Arquivo e medição da Web portuguesa

Daniel Gomes e João Miranda
{daniel.gomes, joao.miranda} @ fccn.pt

POS_CONHECIMENTO
Programa Operacional Sociedade do Conhecimento

UMIC
Agência para a 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

Arquivos de Web nacionais

- 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
- Sociólogo estuda proliferação das redes sociais
- Jurista obtém provas para caso (?)

...

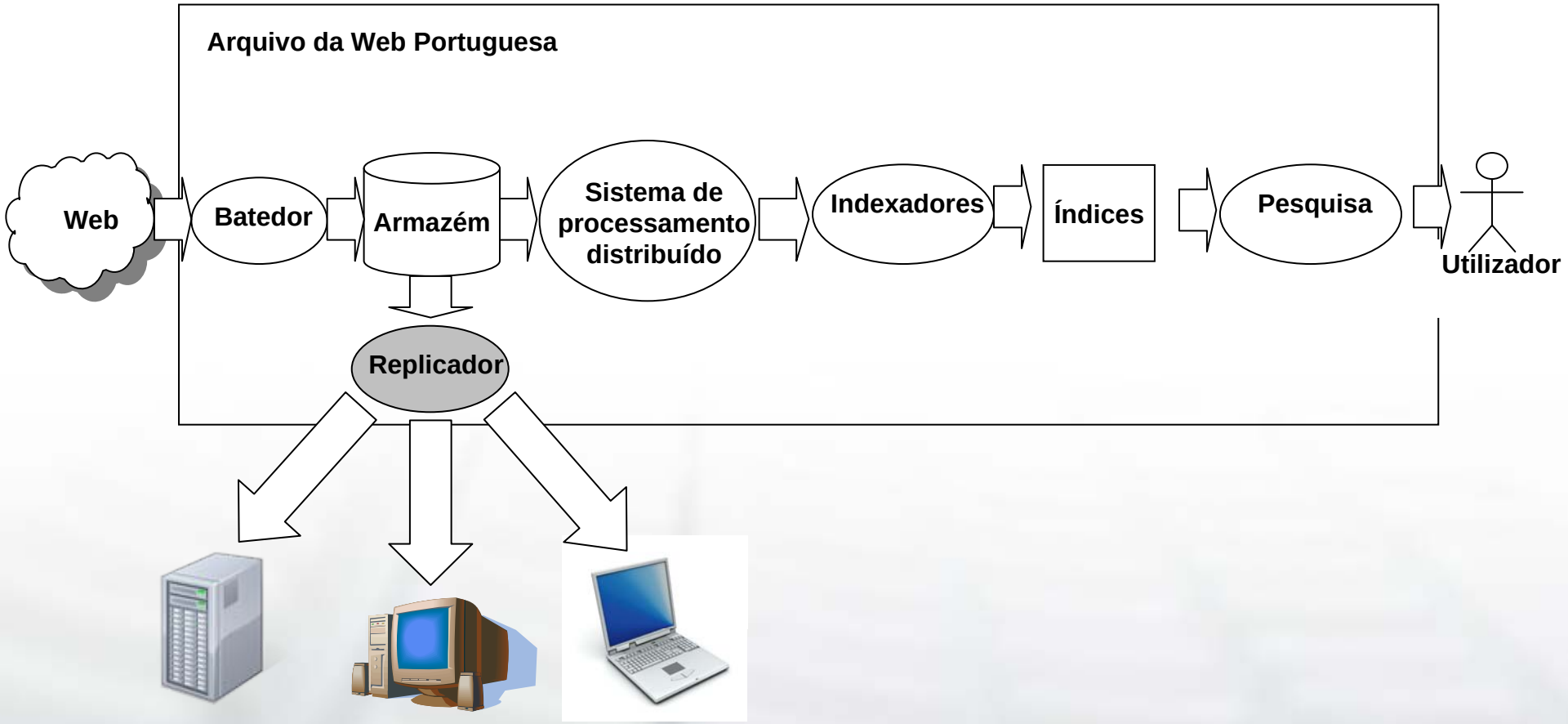
Estrutura da apresentação

  - Motivação
- Sistema de Arquivo da Web Portuguesa
- Resultados de uma medição da Web portuguesa
- Conclusões

