ARQUIVO E MEDIÇÃO DA WEB PORTUGUESA

Daniel Gomes e João Miranda
Fundação para a Computação Científica Nacional
1708-001 Lisboa, Portugal
{daniel.gomes, joao.miranda}@fccn.pt

RESUMO

Este artigo apresenta o projecto de Arquivo da Web Portuguesa em curso na Fundação para a Computação Científica Nacional. O projecto visa preservar a informação publicada na Web para as gerações vindouras à semelhança do que é feito com as publicações impressas nacionais. A disponibilização de serviços eficientes de pesquisa e análise da informação arquivada é essencial para que o Arquivo se torne uma ferramenta usada por todos os cidadãos.
Em Fevereiro de 2008 realizou-se a primeira recolha da Web portuguesa, tendo sido realizadas medições quantitativas. Segundo os resultados obtidos, a Web portuguesa é constituída pelo menos por 56 milhões de conteúdos, o que corresponde a 2,8 TB de informação.

PALAVRAS-CHAVE

Projecto de Arquivo da Web Portuguesa, preservação digital, medições da Web, ferramentas para arquivo da Web

1 INTRODUÇÃO

A Web possibilita que cada um de nós disponibilize informação acessível a todos de uma forma rápida e económica. Diariamente, são publicados milhões de conteúdos na Web como textos, fotografias ou vídeos. A quantidade de informação que é publicada exclusivamente na Web tem vindo a aumentar rapidamente nos últimos anos. No entanto, passado relativamente pouco tempo, a grande maioria desta informação deixa de estar acessível e perde-se irremediavelmente.

O Internet Archive é uma organização norte-americana sem fins lucrativos que recolhe e arquiva conteúdos da Web à escala mundial. 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. Acontecimentos de grande importância para a História dos Estados Unidos da América, como por exemplo o Furacão Katrina, originaram acções de arquivo extraordinárias por parte do Internet Archive. No entanto, 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 perde-se irremediavelmente. Este problema é sentido igualmente por outras comunidades nacionais e pelo menos 16 países já iniciaram as suas próprias iniciativas de arquivo da Web [16].

O projecto de Arquivo da Web Portuguesa (AWP) da Fundação para a Computação Científica Nacional visa a criação de um sistema que terá como missão recolher, armazenar e preservar a informação publicada na Web portuguesa, proporcionando uma cobertura mais exaustiva da informação relacionada com Portugal. A título de exemplo, a quantidade total de informação relativa a sítios web sob o domínio .PT arquivada pelo Internet Archive entre 2000 e 2007 foi de aproximadamente 4 TB, enquanto que em apenas duas recolhas exaustivas do mesmo domínio realizadas pelo AWP foram arquivados 5,3 TB de informação. 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 enquanto língua para comunicação na Web;
- Disponibilizar conteúdos de interesse a diversas comunidades científicas, por exemplo, na área da História, Sociologia ou Processamento Computacional da Língua Portuguesa;
- Contribuir para o desenvolvimento da capacidade local de tratamento e prospecção de informação publicada na Web, reduzindo a dependência de serviços estrangeiros;

- Fornecer provas em casos judiciais que tenham como base informação publicada na Web.

A primeira fase do desenvolvimento do Arquivo teve início em Janeiro de 2008 e prevê-se que termine no prazo de 2 anos. Contudo, a manutenção de um sistema desta natureza e a preservação da informação arquivada é uma tarefa que deverá ser perpetuada posteriormente.

Este artigo descreve o trabalho em curso, incluindo resultados de uma medição da Web portuguesa. O artigo apresenta a seguinte estrutura: no restante desta Secção é apresentada a definição adoptada de Web portuguesa. A Secção 2 descreve os principais objectivos a atingir. Na Secção 3 é apresentada a arquitectura e tecnologia adoptada para a concretização do sistema de Arquivo. A Secção 4 apresenta os resultados obtidos a partir da primeira recolha da Web portuguesa e a Secção 5 conclui o artigo e apresenta trabalho futuro.

