

# Modelação da Web para Desenho de Armazéns de Dados

Daniel Gomes

16 de Setembro de 2010

## **Classificação ACM:**

H.3.1 [Information Systems]: Information Storage and Retrieval: Content Analysis and Indexing;

H.3.4 [Information Systems]: Systems and Software: Distributed systems;

H.3.5 [Information Systems]: Information Storage and Retrieval: Online Information Services-Web-based services.

**Palavras-Chave:** Armazenamento de Dados Web, Recolha de Dados da Web, Caracterização da Web.

## **Resumo**

Os utilizadores da Web recorrem a ferramentas que os ajudem a satisfazer as suas necessidades de informação. Contudo, as características específicas dos conteúdos provenientes da Web dificultam o desenvolvimento destas aplicações. Uma aproximação possível para a resolução deste problema é a integração de dados provenientes da Web em Armazéns de Dados que disponibilizem métodos de acesso uniformes e facilitem o processamento automático da informação. Um Armazém de Dados provenientes da Web (ADW) é conceptualmente semelhante a um Armazém de Dados relacionais. No entanto, a estrutura da informação a carregar a partir da Web, não pode ser controlada ou facilmente modelada pelos analistas. Os modelos da Web existentes não são tipicamente representativos do seu estado presente. Como consequência, os ADW sofrem frequentemente alterações profundas no seu desenho quando já se encontram numa fase avançada de desenvolvimento. Estas mudanças têm custos elevados e podem pôr em causa a viabilidade de todo um projecto. Este trabalho estuda o problema da modelação da Web e a sua influência no desenho de ADW. Para este efeito, foi extraído um modelo de uma porção da Web, e com base nele, desenhado um protótipo de um ADW. Os resultados obtidos mostram que a modelação da Web deve ser considerada no processo de integração de dados da Web. Os resultados desta investigação estão a ser aplicados na criação de um sistema de arquivo da Web portuguesa.

## **1 Introdução**

A Web ainda é vista principalmente como um meio de publicação destinado a ser interpretado por pessoas, mas esta perspectiva é muito limitativa face às suas potencialidades. A criação de aplicações que interpretem automaticamente dados da Web para extracção

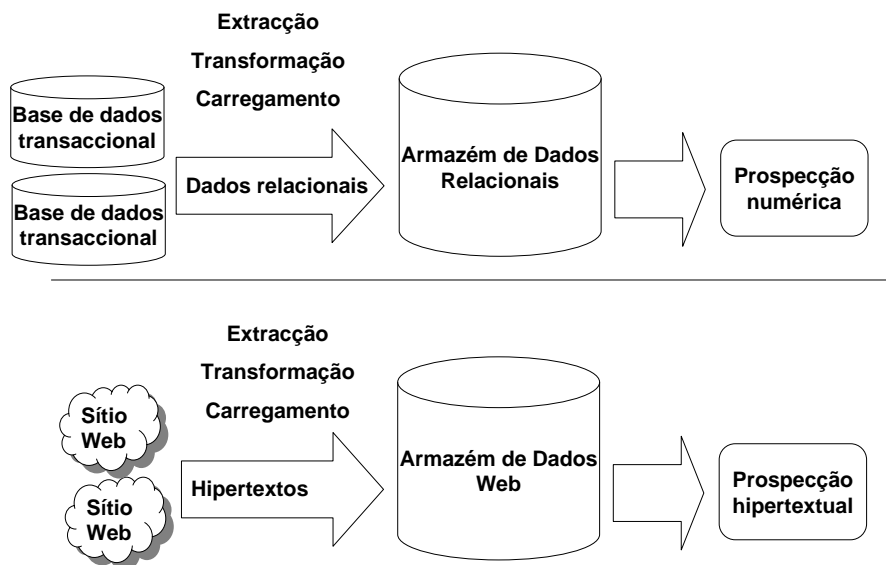


Figura 1: Armazenamento de dados relacionais *vs.* Web.

de conhecimento é possível e necessária. Estas aplicações podem ter finalidades distintas, como originar serviços de pesquisa, estudos científicos ou análises de informação histórica que já não se encontra disponível. No entanto, se construídas de raiz, todas estas aplicações se irão debater com problemas comuns:

- A informação disponível na Web é vasta e encontra-se dispersa, o que dificulta a localização de dados relevantes;
- A heterogeneidade de formatos de publicação e o desrespeito pelas especificações constituem um entrave à interpretação automática;
- A volatilidade da informação faz com que o seu acesso seja pouco fiável.

Uma possível solução para facilitar o desenvolvimento destas aplicações é a integração de dados provenientes da Web em Armazéns de Dados, que disponibilizem métodos de acesso uniformes destinados ao processamento automático. O processo de integração de dados da Web é conceptualmente semelhante ao de integração de dados relacionais e está descrito na Figura 1. No entanto, a abordagem tradicional de integração de dados usada nos sistemas de Armazenamento de Dados (*Data Warehousing*) tem-se revelado inadequada no contexto da Web, uma vez que os pressupostos no desenho deste tipo de sistemas não são aplicáveis. A arquitectura de um sistema de Armazenamento de Dados relacionais pressupõe que:

- As características das fontes de informação são bem conhecidas à partida;
- Se destina a apoiar sistemas de suporte à decisão que actuem sobre informação de pequena granularidade, tipicamente relacional;
- O processo de integração é feito em fases independentes, sendo a fase de recolha de informação pouco complexa.

Em contraposição, o desenho da arquitectura de um Armazém de Dados provenientes da Web (ADW) deverá pressupor que:

- As fontes de informação não são bem conhecidas;
- Destina-se a suportar sistemas de recuperação de informação hipertextual;
- O processo de integração impõe uma estrita cooperação entre as diferentes fases, sendo a recolha de informação dificultada por problemas de acessibilidade à informação.

Os ADW são carregados com dados de uma determinada porção da Web. Vários estudos revelaram que cada porção da Web apresenta as suas características peculiares [BYCE07]. Assim sendo, é importante delimitar estas porções e modelá-las para que um ADW possa ser desenhado considerando as características da informação que irá processar. O processo de integração de dados da Web compreende as fases de modelação da fonte de informação, recolha, transformação, armazenamento e acesso. Num ADW cada uma delas coloca novos desafios:

**Modelação.** A Web não é uma fonte de informação uniforme e apenas uma parte da informação que disponibiliza será integrada num ADW. Impõe-se assim, a definição de critérios de selecção de porções de informação relevante. Após a definição da fronteira de uma porção da Web, esta deverá ser modelada para permitir um desenho adequado do ADW que a irá alojar. No entanto, esta modelação é dificultada pela ausência de estatísticas e caracterizações acerca da Web. Surge assim o interesse em metodologias que permitam efectuar a modelação sistemática de porções da Web;

**Recolha e transformação.** A tradução do critério de selecção numa política de recolha automática afirma-se como o primeiro desafio desta fase. O processo de recolha e transformação de dados deverá ser eficiente e ao mesmo tempo robusto. A vastidão e diversidade da Web, impossibilitam a previsão e teste de todas as situações que poderão vir a ser encontradas;

**Armazenamento e acesso.** O armazenamento de dados requer estruturas de dados diferentes das usadas na fase de carregamento para poder proporcionar um modelo de acesso eficiente e uniforme. O grande volume de informação impõe que esta esteja acessível a pessoas e máquinas, pelo que os mecanismos de acesso deverão suportar processamento paralelo e gestão escalável da informação.

Este trabalho foca o problema da modelação da Web e a sua influência no desenho de ADW. A metodologia adoptada foi principalmente experimental. Um modelo de uma porção da Web foi extraído, designadamente a Web portuguesa, definida como o conjunto de documentos de interesse cultural e sociológico para os portugueses. Com base neste modelo, foi desenhado um ADW que foi iterativamente desenvolvido e avaliado através da sua utilização em diversos contextos distintos. Por exemplo, o sistema foi usado como plataforma para o estudo de mecanismos de indexação e pesquisa de conteúdos na Web sendo um dos principais componentes do motor de busca tumba! [Sil03, Cos04, Mar04], foi usado como plataforma para análises linguísticas [MS04, SMC<sup>+</sup>06] e como repositório de artigos científicos [Jul05]. Os resultados obtidos mostram que a modelação tem impacto no desenho de ADW e deve ser considerada no processo de integração de dados da Web.

Os principais contributos deste trabalho são os modelos da estrutura e persistência de informação na Web portuguesa e uma análise do impacto que estes tiveram no desenho do sistema oficial que fará seu arquivo, o Arquivo da Web Portuguesa. Estes contributos

científicos resultaram da investigação desenvolvida pelo autor ao longo do seu doutoramento e no trabalho realizado posteriormente, com a aplicação dos resultados obtidos no desenvolvimento do Arquivo da Web Portuguesa.

O objectivo da investigação foi encontrar respostas para as seguintes questões:

- Quais as características que deverão ser consideradas num modelo da Web?
- Como podem ser definidas as fronteiras de uma porção da Web?
- O que pode influenciar um modelo da Web?
- Qual é o grau de persistência da informação disponível na Web?
- Qual a influência das características da Web no desenho de ADW?

Este documento está estruturado da seguinte forma: a secção 2 apresenta um estudo de caracterização da estrutura da Web portuguesa e a secção 3 propõe modelos para estimar a persistência de informação na Web. A secção 4 descreve sumariamente o sistema de arquivo da Web portuguesa e o impacto que os resultados apresentados nas secções anteriores tiveram no seu desenho. Finalmente, a secção 5 apresenta as principais conclusões retiradas, sintetizando as respostas obtidas para as questões que originaram a investigação.

## 2 Caracterização estrutural de uma Web nacional

A Web é composta por porções com características peculiares que são de interesse para grandes comunidades, tais como, a comunidades nacionais. As características das Webs nacionais podem não ser visíveis em caracterizações da Web global devido à sua dimensão relativamente pequena.

Um ADW é carregado com conteúdos de uma porção da Web relevante para uma determinada comunidade de utilizadores. Assim sendo, é fundamental modelar a porção da Web a armazenar para que se possa desenhar e gerir o ADW de forma eficiente. Os ADW são normalmente carregados através da recolha automática de conteúdos realizada por sistemas conhecidos como *batedores* (*crawlers*). Um batedor recolhe e armazena iterativamente conteúdos da Web, seguindo as ligações contidas para encontrar novos conteúdos. Cada nova recolha é iniciada a partir de um conjunto de endereços denominados *raízes*. A Web é caracterizada através da análise de amostras de informação que são posteriormente processadas para extracção de modelos estatísticos. Neste estudo, a metodologia adoptada para extrair amostras para caracterização foi a recolha automática realizada por um batedor para que o modelo derivado fosse o mais representativo possível da informação armazenada em ADW.

Esta secção apresenta uma caracterização detalhada da Web portuguesa extraída a partir de 3,2 milhões de conteúdos recolhidos em 2003. As raízes desta recolha foram derivadas a partir de uma lista de domínios registados sob a hierarquia *.pt*. O autor demonstrou em trabalho posterior que as analisadas características da Web portuguesa não se alteraram significativamente até 2008. Consequentemente, a caracterização apresentada mantêm-se actual. Os resultados obtidos são interessantes para quem necessite de processar informação proveniente da Web portuguesa e permitirão comparações com caracterizações futuras para derivação de tendências de evolução. Adicionalmente, a identificação de métricas significativas para a caracterização de Web nacionais e os métodos adoptados para obter e analisar os resultados obtidos, são úteis para uma audiência mais vasta do que a comunidade interessada na Web portuguesa em particular.

