Arquivo da Web Portuguesa

Daniel Gomes
daniel.gomes arroba 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
  - Gerações futuras poderão testemunhar uma "Idade das Trevas" digital
- Temos que começar a arquivar
  - Para que a História não se perca

Arquivos de web nacionais

- Internet Archive: 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

Trabalho passado em Portugal

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

Arquivo da Web Portuguesa

Iniciativa oficial

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
- Serviços públicos de acesso à informação arquivada
- Prestação de serviços à comunidade científica
  - História, Linguística, Sociologia, ...
- Formação de recursos humanos
- Publicação de artigos científicos e técnicos
  - Divulgação, partilha de conhecimento e obtenção de críticas por parte dos especialistas.

Fases do arquivo da web

- Selecção e obtenção
  - Aquisição da informação
- Armazenamento e replicação
  - Integridade da informação
- Processamento e acesso
  - Manutenção da informação acessível

Preservação

Selecção e obtenção

Discussão de critérios

Critério de selecção para um arquivo web nacional

- Objectivo: seleccionar conteúdos interessantes para preservar
- 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?

Que domínios/sítios seleccionar?

- Country code Top Level Domains têm um âmbito nacional
  - Recolher apenas o .PT?
- Implementação fácil e "leve"
- Os portugueses usam gTLDs (.com, .net, .org): razões comerciais, baratos, registo e administração rápidos.
- Menos de 49% dos conteúdos da web portuguesa estão alojados sob .PT

Que tipos preservar?

- Formatos de publicação mudam mas a informação tem de ser preservada
  - TXT->HTML->XHTML->?
- Estratégias de preservação de acordo com os tipos dos conteúdos
  - Formatos abertos: conversão
  - Formatos proprietários: emulação
- Custos de preservação de acordo com a diversidade de formatos

Distribuição de formatos

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)

Os blogs devem ser arquivados?

- Relevância histórica?
  - Adolescentes usam-nos como meio de comunicação
  - Um deles poderá ser o próximo Presidente
- 15.3% são blogs (03/2008)
- Blog = Meio fácil de publicar na web
  - Programas de TV, rádio, apoio técnico, comunicados de empresas,...
- E o Web Spam?
  - Páginas geradas automaticamente para enganar os motores de busca
  - São um espelho dos nossos tempos
  - Dados foram usados para detectá-lo.

Depois de seleccionar, como obter?

- Entrega: publicadores enviam conteúdos para o arquivo
  - Inspirado no depósito legal tradicional
  - Caro para os publicadores
  - Imposição difícil
  - Escassez de ferramentas e normas
- Recolha: arquivo selecciona e recolha automaticamente os conteúdos dos sítios web dos publicadores
  - Intervenção humana mínima
  - Mais carga no arquivo
  - Dispendiosa em larga escala

Decisões de selecção e obtenção no AWP

- 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)
- A recolha anterior fornece as raízes da próxima
- No futuro todos os conteúdos em português?

Resultados 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%

- Março de 2008
- 90% em cerca de 1 semana
- Estudo detalhado de caracterização está em progresso
- Analisar evolução da web portuguesa
  - Comparação com estudos anteriores

Fontes de informação

- Recolhas trimestrais do AWP (2008-...)
  - 3 realizadas em 2008: 7,2 TB (155 milhões de conteúdos)
- Colecção adquirida ao Internet Archive (2000-2007)
  - 1996-2000 não está disponível
  - 1,4 TB de informação recolhida de .PT (comprimido)
  - Indexação em curso
- Colecções do tumba! (2001-2006)
  - 1,5 TB de informação textual (57 milhões de conteúdos)
  - Exportação em curso

