling* (2nd ed.). New York: Wiley.

Kohavi, R. (1995). A study of Cross-Validation and Bootstrap for accuracy estimation and model selection. *International Joint Conferences on Artificial Intelligence, 2*, 1137-1145.

Krijgsman, G. (2005). Emulator Zone. Retrieved 2005-12-09, from http://www.emulator-zone.com

Lavoie, B., & Gartner, R. (2005). *Technology Watch Report - Preservation Metadata* (No. 05-01): Online Computer Library Center Inc., Oxford University Library Services and Digital Preservation Coalition.

Lavoie, B. F. (2004). *The Open Archival Information System Reference Model: Introductory Guide* (Technology Watch Report No. Watch Series Report 04-01). Dublin, USA: Digital Preservation Coalition.

Lavoie, B. F. (2008). PREMIS With a Fresh Coat of Paint - Highlights from the Revision of the PREMIS Data Dictionary for Preservation Metadata. *D-Lib Magazine, 14*(5/6).

Lavoie, B. F., & Dempsey, L. (2004). Thirteen Ways of Looking at... Digital Preservation. *D-Lib Magazine, 10*(7/8).

Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H., & Kenney, A. R. (2000). *Risk Management of Digital Information: A file format investigation*. Washington, DC: Council on Library and Information Resources.

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.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady, 10*(1966), 707-710.

Li, J., Chen, G., Chi, Z., & Lu, C. (2004). Image coding quality assessment using fuzzy integrals with a three-component image model. *IEEE Transactions on Fuzzy Systems, 1*(12), 99-106.

Library of Congress. (2004a). Sustainability of Digital Formats - Planning for Library of Congress Collections. Retrieved 2008/06/18, from http://www.digitalpreservation.gov/formats

Library of Congress. (2004b). Sustainability of Digital Formats - Planning for Library of Congress Collections. Retrieved 2008-06-18, from http://www.digitalpreservation.gov/formats

Lorie, R. A. (2001). *Long Term Preservation of Digital Information.* Paper presented at the First ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL'01), Roanoke, Virginia, USA.

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.

Lupovici, C., & Masanès, J. (2000). *Metadata for the Long Term Preservation of Electronic Publications* (No. 2). The Hague, The Netherlands: NEDLIB Consortium.

Lynch, C. (1999). Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. *D-Lib Magazine, 5*(9).

Lynch, C. (2000). Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust. In *Authenticity in a Digital Environment.* Washington, DC: Council on Library and Information Resources.

Lynch, C. A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in Digital Age. *ARL Bimonthly Report*(226).

MacNeil, H., Wei, C., Duranti, L., Gilliland-Swetland, A., Guercio, M., Hackett, Y., et al. (2001). *Authenticity Task Force Report.* Vancouver, Canada: InterPARES Project.

Mellor, P., Wheatley, P., & Sergeant, D. M. (2002). *Migration on Request, a Practical Technique for Preservation.* Paper presented at the ECDL '02: 6th European Conference on Research and Advanced Technology for Digital Libraries, London, UK.

Menascé, D. A. (2002). QoS Issues in Web Services. *IEEE Internet Computing, 6*(6), 72-75.

Microsoft Corporation. (1981). MS-DOS (Version 1.0) [Operating System].

Millar, L. (2004). *Authenticity of electronic records: a report prepared for UNESCO and the International Council on Archives.* London, UK: International Council on Archives.

Musgrove, M. (2006, January 12). Nikon Says It's Leaving Film-Camera Business. *The Washington Post,* p. D01. Retrieved 2007-12-12, from http://www.washingtonpost.com/wp-dyn/content/article/2006/01/11/AR2006011102323.html

National Library of Australia. (1999). Preservation Metadata for Digital Collections. Retrieved 2005-12-12, from http://www.nla.gov.au/preserve/pmeta.html

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*(1), 31-88.