## 2.1 Identificação das fronteiras

A delimitação da fronteira de uma Web nacional não é óbvia porque a Web é uma rede de conteúdos à escala mundial criada para ultrapassar barreiras geográficas. É necessário estabelecer critérios que determinem quais os conteúdos a incluir como parte de uma Web nacional. Uma Web nacional é interessante para quem necessita de informação com um âmbito nacional. Por exemplo, quando os utilizadores da Web procuram uma página acerca da ‘Biblioteca Nacional’, estão interessados em encontrar informação acerca da biblioteca do seu país. Empiricamente, uma Web nacional é o conjunto de conteúdos que contém informação relativamente a um determinado país. Contudo, esta definição é muito subjectiva e por isso, difícil de implementar como um conjunto de regras interpretáveis por máquinas que permitam realizar recolhas automáticas de informação para ser integrada num ADW.

A gestão dos domínios de topo de cada país (ccTLD: *country-code Top Level Domains*) é delegada para gestores nacionais com o objectivo de permitir uma satisfação mais eficiente das necessidades locais. Por definição, os conteúdos alojados em sítios Web referidos por um nome sob um domínio de topo nacional têm um âmbito local [Pos94]. Assim sendo, estes conteúdos deverão ser incluídos numa Web nacional. Contudo, muitos sítios Web de interesse nacional são registados em domínios de âmbito genérico (gTLD: *general-purpose Top Level Domains*), tais como `.com` ou `.net`, devido a restrições legais ou custos impostos em certos países.

Existem várias hipóteses para tentar ultrapassar este problema e identificar automaticamente conteúdos de interesse nacional alojados fora do domínio de um país. Uma Web nacional poderia incluir conteúdos alojados em servidores com endereços da Internet que estivessem fisicamente atribuídos ao país [BYCL05]. No entanto, a correspondência entre um endereço da Internet e a sua localização geográfica é feita através de bases de dados, tais como o RIPE Network Management Database [RAR92], que por vezes se encontram desactualizadas e fornecem localizações erradas. Além disso, os publicadores alojam os seus sítios web nos serviços que oferecem melhores condições, independentemente da localização geográfica dos servidores. Existem línguas que são faladas num único país do mundo, como por exemplo o sueco. Nestes casos, uma Web nacional poderá ser identificada com base no conjunto de conteúdos textuais escritos numa determinada língua. No entanto, esta aproximação falha quando se tratam de línguas que são faladas em vários países no mundo, como o inglês ou o português. As bases de dados WHOIS contêm informação de contacto acerca dos detentores e gestores de domínios e endereços da Internet e poderiam ser usadas para incluir sítios Web pertencentes a cidadãos nacionais [HSF85]. Porém, existem detentores de domínios que fornecem subdomínios para alojar outros sítios Web, independentemente do âmbito do seu conteúdo. Por exemplo, os blogs de autores portugueses alojados sob o domínio `blogspot.com` deveriam ser considerados como parte da Web portuguesa, mas o registo WHOIS do domínio `blogspot.com` não contém qualquer informação relativa a Portugal.

Provavelmente, atingir-se-ia uma boa cobertura de uma Web nacional reunindo os conteúdos identificados por todas as heurísticas apresentadas mas esta abordagem também incluiria muitos conteúdos irrelevantes. Não obstante de possíveis limitações, para realizar uma caracterização de uma Web nacional é necessário estabelecer um critério de selecção de conteúdos, que possa ser implementado como uma política de recolha automática e permita obter uma boa cobertura da Web portuguesa. Neste trabalho, foi considerado que um conteúdo seria parte da Web portuguesa se satisfizesse uma das seguintes condições:

**Condição 1:** estivesse alojado num sítio Web sob domínio `.pt`;

**Condição 2:** estivesse alojado num sítio Web sob os domínios de uso genérico `.com`, `.net`, `.org` ou `.tv`, escrito em língua portuguesa e recebesse pelo menos uma ligação directa a partir de uma página alojada sob o domínio `.pt`.

O objectivo da Condição 1 é incluir os sítios Web considerados como o núcleo da Web portuguesa. Uma lista dos sítios Web mais acedidos a partir das casas de utilizadores da Internet portuguesas mostrou que 49,5% deles estavam alojados sob o domínio `.pt` [Mar03]. O principal objectivo da Condição 2 é incluir o crescente número de sítios Web cujo domínio está registado fora de `.pt` [Zoo00]. A estrutura de ligações entre os conteúdos da Web pode ser usada para identificar comunidades na Web [FLG00, GKR98]. Com base nestes resultados, a Condição 2 assume que a probabilidade de um sítio Web alojado fora do domínio `.pt` pertencer à Web portuguesa, diminui à medida que o número de ligações a partir do núcleo constituído pelos conteúdos sob `.pt` aumenta. Portanto, o número de ligações de distância ao núcleo foi limitado a 1, ou seja, apenas são aceites ligações directas, para que a inclusão de sítios Web fora de `.pt` se limite àqueles com uma probabilidade elevada de pertencerem à Web portuguesa. No entanto, esta definição não é perfeita. Por exemplo, os conteúdos brasileiros alojados sob `.com` que recebem ligações directas a partir de páginas sob `.pt`, serão considerados como parte da Web portuguesa. A definição de uma Web nacional tem um contexto geográfico associado implicitamente. Ferramentas de localização geográfica que podem ser usados para colmatar as limitações da definição de Web portuguesa proposta. Esta hipótese foi avaliada através da realização de uma experiência para comparação de heurísticas de identificação de sítios Web portugueses fora do domínio nacional, recorrendo a ferramentas que lhes atribuísem um âmbito geográfico:

**DNS LOC:** a informação geográfica acerca de sítios Web pode ser obtida através de uma pesquisa sobre um registo DNS especial denominado LOC [DVGD96]. Um sítio Web foi considerado como parte da Web portuguesa se o registo LOC do seu domínio o localizasse em Portugal;

**Ferramentas geográficas:** duas ferramentas comerciais denominadas Ip2location [Hex03] e Maxmind [Max03]. Um sítio Web foi considerado como parte da Web portuguesa se as ferramentas o localizassem em Portugal. Excepto para um caso, ambas as ferramentas devolveram resultados coincidentes, o que sugere que se baseiem nos mesmos dados para extrair âmbitos geográficos;

**WHOIS:** um sítio Web foi considerado como parte da Web portuguesa se o seu registador tivesse fornecido uma morada de contacto em Portugal.

Para avaliar a eficácia destas heurísticas foi compilada uma lista de sítios Web portugueses alojados fora de `.pt` sugeridos por um painel de utilizadores nacionais. Estes sítios Web foram analisados para verificar se continham conteúdos interessantes para a comunidade portuguesa. Todos estavam escritos em língua portuguesa e os seus conteúdos abarcavam temas diversos tais como desporto, humor ou programas de rádio.

A Tabela 1 apresenta os resultados obtidos usando as 4 heurísticas propostas. Nenhum dos sítios Web analisados tinha um registo DNS LOC associado. As ferramentas geográficas identificaram apenas 44% dos sítios Web portugueses. A heurística baseada na informação do WHOIS atingiu 76% de eficácia sendo que, para os restantes sítios Web não foi encontrado um registo WHOIS. Esta heurística revelou ser um método preciso de identificação de sítios Web portugueses. Porém, as bases de dados WHOIS proíbem o acesso automático à sua informação e utilizam diferentes formatos para devolverem resultados, o que colide com o objectivo de definir as fronteiras da Web portuguesa através

Heurística	% sítios identificados
DNS LOC	0%
Ferramentas geográficas	44%
Morada do registador no WHOIS	76%
Língua e ligações (Condição 2)	82%

Tabela 1: Heurísticas para identificação de sítios Web portugueses alojados fora de .pt.

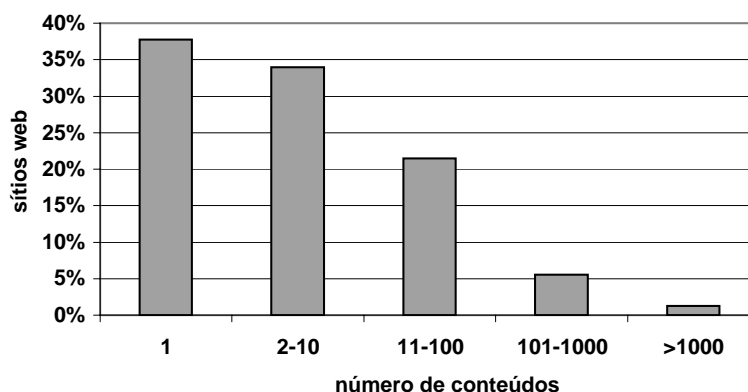


Figura 2: Distribuição de conteúdos por sítio Web.

de heurísticas que permitam realizar a sua recolha automática. A heurística adoptada na definição de Web portuguesa proposta atingiu o melhor resultado, identificando 82% dos sítios Web. Foi identificado se estes sítios Web recebiam pelo menos uma ligação directa de um sítio Web sob .pt através dos motores de busca Google e AllTheWeb.

Os resultados obtidos mostram que a definição de Web portuguesa adoptada é a mais adequada apesar de em alguns casos poder vir a incluir conteúdos não pertencentes à Web portuguesa.

## 2.2 Caracterização de sítios Web

Um sítio Web agrupa informação interligada. As características dos sítios Web influenciam o processo de recolha e armazenamento da informação num ADW. Foi considerado que cada nome de domínio completo (*Fully Qualified Domain Name*) identificava um sítio Web distinto. Esta secção apresenta as principais características dos sítios Web portugueses.

A Figura 2 apresenta a distribuição do número de conteúdos alojados por sítios Web. Cada um aloja em média 70 conteúdos e 93% aloja menos de 101 conteúdos. Porém, verificou-se que 38% dos sítios Web continham uma única página e a maioria delas informava que o sítio web estava em construção ou tinha sido mudado para uma nova localização.

Um sítio Web foi considerado que seria português se alojasse pelo menos um conteúdo pertencente à Web portuguesa segundo a definição adoptada. Foram identificados 46 457 sítios web portugueses sob diversos domínios, como se pode ver na Figura 3. A maioria deles estão alojados sob os domínios .pt (84,2%) e .com (12,5%). Embora seja comum que o nome de um sítio Web comece pelo prefixo "WWW", 40% deles não apresentavam esta característica.

A Figura 4 apresenta a distribuição de software de servidor Web obtida a partir da

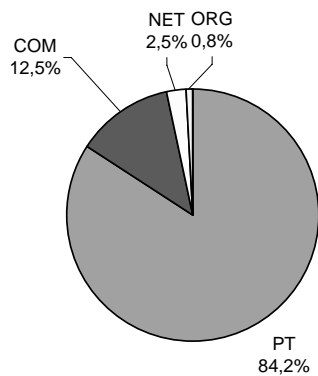


Figura 3: Distribuição de sítios Web por domínio de topo.

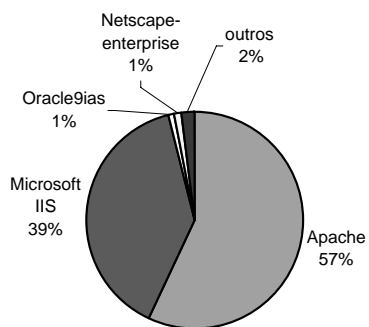


Figura 4: Distribuição de software de servidor pelos sítios Web.

