



Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Arquivo da Web Portuguesa

Daniel Gomes

{daniel.gomes} @ fccn.pt

Universidade Lusófona, 14 de Janeiro de 2009

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento



A era digital começou

- A Web é a maior fonte de informação construída
 - Jornais, livros, documentação técnica
 - Informação publicada exclusivamente na Web
- A informação na Web é efémera
 - Análise de 150 sítios populares após 1 ano (Ntoulas, 2004)
 - Apenas 20% dos URLs permanecem válidos
 - Só 10% dos conteúdos mantêm-se inalterados
 - Web portuguesa é semelhante (Gomes, 2006)
- Temos que começar a arquivar
 - Para que a História não se perca

- Internet Archive arquiva a Web “toda” desde 1996
- Dividir para conquistar: cada país arquiva a sua Web
 - **11 da U. E.:** Alemanha, Áustria, Dinamarca, Finlândia, França, Grécia, Lituânia, Holanda, Suécia, Reino Unido e República Checa.
 - **6 externos:** Austrália, Canadá, Estados Unidos da América, Japão, Nova Zelândia e Noruega.
- Necessários sistemas para suportar o arquivo da Web

Alguns casos de uso para um Arquivo da Web

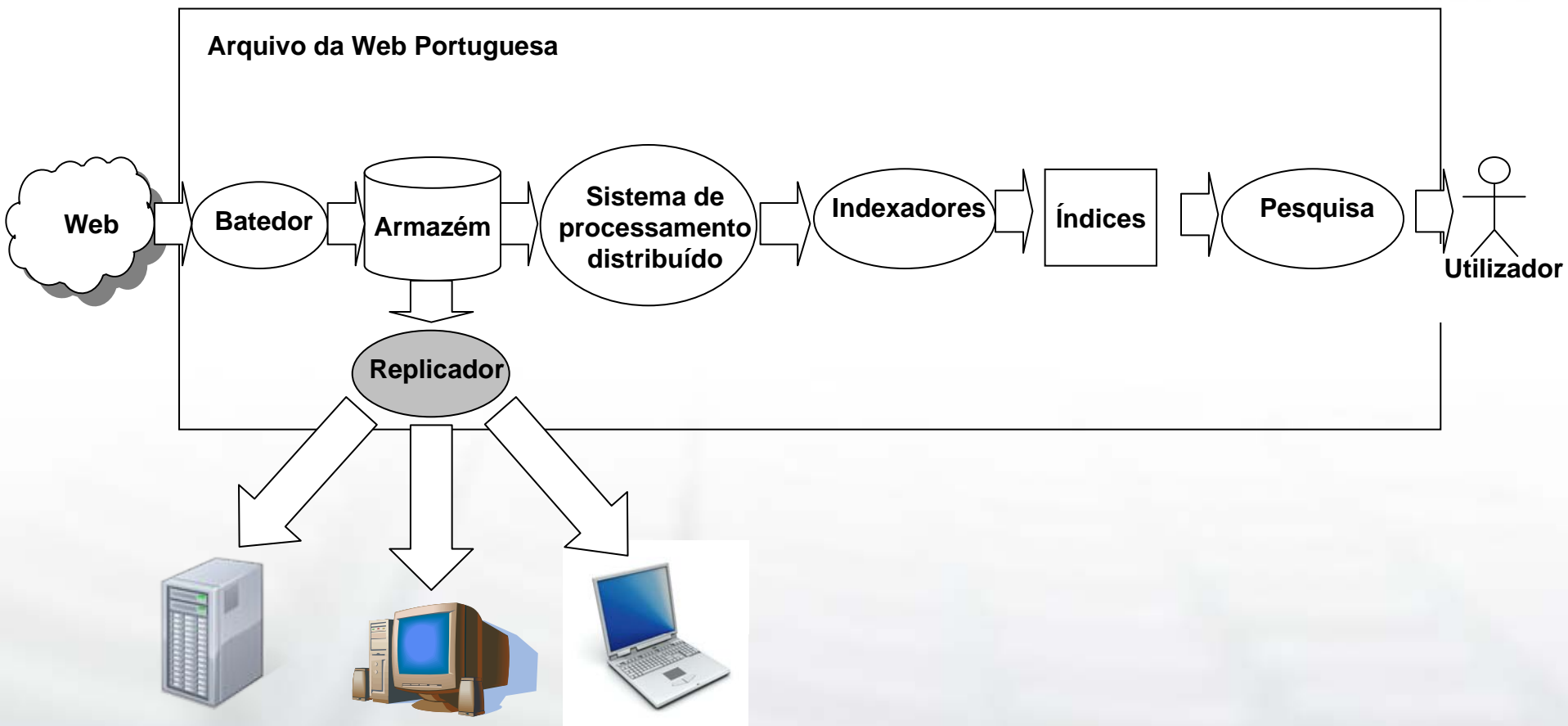
- Jornalista documenta artigo
- Webmaster recupera versão perdida de página
- Historiador analisa documentos digitais
- Utilizador da Web visita *Favorito* quebrado
- Jurista obtém provas para caso (?)
- ...