Estrutura da apresentação

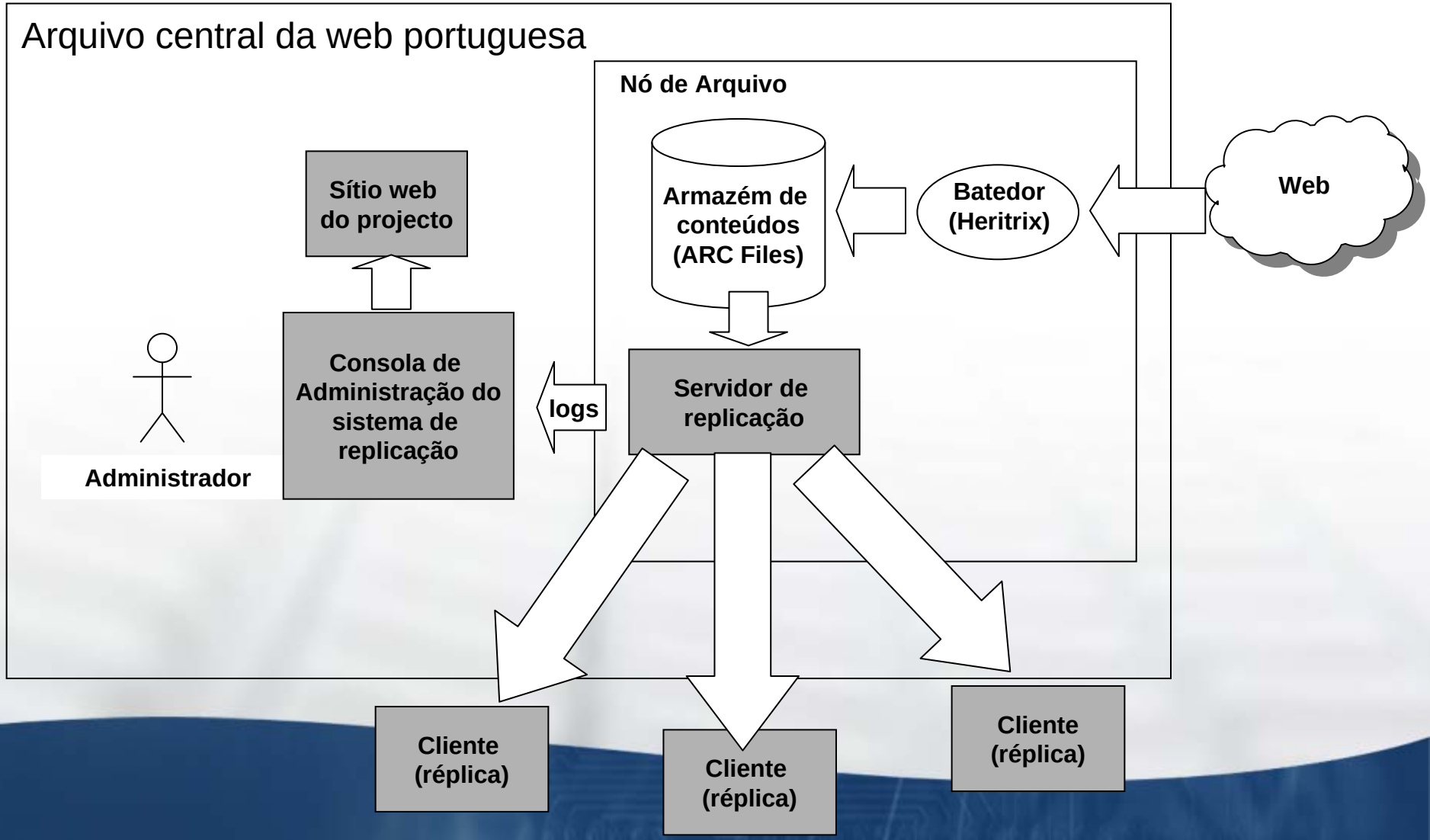
  - Motivação
  - Selecção e obtenção
- Armazenamento e replicação
- Processamento e acesso
- Tecnologias
- Conclusões

Armazenamento e replicação

Armazenamento

- Formato ARC (Internet Archive)
  - wARC será uma norma (ISO/DIS 28500)
- Redundância interna
  - SAN: 25.6 TB em RAID 5
  - 56 discos (500 GB e 1 TB)
  - FATA e Fibre Channel
  - Tape library: 12 TB