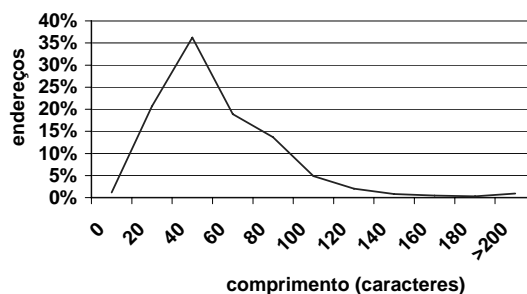


Figura 5: Distribuição do comprimento dos endereços.

análise das respostas contidas no campo *Server* do cabeçalho HTTP. Foram identificados 172 tipos de software de servidor Web distintos mas o Apache HTTP Server (57%) e o Microsoft IIS (39%) são dominantes. Os resultados obtidos à escala global são semelhantes para o Apache (62,5%) mas menos representativos para o IIS (27,4%) [Net04]. Por outro lado, a distribuição obtida para a Web portuguesa contrasta com os resultados obtidos para a Web africana, onde existe um domínio do IIS (56,1%) sobre o Apache (37,9%) [BCSV02]. As implementações do software de servidor Web deveriam seguir normas estabelecidas mas isto por vezes não se verifica devido, por exemplo, a razões comerciais. Assim sendo, um ADW necessita de afinações específicas de acordo com o software usado nos servidores para assegurar o seu bom funcionamento durante a fase de recolha de informação da Web. Os resultados apresentados contribuem para decidir quais as afinações que serão necessárias para processar correctamente a maioria dos sítios Web portugueses.

### 2.3 Caracterização de conteúdos

Um ADW necessita de estruturas de dados eficientes que permitam fazer a correspondência entre os conteúdos armazenados e os endereços na Web de onde foram extraídos. Por exemplo, para garantir que uma estrutura de dados é acedida rapidamente é necessário garantir que um computador tem memória suficiente para alojá-la. Um modelo que permita estimar o volume de dados ocupado pelos endereços contribui para estimar a memória necessária. A Figura 5 apresenta a distribuição do comprimento dos endereços dos conteúdos. Os endereços que apontavam para conteúdos acessíveis apresentam um comprimento entre os 5 e 1 368 caracteres. Em média, um endereço tem 62 caracteres de comprimento e a maioria deles tem um comprimento entre os 20 e 100 caracteres. Uma análise dos endereços revelou que 47,2% continham parâmetros, o que demonstra a crescente popularidade dos conteúdos gerados dinamicamente.

Por definição, o campo *Last-Modified* dos cabeçalhos HTTP fornece a data da última modificação realizada a um conteúdo referido por um determinado endereço [FGM<sup>+</sup>99]. Esta informação é importante permite que um ADW identifique se um conteúdo previamente armazenado necessita de ser recolhido novamente, sem ser necessário realizar a sua descarga. Porém, os resultados apresentados na Figura 6 mostram que para a maioria dos endereços, os servidores Web correspondentes não devolveram um valor no campo *Last-Modified* (53,5%).

O tamanho dos conteúdos da Web é fundamental para estimar o espaço de armazenamento necessário para alojar um ADW. A Figura 7 apresenta a distribuição dos tamanhos dos conteúdos e mostra que a maioria se situa entre os 4 e 64 KB. O tamanho médio de um conteúdo é de 32,4 KB, ao passo que o do texto extraído é de 2,8 KB. No total, verificou-se

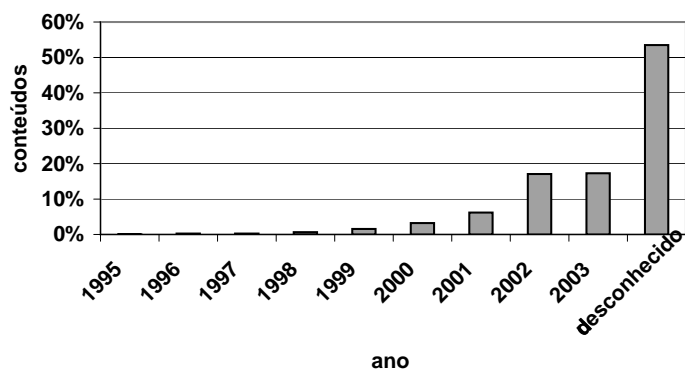


Figura 6: Distribuição de datas de última modificação (*Last modified dates*).

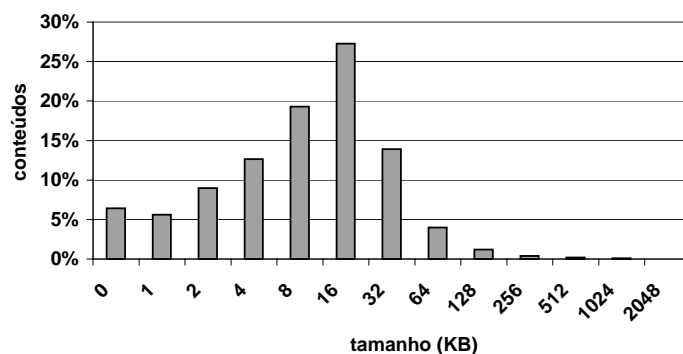


Figura 7: Distribuição dos tamanhos dos conteúdos.

Tipo de media	Tamanho médio conteúdo (KB)	Tamanho médio do texto extraído (KB)	Relação texto/conteúdo
powerpoint	1 054,9	7,0	0,7
text/rtf	475,6	1,2	0,3
application/pdf	207,4	13,6	6,6
application/rtf	121,3	4,7	3,9
application/msword	118,6	9,9	8,3
excel	50,4	21,9	43,4
application/x-shockwave-flash	43,9	0,3	0,7
text/html	20,5	2,5	12,2
text/richtext	16,3	16,2	99,2
application/x-tex	16,1	14,7	91,2
text/plain	10,5	7,8	74
text/tab-separated-values	3,9	3,8	97,5

Tabela 2: Tamanho médio dos conteúdos, texto extraído e relação entre o tamanho do texto e o conteúdo original.

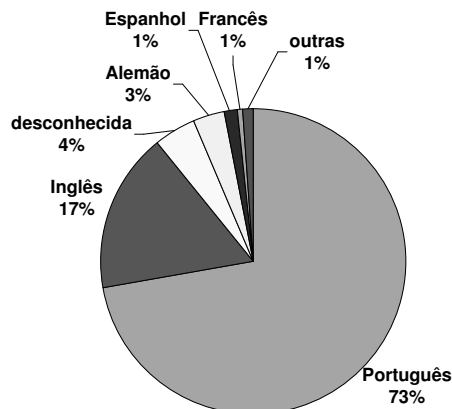


Figura 8: Distribuição de línguas sob o domínio .pt.

que o volume de dados dos textos corresponde a 11% do volume dos conteúdos originais. Assim sendo, um ADW que necessite apenas dos textos extraídos, como por exemplo um motor de busca sobre a Web, terá requisitos de armazenamento menos exigentes do que um que armazene os conteúdos originais. A Tabela 2 apresenta para cada tipo de conteúdo recolhido o seu tamanho médio, o tamanho médio do texto extraído e a relação entre eles. Os conteúdos mais pequenos tendem a ser constituídos por uma maior percentagem de texto. Os conteúdos do tipo *text/plain* são compostos apenas por 74% de texto porque existem servidores Web que devolvem o tipo *text/plain* quando não reconhecem a extensão de um determinado ficheiro. Daí que, ficheiros de tipos não completamente textuais, tais como PowerPoint Presentation (.PPS) ou Java Archives (.JAR), foram incorrectamente identificados como sendo textos planos e resultaram numa percentagem de texto extraído relativamente baixa. Durante uma recolha automática, é necessário estipular tamanhos máximos para os ficheiros a serem descarregados para garantir a robustez de um batedor contra situações anómalas, tais como, transmissões de rádio em fluxo contínuo (*streaming*) publicadas como ficheiros de comprimento infinito. Porém, os tamanhos máximos deverão ser estipulados consoante os tipos dos conteúdos. Por exemplo, ao ser identificado uma página HTML com um tamanho superior a 1 MB é provável que um batedor esteja perante uma situação anómala e deva cancelar a recolha do conteúdo, mas caso se trate de uma apresentação do tipo Powerpoint, o valor já é aceitável. Os resultados apresentados permitem definir limites de acordo com os tipos dos conteúdos, reduzindo assim a hipótese de exclusão de conteúdos por terem sido estipulados limites irrealistas face às características da Web. Por outro lado, os resultados obtidos permitem estimar os recursos necessários para implementar um ADW de acordo com os tipos de conteúdos que irá alojar.

A língua empregue nos conteúdos foi identificada através de uma ferramenta baseada num algoritmo de análise de n-gramas [MS05]. A Figura 8 apresenta a distribuição de línguas em conteúdos alojados sob o domínio .pt. As línguas dominantes são o português (73%) e o inglês (17%). As restantes línguas são usadas em 6% dos conteúdos. A ferramenta não conseguiu identificar a língua de 4% dos conteúdos porque, por exemplo, estes conteúdos continham pouco texto ou estavam escritos usando várias línguas. Tratando-se da Web portuguesa, estes resultados naturalmente contrastam com os obtidos por O'Neill à escala mundial, em que 72% das páginas estão escritas em inglês e apenas 2% em português [OLB03].

Nº de duplicados	Nº de conteúdos	% total dos conteúdos
0	2 462 490	90,0%
1	205 882	7,5%
2	33 468	1,2%
3	12 814	0,5%
4	6 086	0,2%
5	5 272	0,2%
6-10	6 453	0,2%
11-100	2 318	0,1%
101-1 000	154	0,0%
>1 000	5	0,0%
Total	2734942	100,0%

Tabela 3: Distribuição do número de duplicados.

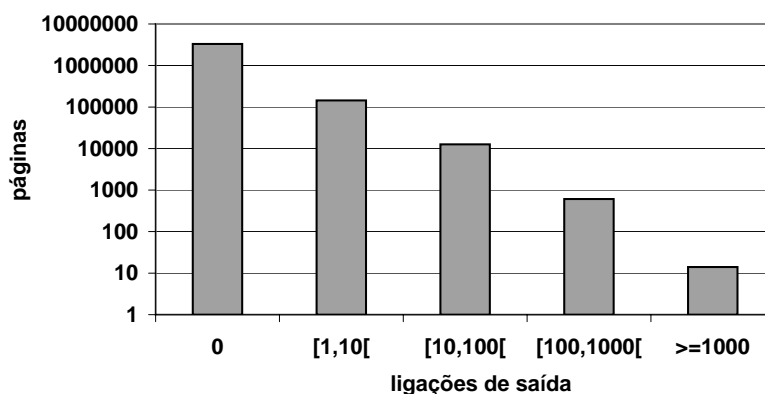


Figura 9: Distribuição do número de ligações contidas por página.

## 2.4 Duplicação e ligações

Esta secção apresenta medições da duplicação de conteúdos e estrutura de ligações na Web portuguesa. Cada conteúdo foi identificado pelo seu resumo criptográfico gerado usando o algoritmo MD5. Observou-se que 15,5% dos endereços apontavam para conteúdos duplicados, isto é, diferentes endereços apontavam para conteúdos exactamente iguais. 42% dos duplicados eram originados por replicação de um conteúdo alojado no mesmo sítio Web e 60% por replicação de um conteúdo alojado noutra sítio Web. A Tabela 3 apresenta a distribuição do número de duplicados encontrados por cada conteúdo. A grande maioria dos conteúdos são únicos (90%) e cerca de 7,5% apresentam um único duplicado. Os conteúdos com muitos duplicados são raros. Os 5 conteúdos que foram replicados mais do que 1 000 vezes, foram causados por servidores Web com problemas de funcionamento que estavam a devolver a mesma página de erro para todos os pedidos que lhes eram realizados.

