Título da comunicação: Curadoria digital na preservação da Web.

Resumo:

A Web é atualmente um dos principais meios de comunicação nas sociedades desenvolvidas e tem crescido exponencialmente desde a sua criação, tanto no número de utilizadores como no volume de conteúdos disponíveis. No entanto, a informação publicada na Web é efémera, sendo que, passado 1 ano, 80% das páginas são modificadas, ou desaparecerem. Nas últimas duas décadas, surgiram várias iniciativas de preservação de conteúdos publicados na Web, a maior parte com foco regional ou nacional. O Arquivo.pt é uma infraestrutura de investigação pública, gratuita e disponível *online*, que preserva páginas Web de interesse nacional.

O objetivo desta apresentação será descrever o processo de curadoria digital aplicado no Arquivo.pt, nomeadamente ao nível da seleção, recolha, preservação, coleção e arquivo das páginas Web.

O Arquivo.pt recolhe as páginas do domínio.PT, como também outras páginas consideradas relevantes para a comunidade portuguesa. Qualquer pessoa pode sugerir um endereço para ser preservado.

São efetuados 5 tipos de aquisição de conteúdos para preservação, nomeadamente: (i) recolhas diárias de páginas que são alteradas com muita frequência tais como as publicadas por *sites* noticiosos; (ii) recolhas trimestrais de páginas alojadas sob o domínio .PT e outras páginas relevantes sugeridas pelos utilizadores; (iii) recolhas colaborativas de páginas Web realizadas com o apoio da comunidade para a seleção de conteúdos a preservar, como foram os casos das recolhas das eleições legislativas de 2015 e das eleições presidenciais de 2016; (iv) recolhas especiais relacionadas com eventos esporádicos, como por exemplo o projeto de preservação de sites de projetos de Investigação e Desenvolvimento e (v) integração de conteúdos doados por entidades externas, como foi o caso de uma coleção de conteúdos da autoria de um fotojornalista.

A informação preservada é guardada em discos rígidos, de forma redundante de modo a que não se perca informação em caso de avaria. A informação é também replicada para o *Internet Archive*, na Califórnia, estando segura mesmo no eventual caso de perda total da informação do *Data Center* do Arquivo.pt. Os dados são também armazenados em

cassetes, uma vez que as mesmas possuem uma maior durabilidade em comparação com os discos rígidos. As cassetes são armazenadas em diferentes localizações geográficas.

Não existe preservação digital sem a garantia de acessibilidade à informação armazenada. O Arquivo.pt armazena informação publicada na Web para acesso futuro e disponibiliza um serviço para a sua pesquisa e acesso através da Internet. No Arquivo.pt existem 3 métodos de pesquisa de dados históricos: pesquisa por endereço; pesquisa por texto livre e pesquisa avançada refinada através da especificação de metadados.

A apresentação descreverá de uma forma geral, como são colecionadas e arquivadas as páginas Web. Será apresentada uma descrição geral da arquitetura e funcionamento do sistema que suporta o Arquivo.pt e dos respetivos fluxos de trabalho de recolha, indexação e pesquisa. A geração de metadados relativa aos artefactos preservados é principalmente processada de forma automática devido ao enorme volume de dados envolvidos. Contudo, existem também metadados produzidos por especialistas. Ambos os processos de geração de metadados para os conteúdos web preservados serão debatidos.

Por fim, serão sugeridas possibilidades de colaboração com o Arquivo.pt para a melhoria do processo de curadoria digital, por exemplo através de um maior envolvimento da comunidade em ações colaborativas de seleção de conteúdos para arquivo.

Fernando Melo, Daniel Gomes
fernando.melo@fccn.pt, daniel.gomes@fccn.pt
FCT-FCCN: Arquivo.pt

Nota biográfica:

F*ernando Melo*
Mestre em Engenharia Informática, pelo Instituto Superior Técnico

Colaborador no Arquivo.pt. É responsável pela qualidade de reprodução das páginas arquivadas.
Os seus principais interesses são georreferenciação automática, processamento de linguagem natural, usabilidade e desenvolvimento Web.

Em particular na sua tese de mestrado propõe um novo classificador para encontrar de forma automática as coordenadas geoespaciais de documentos com base apenas no seu texto.