1.1 Definição de Web portuguesa

Intuitivamente entende-se como Web portuguesa o conjunto de conteúdos publicados na Web de interesse para a comunidade de Portugal. Esta definição é subjectiva e difícil de definir como um critério automático de selecção [1, 8]. No entanto, um sítio web referenciado por um nome sob o domínio .PT está por definição relacionado com Portugal [18]. Assim sendo, é assumido que um conteúdo pertence à Web portuguesa se o domínio de topo do nome do seu sítio web respeita uma das seguintes condições:

1. Está sob a hierarquia .PT;
2. Não está sob a hierarquia .PT mas o conteúdo está embebido numa página alojada sob .PT. O objectivo é recolher a informação necessária para que as páginas portuguesas arquivadas possam vir a ser apresentadas de forma completa. Por exemplo, uma imagem alojada no sítio web de partilha de fotografias *www.flickr.com* que esteja embutida numa página de um sítio web de blogs como por exemplo o *blogs.sapo.pt*, será considerada como parte da Web portuguesa;
3. Não está sob a hierarquia .PT mas existe um redireccionamento a partir de um nome sob .PT. Por exemplo, uma firma multinacional que regista o domínio .PT com a sua marca mas que o coloca a apontar para o sítio web principal da empresa que está sob a hierarquia .COM.

Posteriormente, esta definição poderá ser alterada para abranger todas as páginas escritas na língua portuguesa independentemente dos domínios dos sítios web que as alojam. Porém, para permitir este alargamento do critério de selecção serão necessários recursos adicionais significativos, pois toda a Web terá de ser percorrida para encontrar as páginas escritas em português.

2 OBJECTIVOS

Com a criação de uma infra-estrutura que suporte recolhas periódicas da Web portuguesa, assim como o seu arquivo e acesso a longo prazo, pretendem-se disponibilizar os seguintes serviços:

- **Pesquisa histórica por endereço da Web.** Esta função permitirá aceder a conteúdos arquivados ao longo do tempo que tenham sido recolhidos a partir de um determinado endereço (URL). Este método de pesquisa apresenta a limitação de impor que os utilizadores conheçam o endereço da página onde pensam encontrar a informação que procuram, o que normalmente não acontece;
- **Pesquisa histórica por termo.** Permitirá identificar páginas arquivadas ao longo dos anos que contenham determinados termos através de uma interface de pesquisa semelhante à disponibilizada pelos motores de busca sobre a Web, como por exemplo o Google. Os resultados serão apresentados aos utilizadores por ordem decrescente de relevância para a pesquisa realizada;
- **Pesquisa sobre a Web portuguesa actual.** O arquivo irá permitir pesquisar sobre várias recolhas da Web portuguesa, logo, a disponibilização de um serviço de pesquisa apenas sobre a recolha mais recente, como acontece nos motores de busca actuais, é um contributo para a comunidade portuguesa relativamente fácil de atingir;
- **Disponibilização de colecções históricas de conteúdos Web para fins de investigação.** A Web contém informação sobre os mais diversos assuntos e investigadores de diversas áreas usam-na como fonte de informação para os seus estudos. A disponibilização de colecções de conteúdos da Web permitirá que os investigadores possam processar informação nos seus computadores sem terem de realizar recolhas da Web;

- **Processamento paralelo dos dados arquivados.** O projecto contribuirá para a computação científica nacional permitindo que investigadores executem os seus programas sobre os dados Web arquivados usando várias máquinas em paralelo, mesmo sem serem especialistas em sistemas informáticos distribuídos;
- **Publicação de relatórios periódicos de caracterização da Web portuguesa.** O desenho de sistemas para processamento de dados provenientes da Web depende das características da informação a tratar. Estes permitirão fornecer estatísticas acerca da quantidade e qualidade da informação publicada na Web portuguesa;
- **Salvaguarda da informação arquivada.** Será desenvolvido um sistema que permitirá a um utilizador da Internet disponibilizar espaço em disco no seu computador para armazenar uma cópia de segurança de parte dos dados arquivados, recorrendo à instalação de uma pequena aplicação no seu computador. Assim sendo, qualquer indivíduo ou instituição poderá colaborar para a preservação da Web portuguesa.