A análise de ligações realizada sobre a Web portuguesa teve como principal objectivo medir a interligação de informação entre diferentes sítios Web. Assim sendo, as ligações internas a cada sítio Web não foram consideradas na obtenção dos resultados apresentados. A Figura 9 mostra que 95% das páginas não apontam para qualquer conteúdo alojado noutra sítio Web português e a média de ligações por página é de 0,23. Porém, existem também páginas que apontam para um grande número de sítios Web, como é o caso das páginas de portais ou directórios. Das ligações existentes, 44% apontavam para um sítio

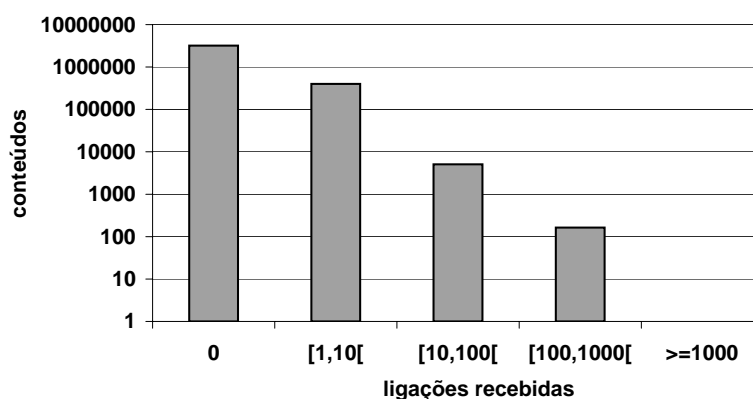


Figura 10: Distribuição do número de ligações recebidas por conteúdo.

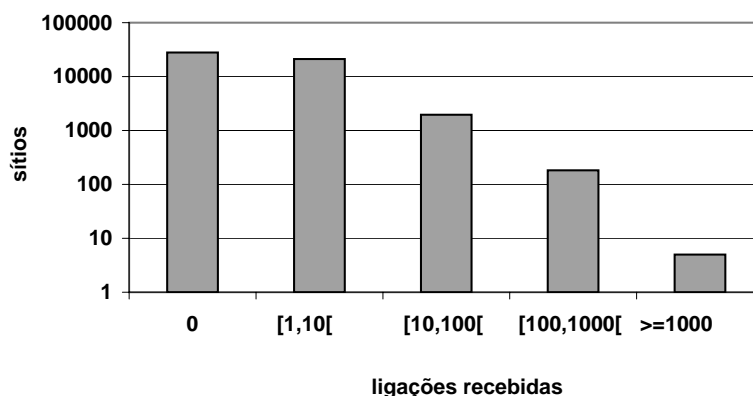


Figura 11: Distribuição do número de ligações recebidas por sítio Web.

Web português sendo que, 90% delas apontavam para a página de entrada.

O número de ligações que um conteúdo ou um sítio Web recebe é indicativo da sua importância [EMT04]. A Figura 10 mostra que 89% dos conteúdos não foram referenciados por uma ligação originada noutro sítio Web português. Este resultado é coerente com o fenómeno observado à escala mundial, em que um pequeno número de páginas recebe a grande maioria das ligações [Kle99]. Por outro lado, contrasta com o resultado obtido por Broder et al. e mostra que a conectividade do grafo da Web decresce dentro de porções mais pequenas da Web [BKM+00]. A Figura 11 apresenta a distribuição do número de ligações recebidas por sítio Web e cerca de 45% deles receberam pelo menos uma ligação de outro sítio. É de notar que 55% dos sítios Web foram recolhidos porque pertenciam ao conjunto de raízes com que se iniciou a recolha e não seriam encontrados pelo batedor através do seguimento de ligações. Este facto enfatiza a necessidade de dispor de um vasto conjunto de raízes para recolher uma Web nacional com eficácia.

### 3 Persistência de informação na Web

A Web está em permanente mutação. Contudo, existe informação que persiste durante longos períodos de tempo. A existência de modelos que permitam estimar a persistência

Identificador de recolha	Data mediana	Volume (GB)	Nº de endereços (milhões)	Nº de sítios
1	2002-11-06	44	1.2	19 721
2	2003-04-07	129	3.5	51 208
3	2003-12-20	120	3.3	66 370
4	2004-07-06	170	4.4	75 367
5	2005-04-12	259	9.4	83 925
6	2005-05-28	212	7.3	81 294
7	2005-06-18	288	10	94 393
8	2005-07-21	299	10.2	106 841

Tabela 4: Estatísticas acerca das recolhas analisadas.

da informação na Web é fundamental para o desenho de ADW. Caso estes modelos não existam, algumas decisões de desenho terão de ser tomadas em fases demasiado tardias do desenvolvimento de um ADW, acarretando custos adicionais significativos. Por exemplo, um ADW poderia ser desenhado usando um mecanismo de armazenamento, que guardasse apenas as alterações que ocorrem num conteúdo ao longo do tempo (*delta storage*), para poupar espaço em disco. Este tipo de mecanismo assume que o identificador de cada conteúdo se mantém ao longo do tempo para que seja possível acompanhar alterações. Porém, os endereços da Web que identificam os conteúdos não são permanentes e esta opção de desenho poderia revelar-se inadequada quando o ADW entrasse em produção.

Esta secção propõe modelos para estimar a persistência de informação na Web, analisando endereços e conteúdos. Estes modelos foram extraídos a partir da análise de 8 recolhas da Web portuguesa realizadas ao longo de 3 anos (2002-2005). A Tabela 4 apresenta a mediana das datas de recolha dos conteúdos, uma vez que cada recolha demorou vários dias a ser realizada, o volume total dos dados recolhidos, o número total de endereços e sítios Web visitados. Cada recolha foi iniciada usando como raízes as páginas de entrada dos sítios Web visitados na recolha antecedente. Idealmente, as recolhas apresentariam um volume crescente de informação, acompanhando o crescimento da Web. Contudo, a Recolha 6 teve de ser terminada prematuramente devido a problemas técnicos. Por forma a evitar que este facto influenciasse os resultados obtidos, esta recolha não foi comparada com a anterior.

### 3.1 Persistência de endereços

Os endereços identificam os conteúdos disponíveis na Web e são a base da sua estrutura. Assim sendo, o desenho de um ADW deverá incluir estruturas de dados e algoritmos que permitam gerir endereços eficientemente, considerando as suas características de persistência. Existem várias situações que desencadeiam o desaparecimento em bloco de endereços. Tais como, migrações de plataforma tecnológica, desactivação completa de sítios Web ou endereços que contêm identificadores de sessão que desaparecem após uma única visita de um utilizador.

Neste estudo, foi considerado que um endereço persistiu entre recolhas se este referiu conteúdos que foram descarregados com sucesso em ambas, independentemente da ocorrência de alterações no conteúdo referenciado. A Figura 12 apresenta a relação entre a persistência de endereços e a sua idade. A idade de um endereço persistente entre cada par de recolhas foi obtida a partir da diferença entre as datas medianas das recolhas medida em número de dias. Por exemplo, considerando que a data mediana dos conteúdos da Recolha 1 foi dia 6 de Novembro de 2002 e a da Recolha 3 foi dia 20 de Dezembro de 2003,

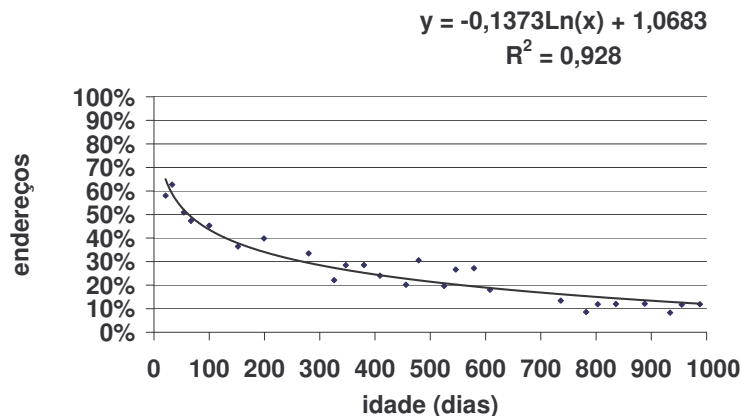


Figura 12: Tempo de vida dos endereços.

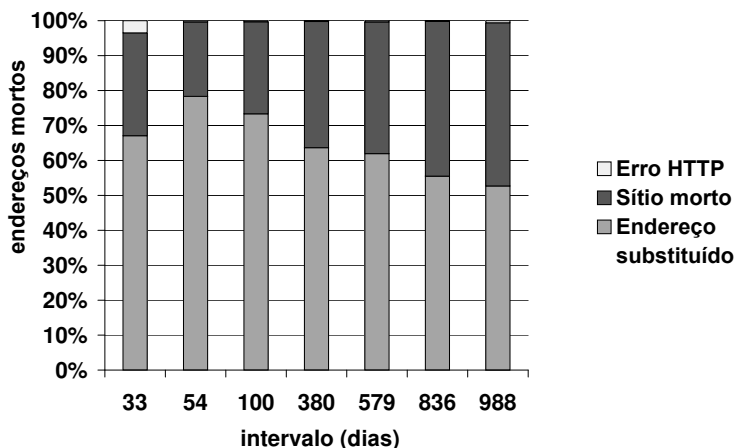


Figura 13: Razões para a morte de endereços.

24% dos endereços da Recolha 1 persistiram até à Recolha 3 e tinham uma idade aproximada de 409 dias. O gráfico da Figura 12 mostra que a maioria dos endereços têm um tempo de vida curto e que o índice de mortalidade é mais acentuado durante os primeiros meses. Contudo, existe uma minoria de cerca de 10% dos endereços que persiste por longos períodos de tempo. Os resultados obtidos mostram que o tempo de vida dos endereços pode ser modelado por uma função logarítmica que apresenta um valor de R-quadrado de 0,928 em relação aos valores da amostra. Esta função permite estimar a probabilidade de um endereço se manter válido dada a sua idade. Segundo os resultados obtidos, metade dos endereços de uma colecção de dados provenientes da Web tornam-se inválidos após 60 dias. Anteriormente, foi proposto um modelo para estimar a frequência das alterações em páginas da Web, segundo o pressuposto de que os seus endereços persistiriam ao longo do tempo [CGM03]. O modelo proposto na Figura 12 complementa esse trabalho, uma vez que permite estimar o intervalo de tempo durante o qual esse pressuposto é válido.