Nayak, P. R., & Ketteringham, J. M. (1994). *The VCR: A Miracle at JVC "Be Very Polite and Gentle", Breakthroughs!* : Pfeiffer & Company.

Newcomer, E., & Lomow, G. (2005). *Understanding SOA with Web Services*: Addison Wesley.

Newspaper Association of America, & International Press Telecommunications Council. (1999). *Information Interchange Model Version No. 4.1*. Windsor, UK.

OASIS. (2005). Universal Description, Discovery and Integration (UDDI). Retrieved 2008-04-21, from http://www.uddi.org/

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

Ockerbloom, J. M. (2003). TOM Conversion Service. Retrieved 2006-12-10, from http://tom.library.upenn.edu/convert/

OCLC/RLG Preservation Metadata Working Group. (2002). *A Metadata Framework to Support the Preservation of Digital Objects*. Dublin, USA: OCLC Online Computer Library Center, Inc.

Oliveira, C. (2001). *Classificação de imagens colectadas na web*. Universidade Federal de Minas Gerais, Belo Horizonte.

Oltmans, E., Diessen, R. J. v., & Wijngaarden, H. v. (2004). *Preservation Functionality in a Digital Archive*. Paper presented at the Joint ACM/IEEE Conference on Digital Libraries (JCDL'04).

Parallels. (1995). Parallels Desktop Web site. Retrieved 2006-10-12, from http://www.parallels.com

Pearson, D. (2008). AONS II: continuing the trend towards preservation software 'Nirvana'. *New Technology of Library and Information Service*(1), 42-49.

Petrov, O., Vatolin, D., Parshin, A., & Titarenko, A. (2006). *MSU Subjective Comparison of Modern Video Codecs*. Moscow, Russia: CS MSU GRAPHICS & MEDIA LAB VIDEO GROUP.

Portuguese National Archives, & University of Minho. (2006). RODA Web site. Retrieved 2006-04-21, from http://portal.roda.dgarq.gov.pt

PREMIS Working Group. (2005). *Data dictionary for preservation metadata: final report of the PREMIS Working Group* (Final report). Dublin, Ohio, USA: OCLC Online Computer Library Center & Research Libraries Group.

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

Ramalho, J. C., Ferreira, M., Castro, R., Faria, L., Barbedo, F., & Corujo, L. (2007). *XML e Preservação Digital*. Paper presented at the XATA - XML: Aplicações e Tecnologias Associadas, FCUL, Lisboa, Portugal.

Ramalho, J. C., Ferreira, M., Faria, L., & Castro, R. (2007). *Relational Database Preservation through XML modelling*. Paper presented at the Extreme Markup Languages, Montréal, Québec, Canada.

Ramalho, J. C., Ferreira, M., Faria, L., Castro, R., Barbedo, F., & Corujo, L. (2008). *RODA and CRiB - A Service-Oriented Digital Repository*. Paper presented at the International Conference on Preservation of Digital Objects (iPRES), London, UK.

Ramalho, J. C., Ferreira, M., Ferros, L., Lima, M. J. P., & Sousa, A. (2006). *Digitarq 2 - Nova arquitectura aplicacional para gestão de Arquivos Definitivos*. Paper presented at the 2nd International Conference on Enterprise Archives, Seixal, Portugal.

Rauber, A., & Aschenbrenner, A. (2001). Part of Our Culture is Born Digital - On Efforts to Preserve it for Future Generations. *TRANS - On-line Journal for Cultural Studies., 10*.

Rauch, C. (2004). *Preserving Digital Entities - A Framework for Choosing and Testing Preservation Strategies*. Unpublished Master Thesis, Vienna University of Technology, Vienna.

Rauch, C., Krottmaier, H., & Tochtermann, K. (2007). *File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats*. Paper presented at the International Conference on Electronic Publishing, Vienna, Austria.

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.