3 DESCRIÇÃO DO SISTEMA

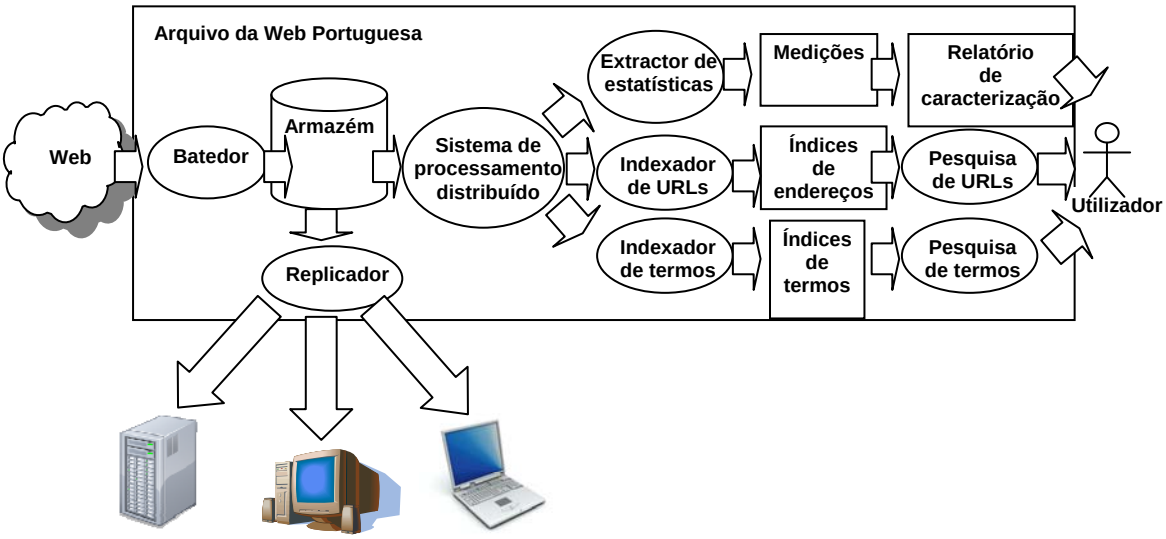

Figura 1. Arquitectura do sistema de Arquivo da Web Portuguesa.

A Figura 1 apresenta uma descrição da arquitectura do sistema do AWP. O *Batedor* percorre a Web, recolhe conteúdos e guarda-os em formato ARC [2] no *Armazém*. Após a informação recolhida da Web estar armazenada é necessário preservá-la e mantê-la acessível. O *Replicador* tem a função de preservar a informação arquivada através da criação de cópias de segurança dos ficheiros arquivados em diferentes computadores espalhados pela Internet, para que, em caso de destruição do Armazém, a informação arquivada possa ser recuperada a partir das cópias de segurança e não se perca irremediavelmente. O *Sistema de processamento distribuído* tem a capacidade de executar tarefas de computação sobre grandes quantidades de informação. A principal vantagem deste sistema é que permite desenvolver aplicações de processamento de dados sem preocupações ao nível da sua gestão e execução distribuída. O Arquivo inicialmente incluirá três aplicações de processamento dos dados arquivados: o *Extractor de estatísticas* que gera medições da Web portuguesa, o *Indexador de endereços* que cria índices que serão usados para suportar a pesquisa histórica por endereço e o *Indexador de termos* que cria índices que serão usados para suportar a pesquisa histórica por termo.

As tecnologias usadas para a concretização do AWP são exclusivamente de código-aberto porque dão maior garantia de preservação do sistema e consequentemente dos dados arquivados a longo prazo, em comparação com soluções de código-fechado que são normalmente dependentes das empresas proprietárias. O facto de o código ser aberto permite também que as ferramentas sejam alteradas para responder aos requisitos específicos do AWP.

Inicialmente, cada entidade empenhada na preservação da informação publicada na Web desenvolveu individualmente as suas ferramentas. Esta situação levou ao desperdício de recursos porque os mesmos problemas estavam a ser resolvidos recorrentemente cada vez que surgia uma nova iniciativa de arquivo da Web. Por forma a fazer face a este problema, foi criado em 2004 o projecto Archive-access que, liderado pelo Internet Archive, reúne e disponibiliza gratuitamente ferramentas para o arquivo da Web [13]. Este projecto permitiu uma grande evolução nesta área pois as ferramentas passaram a ser desenvolvidas em colaboração internacional.

