Preservar a Web: um desafio ao alcance de todos

Daniel Gomes
Arquivo da Web Portuguesa
Fundação para a Computação Científica Nacional
Avenida do Brasil, 101
1700-066 Lisboa
Tel: +351 218440100
E-mail: daniel.gomes@fccn.pt

RESUMO
A Internet permitiu ligar computadores a nível mundial. A Web é um dos meios de comunicação criados sobre a Internet e é composta por páginas e outros conteúdos ligados entre si (ex. imagens, vídeos). A Web tem vindo a afirmar-se como o principal meio de publicação das sociedades desenvolvidas. Porém, a informação nela publicada é extremamente efémera, tornando-se inacessível rapidamente. Os Arquivos da Web são sistemas informáticos criados com o objectivo de preservar e manter acessível a informação publicada na Web após esta deixar de estar disponível em-linha, analogamente ao que se tem feito ao longo dos séculos com as publicações impressas que deixam de ser editadas. Contudo, o arquivo da Web levanta grandes desafios e a colaboração entre comunidades é fundamental para garantir a preservação de conhecimento para as gerações futuras. Este artigo descreve o estado actual do Arquivo da Web Portuguesa e propõe iniciativas de preservação da Web em colaboração com a comunidade.

PALAVRAS-CHAVE: Arquivo da Web Portuguesa, preservação digital, pesquisa histórica

ABSTRACT
The Internet connected computers world-wide. The Web is one of the means of communication created over the Internet and it is composed by inter-connected pages and other types of contents, such as images or videos. The Web became the main mean of communication in developed societies. However, the information published on the Web is extremely ephemeral and becomes quickly unavailable. Web archives are computer systems that preserve and maintain accessible the information published on the Web, after it becomes inaccessible online. In a nutshell, Web archives aim to preserve Web publications in the same way that it has been done with printed publications for centuries. However, Web archiving raises new challenges and the collaboration among communities is mandatory to achieve the great goal of preserving knowledge for the future. This article describes the current state of the Portuguese Web Archive and proposes cooperative initiatives for Web preservation.

KEYWORDS: Portuguese Web Archive, digital preservation, temporal search

INTRODUÇÃO

A Web possibilita que qualquer pessoa disponibilize informação acessível a todos de uma forma rápida e económica. Diariamente, são publicados milhões de conteúdos exclusivamente na Web, como por exemplo textos, fotografias ou vídeos. No entanto, passado relativamente pouco tempo, a grande maioria desta informação deixa de estar acessível e perde-se irremediavelmente.

Após 1 ano, apenas cerca de 20% de um conjunto de endereços da Web ainda apontam para um conteúdo válido (NTOULAS et al., 2004). Ou seja, após 1 ano é provável que 8 em cada 10 dos Favoritos que um utilizador da Web tem no seu navegador (*browser*) se tenham perdido.

O Internet Archive é uma organização norte-americana que recolhe e arquiva conteúdos da Web à escala mundial (KAHLE, 2002). É difícil para uma única organização fazer um arquivo exaustivo de todos os conteúdos publicados porque a Web está em permanente mutação e muita informação desaparece antes de poder arquivada. Além disso, a documentação de acontecimentos históricos de relevância nacional para Portugal não é prioritária para o Internet Archive e grande parte da informação publicada na Web portuguesa não é arquivada. Este problema é sentido igualmente por outras comunidades nacionais e pelo menos 17 países já iniciaram as suas próprias iniciativas de arquivo da Web (NATIONAL LIBRARY OF AUSTRALIA, 2006).

O Arquivo da Web Portuguesa (AWP) da Fundação para a Computação Científica Nacional é um serviço que tem como missão recolher, armazenar e preservar a informação publicada na Web portuguesa, proporcionando um arquivo mais exaustivo da informação publicada na Web de interesse para a comunidade portuguesa.

Os serviços a serem prestados pelo AWP ultrapassam o âmbito histórico-cultural da preservação de informação digital. Este projecto permitirá, por exemplo, fornecer medições da evolução da Web portuguesa, contribuir para a expansão do uso do português na Web ou disponibilizar conteúdos de interesse a diversas comunidades científicas.

O arquivo da Web portuguesa requer um esforço à escala nacional e a comunidade de bibliotecários, arquivistas e documentalistas poderá desempenhar um papel fundamental. Este artigo introduz a temática do arquivo da Web, descreve o funcionamento geral do AWP e quais os principais desafios que uma iniciativa desta natureza tem de enfrentar. Finalmente, apresenta um conjunto de iniciativas em que a comunidade pode assumir um papel fundamental no arquivo e preservação da Web portuguesa.

A WEB COMO FERRAMENTA DE PUBLICAÇÃO

A Internet é uma infra-estrutura de comunicação que liga computadores do mundo entre si. Sobre a Internet existem vários serviços, como por exemplo, o correio electrónico, as mensagens instantâneas ou serviços de voz (VOIP). A Web é mais um serviço que foi criado sobre a Internet e é composta pelas páginas e ficheiros ligados entre si através de hiperligações. Pode-se dizer que a Internet é o equivalente às estradas, e a Web, correio electrónico e outros serviços são os diferentes veículos que nela circulam.

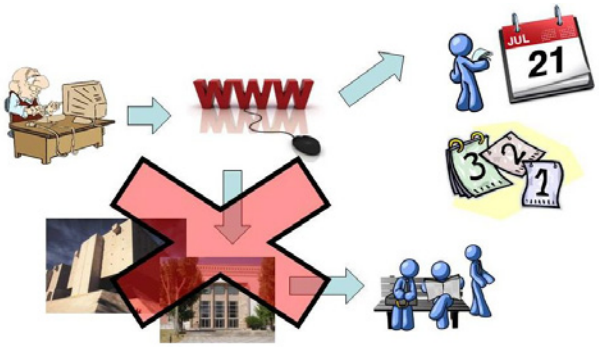

Figura 1. A Web carece de mecanismos que garantam a persistência de acesso à informação ao longo do tempo.