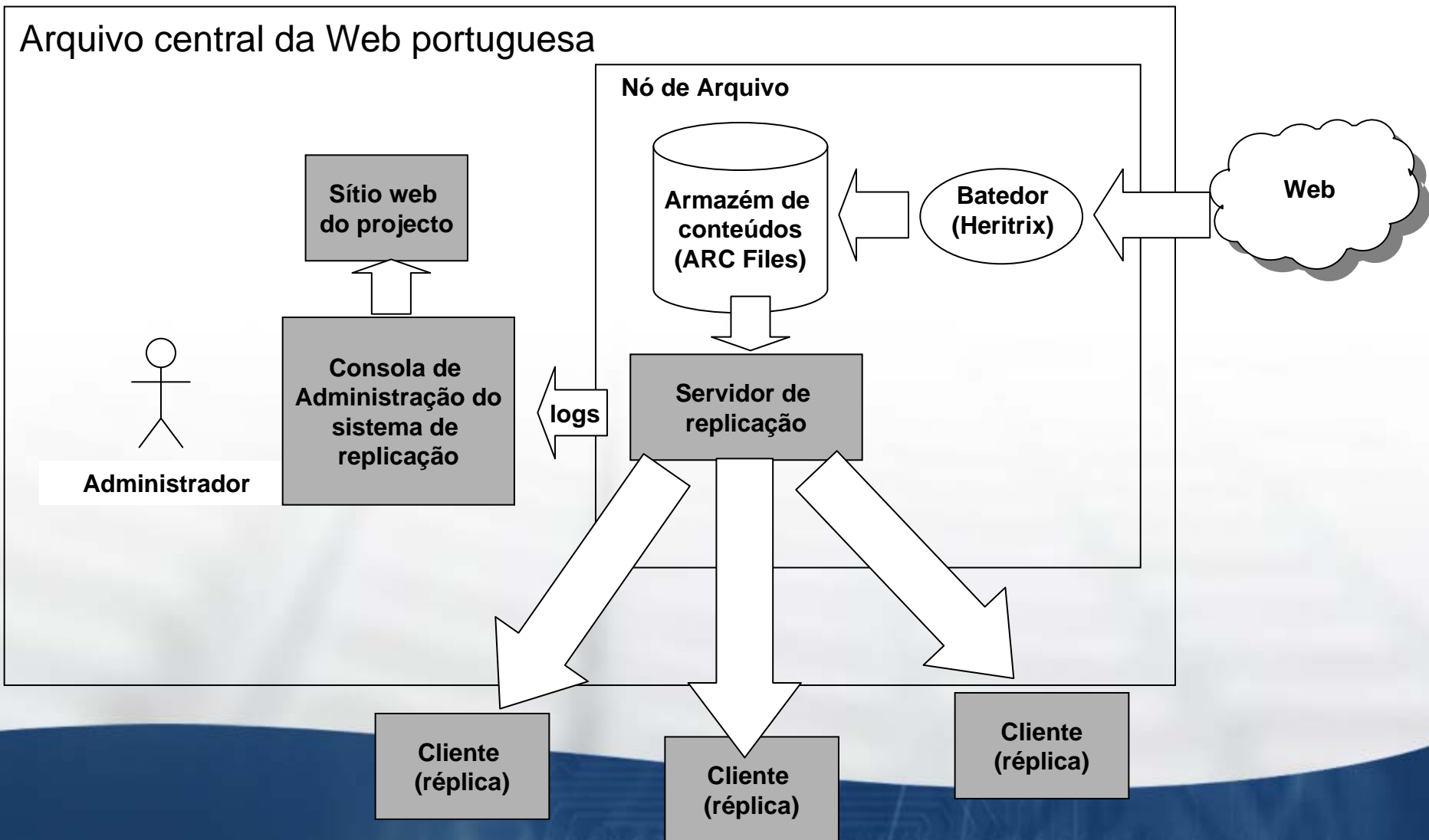
- Motivação
- Contexto do projecto de Arquivo da Web Portuguesa
- Principais componentes do sistema
 - Armazenamento e replicação
 - **Processamento paralelo**
 - Pesquisa
- Resultados de uma medição da Web portuguesa
- Conclusões



- Digital Deposit (2001)
 - FCUL/BN
 - Recolha selectiva
- Tomba (2006)
 - FCUL/FCCN
 - Recolhas do tumba! (2002-2006)
 - Textos principalmente

- Iniciar o “depósito legal” da Web portuguesa
- Serviços públicos de acesso à informação arquivada
 - Pesquisa por termo e endereço
- Prestação de serviços à comunidade científica
 - Disponibilização de colecções de dados
 - Partilha de plataforma de computação
- Formação de recursos humanos
 - Em colaboração com universidades e laboratórios
- Forte componente de investigação e desenvolvimento
 - Estudo da Web portuguesa
 - Melhoria de mecanismos de pesquisa