O sistema de AWP baseia-se principalmente em tecnologia disponibilizada pelo projecto Archive-access:

- O Batedor usa o sistema de recolha Heritrix [15];
- O Sistema de processamento distribuído baseia-se no Hadoop, uma poderosa plataforma para processamento paralelo que implementa o paradigma de programação MapReduce criado pela empresa Google [5, 22]. A empresa Yahoo! lançou uma aplicação baseada em 10 000 processadores que cooperam através do Hadoop [12];
- A indexação e pesquisa por endereço é concretizada através da Wayback Machine [23];
- A indexação e pesquisa por termo é baseada no NutchWax, uma adaptação do motor de busca sobre a Web denominado Nutch, que permite processar os ficheiros de arquivo no formato ARC [3];
- O Replicador está a ser desenvolvido pela equipa do AWP e será disponibilizado como um projecto livre de código-aberto denominado *rARC* (replicador de ARCs) para que possa vir a ser utilizado por outras iniciativas de arquivo da Web.

Estas tecnologias são uma valiosa base para o desenvolvimento do AWP mas a sua utilização não é imediata porque, sendo tecnologia de ponta desenvolvida colaborativamente, apresentam sinais de imaturidade e instabilidade. Frequentemente, os processos de instalação e operação não estão documentados e existem erros e incompatibilidades entre versões. A decisão de usar ferramentas do Archive-access exige um envolvimento na sua melhoria.

Um projecto de arquivo da Web apresenta requisitos de maquinaria consideráveis devido ao elevado volume de dados a tratar. A infra-estrutura actual do AWP é constituída por 7 servidores com 8 GB de memória e 2 processadores Quad-core, 1 sistema de gestão de armazenamento com capacidade para 27,2 TB em RAID 5 e 1 robot de cassetes com 30 posições (12 TB) para a realização de cópias de segurança. A maquinaria está a ser testada por forma a definir quais as características adequadas para a criação do ambiente de produção.

4 MEDIÇÃO DA WEB PORTUGUESA

Medir uma Web nacional e compará-la com a de outros países é fundamental para avaliar o desenvolvimento de um país em termos da difusão e utilização das tecnologias de informação. O número de registos de nomes sob um domínio de topo nacional (*country-code Top Level Domain*) está relacionado com a dimensão de uma Web nacional porque os sítios web são referenciados por estes nomes [18]. No entanto, apenas este número não é representativo da dimensão de uma Web nacional porque muitos deles não resultam na criação de conteúdos na Web, sendo criados por exemplo para identificação de equipamentos como servidores de correio electrónico ou *routers*. Por outro lado, existem casos em que apenas um único nome de domínio refere uma grande quantidade de conteúdos na Web, como é o caso do domínio SAPO.PT que inclui milhares de nomes de sítios web como seus subdomínios.

Existe um grande risco em usar caracterizações da Web mundial para descrever a Web portuguesa porque esta apresenta características peculiares. Por exemplo, a grande maioria dos textos da Web portuguesa estão escritos em português mas à escala mundial a língua dominante é o inglês [10, 17]. A Web portuguesa é relativamente pequena e pode ser recolhida exaustivamente e caracterizada mas os resultados obtidos podem diferir consoante a metodologia usada [10]. Contudo, mantendo a mesma metodologia para a realização de várias medições da Web é possível analisar a sua evolução com um elevado rigor.

4.1 Metodologia

O Batedor ciclicamente recolhe conteúdos referidos por endereços da Web e extrai ligações para novos endereços. Uma recolha é iniciada a partir de um conjunto de endereços denominados *raízes*.