Replicação externa: rARC



Características do rARC

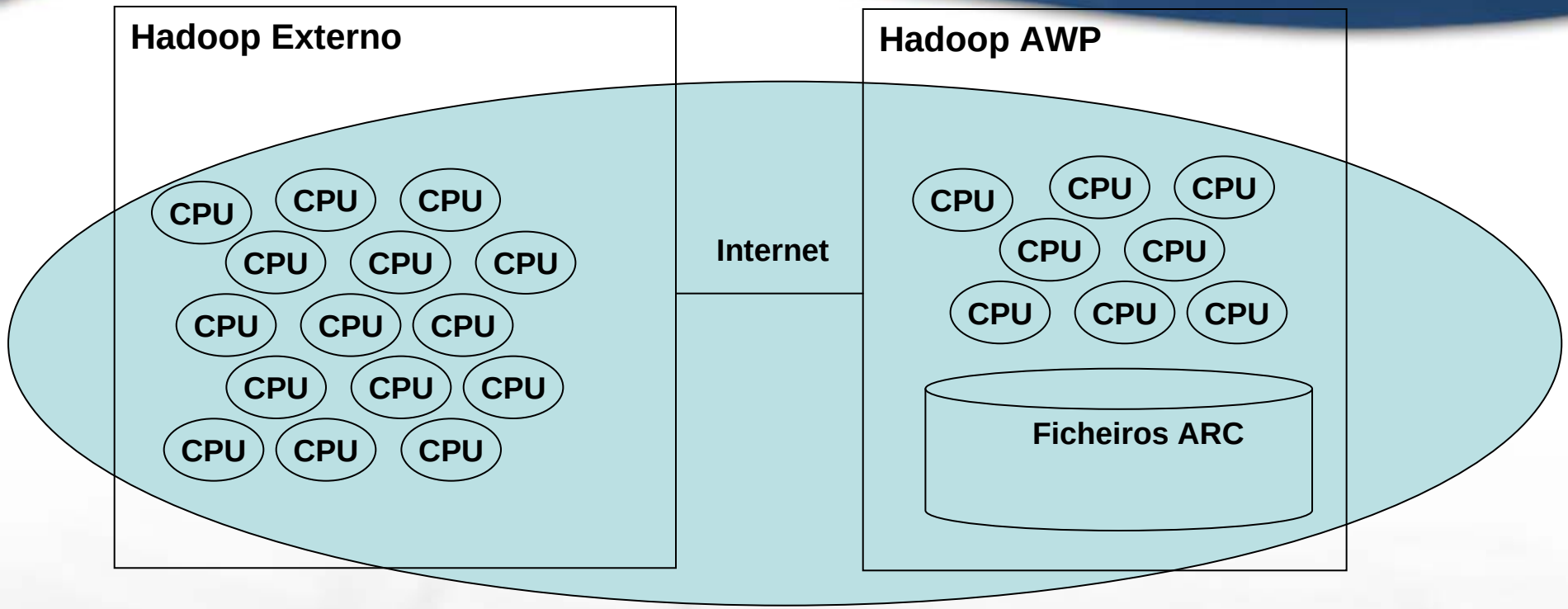
- Permite pequenas e grandes contribuições de espaço
- Não é intrusivo, não carrega o computador do cliente
- Fácil de instalar
- É independente de plataforma
- Confidencialidade
  - Cópias de segurança cifradas
- Integridade
  - Protecção contra clientes maliciosos que tentem adulterar as cópias para inserir conteúdos maliciosos no arquivo.