Um endereço foi considerado morto se não referenciava um conteúdo na última recolha realizada (Recolha 8) mas fazia-o nas anteriores. Um sítio Web foi considerado morto se não continha pelo menos um endereço vivo. A Figura 13 apresenta as principais razões identificadas para a morte de endereços. O eixo dos  $xx$  representa o tempo em dias passado entre os pares de recolhas analisadas. Considerando um intervalo de 54 dias entre recolhas, a série *Endereço substituído* indica que para 78% dos endereços mortos, os sítios Web

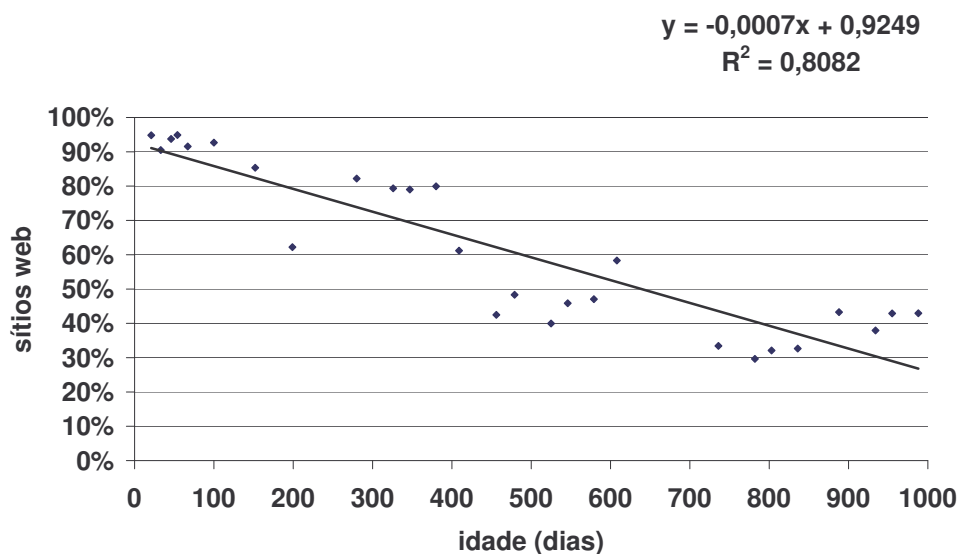


Figura 14: Persistência de sítios Web.

correspondentes mantinham-se activos mas não tinham ligações para eles, o que sugere que os endereços que morreram foram substituídos por outros novos. Por outro lado, para 21% dos endereços mortos o sítio Web também tinha desaparecido. A percentagem de endereços mortos devido ao desaparecimento completo do sítio Web aumenta ao longo do tempo. A série *Erro HTTP* contabiliza os endereços mortos, para os quais ainda existiam ligações a partir de páginas da Web ou faziam parte das raízes da recolha. O batedor ao visitar endereços referidos recebeu um erro HTTP como resposta do servidor correspondente (ex. 404 File Not Found). Este tipo de causa de morte torna-se visível apenas no intervalo de tempo mais curto (33 dias) representando uma percentagem de 3,5%. Empiricamente, as páginas Web tendem a ser corrigidas para não referenciar endereços mortos. Por isso, faz sentido que esta causa de morte de endereços não seja visível em intervalos de tempo mais alargados. É de notar que durante um processo de recolha automática da Web podem ocorrer problemas operacionais, como por exemplo quebras de rede, que podem inibir o acesso a endereços válidos, sendo estes indevidamente considerados mortos. A ocorrência destes problemas foi analisada e concluiu-se que poderão no máximo ter influenciado os resultados em 0,4%. Além disso, a maioria destes endereços foram considerados mortos, após ter passado um minuto de espera para descarga do conteúdo correspondente, o que indica que provavelmente estaria inacessível.

O tempo de vida dos sítios Web foi também estudado. Para cada recolha, foi medida a percentagem de sítios que se mantinham válidos nas recolhas seguintes. A Figura 14 mostra que 90% dos sítios Web se mantêm vivos após 100 dias, mas esta percentagem desce para cerca de 30-40% quando os sítios Web atingem uma idade superior a 700 dias. Segundo os resultados obtidos, metade dos sítios de uma amostra da Web morrem passados 556 dias. Os sítios Web são mais persistentes que os endereços e um ADW poderá tirar maior partido da reutilização de informação acerca de sítios Web do que acerca de endereços individuais. Por exemplo, considere-se um caso em que a informação de um sítio Web que estava a ser publicada recorrendo a ficheiros em formato HTML, é migrada para um sistema de gestão de conteúdos. Nestes casos é aconselhável que os endereços antigos passem a apontar para os novos alojados no sistema de gestão de conteúdos para evitar

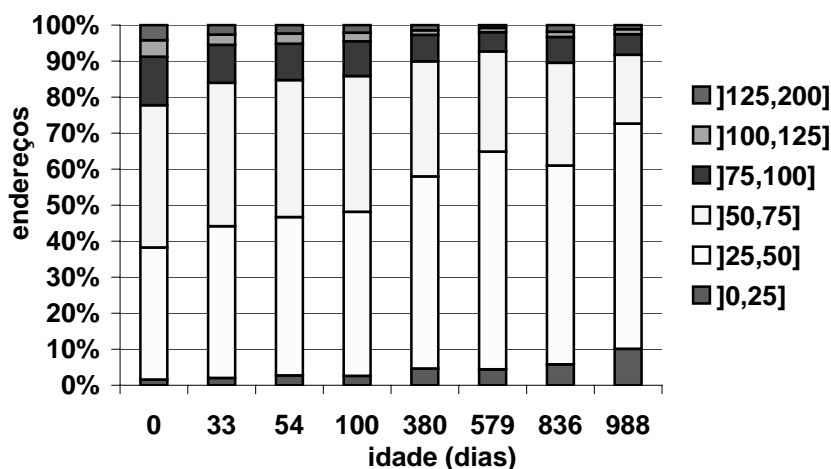


Figura 15: Distribuição do comprimento dos endereços persistentes medida em número de caracteres.

a ocorrência de ligações quebradas e conseqüente perda de visitantes. Porém, esta boa prática raramente é seguida, dando-se o desaparecimento da maioria dos endereços antigos. Os meta-dados acerca do sítio Web mantidos no ADW não necessitam de actualizações, ao passo que os endereços antigos do sítio Web foram desactivados e é necessário inserir os novos endereços e respectivos meta-dados.

### 3.1.1 Características dos endereços persistentes

Na secção anterior apresentou-se um modelo que permite estimar a persistência de endereços no geral. No entanto, existem endereços com maior probabilidade de persistirem ao longo do tempo. Nesta secção são analisadas as características dos endereços persistentes, para identificar quais as que poderão ser indicadoras de que um endereço terá maior probabilidade de persistência a longo prazo. A última recolha foi usada como base de comparação (Recolha 8). Os endereços que persistiram das recolhas anteriores até à última foram analisados para identificar características sugestivas da sua persistência. A idade dos endereços é a diferença em dias entre a data de cada recolha e a Recolha 8 que tem a idade 0.

A Figura 15 apresenta a relação entre os endereços persistentes e o seu comprimento e mostra que os endereços com comprimento inferior a 50 caracteres têm tendência para ser mais persistentes do que outros mais longos. Uma análise mais detalhada revelou que os endereços curtos referem principalmente páginas de entrada de sítios Web, o que é consistente com os resultados que mostraram que os sítios Web têm tempos de vida mais longos do que os endereços. Os endereços persistentes curtos tendem a referir endereços que não contêm parâmetros embebidos. Por outro lado, uma visita a uma amostra dos endereços mais longos revelou que, estes foram usados em sítios Web com um desenho débil que acabaram por ser remodelados ou desactivados.

Os autores de páginas da Web usam ligações para referir informação relacionada com as suas publicações. O número de ligações que um endereço recebe de sítios Web externos é indicativo da sua importância, ao passo que as ligações recebidas internamente são navegacionais. A Figura 16 descreve a distribuição de endereços persistentes que receberam

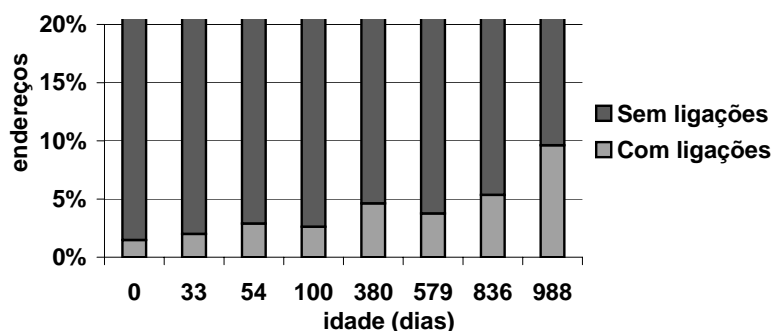


Figura 16: Distribuição de endereços persistentes com ligações a partir de outros sítios Web.

pelo menos uma ligação a partir de uma página alojada noutra sítio Web. É de salientar que o número de endereços que receberam ligações externas é de apenas 1,5% do total dos endereços da base de comparação (idade 0). Porém, nota-se que os endereços que recebem ligações tendem a persistir ao longo do tempo, aumentando de 2% entre endereços com 33 dias de idade para 9,6% entre os endereços com 988 dias de idade. Enumeram-se duas explicações possíveis para este facto. A primeira é que os endereços persistentes tendem a acumular mais ligações, visto o seu tempo de vida ser mais alongado. A segunda é que como o número de ligações recebidas aumenta a visibilidade dos endereços nos motores de busca, estes ganham valor comercial e os seus detentores têm maior cuidado em preservá-los.

### 3.2 Tempo de vida dos conteúdos

Um ADW necessita de periodicamente recolher nova informação a partir da Web. Um modelo para estimar o tempo de vida dos conteúdos permite otimizar o agendamento de novas recolhas para refrescamento da informação. Fetterly et al. observaram um conjunto de páginas durante 11 semanas e concluíram que a idade de um conteúdo é um bom estimador da sua persistência futura [FMNW03]. Esta secção apresenta um modelo extraído a partir de uma análise num intervalo de tempo mais alongado.

O estabelecimento de uma fronteira que permita decidir se uma página sofreu alterações suficientes que justifiquem que o seu conteúdo seja actualizado num ADW é muito subjectivo. Assim sendo, neste estudo foi assumido que qualquer alteração a uma página originava um novo conteúdo. Cada conteúdo foi identificado pelo seu resumo criptográfico e foram feitas comparações entre recolhas para estimar a persistência de conteúdos ao longo do tempo, independentemente dos endereços que os referenciam. Para cada recolha, foi calculada a percentagem de conteúdos que persistiam nas recolhas seguintes. A Figura 17 sumariza as percentagens de persistência encontradas. Apenas 34% dos conteúdos persistem após 33 dias mas 13% sobrevivem aproximadamente 1 ano. O tempo de vida dos conteúdos na Web pode ser modelado por uma função logarítmica com um valor de R-quadrado de 0,8661. Esta função permite estimar a percentagem de conteúdos que persistem numa colecção de dados a partir da sua idade. Os resultados obtidos mostram que passados 2 dias de ser realizada uma recolha da Web, 50% dos conteúdos já foram alterados ou deixaram de estar disponíveis. Os resultados obtidos mostram que embora a maioria dos conteúdos apresente um tempo de vida curto, existe uma minoria que persiste ao

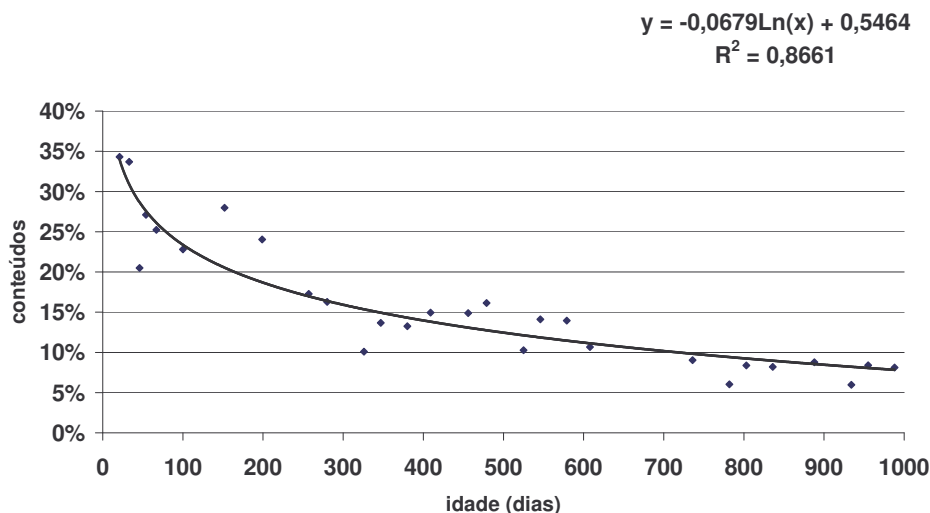


Figura 17: Tempo de vida dos conteúdos na Web.

longo do tempo.