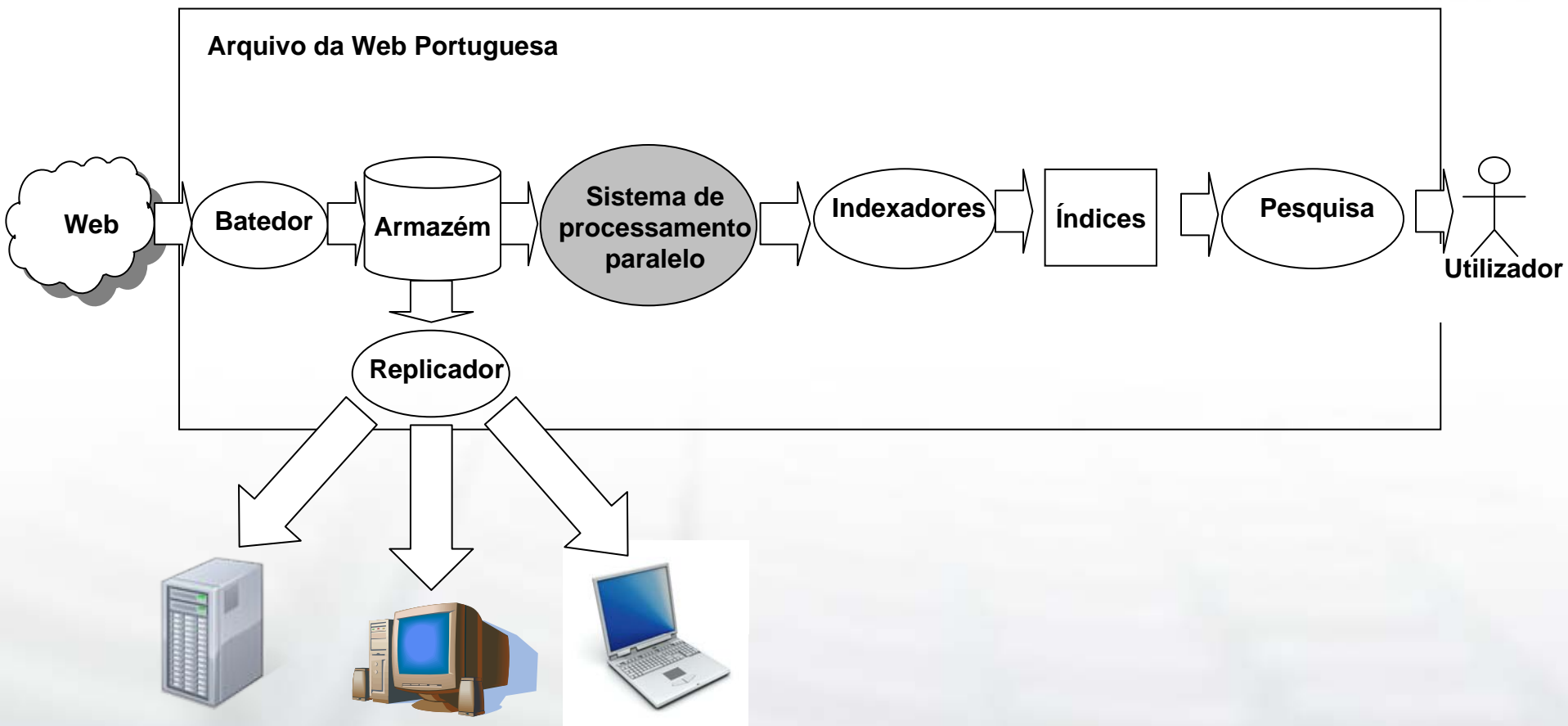


- Não é intrusivo, não carrega o computador do cliente
- Fácil de instalar
- Permite controlar espaço e largura de banda
- Confidencialidade
 - Cópias de segurança cifradas
- Integridade
 - Protecção contra clientes maliciosos que tentem adulterar as cópias para inserir conteúdos maliciosos no arquivo.
- Disponível em breve...



Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

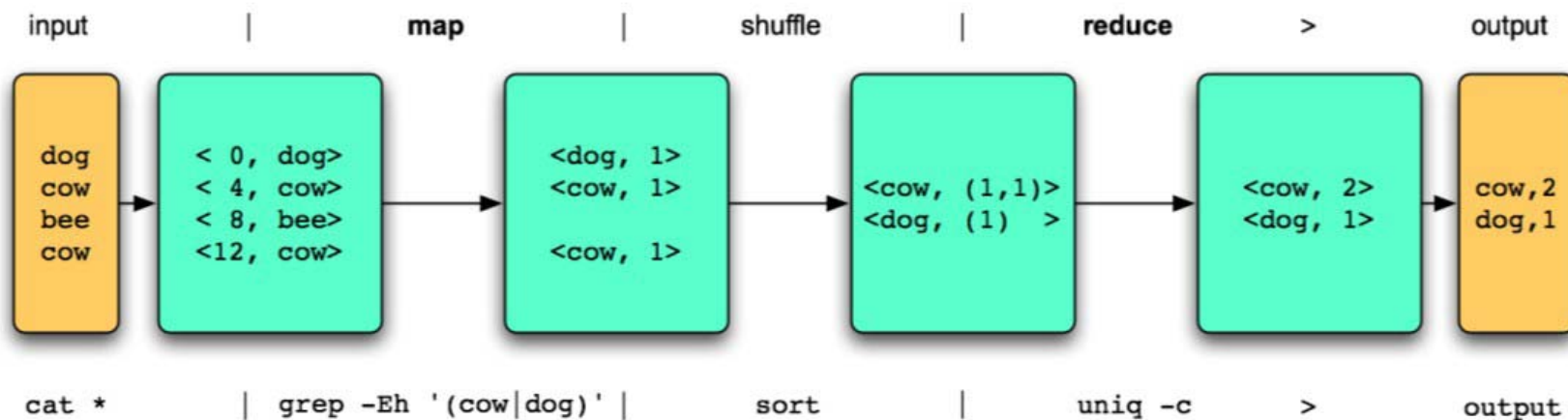
Processamento e pesquisa no AWP



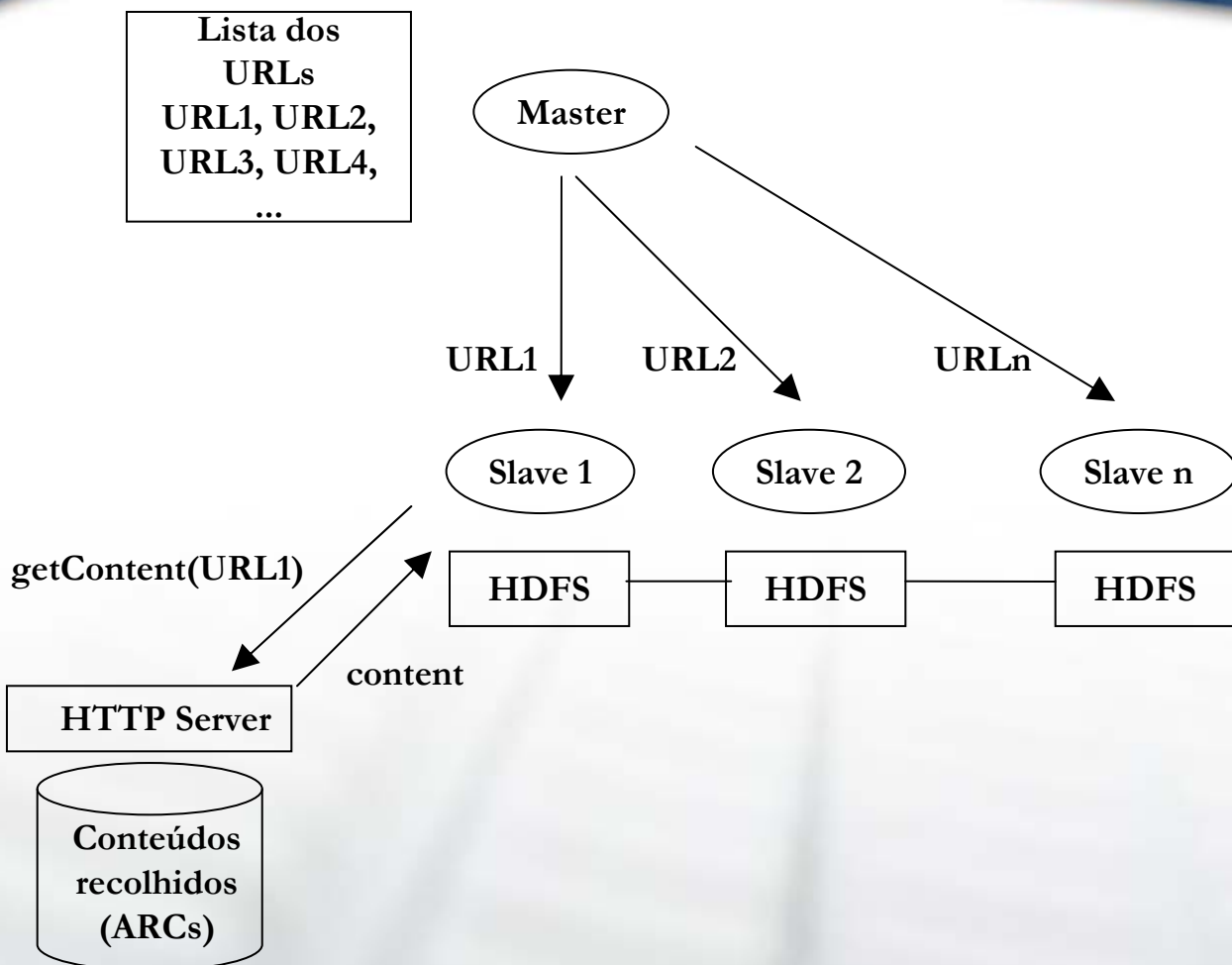
Hadoop: plataforma de computação paralela

- Hadoop.apache.org
- Permite facilmente executar rotinas em larga escala
- Implementa map-reduce do Google File System
- Apenas é necessário escrever 2 rotinas. Exemplo:
 - Map: processa cada conteúdo individualmente
 - Reduce: analisa os resultados do processamento de todos os conteúdos
- Focado no processamento de grandes quantidades de dados
 - Minimiza escritas em disco usando *pipeline* de dados entre *maps* e *reduces*
- Adoptado pela Yahoo em 10 000 servidores