Idealmente, todos os endereços de cada sítio web seriam recolhidos. No entanto, isto não é possível devido à existência de sítios web que geram um número infinito de endereços como é o caso de um calendário onde se possam seguir ligações para ver o próximo mês indefinidamente [4]. É impossível distinguir automaticamente as situações em que um sítio web é infinito das que é invulgarmente grande. Assim sendo, é necessário impor restrições de recolha. O Batedor foi configurado com base em estudos acerca das características da Web portuguesa por forma a evitar situações patológicas para o seu funcionamento, garantir uma boa cobertura da Web portuguesa e não prejudicar o funcionamento dos servidores Web visitados [9]. As principais restrições impostas foram as seguintes:

- Máximo de 10 000 endereços recolhidos por sítio web;
- Todos os tipos de documentos foram recolhidos até um máximo de 10 MB desde que estivessem a menos de cinco ligações de uma raiz;
- O protocolo de exclusão de robots foi respeitado assim como um intervalo mínimo de dois segundos entre cada pedido a um mesmo sítio web [14].

Foram usados 180 000 endereços sob o domínio .PT como raízes, derivados de uma lista de domínios geridos pela FCCN (ex.: .PT, .GOV.PT, .COM.PT) e de uma recolha realizada anteriormente pelo motor de busca tumba! [20]. Verificou-se que 12 931 (71%) das raízes referiam um sítio web válido. A recolha foi realizada durante o mês de Fevereiro de 2008.

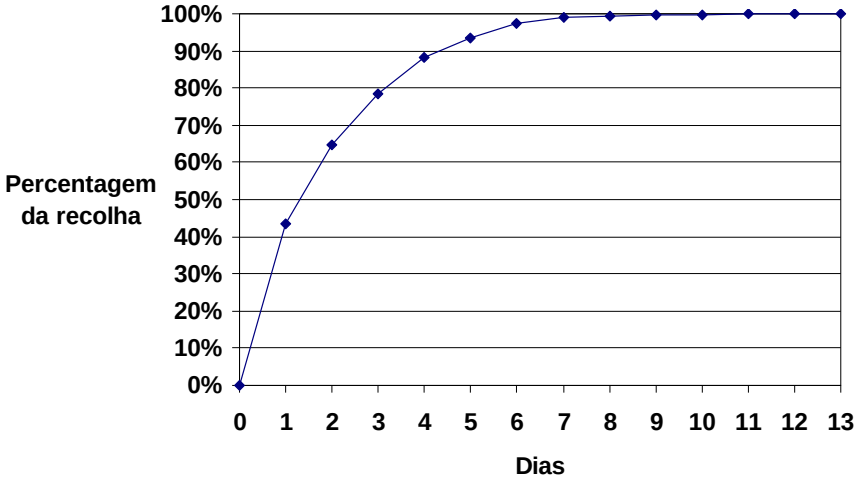

Figura 2. Evolução da recolha da Web portuguesa.

A Figura 2 apresenta a evolução da recolha e mostra que ao fim do sétimo dia, 99% dos conteúdos já tinham sido descarregados. Os restantes seis dias foram ocupados a terminar a recolha de sítios web com tempos de resposta lentos ou invulgarmente grandes. A média de endereços visitados por segundo foi de 49,24 usando uma máquina para alojar o Batedor. Se necessário, as recolhas poderão ser aceleradas usando mais máquinas, uma vez que a largura de banda disponível não foi esgotada. As acções do Batedor não foram lesivas para a grande maioria dos servidores Web, tendo sido recebida apenas uma queixa acerca de uma situação que foi resolvida.

4.2 Resultados

Tabela 1. Recursos visitados e informação recolhida da Web portuguesa.

Métrica	Volume
Endereços visitados	72 milhões
Sítios Web visitados	455 mil
Conteúdos recolhidos	56 milhões
Volume de dados recolhidos	2,8 TB
Volume de dados arquivados em formato comprimido	2 TB

Durante o varrimento da Web portuguesa realizado pelo Batedor foram visitados no total cerca de 72 milhões de endereços alojados em 455 mil sítios web (Tabela 1). Foram recolhidos efectivamente 56 milhões de conteúdos (2,8 TB), que foram armazenados em 2 TB de disco no formato ARC comprimido.