### 3.2.1 Características de conteúdos persistentes

Esta secção apresenta as características dos conteúdos persistentes ao longo do tempo. A última recolha foi usada como base de comparação (Recolha 8). Os conteúdos que persistiram das recolhas anteriores até à última foram analisados para identificar características sugestivas da sua persistência. A idade dos endereços é a diferença medida em dias entre a data de cada recolha e a Recolha 8 que tem 0 dias de idade.

Os conteúdos que se encontram guardados em disco no formato com que serão publicados na Web denominam-se *estáticos*. Por outro lado, os conteúdos *dinâmicos* são gerados em tempo de execução quando um servidor Web recebe um pedido. Este paradigma tem ganho popularidade porque permite uma gestão eficiente da informação em bases de dados, de forma independente do formato usado para a sua publicação na Web. Assumiu-se que os endereços contendo parâmetros referiam conteúdos dinâmicos e os restantes seriam estáticos. Um ADW poderá tirar partido de aplicar diferentes políticas de refrescamento para conteúdos estáticos e dinâmicos. A Figura 18 mostra que 55% dos conteúdos da Recolha 8 realizada em Julho de 2005 eram dinâmicos (idade 0), um número superior ao testemunhado em Maio de 2004 por Castillo [Cas04], o que comprova a crescente prevalência deste paradigma de publicação na Web. A presença de conteúdos dinâmicos decresceu de 32% entre os conteúdos que persistiram 33 dias, para 9% entre os que persistiram mais do que 579 dias. Os resultados obtidos mostram que os conteúdos estáticos tendem a ser significativamente mais persistentes a longo prazo do que os gerados dinamicamente.

A data de última modificação fornecida no cabeçalho HTTP (Last-Modified date) permite detectar se um conteúdo referenciado por um endereço sofreu alterações desde uma visita prévia sem ser necessário realizar a sua descarga. Contudo, os administradores de servidores Web são encorajados a desactivar o fornecimento desta data para conteúdos que mudem frequentemente [The04]. Consequentemente, a simples presença de uma data de última modificação poderá ser um indicador da persistência do conteúdo. Efectivamente,

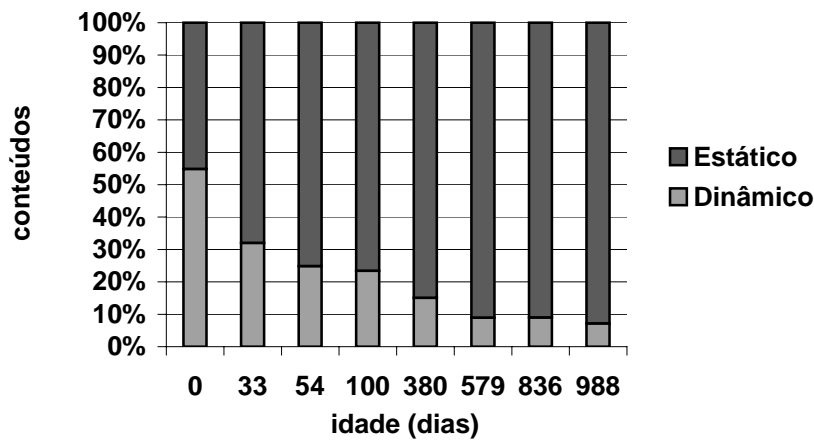


Figura 18: Conteúdos persistentes estáticos e gerados dinamicamente.

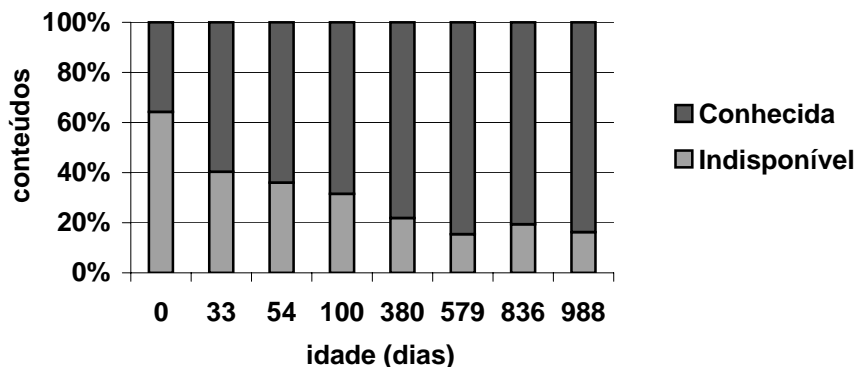


Figura 19: Conteúdos persistentes com data de última modificação.

os resultados apresentados na Figura 19, mostram que os conteúdos que têm uma data de última modificação associada são significativamente mais persistentes do que os restantes.

A Figura 20 apresenta a distribuição dos tamanhos dos conteúdos persistentes por idade. A presença de conteúdos pequenos tende a aumentar com a idade. Cerca de 27% dos conteúdos na Recolha 8 (idade 0) tinham um tamanho inferior a 10 KB, mas este número subiu para 74% entre os conteúdos que persistiram durante 988 dias. Os resultados obtidos mostram que os conteúdos pequenos tendem a ser mais persistentes do que os maiores. Esta conclusão é coerente com a observação feita por Fetterly et al. de que as páginas pequenas tendem a sofrer alterações menos frequentes do que as maiores [FMNW03].

### 3.3 Relação entre a persistência de endereços e conteúdos

Trabalhos anteriores sobre a evolução da Web focaram-se na análise da mudança de conteúdos alojados sob o mesmo endereço, assumindo que o desaparecimento do endereço implicaria o desaparecimento do conteúdo referenciado [CGM03, FMNW03]. Porém, uma mudança no domínio de um sítio Web modifica todos os seus endereços sem implicar al-

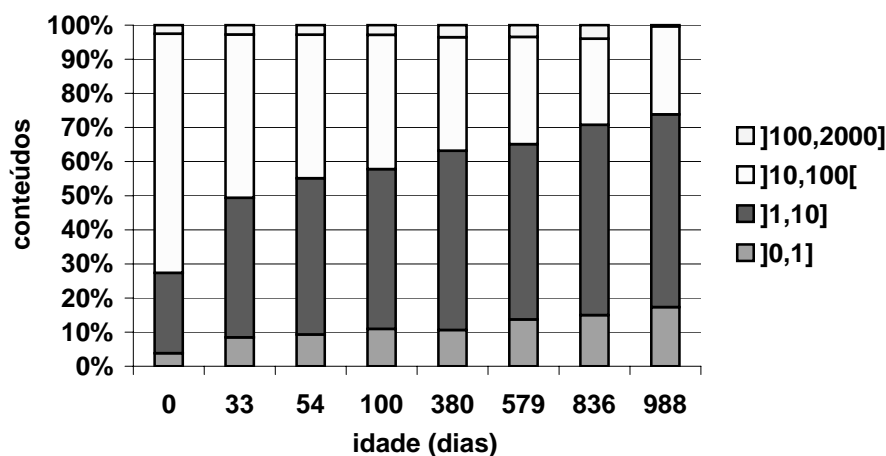


Figura 20: Distribuição do tamanho dos conteúdos persistentes (KB).

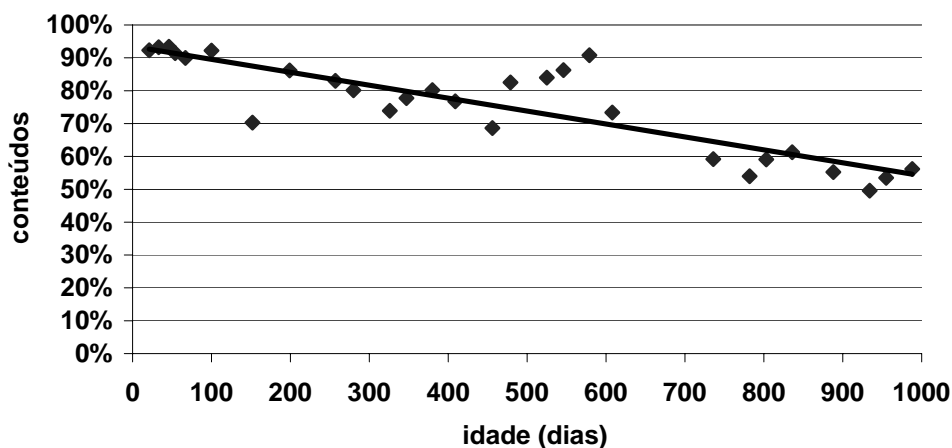


Figura 21: Conteúdos persistentes que mantiveram o mesmo endereço.

terações nos conteúdos disponibilizados. A Figura 21 mostra que ao longo do tempo o número de conteúdos persistentes que mantêm o mesmo endereço tende a decair. Apenas cerca de 60% dos conteúdos com 700 dias de idade estavam disponíveis no seu endereço original. Estes resultados mostram que o pressuposto de que a morte de um endereço implica a morte do conteúdo referenciado é inadequada em análises a longo prazo.

Por outro lado, a permanente evolução da Web poderá sugerir que os endereços que se mantêm activos vão referenciando diferentes conteúdos ao longo da sua vida. A Figura 22 apresenta a percentagem dos endereços persistentes que apontaram sempre para um conteúdo inalterado. Os resultados obtidos mostram que cerca de metade dos endereços que persistem, mantêm-se a apontar para um conteúdo inalterado durante toda a sua vida. Existe assim uma relação significativa entre a persistência de endereços e conteúdos.

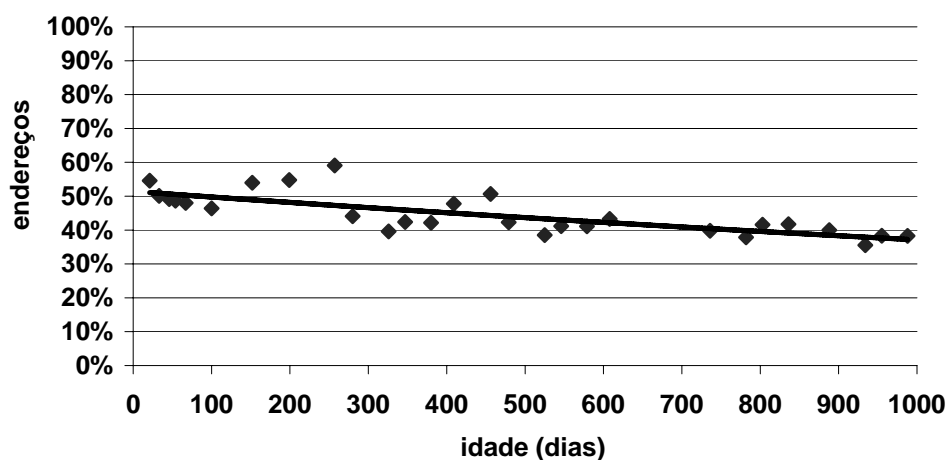


Figura 22: Endereços persistentes que mantiveram o mesmo conteúdo.