A Web foi inventada para transmitir rapidamente informação entre cientistas. Porém, a crescente popularidade deste meio de comunicação fez com que se tornasse numa ferramenta usada pelas massas para publicação de informação. Cada vez mais, publicações que eram publicadas em papel têm vindo a ser substituídas por versões em-linha. Outras novas publicações nascem e existem exclusivamente na Web.

Existem mecanismos para permitir o acesso à informação impressa após a sua publicação (ex. lei do depósito de legal, repositórios institucionais, bibliotecas nacionais). Porém, o modelo actual de publicação na Web não contempla estes mecanismos para preservação e persistência de acesso à informação a longo prazo (Figura 1). Surge assim o interesse na criação de arquivos da Web que garantam o acesso à informação à semelhança do que acontece para as publicações impressas.

ARQUIVO DA WEB VS. PUBLICAÇÕES IMPRESSAS

O arquivo de publicações na Web e impressas partilham o mesmo objectivo: a preservação de conhecimento para o futuro. Contudo, as características específicas de cada um destes meios de comunicação impõem que o seu arquivo seja operacionalizado de forma diferente.

O primeiro passo de um processo de arquivo é a obtenção de informação. Para cada livro publicado existe um número de cópias que são distribuídas por diversas localizações. Cada uma destas cópias terá tipicamente um tempo de vida de vários anos, durante o qual poderá vir a ser arquivada para preservação. Por sua vez, na Web o tempo de vida de uma publicação é muito curto. Um estudo revelou que 50% das páginas da Web portuguesa são alteradas ou desaparecem passados 2 dias e 50% dos sítios da Web desaparecem após cerca de 1 ano e meio (GOMES et al., 2006).

Figura 2. Passados alguns anos a informação publicada na Web torna-se inacessível. Exemplo: página oficial da Expo'98.

Por exemplo, em 1998 Portugal organizou a última exposição mundial do século XX (Expo'98). O sítio Web oficial desta exposição divulgou informações acerca do evento a nível mundial. A Figura 2 mostra que em 2010, a página da Expo'98 não se encontra disponível na Web e se não tivesse sido arquivada, provavelmente ter-se-ia perdido para sempre.

Figura 3. O arquivo de publicações na Web pode ser imposto ao autor ou ser realizado activamente pelo Arquivo.

A obtenção de publicações na Web para arquivo tem de ser muito mais célere do que a de publicações impressas. O arquivo de publicações impressas é feito com uma forte intervenção humana. Os documentos são entregues para arquivo pelos seus autores ou editoras, sendo depois verificados e catalogados por especialistas para assegurar a sua preservação e acesso futuros. O grande volume de dados publicados na Web implica que o seu arquivo seja um processo maioritariamente automatizado. A título de exemplo, em 2008 foram publicados em Portugal cerca de 14 000 títulos impressos. Por sua vez, uma única recolha da Web portuguesa inclui cerca 361 mil sítios da Web compostos por cerca de 28,1 milhões de páginas.

A Figura 3 descreve duas abordagens possíveis para a obtenção de informação proveniente da Web para arquivo. O *depósito imposto ao autor* é uma analogia com o modelo de depósito legal existente para as publicações impressas, em que o autor é legalmente obrigado a entregar cópias das suas publicações para arquivo. Porém, esta abordagem é difícil de operacionalizar porque: impõe custos adicionais para os autores; as imposições legais têm fronteiras que não existem na Internet; não existem normas e ferramentas que permitam aos autores depositar as suas publicações de uma forma eficaz e eficiente.

Por sua vez, na abordagem de *recolha activa pelo Arquivo*, a obtenção de informação é transparente para o autor, existindo um sistema que faz automaticamente a recolha da informação publicada na Web. Consequentemente, a criação de meta-dados associados aos conteúdos arquivados para possibilitar a sua preservação e acesso para o futuro deve também ser automatizada. A recolha activa feita pelos arquivos é idêntica à que é feita por outros sistemas que se alimentam de informação proveniente da Web, como é o caso dos motores de busca (ex. Google), o que torna possível reutilizar tecnologia e normas já existentes. A recolha activa de publicações na Web para arquivo tem um custo mínimo para o autor, que apenas tem de suportar a visita do Arquivo ao seu sítio Web, tal como se tratasse de qualquer outro visitante.

FUNCIONAMENTO DE UM ARQUIVO DA WEB

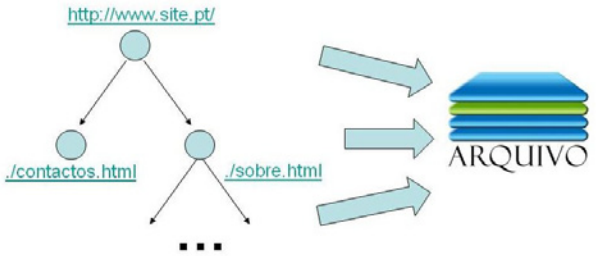

Figura 4. O Arquivo recolhe ciclicamente ficheiros da Web, seguindo as ligações em cada página.

O processo de arquivo da Web divide-se em 3 etapas principais: recolha de informação proveniente da Web, indexação e disponibilização de serviços de pesquisa e acesso. A Figura 4 ilustra o processo de recolha automática de informação. A partir de um conjunto inicial de endereços de sítios da Web, inicia-se um processo automático que consiste em ciclicamente:

1. Recolher um ficheiro da Web e armazená-lo em disco;
2. Extrair os endereços para outros ficheiros a partir das ligações contidas nas páginas;
3. Inserir os novos endereços descobertos para recolha.