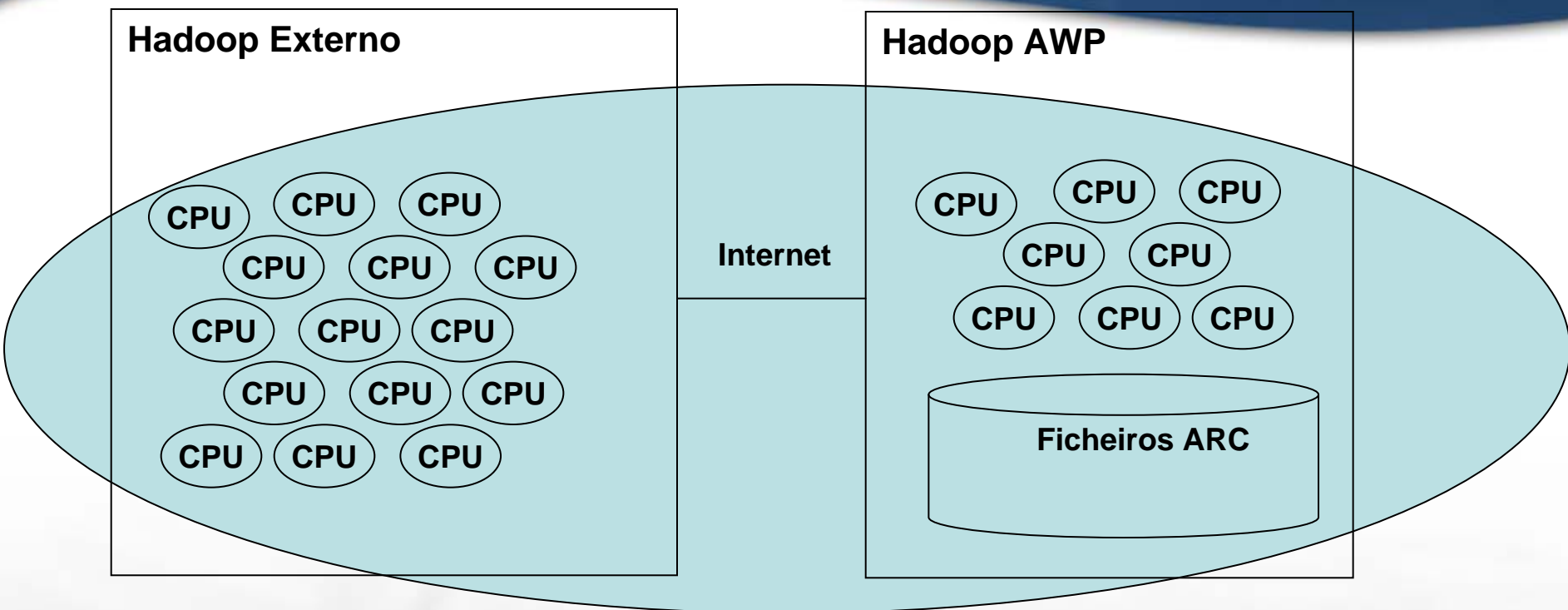
- Objectivo: contar n.º de ocorrências de “cow” e “dog” num texto e dar o resultado em formato CSV



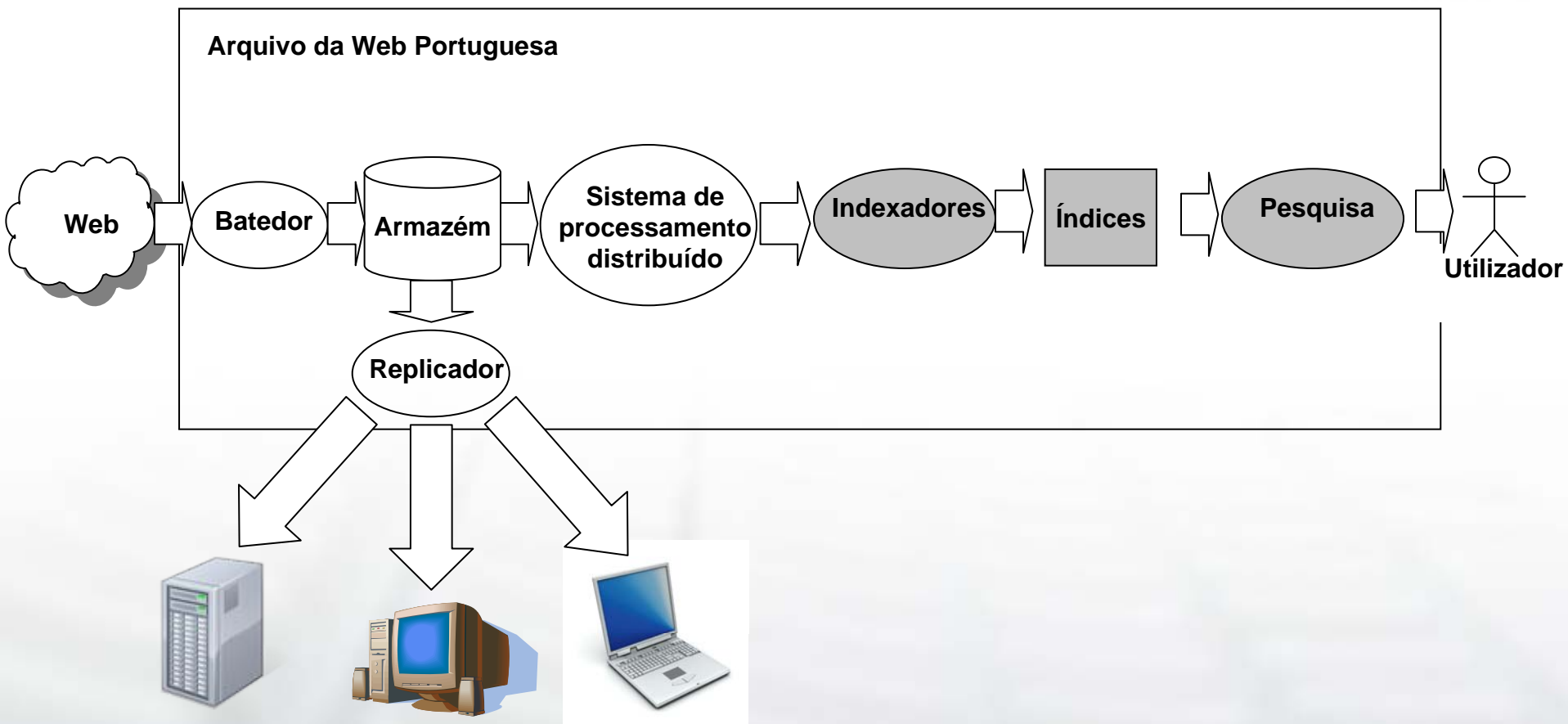
Fonte: Tom White , A Tour of Apache Hadoop

