Título da comunicação: A importância da Informação de Representação na preservação de objetos digitais

Resumo:

Genericamente, as estratégias de preservação digital encontradas na literatura podem ser classificadas segundo três categorias: 1) **preservação de tecnologia**, consistindo na conservação da pilha tecnológica originalmente utilizada na criação ou apresentação do objeto digital, materializando-se essencialmente no recurso a técnicas de emulação que permitem executar software obsoleto em hardware virtualizado; 2) **migração de formatos**, muito provavelmente a estratégia de preservação mais utilizada nos vários domínios de aplicação, consistindo na reorganização dos dados segundo uma nova estrutura, ou formato, que ofereça menor risco de obsolescência ou que assegure um nível de interoperabilidade mais elevado junto da sua comunidade de interesse; e 3) **encapsulamento**, consistindo na conservação de informação apensa ao objeto digital que permita, num futuro longínquo, compreender o objeto e a forma como este deve ser interpretado.

A última destas estratégias remete-nos para a definição de Informação de Representação conforme apresentada no modelo de referência OAIS. De acordo com este referencial normativo, Informação de Representação trata-se de informação que tem como objetivo traduzir um objeto digital em algo inteligível, ou seja, transformar os zeros e uns que o constituem em conceitos que possam ser compreendidos pela sua comunidade de interesse. De grosso modo a Informação de Representação pode ser subdividida em três classes distintas:

1) **Informação estrutural** – informação que descreve o formato e/ou o modelo de dados. Permite a partir de uma sequência de bits, reinterpreta-la sob a forma de conceitos mais próximos dos seres humanos ou das máquinas que têm de os processar, como números, carateres ou pixéis.
2) **Informação semântica** – informação que permite compreender os conceitos que resultam da interpretação dos dados realizada através de Informação estrutural. Se a informação estrutural nos diz que uma dada sequência de bits

se trata de um texto constituído por carateres que formam palavras, a Informação semântica permite-nos compreender o significado dessas palavras, identificando, por exemplo, o idioma em que essas palavras se encontram, e fornecendo dicionários e gramáticas que permitam no futuro compreender esse idioma.

3) **Outra Informação de Representação** – informação sobre software, hardware, algoritmos, documentação, etc., que possa facilitar a interpretação do objeto digital e que não possa ser adequadamente classificada como informação estrutural ou semântica.

Vejamos um exemplo. Imagine-se um ficheiro com o nome "F23AA.TXT". O nome do ficheiro suscita que se trate de um ficheiro de texto. No entanto, ao abrir o ficheiro com um qualquer editor de texto somos confrontados com uma sequência de palavras e números com as seguintes características:

Amares	M	1 473	1 205	5 255	1 806
Barcelos	M	9 393	7 747	35 031	9 921
Braga	M	14 321	10 884	55 343	14 357
Esposende	M	2 768	2 133	10 070	3 022
Terras de Bouro	M	441	418	1 887	1 004
Vila Verde	M	3 821	2 965	13 137	4 956
Ave	M	37 876	30 709	151 246	44 676
Fafe	M	3 691	3 115	14 658	5 151
Guimarães	M	11 736	9 688	47 210	12 754
Póvoa de Lanhoso	M	1 685	1 336	6 209	2 229
Vila Nova de Famalicão	M	10 270	7 856	39 913	10 882
Vizela	M	1 896	1 575	7 019	1 618

À partida, sem mais informação, torna-se difícil saber o que esta informação representa. No entanto, se fornecermos alguma informação adicional, estes dados tornam-se claros. Se informarmos que os dados apresentados representam o número de pessoas que residiam em 2011 nos municípios que figuram na primeira coluna, que a segunda

coluna informa qual o género da população (M para mulher e H para homem), e que as restantes colunas representam grupos etários dos 0 aos 14 anos, dos 15 aos 24, dos 25 aos 64 e mais de 65 anos respetivamente, então a informação previamente apresentada torna-se clara e útil.

A Informação de Representação é particularmente relevante no contexto da curadoria de dados científicos em que a dificuldade não está tanto na interpretação do formato (tipicamente são formatos de texto), mas sim na compreensão semântica dos dados que se pretende consumir.

Nesta comunicação será reforçada a ideia de que o recurso a Informação de Representação se trata de uma estratégia de preservação que, quando utilizada eficientemente se torna, não só viável, como preferencial para certos tipos de objetos digitais. Serão apresentados exemplos e fornecidas estratégias de como aceder, descrever e armazenar este tipo de informação junto dos objetos digitais aos quais dizem respeito.

Nota biográfica:

Miguel Ferreira
Diretor Executivo da KEEP SOLUTIONS

Doutorado pela Universidade do Minho em Tecnologias e Sistemas de Informação, com especialização em Preservação Digital. Trabalhou como investigador na Philips Research em Eindhoven. Desempenhou cargos de consultor e investigador no Arquivo Distrital do Porto e na Universidade do Minho. Em 2007 venceu o Digital Preservation Challenge promovido pelo Digital Preservation Europe (DPE). Em 2008 fundou a empresa KEEP SOLUTIONS onde atualmente desempenha o cargo de Diretor Executivo. Coordenou vários projetos de I&D de âmbito nacional e internacional no âmbito dos quais publicou dezenas de artigos em conferências e revistas internacionais. Mantém uma filiação à Universidade do Minho como investigador colaborador do Centro Algoritmi e acumula o cargo de professor Assistente Convidado no Instituto Politécnico do Cávado e do Ave.