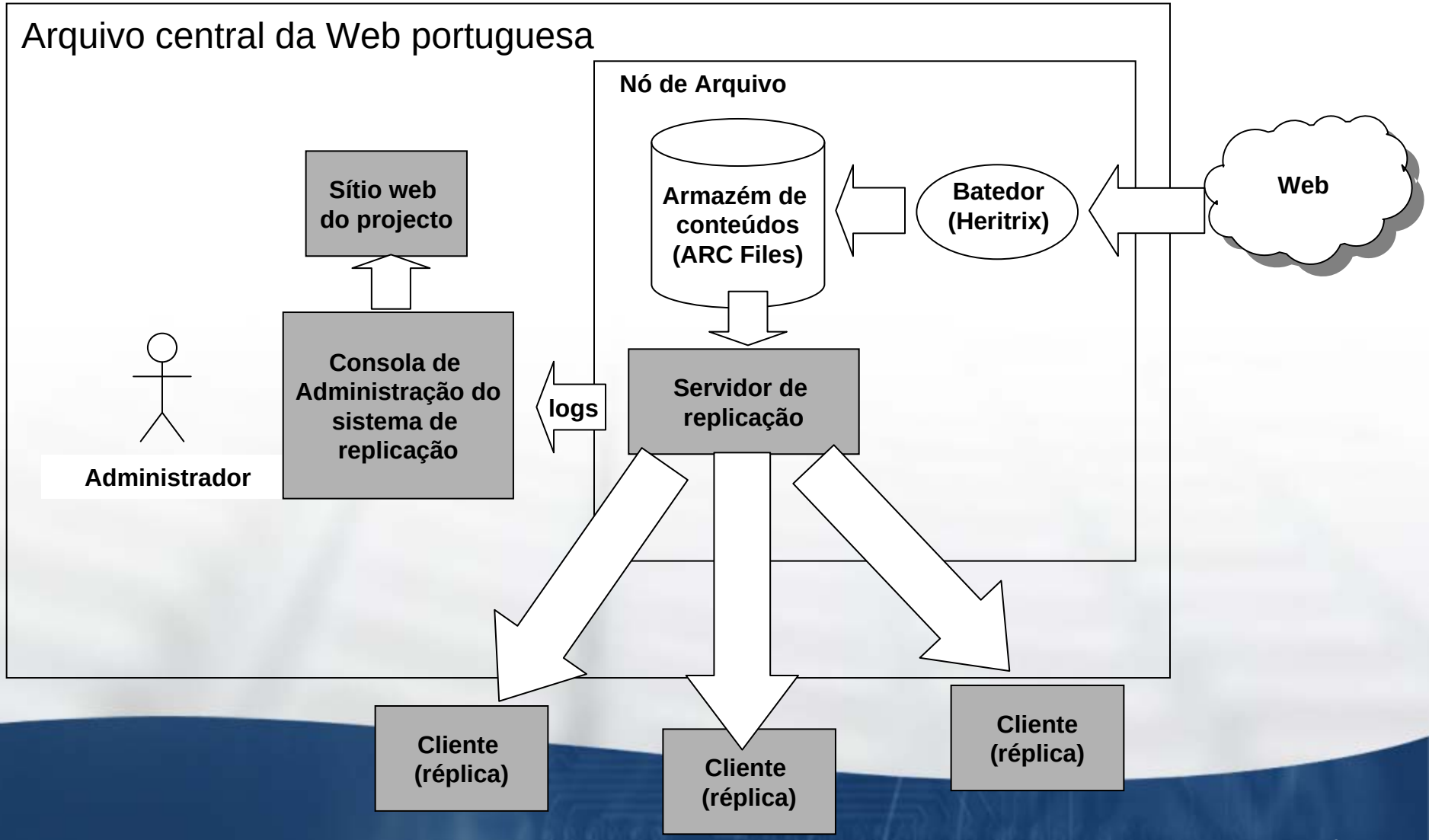
Principais objectivos

- 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
- Publicação de artigos científicos e técnicos
  - Estudo da Web portuguesa