Estima-se que sejam necessários cerca de 10% de espaço adicional por recolha para alojar as estruturas de dados que suportam os sistemas de pesquisa, ou seja, 280 GB para a recolha em causa. Considerando recolhas trimestrais serão necessários 9,12 TB de espaço para armazenamento dos conteúdos arquivados por ano. Porém, existem conteúdos que se mantêm inalterados ao longo do tempo, sendo recolhidos e armazenados repetidamente. No arquivo da Web islandesa, a eliminação de duplicados entre recolhas com cerca de 4 meses de intervalo permitiu reduzir em 42% o tamanho total de uma nova recolha [19]. Com base nestes resultados estima-se que seja possível baixar para 5,76 TB os requisitos anuais de armazenamento para a Web Portuguesa. A ferramenta desenvolvida pela iniciativa de arquivo islandesa para a eliminação de duplicados é de código-aberto e disponibilizada gratuitamente.

Tabela 2. Códigos de resposta HTTP devolvidos pelos servidores Web.

Código de resposta	Número de URLs	%	Descrição
200	56 046 288	85,2%	Conteúdo recolhido com sucesso
302	4 305 265	6,5%	Redireccionamento temporário
404	3 669 855	5,6%	Conteúdo não encontrado
301	789 133	1,2%	Redireccionamento permanente
500	325 225	0,5%	Erro interno do servidor
400	266 318	0,4%	Mau pedido
403	164 241	0,2%	Acesso proibido
303	124 385	0,2%	Redireccionamento alternativo
401	48 334	0,1%	Acesso não autorizado
outros	36 136	0,1%	Outros códigos
Total	65 775 180	100%	-

A Tabela 2 apresenta um sumário dos códigos HTTP respondidos pelos servidores Web durante a recolha [6]. O total de 65 milhões de códigos é inferior aos 72 milhões endereços visitados porque houveram visitas que não originaram códigos de resposta HTTP devido, por exemplo, a erros de ligação ao servidor. O número de ligações quebradas (código 404) encontradas em sítios web é um indicativo da sua qualidade. Uma percentagem elevada de ligações quebradas sugere que os sítios web se encontram abandonados ou são mantidos de forma deficiente. A percentagem de ligações quebradas encontradas na Web portuguesa é semelhante à de outros países [1].

Tabela 3. Os 10 tipos de conteúdo mais usados na Web portuguesa.

Posição	Tipo MIME	Número de URLs	%	Descrição
1	text/html	42 748 509	65,0%	Página Web
2	image/jpeg	11 630 295	17,7%	Imagem
3	image/gif	4 981 051	7,6%	Imagem
4	image/png	1 350 550	2,1%	Imagem
5	text/plain	1 000 333	1,5%	Texto sem formatação
6	application/pdf	905 119	1,4%	Documento PDF
7	no-type	379 884	0,6%	O servidor não devolveu tipo
8	text/xml	359 326	0,5%	Documento em XML
9	application/x-shockwave-flash	348 214	0,5%	Animação Flash
10	application/x-gzip	328 964	0,5%	Arquivo comprimido
11	outros	1 710 156	2,6%	Outros tipos

A Tabela 3 apresenta os 10 tipos de conteúdo mais comuns na Web portuguesa. Os servidores Web visitados responderam com 738 tipos MIME distintos, sendo a maioria destes inválidos [11]. No entanto, 92,4% dos conteúdos pertencem aos quatro tipos mais comuns. Os servidores Web visitados não devolveram a identificação do tipo do conteúdo para 0,6% dos endereços. Esta situação é anómala pois os clientes HTTP, como por exemplo os *browsers*, necessitam desta informação para poderem apresentar os conteúdos correctamente aos utilizadores.

Tabela 4. Os 10 tipos de conteúdo que contribuíram com maior quantidade de dados.