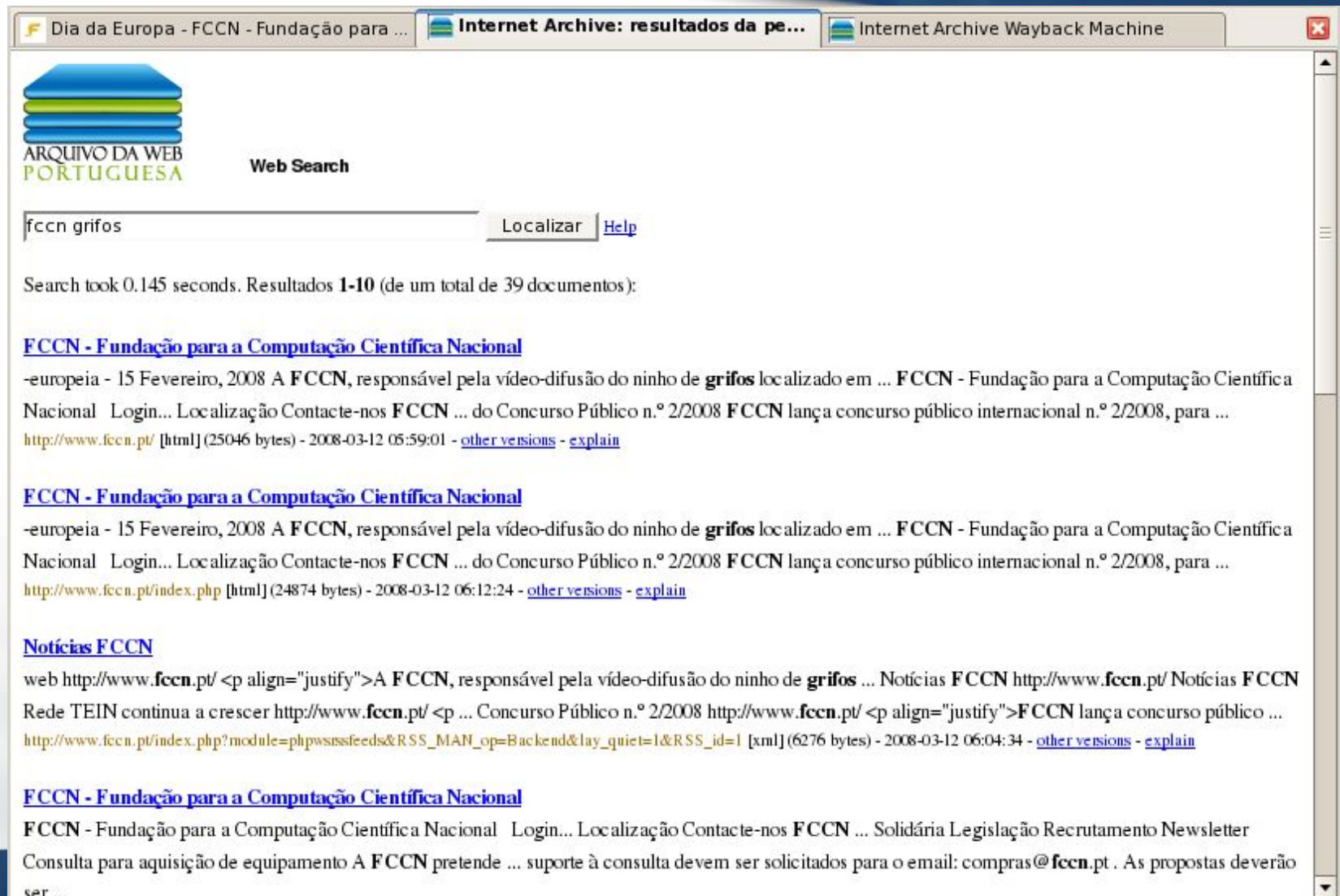


- Processos *masters* e *slaves*
- Cada *slave* está num computador
- Os *slaves* executam *maps* e *reduces*
- *Map* actua sobre um conteúdo de cada vez
- *Reduce* pode actuar sobre os resultados de vários *maps* recebidos por rede
- Resultados dos *map/reduce* são guardados no HDFS




- **GAppA: Grid Appliance para Arquivos**
- **Acesso aos conteúdos arquivados**
- **Partilhar recursos**
- **Instalação simples através de uma máquina virtual**





[Dia da Europa - FCCN - Fundação para ...](#)
[Internet Archive: resultados da pe...](#)
[Internet Archive Wayback Machine](#)


Web Search

[Help](#)

Search took 0.145 seconds. Resultados **1-10** (de um total de 39 documentos):

[FCCN - Fundação para a Computação Científica Nacional](#)
 -europeia - 15 Fevereiro, 2008 A FCCN, responsável pela vídeo-difusão do ninho de **grifos** localizado em ... FCCN - Fundação para a Computação Científica Nacional Login... Localização Contacte-nos FCCN ... do Concurso Público n.º 2/2008 FCCN lança concurso público internacional n.º 2/2008, para ...
<http://www.fccn.pt/> [html] (25046 bytes) - 2008-03-12 05:59:01 - [other versions](#) - [explain](#)