Processamento e acesso

Processamento interno dos conteúdos arquivados

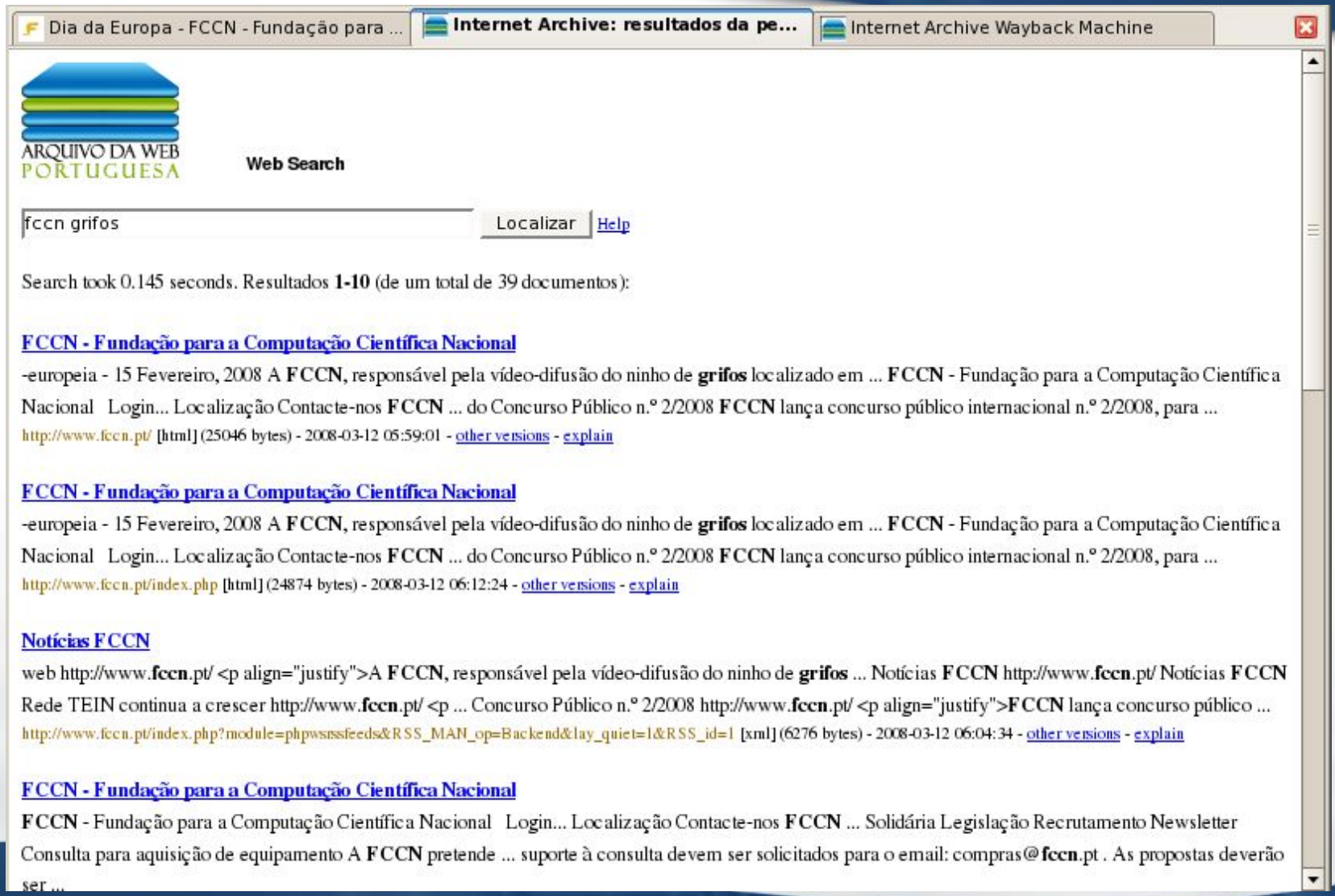
- Permite executar rotinas em larga escala
- Hadoop: plataforma de processamento paralelo
  - Implementa map-reduce do Google File System
  - Apenas é necessário escrever 2 rotinas. Exemplo:
    - Map: separa palavras num texto
      - "Seminário preservação: preservação da web"
      - <Seminário,1>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