Rauch, C., Rauber, A., Hofman, H., Bogaarts, J., Vedegem, R., Pavuza, F., et al. (2005). *A Framework for Documenting the Behaviour and Funcionality of Digital Objects and Preservation Strategies*. Glasgow: DELOS Network of Excellence.

RDF Core Working Group. (2004). *Resource Description Framework (RDF)*: W3C.

Ross, S., & Hedstrom, M. (2005). Preservation research and sustainable digital libraries. *Internation Journal on digital Libraries, 5*(4), 317-324.

Rothenberg, J. (2000). Preserving Authentic Digital Information. In *Authenticity in a Digital Environment*. Washington, DC: Council on Library and Information Resources.

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.

Rusbridge, A. (2003). *Migration on Request* (4th Year Project Report): University of Edinburgh - Division of Informatics.

Russell, K. (2000). *Digital Preservation and the CEDARS Project Experience*. Paper presented at the International Conference on Preservation and Long Term Accessibility of Digital Materials, York, England.

Saltelli, A. (2004). *Sensitivity analysis in practice : a guide to assessing scientific models*. Hoboken, NJ: Wiley.

Saramago, M. d. L. (2004). *Metadados para preservação digital e aplicação do modelo OAIS*. Paper presented at the VIII Congresso da BAD, Estoril, Portugal.

Sarmento, F., Baptista, A. A., & Ramos, I. (2005). *Estudo de comportamento de investigadores face à utilização de um Repositório Institucional*. Paper presented at the Conferência da Associação Portuguesa de Sistemas de Informação (CAPSI), Bragança, Portugal.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. *Psychometrika, 27*(2), 125-140.

Shepard, T., & MacCarn, D. (1998). *The Universal Preservation Format: Background and Fundamentals*. Paper presented at the Sixth DELOS Workshop, Tomar, Portugal.

Shepard, T., & MacCarn, D. (1999). *The Universal Preservation Format: A Recommended Practice for Archiving Media and Electronic Records*. Boston.

Shiraishi, Y. (1985). History of Home Videotape Recorder Development. *SMPTE Journal, 94*(12), 1257-1263.

Shrestha, B., O'Hara, C. G., & Younan, N. H. (2005). *JPEG2000: Image Quality Metrics*. Paper presented at the American Society for Photogrammetry and Remote Sensing Baltimore, USA.

Silva, F. R. (2004). *Uma abordagem para detecção de outliers em dados categóricos*. Universidade Estadual de Campinas Campinas, Brasil.

SOA Reference Model TC. (2008). *Reference Architecture for Service Oriented Architecture Version 1.0*: OASIS.

Sobel, I., & Feldman, G. (1968). A 3x3 Isotropic Gradient Operator for Image Processing. In Stanford Artificial Project (Ed.). Stanford.

Soukoreff, R. W., & MacKenzie, I. S. (2001). *Measuring errors in text entry tasks: an application of the Levenshtein String Distance Statistic*. Paper presented at the ACM Conference on Human Factors in Computing Systems, New York.

Sousa, A. n., Ferros, L. M., Ramalho, J. C., & Lima, M. J. P. d. (2007). *Consulta Real em Ambiente Virtual: implementação de uma sala de referência e leitura virtual num arquivo*. Paper presented at the Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas, Açores, Portugal.

Stanescu, A. (2004). Assessing the Durability of Formats in a Digital Preservation Environment. *D-Lib Magazine, 10*(11).

Stanley, L. G. D., & Stewart, D. L. (2002). *Design sensitivity analysis : computational issues of sensitivity equation methods*. Philadelphia: Society for Industrial and Applied Mathematics.

Swade, D. (1998). Preserving Software in an Object-Centred Culture. In E. Higgs (Ed.), *In History and Electronic Artefacts* (pp. 195-206). Oxford: Clarendon Press.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley.

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.

Technical Standardization Committee on AV, & IT Storage Systems and Equipment. (2002). *Exchangeable image file format for digital still cameras: Exif Version 2.2* (No. JEITA CP-3451): Japan Electronics and Information Technology Industries Association.