## 4 Investigação aplicada na criação do Arquivo da Web Portuguesa

Diariamente, são publicados milhões de conteúdos exclusivamente na Web como textos, fotografias ou vídeos. 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 ser 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 para sempre. Este problema é sentido igualmente por outras comunidades nacionais e pelo menos 16 países já iniciaram as suas próprias iniciativas de arquivo da Web [Nat07].

O projecto oficial de Arquivo da Web Portuguesa (AWP) visa a criação de um sistema que terá como missão recolher, armazenar e preservar a informação publicada na Web, 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. Ao passo que, o AWP arquivou 5,3 TB de informação em apenas duas recolhas exaustivas do mesmo domínio, realizadas durante o primeiro semestre de 2008. Os serviços a serem prestados pelo AWP ultrapassam o âmbito histórico-cultural da preservação de informação digital. Este projecto permitirá, por exemplo, disponibilizar recursos para investigação interessantes para diversas comunidades científicas (História, Sociologia ou Informática), contribuir para o desenvolvimento da capacidade nacional de prospecção de informação publicada na Web, acompanhar a evolução da Web portuguesa ou fornecer provas em casos judiciais que tenham como base informação publicada na Web.

A primeira fase do desenvolvimento do AWP teve início em Janeiro de 2008 e a manutenção de um sistema desta natureza é uma tarefa que a ser perpetuada posteriormente.

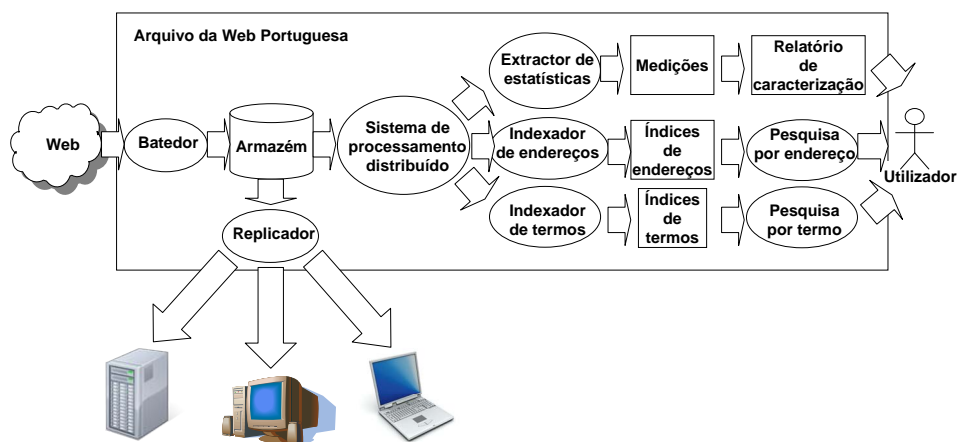


Figura 23: Arquitectura do sistema de Arquivo da Web Portuguesa.

Pretendem-se disponibilizar serviços de pesquisa eficientes sobre a informação arquivada. Publicamente, serão disponibilizados dois métodos distintos de pesquisa histórica. O primeiro, trata-se da pesquisa por endereço da Web que permitirá aos utilizadores acederem a conteúdos arquivados ao longo do tempo, a partir do fornecimento do endereço onde foram publicados. O segundo método, trata-se da pesquisa histórica por termo, que permitirá identificar páginas arquivadas 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. Para fins de investigação, estão a ser desenvolvidos sistemas para acesso e processamento eficiente de grandes quantidades de informação arquivada.

O AWP leva a cabo um esforço permanente de preservação de conteúdos Web históricos detidos por entidades externas. De 2002 a 2006, o projecto de investigação tumba! recolheu cerca de 57 milhões de conteúdos maioritariamente textuais da Web portuguesa 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 para garantia da sua preservação a longo prazo.

#### 4.1 Descrição do sistema

A Figura 23 apresenta uma descrição da arquitectura do sistema que suporta o AWP. O *Batedor* percorre a Web, recolhe conteúdos e guarda-os em formato ARC no *Armazém* [BK96]. 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 em diferentes computadores espalhados pela Internet. Em caso de destruição do Armazém, a informação arquivada poderá ser recuperada a partir das cópias de segurança. O *Sistema de processamento distribuído* tem a capacidade de executar tarefas de computação sobre grandes quantidades de informação. A principal vantagem deste sistema é permitir o desenvolvimento de aplicações de processamento de grandes quantidades de dados sem preocupações ao nível da sua gestão e execução distribuída. O AWP inicialmente inclui três aplicações de processamento dos dados arquivados: o *Extractor de estatísticas* que gera medições da Web portuguesa, o *Indexador de endereços* que cria índices para suporte da pesquisa histórica por endereço e o *Indexador de termos* que cria índices para suporte da pesquisa histórica por termo.

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. Para enfrentar este problema foi criado em 2004 o projecto Archive-access, liderado pelo Internet Archive, que reúne e disponibiliza gratuitamente ferramentas para o arquivo da Web [Int08]. Este projecto permitiu uma grande evolução nesta área pois as ferramentas passaram a ser desenvolvidas em colaboração internacional.

## 4.2 Impacto da investigação no desenho do AWP

O sistema do AWP baseia-se principalmente em tecnologia disponibilizada pelo projecto Archive-access. O ADW desenvolvido pelo autor para a realização da sua investigação (Webhouse), não foi utilizado na implementação do sistema do AWP porque a utilização do software disponibilizado pelo projecto Archive-access, oferece potencialmente uma maior garantia de manutenção do sistema a longo prazo. Contudo, a experiência ganha pelo autor no desenvolvimento do Webhouse e as semelhanças arquitecturais entre as tecnologias, permitiu que os sistemas disponibilizados pelo Archive-access fossem rapidamente adaptados às necessidades do AWP e postos em funcionamento.

A investigação realizada pelo autor mostrou que 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.

O Batedor do AWP ciclicamente recolhe conteúdos referidos por endereços da Web e extrai ligações para novos endereços. Este componente interage directamente com a Web e o seu bom desempenho depende fortemente de ser configurado de acordo com as características peculiares da porção da Web que irá recolher. Por exemplo, o Batedor mantém uma lista de endereços conhecidos para que não repita a inserção e visita aos mesmos endereços. Cada vez que o Batedor extrai um novo endereço a partir de uma ligação contida numa página da Web, verifica se o endereço já existe na lista de endereços. Porém, esta comparação tornar-se-ia ineficiente se fosse feita contra todos os endereços da lista, que facilmente chega a conter milhões de endereços. O Batedor particiona a lista para que o processo de comparação seja feito de forma mais eficiente, contra um número relativamente reduzido de endereços. A função de partição deverá ser escolhida de acordo com as características da Web para que gere um número adequado de partições com dimensão equilibrada que permita o funcionamento óptimo do Batedor. Os resultados obtidos mostraram que particionar a lista dos endereços do Batedor por sítio Web é adequado às características da Web portuguesa.

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 [Cas04]. 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 realizados pelo autor acerca das características da Web portuguesa por forma a evitar situações patológicas para o seu funcionamento, garantir uma boa cobertura e não prejudicar o funcionamento dos servidores visitados. Por exemplo, foram estipulados valores limite para o tamanho máximo dos conteúdos, para evitar a descarga de conteúdos de tamanho infinito e para o número de conteúdos recolhidos por sítio Web, para evitar que o Batedor ficasse preso tentando recolher sítios Web infinitos (*spider traps*).

A Figura 24 apresenta a evolução da recolha e mostra que ao fim do sétimo dia, 99% dos conteúdos já tinham sido descarregados. Os restantes 6 dias foram ocupados a terminar

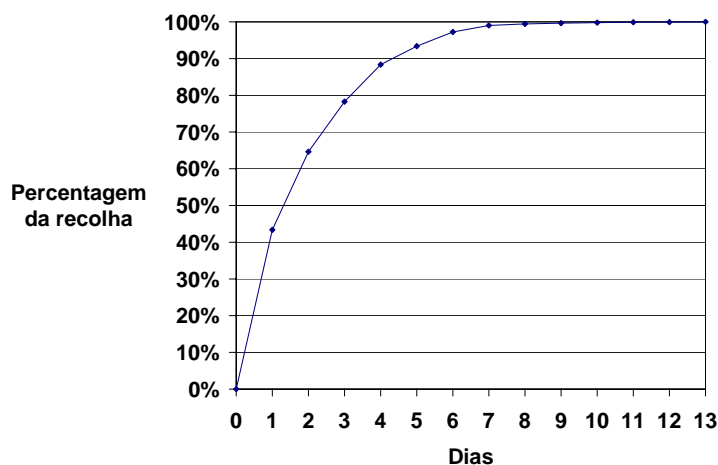


Figura 24: Evolução da recolha da Web portuguesa.

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

Tabela 5: Recursos visitados e informação recolhida da Web portuguesa.

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. O desempenho obtido superou as expectativas e 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, o que demonstra que os valores limite configurados com base nos resultados obtidos de caracterização da Web são adequados.

Durante o varrimento da Web portuguesa realizado pelo Batedor foram visitados no total de 72 milhões de endereços alojados em 455 mil sítios Web (Tabela 5). Foram recolhidos 56 milhões de conteúdos (2,8 TB), que foram armazenados em 2 TB de disco no formato ARC comprimido. Nesta primeira fase do desenvolvimento do AWP foram recolhidos apenas conteúdos alojados sob o domínio .pt. No entanto, como foi demonstrado na secção 2, esta definição exclui grande parte dos conteúdos de interesse para a Web portuguesa que estão alojados sob outros domínios. Assim sendo, está planeado estender a definição de Web portuguesa, através do critério proposto baseado em identificação de língua e estrutura de ligações que provou ser adequado ao processo de recolha automática.

O AWP periodicamente recolhe nova informação da Web portuguesa usando raízes obtidas da recolha anterior. Teoricamente, se fossem usados como raízes todos os endereços visitados anteriormente, o Batedor pouparia recursos na extracção e inserção de endereços para recolha. Porém, dado o curto tempo de vida dos endereços, esta prática resultaria na visita fútil a inúmeros endereços que já não se encontram activos. Uma vez que os endereços dos sítios Web são mais persistentes do que os dos conteúdos, optou-se por usar apenas os endereços das páginas de entrada dos sítios Web como raízes.