Pos.	Tipo MIME	Total de dados	%	Posição na Tabela 3	Descrição
1	text/html	1 133 GB	39,6%	1	Página Web
2	application/pdf	413 GB	14,4%	6	Documento PDF
3	image/jpeg	355 GB	12,4%	2	Imagem
4	text/plain	133 GB	4,7%	5	Texto sem formatação
5	application/x-gzip	124 GB	4,3%	10	Arquivo comprimido
6	application/zip	104 GB	3,6%	16	Arquivo comprimido
7	application/x-tar	79 GB	2,8%	19	Arquivo comprimido
8	application/octet-stream	67 GB	2,4%	15	Octetos
9	app'n/x-shockwave-flash	49 GB	1,7%	9	Animação Flash
10	image/gif	48 GB	1,7%	3	Imagem comprimida
11	Outros	356 GB	12,4%	-	Outros tipos

A Tabela 4 apresenta os 10 tipos de conteúdo que contribuíram com maior quantidade total de dados para a recolha. Comparando a Tabela 3 com a Tabela 4 verifica-se que 7 dos 10 tipos existem em ambas. No entanto, a sua presença relativa varia. Por exemplo, o tipo application/pdf ocupa a sexta posição na Tabela 3 representando apenas 1,4% dos endereços visitados, mas 14,4% do total de dados recolhidos ocupando a segunda posição da Tabela 4.

Verificou-se que 2,4% dos dados recolhidos foram identificados pelos respectivos servidores Web como sendo do tipo application/octet-stream. Este tipo destina-se a identificar genericamente conteúdos digitais, sendo difícil para os clientes HTTP interpretarem correctamente estes conteúdos. Assim sendo, é recomendado que o tipo application/octet-stream seja respondido apenas quando o servidor Web não consegue identificar o tipo do conteúdo que está a servir, sendo a sua utilização desaconselhada na Web ou em mensagens de correio electrónico [7].

5 CONCLUSÕES E TRABALHO FUTURO

O projecto de Arquivo da Web Portuguesa (AWP) visa disponibilizar serviços que permitam que se torne uma importante ferramenta no dia-a-dia de todos os cidadãos. Para atingir este objectivo estão em desenvolvimento métodos de pesquisa que permitam tornar acessível a informação arquivada. A recolha periódica da Web portuguesa foi iniciada em Fevereiro de 2008 tendo sido realizadas medições quantitativas preliminares. Segundo os resultados obtidos, a Web portuguesa é constituída pelo menos por 56 milhões de conteúdos alojados em 455 mil sítios web, o que corresponde a cerca de 2,8 TB de informação. Cerca de 92% dos conteúdos da Web portuguesa são páginas HTML, ou imagens dos tipos JPEG, GIF ou PNG. Um estudo aprofundado de caracterização da Web portuguesa está em fase avançada de concretização. Aquando da publicação deste estudo serão disponibilizados os dados que lhe serviram de base para que possam vir a ser utilizados noutros trabalhos de investigação. Os estudos futuros de caracterização deverão incluir métricas relacionadas com a qualidade da Web portuguesa, como, por exemplo, o nível de acessibilidade das páginas a pessoas com deficiência e o respeito por normas de formato.

De 2002 a 2006, o projecto de investigação tumba! recolheu cerca de 57 milhões de conteúdos maioritariamente textuais da Web portuguesa [21] e o Internet Archive detém cerca de 130 milhões de conteúdos arquivados entre 2000 e 2007. Estes conteúdos estão a ser replicados na infra-estrutura do AWP, sobre os quais serão disponibilizados métodos de pesquisa e acesso.

O AWP está a desenvolver ferramentas inovadoras para o arquivo da Web: o replicador *rARC* que visa a criação de cópias de segurança da informação arquivada espalhadas pela Internet, o *GAppA* que se trata de uma *Grid Appliance* que permitirá a partilha de infra-estruturas e dados arquivados para fins de computação científica e o sistema *WebClass* que tem como objectivo a classificação automática dos conteúdos arquivados por tema (ex.: desporto, ciência, política).

Quando um motor de busca procura um termo dentro de uma única recolha da Web são encontrados milhões de textos mas os utilizadores normalmente analisam no máximo algumas dezenas de resultados. Os motores de busca sobre a Web usam algoritmos de ordenação dos resultados das pesquisas para mostrarem primeiro os resultados mais relevantes. No entanto, o estudo de algoritmos de ordenação para pesquisas de

âmbito histórico sobre várias recolhas realizadas ao longo do tempo é um tópico de investigação pouco aprofundado até à data, mas fundamental para a criação de mecanismos de pesquisa eficientes sobre arquivos da Web. O estudo de algoritmos de ordenação adequados à Web portuguesa está em curso.

