Arquivo da Web Portuguesa

Daniel Gomes
daniel.gomes arroba fccn.pt
Fundação para a Computação Científica Nacional

A era digital começou (já há alguns anos)

- A Web é a maior fonte de informação construída
  - Jornais, livros, documentação técnica
  - Informação publicada exclusivamente na Web
- A informação na Web é efémera
  - Gerações futuras poderão testemunhar uma “Idade das Trevas” digital
- Temos que começar a arquivar
  - Para que a História não se perca

Requisitos de um arquivo da Web

- A forma de arquivo tradicional requer demasiada intervenção humana
  - Não é compatível com a dimensão da Web
- Recolha e armazenamento automático
  - Intervenção humana mínima
- Dispendioso em larga escala
  - Internet Archive (www.archive.org)

Arquivos da web nacionais

- 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 critérios para definir limites das webs nacionais
- Necessárias arquitecturas de sistema e software específico para suportar o arquivo da web

Trabalho passado em Portugal

- Digital Deposit (2001)
  - FCUL/BN
  - Recolha selectiva
- Tomba (2006)
  - FCUL/FCCN
  - Recolhas do tumba! (2002-2006)
  - 54 M; 1,5 TB
  - Textos principalmente
  - Migração para o AWP
- Existe conhecimento nacional acerca do assunto

Estrutura da apresentação

- Introdução
- Arquivo da Web Portuguesa na FCCN
- Conclusões

Contexto

- Projecto de Investigação & Desenvolvimento
  - É necessária investigação para seguir a evolução da web
- Arquitectura e tecnologia diferente do Tomba
- Duração de 2 anos a partir de 2008
- Necessária visão a longo prazo

Principais objectivos

- Iniciar o “depósito legal” da web portuguesa
- Serviço público de acesso ao arquivo
- Formação de recursos humanos
- Disseminação da informação arquivada para preservação
- Publicação de artigos científicos e técnicos
  - Divulgação, partilha de conhecimento e obtenção de críticas por parte dos especialistas.

Benefícios nacionais

- Português como língua da web
- Capacidade local de tratamento de informação da web
  - Segurança nacional não pode depender do estrangeiro
- Exportação do *saber-fazer*
  - Arquivo da web é tecnologia de ponta.
- Dados para a Ciência
  - Sociologia, prospecção de dados, processamento da língua
- Provas judiciais

Principais desafios

- Boa abrangência da web portuguesa
- Pesquisa eficiente num arquivo histórico
  - É um problema em aberto
  - Apresentação de conteúdos
- Preservação da informação
- Continuidade a longo prazo

Imaturidade da web portuguesa

- Ferramentas que funcionam “bem”, falham na web portuguesa
  - Desenvolvidas para a generalidade da web
  - EUA são os maiores produtores de conteúdos e das ferramentas
- Web é recente -> falta *saber-fazer*
  - Web designers instantâneos dominam o mercado
- Acessibilidade
  - Section 508 (1999), RCM 155/2007
  - Textos como imagens
    - Teste simples: copiar/colar-> FF Properties-> ALT text
  - WAI A: apenas 10 dicas rápidas, UMIC-Acesso em português
  - Páginas não aparecem nos motores de busca
- Usabilidade
  - Estudos científicos 1999-2006: web portuguesa está em 1999
- Principais prejudicados são os donos dos sítios web
  - Potencialidade do sítio web limitada

O que arquivar no Arquivo da Web Portuguesa?

- Sites sob .PT (1ª fase)
  - Estamos a perder metade da web portuguesa
  - Alguns utilizadores ficarão insatisfeitos
  - No futuro todos os conteúdos em português
- Tipos GIF, JPEG e HTML
  - 95% dos conteúdos publicados
- Recolhas trimestrais

Como arquivar?

- Meta-dados que permitam preservar e aceder à informação
  - Estratégia de conversão de formatos
- Espaço de armazenamento incremental
- Acessibilidade à informação por pessoas e máquinas
- Ferramentas de gestão e preservação

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!