Terminada a recolha, a informação obtida é processada para construir os índices que permitirão realizar pesquisas rápidas. Dado o grande volume de dados envolvido, este processo implica a utilização de sistemas de processamento sofisticados, compostos por vários computadores que cooperam entre si.

	Web actual	**Web arquivada**
Informação pesquisada	Termos	Termos, intervalo de tempo
Dimensões temporais	1	Várias
Apresentação de resultados	Em-linha	Reproduzidos
Preservação	Não	Sim

Tabela 1. Comparação entre pesquisa sobre a informação da Web actual e arquivada.

Após criados os índices, são disponibilizados serviços de pesquisa e acesso. Aparentemente, um sistema de pesquisa sobre informação arquivada é semelhante a um sistema de pesquisa sobre a Web actual (BRIN et al., 1998). Porém, a Tabela 1 destaca as principais diferenças entre estes dois tipos de sistema. Num motor de busca sobre a Web actual, os seus utilizadores inserem termos a pesquisar para encontrarem páginas disponíveis em-linha que os contenham (ex. Guerra do Iraque). Num arquivo da Web é necessário que os utilizadores identifiquem o intervalo de tempo dentro do qual pretendem encontrar páginas arquivadas (ex. Guerra do Iraque, 1990-1992).

Quando os utilizadores recebem uma lista de resultados proveniente de um motor de busca sobre a Web actual, ao clicarem num deles, são redireccionados para uma página da Web actual disponível em-linha, abandonando o motor de busca. Por sua vez, um arquivo da Web tem de reproduzir a página arquivada apontada na lista de resultados para que os utilizadores lhe possam aceder. Esta reprodução deverá ser feita de forma a proporcionar uma experiência de utilização o mais fiel possível à original. Contudo, com a permanente evolução da tecnologia e dos formatos digitais usados para publicação, esta reprodução pode tornar-se impossível a longo prazo se não forem tomadas medidas de preservação.

A preservação de informação em formatos digitais pode ser automatizada mais facilmente do que a preservação de publicações impressas. Porém, a preservação de conteúdos em formato digital tem de ser iniciada mais cedo porque a informação torna-se inacessível rapidamente. Por exemplo, em 1989 a Microsoft lançou a primeira versão do Microsoft Word para Windows (MICROSOFT, 2010). Em 2010 este formato já não é usado e é extremamente difícil aceder a informação guardada neste tipo de documentos. Dada a relativamente curta existência da Web, os principais formatos usados no seu início ainda são populares em 2010. Contudo, a Web permite que seja publicada informação usando qualquer formato digital e é previsível que muitos deles se tornem inacessíveis no futuro.

O ARQUIVO DA WEB PORTUGUESA

O projecto de Arquivo da Web Portuguesa foi iniciado em 2007 na Fundação para a Computação Científica Nacional. Desde 2008 que o AWP faz recolhas exaustivas da Web portuguesa 4 vezes por ano, incluindo ficheiros de todos os tipos (ex. HTML, imagens, PDF). Em breve está planeado iniciar a recolha mais frequente de publicações portuguesas seleccionadas. A recolha é feita com restrições para garantir o bom funcionamento do Arquivo e não

prejudicar o normal funcionamento dos sítios Web visitados. Por exemplo, é respeitada uma pausa entre pedidos a um mesmo sítio da Web para não o sobrecarregar. Apenas é arquivada informação pública e não são preenchidos formulários. Como tal, todas as páginas protegidas por palavra-passe ou outros mecanismos de restrição de acesso não são recolhidas. Cerca de 90% dos ficheiros da Web portuguesa são recolhidos ao fim de 7 dias. No entanto, a recolha continua para os sítios mais lentos ou com maior número de conteúdos.

Entre 1996 e 2004, existiu pouca colaboração entre iniciativas de arquivo da Web. O resultado foi que as várias iniciativas debateram-se repetidamente com os mesmos problemas, investindo individualmente em tentativas de solução. A principal conclusão destes esforços iniciais foi que arquivar eficazmente a Web é um desafio cuja dimensão requer um esforço conjunto à escala mundial. Em 2004, foi criado o projecto Archive-Access, liderado pelo Internet Archive, que aglutinou e desenvolveu ferramentas de software gratuitas e úteis para arquivar a Web (INTERNET ARCHIVE, 2007).

O AWP adoptou três sistemas pertencentes ao projecto Archive-Access: o sistema de recolha Heritrix, o sistema de pesquisa NutchWAX e o sistema de acesso a conteúdos arquivados Wayback. O sistema de recolha foi adoptado sem alterações significativas. Já os sistemas de pesquisa e acesso tiveram de modificados para que os objectivos do AWP fossem alcançados e satisfeitas as necessidades da comunidade portuguesa:

- A interface de utilização foi traduzida para português;
- A usabilidade e grafismo das páginas foram melhorados;
- Os índices foram reestruturados para permitirem respostas mais rápidas;
- A ordenação dos resultados foi melhorada para permitir fornecer resultados mais relevantes.

Figura 5. Página de resultados para uma pesquisa realizada sobre o Arquivo da Web Portuguesa.

A Figura 5 apresenta um exemplo de uma página de resultados para uma pesquisa realizada sobre o AWP. O utilizador preencheu o formulário no topo da página para pesquisar por páginas contendo as palavras *associação apbad* que tivessem sido arquivadas entre os dias *1 de Janeiro de 1996* e *28 de Fevereiro de 2003*.

Cada resultado da lista apresenta:

- O título da página;
- A sua data de arquivo;
- Excertos da página onde ocorrem as palavras pesquisadas;
- O endereço original a partir do qual a página foi arquivada;
- Uma ligação para uma página do *histórico* de todas as versões arquivadas da página.