[FCCN - Fundação para a Computação Científica Nacional](#)
 -europeia - 15 Fevereiro, 2008 A FCCN, responsável pela vídeo-difusão do ninho de **grifos** localizado em ... FCCN - Fundação para a Computação Científica Nacional Login... Localização Contacte-nos FCCN ... do Concurso Público n.º 2/2008 FCCN lança concurso público internacional n.º 2/2008, para ...
<http://www.fccn.pt/index.php> [html] (24874 bytes) - 2008-03-12 06:12:24 - [other versions](#) - [explain](#)

[Notícias FCCN](#)
 web <http://www.fccn.pt/> <p align="justify">A FCCN, responsável pela vídeo-difusão do ninho de **grifos** ... Notícias FCCN <http://www.fccn.pt/> Notícias FCCN Rede TEIN continua a crescer <http://www.fccn.pt/> <p ... Concurso Público n.º 2/2008 <http://www.fccn.pt/> <p align="justify">FCCN lança concurso público ...
http://www.fccn.pt/index.php?module=phwsssfed&RSS_MAN_op=Backend&lay_quiet=1&RSS_id=1 [xml] (6276 bytes) - 2008-03-12 06:04:34 - [other versions](#) - [explain](#)

[FCCN - Fundação para a Computação Científica Nacional](#)
 FCCN - Fundação para a Computação Científica Nacional Login... Localização Contacte-nos FCCN ... Solidária Legislação Recrutamento Newsletter Consulta para aquisição de equipamento A FCCN pretende ... suporte à consulta devem ser solicitados para o email: compras@fccn.pt . As propostas deverão ser...



Enter Web Address:

All

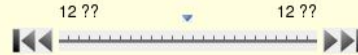
[Adv. Search](#)

1,000 results for <http://www.fccn.pt/>
between 1 01, 1996 and 10 21, 2008

[20080312055901](#) [www.fccn.pt](#) 200 [text/html](#) (new version)

[Home](#) | [Help](#)

Viewing version 1 of 1,000
5:59:01 3 12, 2008



[Help](#)

Wayback - External links, forms, and search boxes may not function within this collection. Url: <http://www.fccn.pt/> (time: 5:59:01 3 12, 2008) [[Hide](#)]

FCCN | Participação Internacional | Documentação | Domínios PT

- English
- RCTS
- Edu.PT
- IPv6
- VoIP
- e-U Campus Virtual
- Biblioteca do conhecimento on-line
- Segurança
- GigaPIX
- Serviços
- Projectos
- Eventos
- Rede Solidária
- Legislação
- Recrutamento
- Newsletter



Receba comodamente a nossa Newsletter



Grifos na Web
Vulture Cam

FCCN responsável pela difusão do sinal de vídeo que permite acompanhar de perto, 24Horas por dia, o processo de nidificação de um casal de grifos.

Um projecto conjunto de:

www.publico.pt/grifosnaweb

Rede TEIN continua a crescer

A rede Trans-Eurasia Information Network (TEIN) - rede regional asiática que interliga as instituições de ensino e investigação deste continente, possuindo uma ligação à rede europeia GÉANT2 - está a crescer, prevendo alargar-se durante o decurso deste ano a novos países do sul da Ásia, como Laos e Camboja.

[Mais...](#)

Lançamento do Concurso Público n.º 2/2008

FCCN lança concurso público internacional n.º 2/2008, para "Fornecimento, instalação e configuração das redes eléctricas, AVAC, segurança e telecomunicações para sala GRID". [Anúncio](#) (não dispensa a consulta das peças concursais).

Megamail celebra aniversário com novo site

O Megamail está a celebrar o seu 8.º ano de funcionamento, com uma nova imagem, um novo interface de serviço mais apelativo, cumprindo as regras de acessibilidade e, sobretudo, disponibilizando novos serviços aos seus utilizadores.

ServerSign EDU

FCCN celebra contrato com a TERENA tendo em vista a implementação nacional do projecto europeu ServerSign EDU. Este projecto tem como principais objectivos o fornecimento, em condições especiais, de certificados de servidor do tipo GlobalSign SureServer EDU Secure Server Certificates a instituições do meio académico e científico, parte da RCTS.

Pequisar

[Pesquisa Avançada](#)

Hora Legal em Portugal

05:00:54

Calendário

◀ Março 2008 ▶

D	S	T	Q	Q	S	S
24	25	26	27	28	29	01
02	03	04	05	06	07	08
09	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	01	02	03	04	05

Eventos

- Eventos
- Sessões de vídeo-difusão
- Outros

Serviços On-Line



Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Software e maquinaria

- Não existe software comercial de arquivo da Web
- Adoptar soluções de código aberto
 - Alteração para o contexto da Web
 - Maior garantia de preservação
 - Gratuitas
 - Existem para o arquivo da Web!
- Archive-access project liderado pelo Internet Archive
 - Poupança de recursos entre iniciativas
 - Heritrix crawler
 - Formatos ARC e WARC
 - NutchWAX (Nutch + Web Archive eXtensions)

- Boa base para o Arquivo da Web Portuguesa mas...
- São tecnologia de ponta
 - Estão em desenvolvimento
 - Pouco maduras e instáveis
 - Documentação com erros ou inexistente
- É necessário contribuir para melhorá-las