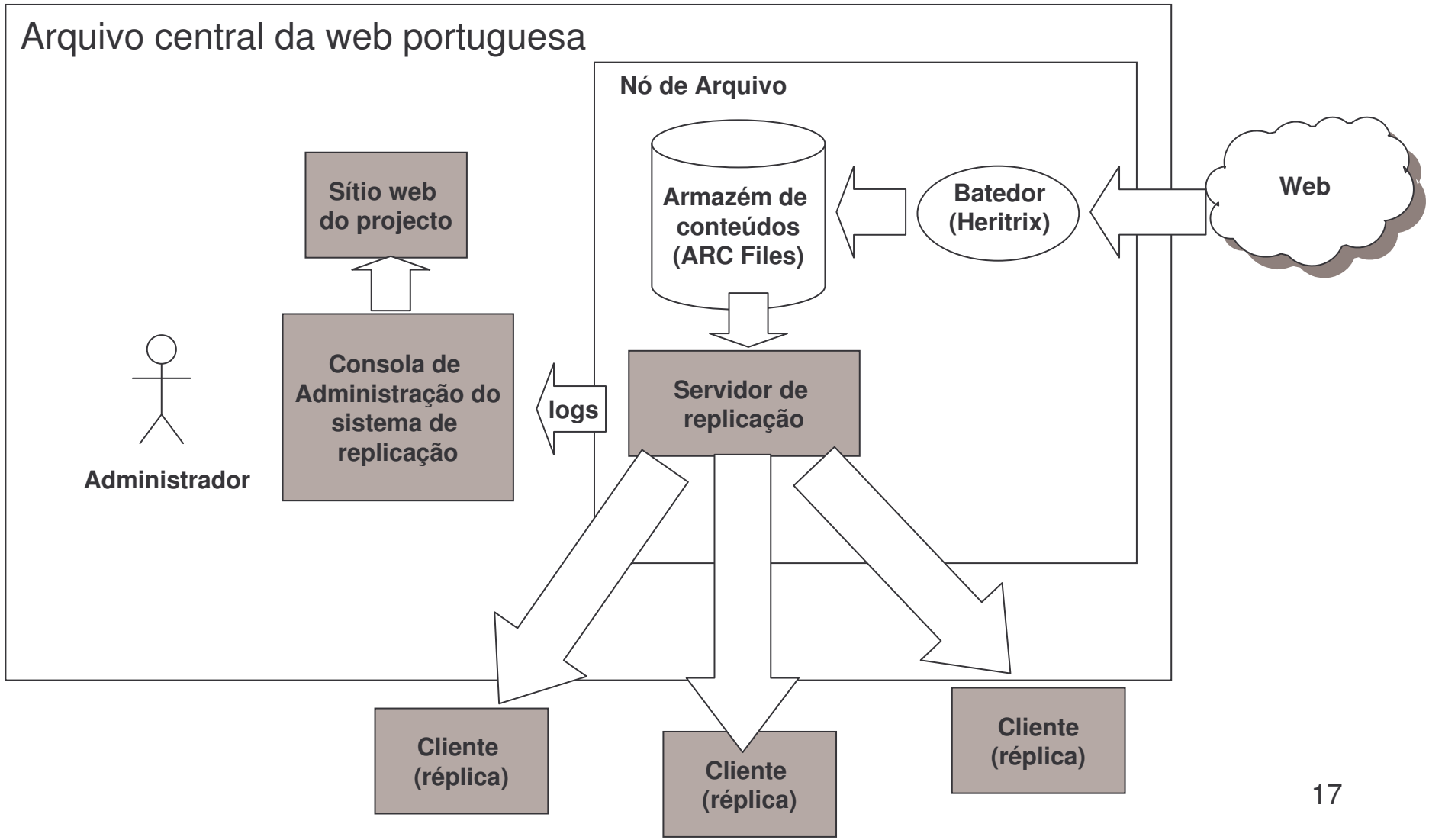
Tecnologias abertas para o arquivo

- Poupança de recursos entre iniciativas
- Internet Archive
  - Archive-access project
  - Heritrix crawler
  - Formatos ARC e WARC
- NutchWAX (Nutch + Web Archive eXtensions)
  - Nutch: motor de busca para a web (índexador: Lucene)
  - Hadoop: plataforma de processamento paralelo
    - Implementa map-reduce do Google File System
    - Adoptado pela Yahoo em 10 000 servidores
- WERA (Web aRchive Access)
  - Antigo Nordic Web Archive
  - Archive viewer application

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
- Queremos contribuir para melhorá-las

Contribuição do AWP: sistema rARC (replicador de ARCs)



Serviços a disponibilizar pelo AWP

- Motor de busca histórico por termo
  - Novo motor de pesquisa sobre a web portuguesa
- Motor de busca histórico por URL
- Colecções históricas de conteúdos web para investigação
- Relatórios de caracterização da web de Portugal
- Infra-estrutura para processamento paralelo dos dados arquivados

A equipa

- Daniel Gomes
  - Coordenador do projecto
  - Interesse na área desde 2001
- André Nogueira
  - rARC
  - Segurança
- João Miranda
  - Recolha de conteúdos
  - Sistemas de publicação na web
- Miguel Costa
  - Ex-tumba!
  - Pesquisa e indexação
- Apoio dos grupos especializados da FCCN

A equipa 2: toda a FCCN

- O Arquivo deverá servir toda a comunidade
  - Utilizadores com diferentes idades, escolaridade e profissões
- Colaboradores da FCCN como 1ª linha de teste
  - Trabalho começa quando um sistema está “pronto”
  - Testar é difícil para quem desenvolveu
    - 25% dos erros escapam aos peritos, 50% não são erros reais
    - Testes de usabilidade
    - Envolvimento de “não-informáticos” é crucial
- Como participar informalmente?
  - Erros de funcionamento
  - Dificuldades de utilização
  - Erros ortográficos ou dúvidas ao ler os textos
  - Sugestão de novas ideias e funcionalidades
  - Divulgação do projecto

Ideias novas

- Importação de conteúdos sob .PT do Internet Archive desde 2001
  - De 1996 a 2001 ainda não estão disponíveis
  - Avaliar recursos necessários
- Grid Appliance
  - Integração com o Hadoop
  - Importação e exportação de capacidade de processamento
- Pesquisa de imagens

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

Obrigado pela atenção.

Daniel Gomes

daniel.gomes arroba fccn.pt

http://arquivo-web.fccn.pt