Teixeira, D., Ferreira, M., & Verhaegh, V. (2003). *An Integrated Framework for Supporting Photo Viewing Activities in Home Environments*. Paper presented at the European Symposium on Ambient Intelligence, Eindhoven, The Netherlands.

Tekli, J., Chbeir, R., & Yetongnon, K. (2006). *Semantic and Structure Based XML Similarity: The XS3 Prototype*. Paper presented at the International Conference on Management of Data, Delhi, India.

Telecommunication Standardization Sector of ITU. (2004). *Objective perceptual assessment of video quality: Full reference television*. Geneva, Switzerland: International Telecommunication Union (ITU).

Tetko, I. V., Livingstone, D. J., & Luik, A. I. (1995). Neural network studies, 1. Comparison of overfitting and overtraining. *Journal of Chemical Information and Computer Sciences, 35*(5), 826-833.

The Cedars Project Team. (2001). *The Cedars Project Report*. UK: Consortium of University Research Libraries.

The Cedars Project Team. (2002). *Cedars Guide to Preservation Metadata*: The Cedars Project.

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.

UK National Archives. (2002). PRONOM - The file format registry. Retrieved 2008-04-21, 2008, from http://www.nationalarchives.gov.uk/pronom/

UK National Archives. (2005). Droid: Digital Record Object Identification (Version 1.0) [Format detector]. Surrey: UK National Archives.

University of Southampton. (2007). Registry of Open Access Repositories (ROAR). Retrieved 2007-11-22, from http://roar.eprints.org/

VMWare. (1998). VMWare Workstation Web site. Retrieved 2006-10-11, from http://www.vmware.com/

W3C. (2002). Web Services Activity. Retrieved 2008-06-21, from http://www.w3.org/2002/ws/

Walker, F. L., & Thoma, G. R. (2003). *A SOAP-Based Tool for User Feedback and Analysis*. Paper presented at the InfoToday, Medford N.J., USA.

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.

Walker, F. L., & Thoma, G. R. (2005). *Image Preservation Through PDF/A*. Paper presented at the IS&T's 2005 Archiving Conference, Washington, D.C., USA.

Wang, L. W., Zhang, Y., & Feng, J. F. (2005). On the Euclidean distance of images. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1334-1339.

Wang, Y. (2006). *Survey of Objective Video Quality Measurements* (No. WPI-CS-TR-06-02). Massachusetts, USA: EMC Corporation Hopkinton.

Wang, Z., & Bovik, A. C. (2002). A universal image quality index. *Ieee Signal Processing Letters, 9*(3), 81-84.

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING, 13*(4), 600-612.

Waters, D. (2002). *Good Archives Make Good Scholars: Reflections on Recent Steps Toward the Archiving of Digital Information*. Paper presented at the The State of Digital Preservation: An International Perspective, Washington D.C.

Waugh, A., Wilkinson, R., Hills, B., & Dell'oro, J. (2000). *Preserving Digital Information Forever*. Paper presented at the Fifth ACM Conference on Digital Libraries, San Antonio, Texas.

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.

Werf, T. v. d. (2002). *Our digital heritage: how authentic should it be?* Paper presented at the Victorian Association for Library Automation Inc., Melbourne.

Wikipedia contributors. Jean-François Champollion. Retrieved 2005-01-23, from http://en.wikipedia.org/wiki/Jean-Fran%E7ois_Champollion

Wikipedia contributors. (2005). Rosetta Stone. 2005, from http://en.wikipedia.org/wiki/Rosetta_stone

Wikipedia contributors. (2006a). Color depth. Retrieved 2008-04-21, from http://en.wikipedia.org/w/index.php?title=Color_depth&oldid=86738648

Wikipedia contributors. (2006b). Image compression. Retrieved 2008-04-21, from http://en.wikipedia.org/w/index.php?title=Image_compression&oldid=83896661

Wikipedia contributors. (2007). Digital camera. Retrieved 13 December 2007 12:24 UTC, from http://en.wikipedia.org/w/index.php?title=Digital_camera&oldid=177619169

Winkler, W. E. (1999). *The state of record linkage and current research problems*. Wachington, DC, USA: U.S. Bureau of the Census.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (Second ed.): Morgan Kaufmann.