Para o sucesso do projecto do AWP nas suas múltiplas vertentes são necessários conhecimentos profundos sobre diferentes áreas de conhecimento. Por isso, é necessário cativar a participação de entidades externas à comunidade do arquivo da Web, principalmente a nível nacional. Pretende-se que o AWP, além de ser um fornecedor de recursos para investigação, possa contribuir para fomentar a colaboração dentro da comunidade científica nacional e servir como prova de conceito para resultados de investigações.

REFERÊNCIAS

[1] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis Efthimiadis. Characterization of national web domains. *ACM Transactions on Internet Technology*, 7(2), 2007.

[2] Mike Burner and Brewster Kahle. WWW Archive File Format Specification. http://pages.alexa.com/company/arcformat.html, September 1996.

[3] Mike Cafarella and Doug Cutting. Building Nutch: Open Source Search. *Queue*, 2(2):54–61, 2004.

[4] Carlos Castillo. *Effective Web Crawling*. PhD thesis, University of Chile, November 2004.

[5] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Procedings of the 6th Symposium on Operating System Design and Implementation*, December 2004.

[6] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. *Hypertext Transfer Protocol – HTTP/1.1*, June 1999.

[7] N. Freed and N. Borenstein. *Multipurpose Internet Mail Extensions (MIME) Part Two: Media Types*, November 1996.

[8] Daniel Gomes, Sérgio Freitas, and Mário J. Silva. Design and selection criteria for a national web archive. In Julio Gonzalo, Costantino Thanos, M. Felisa Verdejo, and Rafael C. Carrasco, editors, *Proc. 10th European Conference on Research and Advanced Technology for Digital Libraries, ECDL*, volume 4172. Springer-Verlag, September 2006.

[9] Daniel Gomes and Mário J. Silva. The viúva negra crawler: an experience report. *Softw. Pract. Exper.*, 38(2):161–188, 2008.

[10] Daniel Gomes and Mário J. Silva. Characterizing a national community web. *ACM Transactions on Internet Technology*, 5(3):508–531, 2005.

[11] Internet Assigned Numbers Authorrity (IANA). *IANA: MIME Media Types*, November 2004.

[12]Yahoo! Inc. Yahoo! launches world's largest hadoop production application. http://developer.yahoo.com/blogs/hadoop/2008/02/yahoo-worlds-largest-production-hadoop.html, February 2008.

[13] Internet Archive. Nutchwax - Home Page. http://archive-access.sourceforge.net/, March 2008.

[14] Martijn Koster. A standard for robot exclusion. http://www.robotstxt.org/wc/norobots.html, June 1994.

[15] Gordon Mohr, Michele Kimpton, Micheal Stack, and Igor Ranitovic. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04)*, Bath, UK, September 2004.

[16] National Library of Australia. PADI - Web archiving. http://www.nla.gov.au/padi/topics/92.html, August 2007.

[17] E. T. O'Neill, B. F. Lavoie, and R. Bennett. How "world wide" is the web?: Trends in the evolution of the public web. *D-Lib Magazine*, 9(4), April 2003.

[18] J. Postel. *Domain Name System Structure and Delegation*, 1994.

[19] Kristinn Sigursson. Managing duplicates across sequential crawls. In *6th International Web Archiving Workshop (IWAW06)*, Alicante, Spain, September 2006.

[20] Mário J. Silva. The case for a portuguese web search engine. In Pedro Isaias, editor, *Proceedings of IADIS International Conference WWW/Internet 2003*, Algarve, Portugal, November 2003.

[21] Mário J. Silva. Searching and archiving the web with tumba! In *CAPSI 2003 - 4a. Conferência da Associação Portuguesa de Sistemas de Informação*, Porto, Portugal, November 2003.

[22] The Apache Software Foundation. About Hadoop. http://lucene.apache.org/hadoop/about.html, August 2006.

[23] Brad Tofel. Preserving the bits of the danish internet. In *7th International Web Archiving Workshop (IWAW07)*, Viena, Austria, September 2007.