# Proposta de projecto de colaboração com o Arquivo da Web Portuguesa

## *ArqTag: Etiquetagem comunitária de conteúdos arquivados*

A FCCN tem em curso o projecto de Arquivo da Web Portuguesa e procura colaborar com entidades externas de Investigação e Desenvolvimento, que tenham interesse em participar na realização de projectos inovadores.

Periodicamente a web portuguesa é recolhida e armazenada para preservação futura. Esta grande quantidade de dados requer mecanismos que permitam aceder à informação arquivada, restringir o espaço de procura e extrair documentos relevantes para as pesquisa realizadas.

O arquivo de publicações impressas é sujeito a um valioso processo de catalogação por parte de bibliotecários especializados, o que facilita a sua posterior pesquisa. A catalogação de conteúdos da Web por parte de profissionais seria valiosa para a melhoria dos resultados de pesquisas sobre a informação arquivada, principalmente para conteúdos não textuais, como é caso das imagens, mas que são procurados através de linguagem textual pelos utilizadores.

Infelizmente, a dimensão da Web não permite que o processo de catalogação tradicional seja aplicado num arquivo da Web. No entanto, é possível distribuir a tarefa da geração de anotações por comunidades de utilizadores, permitindo que estes indexem os conteúdos a que acedem, associando-lhes etiquetas descritivas (*tags*).

Note-se que os profissionais, como arquivistas ou bibliotecários, poderiam continuar a dar um contributo acrescido através da geração de etiquetas de maior qualidade, inclusivamente através de processos de indexação formais.

A etiquetagem livre poderá parecer caótica em comparação com as normas usadas para catalogação de publicações impressas. Mas por outro lado, à medida que as normas de catalogação se tornam obsoletas perdem o seu valor acrescido em relação às etiquetas livres.

As etiquetas são usadas para melhorar os resultados dos motores de busca sobre a Web. Num arquivo da web têm um papel adicional porque permitem adaptar as suas descrições textuais à evolução da linguagem ao longo dos anos.

Por exemplo, quem pesquisasse por "Guerra no Iraque" na década de 80 (1980-1988), provavelmente gostaria de encontrar informação relativa ao conflito Irão- Iraque; em 1990 em relação à guerra causada pelo invasão do Kuwait por parte do Iraque; e em 2008 à guerra causada pela invasão do Iraque pelos EUA.

Ao longo dos anos, as notícias que foram publicadas com o título "Guerra no Iraque" eram inequívocas na data da sua publicação e os motores de busca devolveram respostas satisfatórias (assumindo que os havia em 1980).

Em 2010 quando alguém fizer a pesquisa "Guerra no Iraque", será difícil determinar que páginas estará à espera de encontrar. Mas com um refinamento da pesquisa para "Primeira guerra EUA-Iraque" referindo-se à guerra de 90, o objectivo da pesquisa torna-se mais claro.

Por outro lado, as notícias que foram publicadas na época não usavam esta terminologia. Em 1990, não se sabia que iria haver uma segunda guerra EUA-Iraque e por isso as notícias nessa altura não continham os termos "Primeira guerra EUA-Iraque".

Um sistema de etiquetagem poderá ajudar a desambiguar estas pesquisas porque as anotações das páginas são enriquecidas ao longo do tempo. Assim sendo, em 2010 se alguém etiquetasse as páginas das três notícias descritas anteriormente da seguinte forma:

- Notícia relativa a "Guerra no Iraque" (‘80). Etiquetas: Guerra Irão-Iraque, Ayatollah Khomeini, fundamentalismo islâmico.
- Notícia relativa a "Guerra no Iraque" (’90). Etiquetas: Primeira guerra do Golfo, guerra EUA-Iraque, Tempestade no deserto, Kuwait, petróleo.
- Notícia relativa a "Guerra no Iraque"(’07). Etiquetas: Segunda guerra do Golfo, invasão do Iraque, armas de destruição maciça.

Se as etiquetas fossem usadas como fonte de informação para as pesquisas sobre o arquivo, o sistema poderia agora mais facilmente

encontrar a notícia desejada para os termos "Primeira guerra EUA-Iraque".

Um conteúdo arquivado poderá também ser etiquetado com termos em várias línguas e esta informação pode ser usada para suportar pesquisas multilíngue sobre os conteúdos da Web portuguesa.

O objectivo deste projecto é desenvolver um sistema que permita aos utilizadores do Arquivo da Web Portuguesa colocar etiquetas sobre os conteúdos arquivados, à semelhança do que é feito através do del.icio.us para os conteúdos actuais da web.

Estas etiquetas deverão ser armazenadas para que possam ser acedidas rapidamente para, por exemplo, serem usadas na criação de índices. O uso da tecnologia RDF, OWL, ou outras técnicas comuns do domínio da “Web semântica” serão aqui recomendadas
O sistema deverá incluir mecanismos de defesa contra utilizadores mal intencionados, como é o caso dos *web spammers*.

O sistema poderá ser implementado de raiz de preferência em JAVA ou usando como base software de código-aberto gratuito, com codificação da informação em XML ou outras técnicas relacionadas.