Woodyard, D. (1998). Farewell my Floppy: a strategy for migration of digital information. *Electronic Preservation* Retrieved 2009-01-05, from http://www.nla.gov.au/nla/staffpaper/valadw.html

Woodyard, D. (2000). *Digital Preservation: The Australian Experience*. Paper presented at the Third Conference Digital Library: Positioning the Fountain of Kowledge, Malaysia.

Xiao, C., Wang, W., Lin, X., & Yu, J. X. (2008). *Efficient Similarity Joins for Near Duplicate Detection*. Paper presented at the WWW 2008, Beijing, China.

Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., & Sheng, Q. Z. (2003). *Quality Driven Web Services Composition*. Paper presented at the 12th International Conference on the World Wide Web (WWW), Budapest, Hungary.

ÍNDICE REMISSIVO

A

Acesso, 21
actualização de versões, 27
Agente, 42
agentes, 30, 31, 36, 41, 83
Análise de Utilidade, 63
aplicações, xxvii, 13, 14, 24, 25, 27, 34
arqueologia digital, 34
árvore-objectivo, 63, 64, 65
ASCII, 39
áudio, xxvi, 14, 100
autenticidade, 37, 38, 40, 43, 44

B

bases de dados, xxvi, 14
Biblioteca do Congresso, 36, 62, 105

C

canonização, 39
características essenciais, 39
CCSDS, 18
CD, xxv, xxix, 15, 23
Coeficiente de Similaridade de Jaccard, 157, 158, 210, 211
comunidade de interesse, 21
controlo de qualidade, viii, 5, 6, 8, 50, 51, 52, 67, 68, 71, 74, 75, 80, 81, 97, 99, 100, 127, 129, 132, 133, 134, 168, 183, 184, 185, 189, 194, 213
conversores, viii, xxvi, 3, 6, 27, 29, 30, 31, 32, 36, 51, 52, 59, 60, 63, 71, 73, 80, 81, 82, 87, 91, 94, 95, 124, 127, 155, 164, 167, 168, 184, 186, 193
correlação de Pearson, 151, 159, 169, 170, 187
correlação de Spearman, 151, 159, 169
custo, 23, 46, 51, 65, 67, 79, 83, 85, 87, 92, 94, 95, 96
custo de utilização, 83, 85, 92, 94

D

Data warehousing, 131
diagramas vectoriais, xxvi, 14
Digital Curation Centre, 36
direitos, 41, 43
disco rígido, 15, 17, 23
disponibilidade, 65, 79, 92, 143
disquete, 15, 23
Documentos de texto, xxvi, 14
DVD, xxv, xxix, 13, 15, 17

E

emulação, 23
emulador, 23, 24, 25
encapsulamento, 21, 32
Entidade Intelectual, 41
entidades intelectuais, 41
estabilidade, 65, 92, 93
Evento, 42
eventos, 41
exactidão, 169
extractor de propriedades, 97

F

Ficheiro, 43
formato, 15
formato canónico, 39
formato de preservação, 29
Formato Universal de Preservação, 32
fotografias digitais, xxvi, 14
funções de similaridade, 7, 99, 128, 172, 186, 193, 197

G

Global Digital Format Registry, 35

H

hardware, xxv, xxvi, 15, 22, 23, 24, 25, 26, 29, 33, 38, 40
HTTPS, xxix, 130
Hypertext Transfer Protocol sobre Secure Socket Layer. *See* HTTPS