Figura 6. Página arquivada em Fevereiro de 2002.

A Figura 6 apresenta uma página arquivada em 2002, reproduzida pelo AWP em 2010. A barra amarela no topo da página informa o utilizador de que se trata de uma página arquivada a 3 de Fevereiro de 2002 pelas 23 horas, 57 minutos e 57 segundos, a partir do endereço http://www.apbad.pt/index.htm. Não se tratando da página original, o utilizador é alertado para o facto de poderem existir funções não reprodutíveis na versão arquivada, como por exemplo, formulários. Porém, os utilizadores podem normalmente seguir ligações para outras páginas, realizando uma experiência de navegação no passado.

As páginas arquivadas ficam acessíveis ao público 1 ano após o seu arquivo. O objectivo deste período de embargo de acesso é evitar a concorrência com os sítios Web que os publicaram. Os seus autores podem dar indicações para que uma página não seja arquivada através do Robots Exclusion Protocol (KOSTER, 1994). Após ter sido arquivado um ficheiro, os autores podem solicitar o bloqueio do seu acesso através do AWP. Para este fim, os autores devem enviar a lista dos endereços originais dos ficheiros a bloquear, a data correspondente e os respectivos comprovativos da detenção dos direitos de autor. Os endereços de todos os ficheiros a bloquear são necessários porque os conteúdos de uma página podem ser da autoria de várias entidades. Por exemplo, numa página da Web o direito autoral do texto pode pertencer a um escritor mas a imagem contida na página pertencer a um fotógrafo. A apresentação dos comprovativos é necessária para impedir que sejam feitos pedidos de bloqueio ilegítimos. Por exemplo, se não fossem solicitados quaisquer comprovativos de detenção de direitos sobre os conteúdos arquivados, um pirata informático poderia solicitar o bloqueio de acesso aos conteúdos de uma faculdade ou do Diário da República.

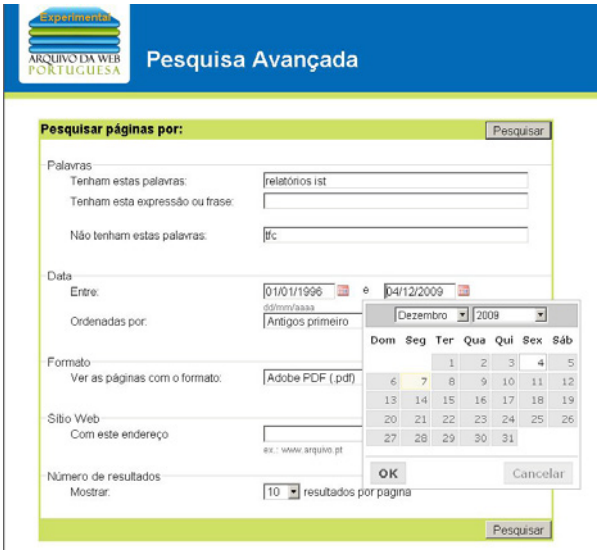

Figura 7. Pesquisa avançada.

A Figura 7 apresenta o formulário de pesquisa avançada, acessível a partir da ligação *Pesquisa Avançada* disponível na página principal de pesquisa (Figura 5). Estas novas funções foram desenvolvidas pelo AWP e permitem, por exemplo, alterar a ordem cronológica da apresentação dos resultados ou seleccionar os tipos de ficheiro a pesquisar (ex. PDF, DOC, PowerPoint).

Resultados

Em Fevereiro de 2010, o AWP detém cerca de 630 milhões de ficheiros arquivados (17,4 Terabytes[1] em formato comprimido). Cerca de 55% do acervo encontra-se indexado e está pesquisável.

O AWP mantém um sítio da Web informativo disponível em http://www.arquivo.pt onde se podem encontrar notícias, explicações acerca do funcionamento do sistema e publicações científicas de caracterização da Web portuguesa. Este sítio tem registado um número crescente de visitas que é ilustrativo do interesse pela temática do arquivo e preservação da Web. Em 2009, foram visitadas um total de 149 257 páginas (média de 408 páginas/dia).

Em Janeiro de 2010, foi lançada uma versão experimental de um serviço de pesquisa sobre cerca de 130 milhões de ficheiros arquivados. Sem ter sido feito qualquer investimento de divulgação, este serviço registou no mês em que foi lançado uma média de 1 229 páginas servidas por dia.

FORNECIMENTO DE DADOS HISTÓRICOS

O AWP iniciou as suas actividades em 2008. A única maneira de preservar conteúdos publicados na Web em data anterior, é obtendo-os a partir de entidades externas que os tenham em sua posse. Assim sendo, foi iniciado um esforço para a obtenção de conteúdos históricos.

O AWP arquiva informação proveniente da Web em ficheiros com o formato ARC (BURNER et al., 1996). Porém, aceita conteúdos fornecidos em qualquer formato e converte-os para este formato para que possam ser indexados. Este processo de conversão é mais eficiente se os fornecedores disponibilizarem as seguintes informações acerca dos dados históricos:

- Os endereços originais do sítio Web e de cada ficheiro (URL). Se estiver a ser fornecida uma cópia de um sítio da Web recomenda-se que os nomes originais dos ficheiros sejam mantidos inalterados;
- As datas de publicação dos ficheiros, mesmo que sejam aproximadas. Caso os ficheiros sejam gravados, por exemplo, para um DVD para serem fornecidos, é necessário ter o cuidado de manter as suas datas originais de criação;
- Os formatos dos ficheiros. É necessário ter o cuidado de manter as extensões originais dos ficheiros (ex. .gif, .html, .jpg) e identificar o tipo dos conteúdos que não contenham extensões no nome dos ficheiros.