Arquitectura geral



Todos podem contribuir para a preservação da Web



Características do rARC

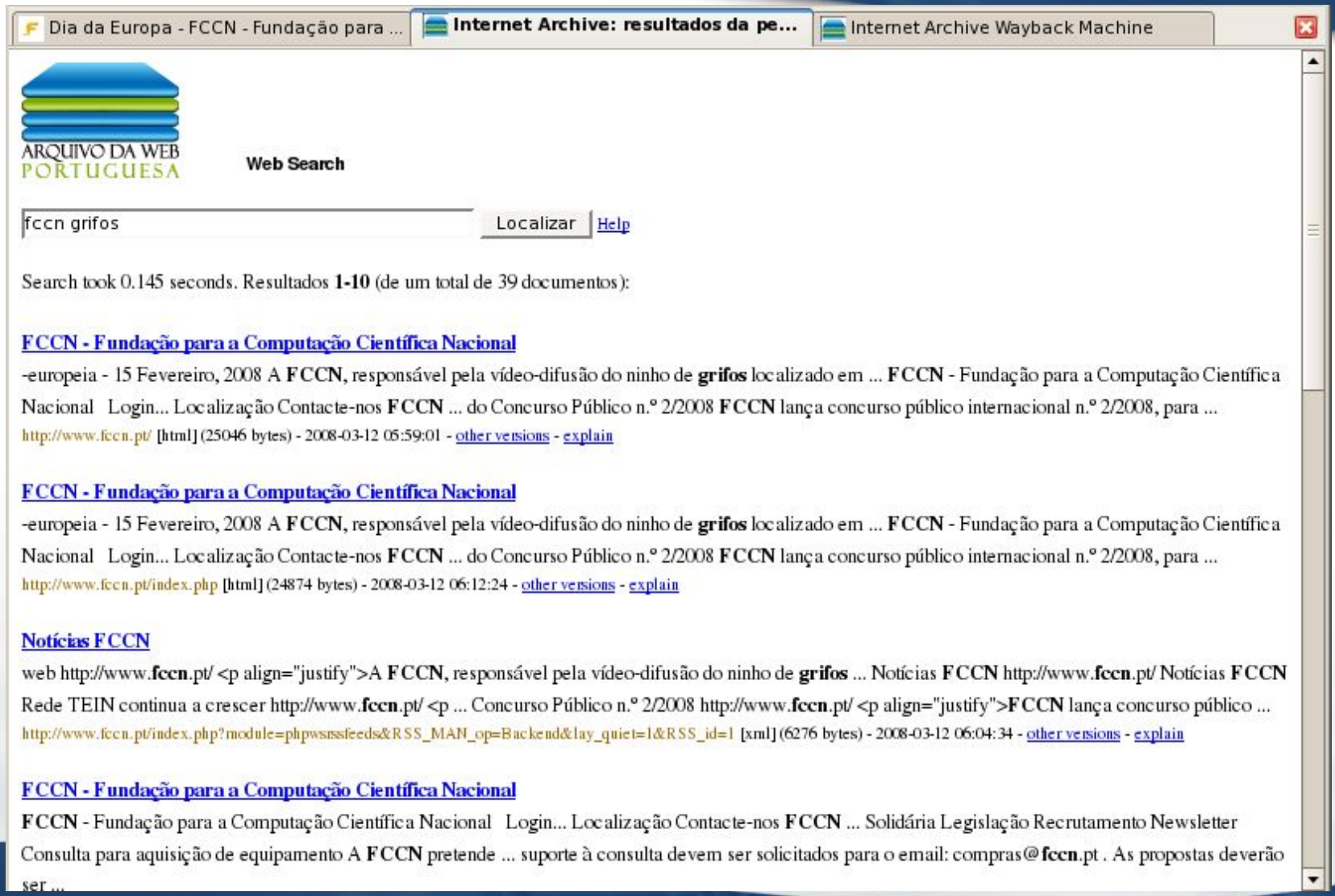
- Não é intrusivo, não carrega o computador do cliente
- Fácil de instalar
- Permite controlar espaço e largura de banda
- Confidencialidade
  - Cópias de segurança cifradas
- Integridade
  - Protecção contra clientes maliciosos que tentem adulterar as cópias para inserir conteúdos maliciosos no arquivo.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