I

incorporação, 19, 40
informação, xxv, xxvi, xxvii, 13, 14, 18, 19, 20, 21, 23, 24, 26, 29, 32, 33, 34, 35, 38, 40, 41, 42, 43, 44, 45, 46
Ingestão, xxvi, 19
Internet, xxvi, 30, 32, 35, 66, 91, 130, 165, 179, 189, 192
ISO, 18

J

Java Virtual Machine, 32
JPEG, xxvi, xxix, 17, 28

L

LDAP, 36

M

máquina virtual universal, 32
Média do Quadrado do Erro, 159, 162
metainformação de preservação, 40
migração, 21, 26, 27, 28, 29, 30, 31, 32, 39, 44, 45, 46
migração a-pedido, 27, 29, 31, 46
migração para suportes analógicos, 27
Migration Advisor, 72, 74, 97, 110, 118, 119, 120, 121, 122, 124, 126, 128, 131, 132, 133, 134, 162, 163, 164, 165, 166, 167, 168, 169, 170, 171, 172, 173, 184, 185, 186, 187, 194, 195
Migration Broker, 89
Migration Knowledge Base, 132, 194
MIME, 35
Mime Media Types, 34, 35

N

normalização, 27, 28, 29, 31, 46

O

OAIS, 18, 19, 20, 40, 44
Object Evaluator, 96, 134
Objecto, 43
objecto conceptual, 15, 17, 21, 22, 26
objecto digital, xxvi, xxvii, 2, 3, 8, 14, 15, 17, 18, 23, 24, 26, 27, 32, 33, 37, 38, 39, 43, 44, 79, 93, 94, 182
objecto experimentado, 16
objecto físico, 15, 21, 38, 45
objecto lógico, 15
objecto semântico, 15
objectos, 41
objectos conceptuais, 99
objectos digitais, xxvi, 15, 21, 22, 24, 26, 27, 28, 29, 32, 33, 34, 39, 44, 46, 64
OCLC/RLG, 41

P

Pacotes de Informação de Disseminação, 21
PDF, xxvii, xxx, 31, 35, 39, 42, 58, 67, 155, 200, 208, *See* Portable Document Format
Pedra de Rosetta, 33, 34
Planeamento de Preservação, 20
PNG, xxvii, xxx, 28
políticas, 20, 29, 39, 44, 46
população potencialmente utilizadora, 20
Portable Document Format, xxvii, xxx, 102, 131, 201
precisão, 169
PREMIS, xxx, 14, 40, 41, 43, 44, 77, 97, 179
preservação digital, vii, xxvi, 2, 5, 8, 13, 14, 17, 18, 21, 23, 26, 34, 40, 41, 49, 182
propriedades significativas, 38, 39, 65, 97, 99, 100, 134, 135, 136, 163
proveniência, 40

R

realidade virtual, xxvi, 14
refrescamento, 23
repositório, 19, 21, 28, 32, 39, 40
Repositório de Dados, 20
Representação, 43
Representation Information Registry Repository, 36
royalties, 29, 110, 113

S

Sequência de bits, 43
Service Registry, 83, 84
Serviços, 31
similaridade, 99, 101, 107, 108, 109, 110, 134, 135, 136, 139, 145, 147, 149, 150, 151, 152, 157, 158, 159, 160, 162, 203, 204, 205, 207, 208, 211, 218
software, xxv, xxvi, xxvii, 13, 15, 22, 23, 24, 26, 27, 28, 29, 31, 32, 35, 38, 40, 42, 46
submissão, 19
suporte físico, xxvii, 15, 17, 23, 32

T

taxionomia de avaliação, 100, 119
Thibodeau, 21, 45
TIFF, xxvii, xxx, 17, 28
TOM, 30, 35, 91, *See* Types Object Model
Typed Objects Model, 30

V

vídeo, xxvi, 12, 45

W

Web, xxv, xxvii, 14
Web service, 31
Web services, xxx, 36, 52, 53, 75, 129, 188, 189, 190, 195
Word Object Model, 199
WS-BPEL, xxx, 129, 195