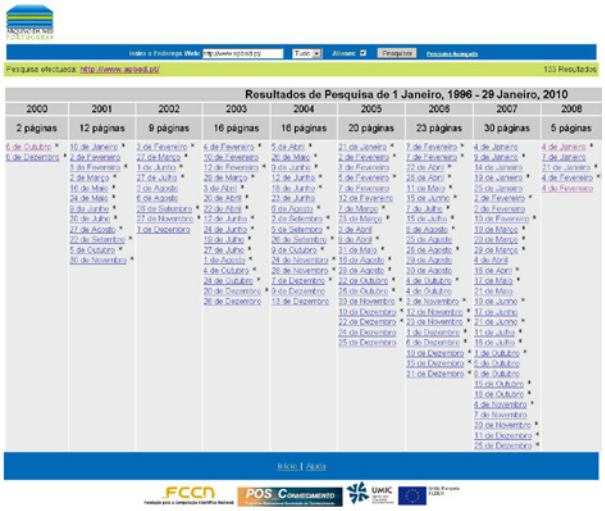

Figura 8. A integração de uma colecção de dados arquivada pelo Internet Archive permite pesquisar informação anterior à existência do Arquivo da Web Portuguesa.

Em 2010, tinham sido convertidos com sucesso 171 milhões de ficheiros históricos fornecidos por exemplo pela Biblioteca Nacional e pelo Internet Archive. A Figura 8 mostra uma lista das versões das páginas arquivadas a partir do endereço http://www.apbad.pt/ entre 2000 e 2008. Teria sido impossível apresentar esta informação no AWP caso esta não tivesse sido fornecida por entidades externas.

Qualquer pessoa ou instituição pode fornecer conteúdos históricos que tenha em sua posse, como por exemplo, cópias de segurança de um sítio Web feitas ao longo do tempo. Existe interesse em arquivar todos os conteúdos que já não estejam disponíveis na Web, independentemente da sua data de publicação.

SELECÇÃO DE SÍTIOS DA WEB PORTUGUESA INTERESSANTES PARA ARQUIVO

Uma recolha da Web portuguesa é iniciada a partir de um conjunto de endereços de sítios da Web interessantes para arquivo. Uma selecção cuidada destes endereços é fundamental para arquivar exaustivamente a Web portuguesa. Actualmente, o processo de selecção

[1] Nota: 1 Terabyte = 1 000 Gigabyte = 1 000 000 Megabyte

de sítios da Web para arquivo é feito de forma exclusivamente automática, baseando-se numa lista de endereços sob o domínio .PT. Porém, existem numerosos sítios da Web com manifesto interesse para a comunidade portuguesa alojados fora do domínio nacional, como por exemplo a Wikipedia portuguesa.

Figura 9. Formulário de sugestão de sítios Web portugueses para arquivo.

O AWP permite que qualquer pessoa sugira um sítio da Web para arquivo. A Figura 9 apresenta o formulário disponível para este efeito. Assim sendo, a comunidade pode contribuir para um arquivo mais exaustivo da Web portuguesa, sugerindo sítios da Web interessantes e fornecendo informações adicionais como descrição do conteúdo, palavras-chave ou periodicidade estimada de actualização dos conteúdos.

A sugestão de sítios da Web permite documentar com maior exaustão eventos de interesse histórico, como por exemplo as eleições. Os profissionais da área de informação e documentação têm uma formação que lhes permite determinar com maior rigor a importância do arquivo de uma informação para acesso futuro. Assim sendo, poderão ser valiosos contribuidores para a selecção de sítios da Web para arquivo periódico.

Por outro lado, é necessário validar segundo critérios de selecção concisos as sugestões feitas pelos cidadãos em geral, antes de incluí-las na lista de raízes. Os especialistas das áreas de ciências da informação e documentação poderiam definir directrizes que permitiriam que um não-especialista identificasse se um determinado sítio da Web deveria ser arquivado.

CONTRIBUIÇÃO DE ESPAÇO EM DISCO PARA PRESERVAR A WEB PORTUGUESA

O rARC é um sistema que permite a qualquer utilizador da Internet contribuir para a preservação da Web. A Figura 10 ilustra o funcionamento do rARC. Cada contribuidor instala uma aplicação e cede espaço em disco do seu computador para guardar uma cópia de parte da informação arquivada no repositório central de um arquivo da Web. Em caso de perda dos conteúdos arquivados no repositório central, devido por exemplo, a uma catástrofe natural, a informação será recuperada a partir das cópias guardadas nos computadores dos contribuidores.

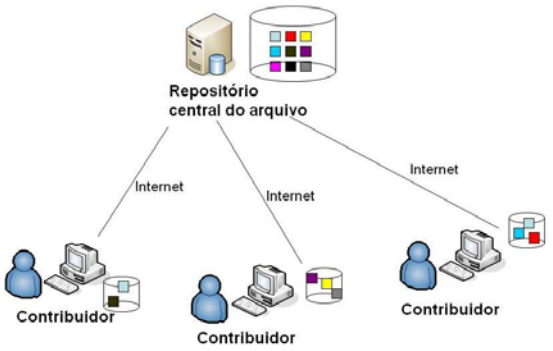

Figura 10. RARC–sistema de replicação colaborativa. Em caso de destruição do repositório central, os dados arquivados são recuperados a partir das cópias dos computadores dos contribuidores.

Apenas será possível criar uma cópia integral do repositório central se for cedido espaço suficiente pelos contribuidores. No entanto, mesmo que não seja possível realizar esta cópia, a salvaguarda parcial da informação arquivada é um cenário mais favorável do que a sua total perda. Quanto maior for a adesão da comunidade, maior será a salvaguarda de conhecimento para o futuro.