Processamento e pesquisa no AWP

Processamento dos conteúdos arquivados

- Permite executar rotinas em larga escala
- Hadoop: plataforma de processamento paralelo
  - Implementa map-reduce do Google File System
  - Apenas é necessário escrever 2 rotinas. Exemplo:
    - Map: separa palavras num texto
      - "O tempo perguntou ao tempo"
      - <O,1>< tempo,1>< perguntou,1><ao,1><tempo,1>
    - Reduce: conta quantas vezes ocorre cada termo
      - <O,1>< tempo,2>< perguntou,1><ao,1>
  - Adoptado pela Yahoo em 10 000 servidores

Pesquisa por termo



Pesquisa por endereço

Fundação para a Computação Científica Nacional

Conteúdo arquivado

Software e maquinaria

Que tecnologia usar?

- 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)

Problemas com as tecnologias de arquivo

- 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

Maquinaria

- 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

Medição da Web portuguesa

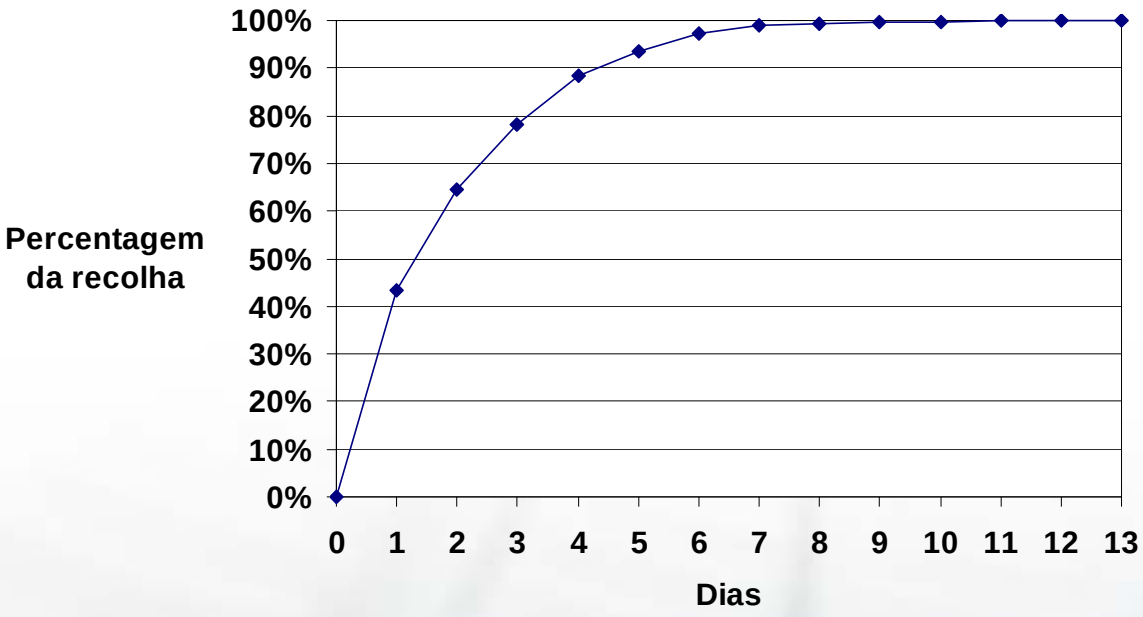
Critério de selecção para um arquivo Web nacional

- 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 português?

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

Caracterização de uma recolha

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

Trabalho "futuro"

- 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

Conclusões

- 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

ARQUIVO DA WEB PORTUGUESA

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

http://arquivo-web.fccn.pt