Os modelos de persistência de informação e tamanhos dos conteúdos por tipo, contribuíram para definir a periodicidade das recolhas e estimar o espaço de armazenamento

necessário. Porém, como existem conteúdos que se mantêm inalterados ao longo do tempo, sendo recolhidos e armazenados repetidamente. O arquivo da Web islandesa desenvolveu uma ferramenta para o Batedor que faz a detecção de duplicados durante o processo de recolha. Se o Batedor detectar que um conteúdo foi recolhido anteriormente não o volta a guardar [Sig06]. Esta ferramenta aparentava ter uma grande potencialidade de aplicação no AWP. No entanto, com base no modelo obtido para a Web portuguesa, verificou-se que o seu desempenho seria limitado. A ferramenta apresenta várias opções de funcionamento. Uma delas é usar o campo do cabeçalho HTTP Last-modified para decidir se um conteúdo sofreu uma alteração desde a última recolha. Os resultados apresentados na secção 2.3 mostram que esta aproximação seria pouco eficiente na Web portuguesa porque mais de metade dos conteúdos não têm esta informação associada. Além disso, a ferramenta não permite identificar duplicados dentro de uma recolha. Estes representam 15,5% dos conteúdos de cada recolha da Web portuguesa. Por outro lado, apenas os conteúdos persistentes que mantivessem o mesmo endereço seriam identificados como duplicados, o que como demonstra a Figura 21, não se verifica ao longo do tempo. Face a estes factos, foi decidido que a ferramenta de eliminação de duplicados necessitaria de ser melhorada antes de ser adoptada no AWP. Alternativamente, foi proposto o desenvolvimento de uma aplicação que procurasse conteúdos duplicados após terem sido arquivados. Esta aplicação teria de analisar cada conteúdo e procurar entre toda a informação arquivada por duplicados. Este processo poderá ser agilizado se forem tidas em consideração as características dominantes dos conteúdos persistentes. Por exemplo, em vez da análise para a identificação de duplicados ser executada sobre toda a informação arquivada, poderia focar-se sobre os conteúdos estáticos, com data de última modificação e tamanho pequeno, uma vez que estas são as características dominantes dos conteúdos persistentes ao longo do tempo.

Um dos principais desafios com que os arquivos da Web se debatem é como disponibilizar mecanismos de acesso eficientes sobre a informação arquivada. Actualmente, o Internet Archive disponibiliza apenas pesquisa por endereço, o que dada a fraca persistência dos endereços é um método de acesso muito limitativo. Os utilizadores dos arquivos da Web anseiam por serviços de pesquisa por termo, com uma funcionalidade análoga à dos motores de busca sobre a Web, que lhes permitam encontrar os resultados mais relevantes entre os milhões de textos que contêm os termos pesquisados. Nos motores de busca sobre a Web são usados critérios de relevância para a ordenação dos resultados das pesquisas [PBMW99]. Porém, os algoritmos dos motores de busca foram desenvolvidos para pesquisar informação sobre uma única recolha da Web e não sobre informação histórica. O estudo de critérios de relevância para a ordenação de pesquisas de âmbito histórico é um tópico de investigação recente. Para obter resultados óptimos é importante que os algoritmos sejam adaptados às características da Web portuguesa. Por exemplo, é necessária especial atenção na aplicação de algoritmos de ordenação baseados na análise de ligações porque a Web portuguesa é menos conexas do que a global e estes algoritmos poderão não atingir os resultados esperados sobre esta porção da Web.

## 5 Conclusões

Os resultados obtidos contribuíram para encontrar respostas às questões que originaram o trabalho de investigação apresentado. A caracterização de sítios, conteúdos e estrutura de ligações de uma porção da Web é crucial para desenhar um ADW que a processe. As características derivadas de análises da Web global não são representativas de porções mais pequenas, como por exemplo, as Webs nacionais. Contudo, as Webs nacionais são

de interesse para grandes comunidades de utilizadores. Existem características da Web que só podem ser modeladas a partir de diversas amostras recolhidas ao longo do tempo, como por exemplo, a persistência de informação. Estas características devem ser incluídas num modelo da Web porque descrevem tendências de evolução, que influenciam o desenho e gestão de ADW que guardem colecções Web construídas incrementalmente.

As fronteiras de uma porção da Web deverão ser delimitadas através de um conjunto de critérios de selecção automática. A porção da Web deverá conter informação que satisfaça as necessidades das aplicações que a irão processar. Os critérios de selecção deverão ser facilmente concretizáveis como políticas de recolha. O recurso a algoritmos de classificação de conteúdos e a restrição das recolhas a sítios Web alojados em determinados domínios são opções que se revelaram adequadas.

A metodologia usada para recolher amostras da Web influencia os modelos obtidos. A recolha automática de dados da Web é um método de amostragem adequado ao contexto dos ADW, uma vez que emula o processo de extracção de dados normalmente usado nestes sistemas. Contudo, a configuração e tecnologia usadas nos sistemas de recolha de dados da Web e a existência de situações prejudiciais ao processamento automático de dados, poderão influenciar os modelos obtidos.

Durante esta investigação foram recolhidas amostras da Web portuguesa durante três anos, a fim de modelar a persistência dos seus endereços e conteúdos. Verificou-se que estas duas métricas podem ser modeladas através de distribuições logarítmicas. A maioria dos endereços têm tempos de vida curtos e a taxa de mortalidade é mais elevada nos primeiros meses. Existe porém uma pequena percentagem de endereços que persiste durante vários anos. Os modelos obtidos para a persistência de conteúdos sugerem que passados 2 dias, metade dos conteúdos de uma colecção de dados provenientes da Web, sofrem alterações ou desaparecem. Por outro lado, cerca de metade dos endereços persistentes referenciam um conteúdo que permanece inalterado durante a sua vida.

Os modelos de característica da Web influenciam o desenho de ADW e permitem fazer suposições realistas acerca dos dados a processar durante as fases iniciais dos projectos. A duplicação de conteúdos é frequente na Web. No entanto, é difícil evitar a recolha de duplicados porque são frequentemente referenciados por endereços distintos e aparentemente não relacionados. Os mecanismos adoptados para suportar a eliminação de duplicados deverão ter em consideração a precariedade dos endereços. A recolha de informação é uma tarefa particularmente sensível. O componente de software responsável pela recolha interage directamente com a Web, tendo de estar preparado para enfrentar situações imprevistas prejudiciais ao seu bom funcionamento. Os modelos da Web contribuem para o desenho de sistemas de recolha automática da Web robustos a estas situações. Um ADW deverá apresentar uma arquitectura distribuída que permita tratar de grandes quantidades de dados extraídos da Web. As características da Web influenciam a definição de estratégias de particionamento para distribuir carga entre os processos que compõem um ADW.

O projecto de Arquivo da Web Portuguesa (AWP) visa disponibilizar serviços que o tornem numa importante ferramenta no dia-a-dia dos cidadãos. O trabalho de modelação apresentado contribuiu para o desenho eficiente do sistema que suporta o AWP. Os estudos futuros de modelação 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.

## Referências

- [BCSV02] Paolo Boldi, Bruno Codenotti, Massimo Santini, and Sebastiano Vigna. Structural properties of the African web. In *Proceedings of the 11th International World Wide Web Conference*, Honolulu, Hawaii, May 2002.
- [BK96] Mike Burner and Brewster Kahle. WWW Archive File Format Specification. <http://pages.alexandria.com/company/arcformat.html>, September 1996.
- [BKM<sup>+</sup>00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. In *Proceedings of the 9th International World Wide Web Conference on Computer networks*, pages 309–320. North-Holland Publishing Co., 2000.
- [BYCE07] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis Efthimiadis. Characterization of national web domains. *ACM Transactions on Internet Technology*, 7(2), 2007.
- [BYCL05] Ricardo Baeza-Yates, Carlos Castillo, and Vicente López. Characteristics of the web of Spain. *Cybermetrics - International Journal of Scientometrics, Informetrics and Bibliometrics*, 9(1), 2005.
- [Cas04] Carlos Castillo. *Effective Web Crawling*. PhD thesis, University of Chile, November 2004.
- [CGM03] Junghoo Cho and Hector Garcia-Molina. Estimating frequency of change. *ACM Transactions Internet Technology*, 3(3):256–290, 2003.
- [Cos04] Miguel Costa. Sidra: a flexible web search system. Master’s thesis, Department of Informatics, University of Lisbon, December 2004. DI/FCUL TR-04-17.
- [DVGD96] C. Davis, P. Vixie, T. Goodwin, and I. Dickinson. *A Means for Expressing Location Information in the Domain Name System*, January 1996.
- [EMT04] Nadav Eiron, Kevin S. McCurley, and John A. Tomlin. Ranking the web frontier. In *Proceedings of the 13th international conference on World Wide Web*, pages 309–318, New York, NY, USA, 2004. ACM Press.
- [FGM<sup>+</sup>99] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee. *Hypertext Transfer Protocol – HTTP/1.1*, June 1999.
- [FLG00] Gary Flake, Steve Lawrence, and C. Lee Giles. Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA, August 2000.
- [FMNW03] Dennis Fetterly, Mark Manasse, Marc Najork, and Janet Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th International Conference on World Wide Web*, pages 669–678, New York, NY, USA, 2003. ACM Press.
- [GKR98] David Gibson, Jon M. Kleinberg, and Prabhakar Raghavan. Inferring web communities from link topology. In *Proceedings of the 9th ACM Conference*

- on Hypertext and Hypermedia*, pages 225–234, Pittsburgh, Pennsylvania, June 1998.
- [Hex03] Hexa Software Development Center. Geo targeting IP address to country city region ISP latitude longitude database for Internet developers - ip2location. <http://www.ip2location.com/>, April 2003.
- [HSF85] K. Harrenstien, M. K. Stahl, and E. J. Feinler. *NICNAME/WHOIS*, 1985.
- [Int08] Internet Archive. Nutchwax - Home Page. <http://archive-access.sourceforge.net/>, March 2008.
- [Jul05] Josefa Jul. Calimaco, um repositório de documentos biológicos. Technical report, Department of Informatics, University of Lisbon, Lisbon, Portugal, September 2005. in Portuguese.
- [Kle99] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.
- [Mar03] Marktest. Netpanel. <http://netpanel.marktest.pt/>, April–May 2003.
- [Mar04] Bruno Martins. Inter-document similarity in web searches. Master’s thesis, Department of Informatics, University of Lisbon, October 2004.
- [Max03] Maxmind LLC. Maxmind: How to locate your internet visitors geotargeting IP address to country state city ISP organization latitude longitude. <http://www.maxmind.com/>, April 2003.
- [MS04] Bruno Martins and Mário J. Silva. A statistical study of the WPT 03 corpus. In *Proceedings of EsTAL - España for Natural Language Processing*, Alicante, Spain, October 2004.
- [MS05] Bruno Martins and Mário J. Silva. Language identification in web pages. In *Proceedings of the 20th Annual ACM Symposium on Applied Computing (ACM-SAC-05)*, Santa Fe, New Mexico, March 2005. ACM Press.
- [Nat07] National Library of Australia. PADI - Web archiving. <http://www.nla.gov.au/padi/topics/92.html>, August 2007.
- [Net04] Netcraft Ltd. Netcraft: April 2003 archives. <http://news.netcraft.com/archives/2003/04/index.html>, July 2004.
- [OLB03] 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.
- [PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Database Group, November 1999.
- [Pos94] J. Postel. *Domain Name System Structure and Delegation*, 1994.
- [RAR92] RARE. RIPE ncc - network management database. <http://www.ripn.net/nic/ripe-docs/ripe-078.ps>, September 1992.

- [Sig06] Kristinn Sigurdsson. Managing duplicates across sequential crawls. In *6th International Web Archiving Workshop (IWA06)*, Alicante, Spain, September 2006.
- [Sil03] 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.
- [SMC<sup>+</sup>06] M. J. Silva, B. Martins, M. S. Chaves, N. Cardoso, and A. P. Afonso. Adding Geographic Scopes to Web Resources. *CEUS - Computers, Environment and Urban Systems, Elsevier Science*, 30(4):378–399, July 2006.
- [The04] The Apache Software Foundation. *Apache HTTP Server Version 1.3: Module mod\_include*, November 2004.
- [Zoo00] Matthew Zook. Internet metrics: using host and domain counts to map the Internet, 2000.