- Blade system
 - 7 blades: 2 x Quad-core, 8 GB
 - Espaço para 16 blades
- Armazenamento
 - SAN: 25.6 TB em RAID 5
 - 56 discos (500 GB e 1 TB)
 - Discos SATA (7.2K) e Fibre Channel (15K)
 - Autómato de tapes: 12 TB





Fundação para a Computação Científica Nacional
Foundation for National Scientific Computing

Medição da Web portuguesa

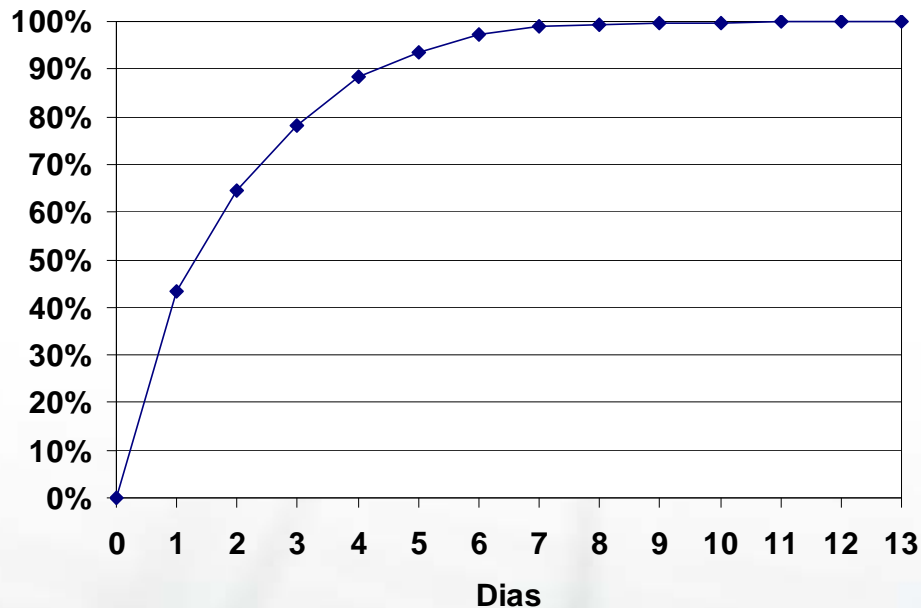
Critério de selecção para um arquivo Web nacional

- Objectivo: seleccionar conteúdos para arquivar
- Critério de relevância histórica?
 - Requer intervenção humana
 - 50 milhões de conteúdos por trimestre
- Que critério de selecção automática adoptar para recolher conteúdos de uma Web nacional?

Definição de Web portuguesa

- Recolha automática
- Sites sob .PT (1ª fase)
 - Noutros domínios: embebidos + redirecções
- Todos os tipos são aceites (máximo de 10 MB)
- 10 000 URLs por sítio Web
- Profundidade máxima de 5 ligações
- Respeito por regras de exclusão de robots (REP e meta-tag ROBOTS) e 2 segundos de pausa
- No futuro todos os conteúdos em portuguêsês?

Evolução da recolha



- Fevereiro de 2008
- 1 computador
- 99% em 7 dias
- Conteúdos alojados na RCTS
 - 1 Gbps ou 100 Mbps
 - 11% dos URLs
 - 18% volume dados

Métrica	Volume
Endereços visitados	72 milhões
Sítios Web visitados	455 mil
Conteúdos recolhidos	56 milhões
Volume de dados recolhidos	2,8 TB
Dados comprimidos	2 TB

Código	# URLs	%
200	56 046 288	85,2%
302	4 305 265	6,5%
404	3 669 855	5,6%
301	789 133	1,2%
500	325 225	0,5%
400	266 318	0,4%
403	164 241	0,2%
303	124 385	0,2%
401	48 334	0,1%
outros	36 136	0,1%
Total	65 775 180	100%

- Estudo detalhado de caracterização está em progresso
- Analisar evolução da Web portuguesa
 - Comparação com estudos anteriores

Distribuição de formatos: número de endereços

MIME	% conteúdos
text/html	65%
image/jpeg	17,7%
image/gif	7,6%
application/pdf	2,1%
text/plain	1,5%
Outros	6,1%

- **Preservar formatos HTML, JPEG e GIF: cobririam 90% da Web portuguesa (03/2008)**

Distribuição de formatos: volume de dados

MIME	% Dados	Posição #URLs
text/html	39,6%	1
application/pdf	14,4%	6
image/jpeg	12,4%	2
text/plain	4,7%	5
application/x-gzip	4,3%	10
Outros	24,6%	

- Posições alteram-se
- Dominância de formatos não é tão evidente

- Integração de colecções externas
 - Internet Archive (2000-2007): 800 GB
 - Recolhas do tumba! (2002-2006): 1,5 TB!
- Investigação acerca de pesquisa temporal sobre a Web
- Novas ferramentas em desenvolvimento
 - rARC: replicador de ARCs
 - GAppA: Grid Appliance para o Arquivo
 - WebClass: classificador de conteúdos
- Medição da acessibilidade da Web portuguesa

- Arquivar a Web tem interesse nacional
- Um arquivo necessita de ser pesquisável ou a informação arquivada “morre” por estar inacessível
- Arquivar a Web portuguesa é possível
- Contamos com a ajuda de todos
 - Colaboramos com projectos de investigação (ex. mestrado, doutoramento)



ARQUIVO DA WEB
PORTUGUESA

**Obrigado pela atenção.
Questões e sugestões são bem-vindas!**

<http://www.arquivo.pt>