A aplicação instalada por um contribuidor tem um impacto mínimo no desempenho do seu computador. Após copiar a informação a partir do arquivo, a aplicação estará a maior parte do tempo inactiva, realizando apenas ligações esporádicas ao repositório central para verificar a integridade das cópias. A aplicação inclui um protector de ecrã que apresentará exemplos de páginas arquivadas. Os contribuidores não podem aceder às páginas contidas nas cópias que estão guardadas nos seus computadores porque estas estão gravadas num formato que não permite que sejam reproduzidas directamente pelos navegadores (*browsers*).

Figura 11. Localização aproximada das cópias guardadas pelos contribuidores do rARC.

No sítio web do AWP é mantida uma lista ordenada dos

contribuidores que doaram mais espaço e a localização geográfica aproximada das cópias (Figura 11). Esta localização permite realizar simulações da quantidade de informação que se salvaria no caso de catástrofes limitadas geograficamente. Por exemplo, poder-se-ia estimar que quantidade de informação se salvaria se houvesse uma catástrofe natural que destruísse o repositório central e as cópias guardadas na zona de Lisboa.

Com o passar do tempo é previsível que existam contribuidores que mudem de computador, reduzam o espaço doado ou decidam desinstalar a aplicação. No entanto, enquanto tiveram as cópias guardadas no seu disco contribuíram para a preservação da História, pois caso tivesse ocorrido um problema com o arquivo central, estas cópias teriam sido muito valiosas para que a informação arquivada não se tivesse perdido.

Todo o *software* do rARC é disponibilizado gratuitamente, podendo ser adaptado para satisfazer as necessidades de diferentes iniciativas de arquivo. O rARC foi desenvolvido no contexto do arquivo da Web para permitir a salvaguarda de ficheiros de forma segura. Porém, pode ser adaptado para distribuir cópias de informação por vários computadores noutros contextos de utilização. Por exemplo, uma organização poderá usar o rARC para guardar cópias de documentos importantes usando o espaço dos computadores dos seus colaboradores, sem que estes tenham acesso à informação.

Estimativas e resultados experimentais

A Web portuguesa é totalmente arquivada 4 vezes por ano. Cada um destes arquivos ocupa cerca de 2 Terabytes de informação (aproximadamente 55 milhões de ficheiros). Considerando que um computador vulgar tem no mínimo um disco de 320 Gigabytes e que 4,5 milhões de portugueses usam a Internet (OCDE, 2008), se cada um destes utilizadores contribuísse com apenas 1% do espaço do seu computador (cerca de 320 Megabytes), seria possível criar cópias de toda a Web portuguesa 720 vezes. Assumindo que a Web portuguesa não cresceria, este espaço permitira guardar cópias para arquivos feitos ao longo de 180 anos.

O principal obstáculo inicialmente identificado ao sucesso do rARC foi a dificuldade em cativar contribuidores. No final de 2009, foi realizada uma experiência em que o rARC foi disponibilizado ao público apenas através de uma ligação no sítio Web do AWP. Não foi feito qualquer investimento em divulgação adicional. Em Janeiro de 2010, o rARC contava com 130 contribuidores que contribuíram com 2,7 Terabytes de espaço em disco, o suficiente para copiar um arquivo total da Web portuguesa. A média registada de espaço doado por contribuidor foi de cerca de 21 000 Megabytes.

Os resultados obtidos são motivantes e mostram que o rARC se trata de uma ferramenta apresenta potencial para ser um forte contributo para a preservação da Web. Contudo, é fundamental continuar a evoluir a ferramenta para angariar e fidelizar novos contribuidores, criando assim uma comunidade activa e ciente da importância da sua contribuição para a preservação da História.

PERFIS DE UTILIZADORES E CASOS DE USO PARA UM ARQUIVO DA WEB

O AWP irá disponibilizar serviços e recursos que poderão ser úteis para diversos perfis de utilizadores em várias situações, como por exemplo:

- Jornalista documentando artigo;
- Gestor de um sítio da Web recuperando versão perdida de página;
- Historiador estudando documentos digitais;
- Utilizador da Web visitando *Favorito* quebrado;
- Jurista obtendo provas para caso.

Existem outros casos de uso e perfis de utilizadores que interessa analisar exaustivamente para que o AWP possa vir a responder às necessidades do máximo número possível de utilizadores. A comunidade de profissionais de informação e documentação poderia contribuir para identificar os perfis e necessidades dos utilizadores do AWP. Os resultados desta investigação contribuiriam para planear com maior solidez o desenvolvimento futuro do serviço, permitindo:

- Identificar perfis de utilizadores de um Arquivo da Web e estimar a sua prevalência no universo de utilizadores;
- Identificar os casos de uso para cada perfil e estimar a percentagem dos casos de uso entre os perfis;
- Identificar necessidades de informação para cada caso de uso;
- Identificar casos de uso e necessidades de informação que não são satisfeitas pelos motores de busca e arquivos da Web actuais.

RESPEITO E DIVULGAÇÃO DE BOAS PRÁTICAS DE PUBLICAÇÃO NA WEB

O AWP publicou um conjunto de recomendações para que a informação publicada na Web possa vir a ser correctamente arquivada, preservada e acedida ao longo dos anos (ARQUIVO DA WEB PORTUGUESA, 2009). Estas recomendações deverão ser respeitadas pelos autores e disseminadas. Além de contribuírem para a preservação das publicações, atraem novos visitantes, pois contribuem para que as suas páginas apareçam em melhor posição nos resultados dos motores de busca.