Processamento colaborativo



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

Pesquisa por termo



Pesquisa por endereço

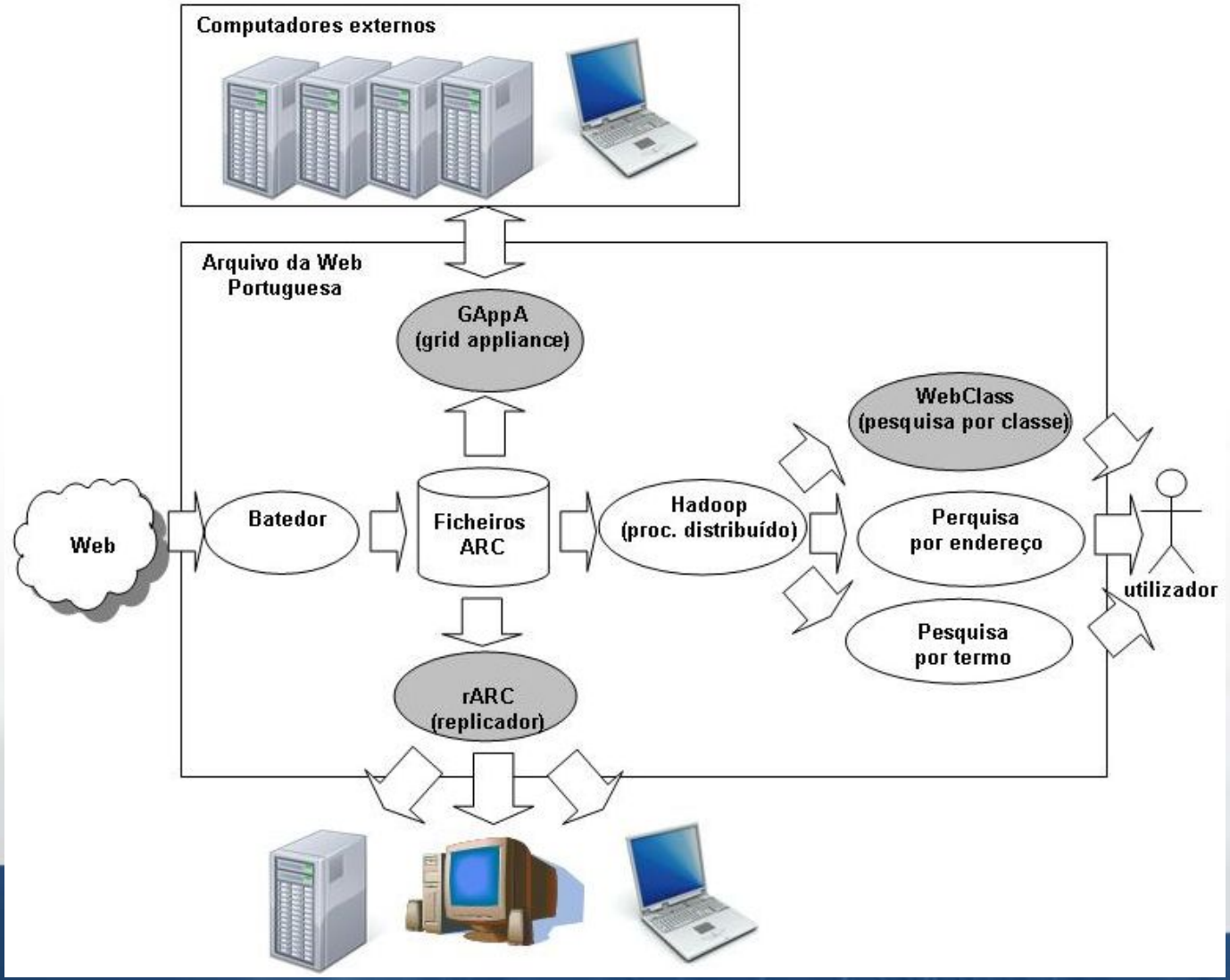
Fundação para a Computação Científica Nacional

Conteúdo arquivado

WebClass: classificação automática de conteúdos

- Atribuição de classes a cada conteúdo
- Classes = Secções de jornal
  - Fácil de obter conjuntos de treino
  - Utilizadores percebem os nomes das classes
- Método alternativo de pesquisa: todas as notícias acerca de Desporto num determinado dia
- Ajuda a desambiguar pesquisas
  - Pesquisa por "Figo": jogador de futebol, fruto ou International Federation of Gynecology and Obstetrics?
  - Pesquisa por "Figo" em Desporto: jogador de futebol!

Em suma, a arquitectura do AWP



Tecnologias

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
- Queremos contribuir para melhorá-las

Contributos

- Comunidade nacional
  - Serviços de pesquisa de acesso público
  - Infra-estrutura para prospecção de dados web
    - Segurança: vírus, xenofobia, roubo de identidade
    - Medição da acessibilidade
    - Computação científica
  - Colecções de dados para investigação
  - Relatórios acerca da evolução da web portuguesa
- Comunidade do arquivo da web
  - Novas ferramentas em desenvolvimento
    - rARC: replicador de ARCs
    - GAppA: Grid Appliance para o Arquivo
    - WebClass: classificador de conteúdos
  - Investigação em curso acerca de pesquisa temporal sobre a web

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. Contacte-nos:

daniel.gomes arroba fccn.pt

http://arquivo-web.fccn.pt