Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

Directrizes para selecção de sítios web relevantes para arquivo

A FCCN tem em curso o projecto de Arquivo da Web Portuguesa (AWP) e procura colaborar com entidades de Investigação e Desenvolvimento, que tenham interesse em participar na realização de actividades inovadoras. Este documento apresenta uma proposta de um projecto com a duração estimada de 1 ano, que poderia fazer parte de um trabalho de mestrado ou de iniciação à investigação.

O AWP tem como grande objectivo, periodicamente recolher e armazenar informação proveniente da web relevante para a comunidade portuguesa. Pretende-se assim preservar a web portuguesa da forma mais exaustiva possível.

O processo de recolha é realizado por um componente denominado batedor que iterativamente recolhe, extrai e segue ligações para novos conteúdos. O batedor inicia a sua actividade a partir de um conjunto de endereços relevantes para arquivo, denominados *raízes*. A selecção de raízes é fundamental para a qualidade do serviço prestado pelo AWP.

Actualmente, o processo de selecção de raízes é feito de forma exclusivamente automática baseando-se numa lista de endereços de sítios web sob o domínio .PT. Porém, existem numerosos sítios web de manifesto interesse para a comunidade portuguesa alojados fora do domínio nacional. A selecção automática de sítios web fora do domínio .PT é complexa, onerosa e apresenta taxas de erro consideráveis.

O AWP permite que qualquer pessoa sugira um sítio web para arquivo. Contudo, é necessário validar estas sugestões antes de incluí-las na lista de raízes, segundo critérios de selecção concisos. Os critérios de selecção automáticos são determinísticos mas limitados às capacidades de interpretação das máquinas. Por outro lado, a relevância percepcionada por humanos é subjectiva.

O principal objectivo do trabalho proposto é investigar critérios de selecção de sítios web relevantes para arquivo que resultarão num conjunto de directrizes que permitirão que uma pessoa não especialista, identifique se um determinado sítio web deverá ser

arquivado. Pretende-se assim validar humanamente as sugestões de sítios web, para evitar o arquivo de conteúdos irrelevantes para a comunidade portuguesa. O principal produto deste projecto será um questionário relativamente curto, cujas respostas permitirão inferir se um sítio web deverá ser arquivado.

Bibliografia

- Julien Masanès, Web Archiving, 2006.
- Michael Day, Collecting and Preserving the World Wide Web, 2003.
- Daniel Gomes and Mário J. Silva, Characterizing a national community web, ACM Transactions on Internet Technology, 2005.
- Daniel Gomes and Sérgio Freitas and Mário J. Silva, Design and Selection Criteria for a National Web Archive, 2006.
- National Library of Australia, Online Australian Publications: Selection Guidelines for Archiving and Preservation by the National Library of Australia, 2005.