As recomendações emitidas foram classificadas em fundamentais e aconselháveis. O desrespeito por uma recomendação fundamental impossibilita o arquivo da informação. Como exemplos de recomendações fundamentais temos:

- A existência de uma ligação para o endereço de cada ficheiro. Todos os ficheiros de um sítio da Web têm de poder ser referidos directa e individualmente através de um endereço (URL), quer sejam imagens, vídeos ou páginas. Por exemplo, o endereço http://www.arquivo.pt/logo.jpg referencia o logótipo do AWP;
- Os textos deverão ser publicados usando formatos textuais. Os arquivos da web processam as palavras contidas nos textos das páginas para as tornarem pesquisáveis. Porém, é difícil extrair texto a partir de formatos não textuais, como imagens, código de

programação JavaScript ou animações Flash. É possível fazer um teste simples para identificar se um texto contido numa página da Web foi publicado usando um formato adequado. Basta *seleccionar* o texto da página, *copiá-lo* e tentar *colá-lo* num editor de texto, como por exemplo o Microsoft Word. Se não for possível executar com sucesso algum destes passos, então o texto foi publicado usando um formato inadequado;

- O formato dos ficheiros e codificação de caracteres deverão estar identificados correctamente. Caso contrário, será difícil identificar qual a ferramenta capaz de interpretar, aceder e preservar a informação contida no ficheiro.

Exemplos de recomendações aconselháveis:

- O endereço para um ficheiro inalterado deverá ser mantido ao longo do tempo. Por vezes, ocorre a mudança de endereços, embora os conteúdos por eles referenciados se mantenham inalterados. Por exemplo, se um sítio se chamava www.site.com e mudou para www.site.pt e o primeiro domínio foi desactivado, todas as ligações existentes noutras páginas ou em Favoritos irão ficar inválidas. Porém, os conteúdos do sítio mantiveram-se inalterados. A manutenção dos endereços ao longo do tempo permite aceder ao histórico de versões de uma página, optimizar a frequência da recolha de conteúdos e tornar a navegação entre páginas do passado mais eficiente;
- Deverão ser fornecidos meta-dados. Esta informação enriquece os conteúdos publicados. O Dublin Core é um esquema que permite fornecer meta-dados como o título, nome do autor ou língua de um conteúdo digital (WEIBEL et al., 1998). Os meta-dados das páginas da Web são descritos através de etiquetas que são interpretadas pelos computadores mas que não são apresentadas visualmente aos utilizadores como sendo parte da página. Os meta-dados são particularmente úteis em páginas que contenham pouco texto e deverão descritivos e adequados a cada uma delas, mesmo em sítios da Web que sejam actualizados frequentemente;
- A data de publicação deverá ser explicitamente identificada. Para facilitar a localização cronológica de uma informação. A data de publicação de uma página deverá ser definida pelo autor e apenas alterada no caso de mudança significativa de conteúdo;
- As recomendações de usabilidade e acessibilidade para pessoas com deficiência deverão ser respeitadas. Um sítio da Web difícil de usar nos dias de hoje, também o será no futuro, quando for acedido através de um arquivo da Web. Com o envelhecimento natural, por exemplo da visão, os autores de páginas pouco acessíveis poderão não conseguir aceder aos seus próprios conteúdos que foram arquivados. As recomendações de usabilidade (NIELSEN et al., 2006), acessibilidade para pessoas com deficiência (W3C, 2008) e preservação estão interligadas.

É impossível preservar a Web sem a colaboração dos autores e é essencial realizar um trabalho de consciencialização e disseminação de boas práticas de publicação na Web.

BANCADA PARA AVALIAÇÃO DE RESULTADOS DE PESQUISAS TEMPORAIS

Tal como nos motores de busca actuais, os arquivos da Web necessitam de ordenar os resultados das pesquisas, para que os mais relevantes surjam nas primeiras posições. Porém, os algoritmos existentes foram desenvolvidos para ordenar resultados relativos à Web actual e os sistemas de pesquisa sobre informação arquivada necessitam de ordenar resultados considerando uma perspectiva temporal, pesquisando sobre várias recolhas da Web realizadas ao longo do tempo (COSTA et al., 2009). O AWP tem desenvolvido novos algoritmos de ordenação. No entanto, é imprescindível que existam mecanismos rigorosos de avaliação do impacto geral dos algoritmos desenvolvidos. É frequente que ao tentar melhorar os resultados para certas pesquisas se prejudiquem os resultados de outras.

Existem bancadas para avaliação de sistemas de pesquisa sobre a Web actual compostas por uma colecção de páginas e um conjunto de pares pesquisa/resposta considerados correctos (VOORHEES et al., 2005). Embora estes conjuntos de dados sejam um boa base de avaliação, não podem ser utilizados directamente para avaliar os resultados devolvidos pelo sistema de pesquisa sobre o AWP porque:

- As tarefas definidas poderão não reflectir as necessidades reais dos diversos utilizadores de um arquivo da Web;
- Não consideram a dimensão temporal da informação arquivada. Os conjuntos de teste são constituídos apenas por uma recolha da Web, existindo apenas uma versão de cada conteúdo;
- A afinação dos mecanismos de ordenação é dependente da língua e as colecções existentes são maioritariamente compostas por textos em inglês, ao passo que no AWP a maioria dos textos estão escritos em português.

Assim sendo, é necessário criar uma nova bancada para avaliação dos resultados de pesquisas temporais que permita avaliar o desempenho de vários algoritmos de ordenação de resultados, tendo em consideração as especificidades do AWP e dos seus utilizadores. Esta bancada incluiria uma colecção de dados arquivados e uma lista de pesquisas de índole histórica e respectivos resultados relevantes. Por exemplo, para a tarefa "Encontre informação acerca da campanha de um determinado partido durante as eleições de 2009", os resultados relevantes poderiam ser o sítio web do partido e os blogs dos seus candidatos, recolhidos pelo AWP em Maio de 2009.

A criação de uma bancada para avaliação de qualidade implica a mobilização de um número significativo de pessoas, para que se reúna um conjunto de pesquisas e resultados representativo de vários tipos de pesquisa. Para atingir este objectivo, poderão ser usadas ferramentas como redes sociais, inquéritos remotos ou sistemas de trabalho colaborativo.

CONCLUSÕES

Existem cada vez mais publicações disponíveis exclusivamente na Web. No entanto, a informação publicada neste meio de comunicação é muito mais volátil do que a publicada em documentos impressos. Os arquivos da Web são sistemas informáticos criados para realizar o arquivo automático da informação proveniente na Web e partilham o mesmo objectivo que os arquivos ou bibliotecas nacionais: a preservação de conhecimento para as gerações futuras. Existem várias iniciativas de arquivo da Web em todo o mundo que colaboram para enfrentar os vários e complexos desafios de preservar a informação publicada na Web.

O Arquivo da Web Portuguesa (AWP) é um serviço da Fundação para a Computação Científica Nacional que arquiva, preserva e dá acesso a publicações em-linha de interesse para a comunidade nacional. Em Fevereiro de 2010, o AWP detinha 630 milhões de ficheiros. Porém, não basta armazenar a informação. É necessário disponibilizar serviços que garantam que esta se mantém acessível. Está disponível em http://experimental.arquivo.pt uma versão experimental de um serviço que permite pesquisar sobre cerca de 130 milhões de ficheiros arquivados entre 1996 e 2007. É possível também aceder a funções de pesquisa avançada e ao histórico de páginas arquivadas.

A comunidade de bibliotecários, arquivistas e documentalistas pode colaborar para a preservação da Web portuguesa seleccionando sítios Web interessantes para arquivo, fornecendo informação histórica proveniente da Web que tenha em seu poder, contribuindo com espaço em disco dos seus computadores para guardar cópias do arquivo, identificando perfis de utilizadores dos arquivos das Web, respeitando e divulgando boas práticas de publicação na Web e participando em iniciativas de avaliação da qualidade dos serviços prestados.

A divulgação do AWP é fundamental para o seu sucesso. O sítio Web http://www.arquivo.pt disponibiliza informação adicional e últimas novidades acerca do serviço.

AGRADECIMENTOS

Este artigo é um resultado do trabalho da equipa do Arquivo da Web Portuguesa composta por André Nogueira, David Cruz, João Miranda, Miguel Costa e Simão Fontes. Agradeço ao João Miranda e Teresa Costa pelas suas preciosas revisões. O Arquivo da Web Portuguesa conta com o apoio do programa POSC/EU e UMIC.

REFERÊNCIAS

ARVIDSON A., LETTENSTROM F. The Kulturarw Project-the Swedish Royal Web Archive. The Electronic Library, 16(2):105–108, 1998.

ARQUIVO DA WEB PORTUGUESA. Recomendações para a criação de conteúdos preserváveis ao longo do tempo. Disponível em: http://arquivo-web.fccn.pt/colaboracoes/recomendacoes-para-autores-de-sitios-web, 2009.

BURNER M., KAHLE, B. WWW Archive File Format Specification. Disponível em: http://pages.alexa.com/company/arcformat.html, Setembro 1996.

BRIN, S., PAGE, L. The anatomy of a large-scale hypertextual Web search engine. Proceedings of the Seventh international Conference on World Wide Web 7 (Brisbane, Australia). P. H. Enslow and A. Ellis, Eds. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, 107-117, 1998.

COSTA, M., SILVA M. J., Towards Information Retrieval Evaluation over Web Archives, SIGIR 2009 Workshop on The Future of IR Evaluation, 2009.

GOMES, D., SILVA, M. J. Modelling information persistence on the Web. Proceedings of the 6th international Conference on Web Engineering (Palo Alto, California, USA, July 11 - 14, 2006). ICWE '06, vol. 263. ACM, New York, NY, 193-200, 2006.

INTERNET ARCHIVE. Nutchwax – Home Page. Disponível em: http://archive-access.sourceforge.net/, Novembro de 2007.

KAHLE, B. The Internet Archive. RLG Diginews, volume 6, número 3, 2002.

KOSTER, M. A standard for robot exclusion. Disponível em: http://www.robotstxt.org/wc/norobots.html, 1994.

MICROSOFT. Microsoft Word Grows Up: Q&A. Disponível em: http://www.microsoft.com/presspass/features/2007/jan07/01-04word.mspx, 2010.

NATIONAL LIBRARY OF AUSTRALIA. PADI - Web archiving. Disponível em: http://www.nla.gov.au/padi/topics/92.html, 2010.

NIELSEN J., LORANGER H.. Prioritizing Web Usability. New Riders, 2006.

NTOULAS, A., CHO, J., OLSTON, C. What's new on the web?: The evolution of the web from a search engine perspective. Proceedings of the 13th international Conference on World Wide Web. ACM Press, New York, NY, Maio 17 - 20, 2004.

OCDE. Ministerial Meeting on the Future of the Internet Economy: A Statistical Profile, Seoul, Coreia, 17-18 de Junho de 2008

VOORHEES, E. M., HARMAN, D. K. TREC: Experiment and Evaluation in Information Retrieval (Digital Libraries and Electronic Publishing). The MIT Press, 2005.

W3C. Web Content Accessibility Guidelines (WCAG) 2.0. Disponível em: http://www.w3.org/TR/WCAG20/, 11 Dezembro de 2008.

WEIBEL S., KUNZE J., LAGOZE C., WOLF M.. Dublin core metadata for resource discovery. RFC 2413, IETF, Setembro 1998.