ARQUIVO.PT

Como publicar conteúdos na Web preserváveis para o futuro

Hugo Viana
hugo.viana@fccn.pt

Sumário

- O que são Motores de Busca
- Como públicar conteúdos Web preserváveis
- Protocolo de exclusão de Robôs
- Como criar um ficheiro Robots.txt

O que são os motores de busca

É um software que varre toda a Internet em busca de informação desejada (documentos ou endereços de páginas web) .

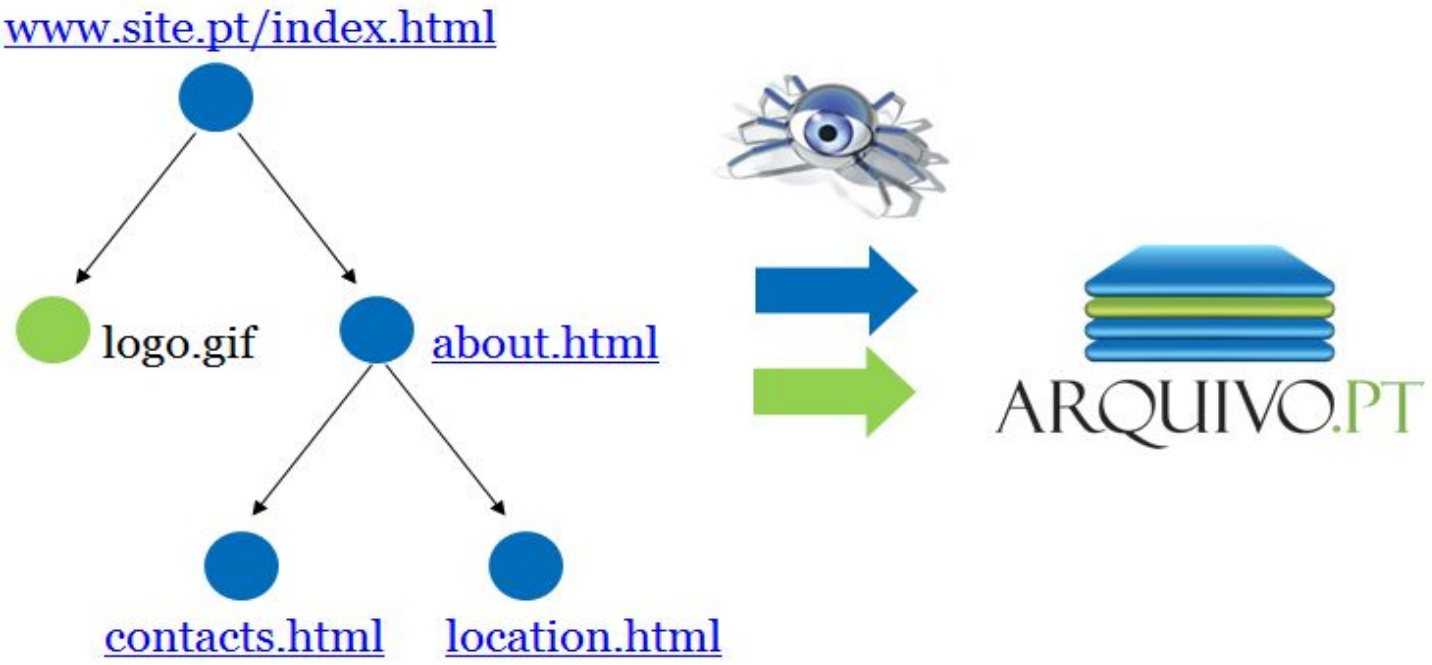
Componentes de um motor de busca Web conventional

1. Batedor
2. Armazenamento
3. Indexador
4. Ordenador
5. Apresentador

Batedores

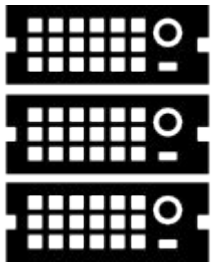
A partir de um conjunto inicial de URLs (raizes), os batedores do motor de busca iniciam uma recolha da Web, percorrendo todos as ligações criadas dentros dos Web sites.

Informação é recolhida automaticamente



Armazenamento

Após a recolha ter terminado, toda a informação recolhida da web fica armazenada no repositório para ser indexada.



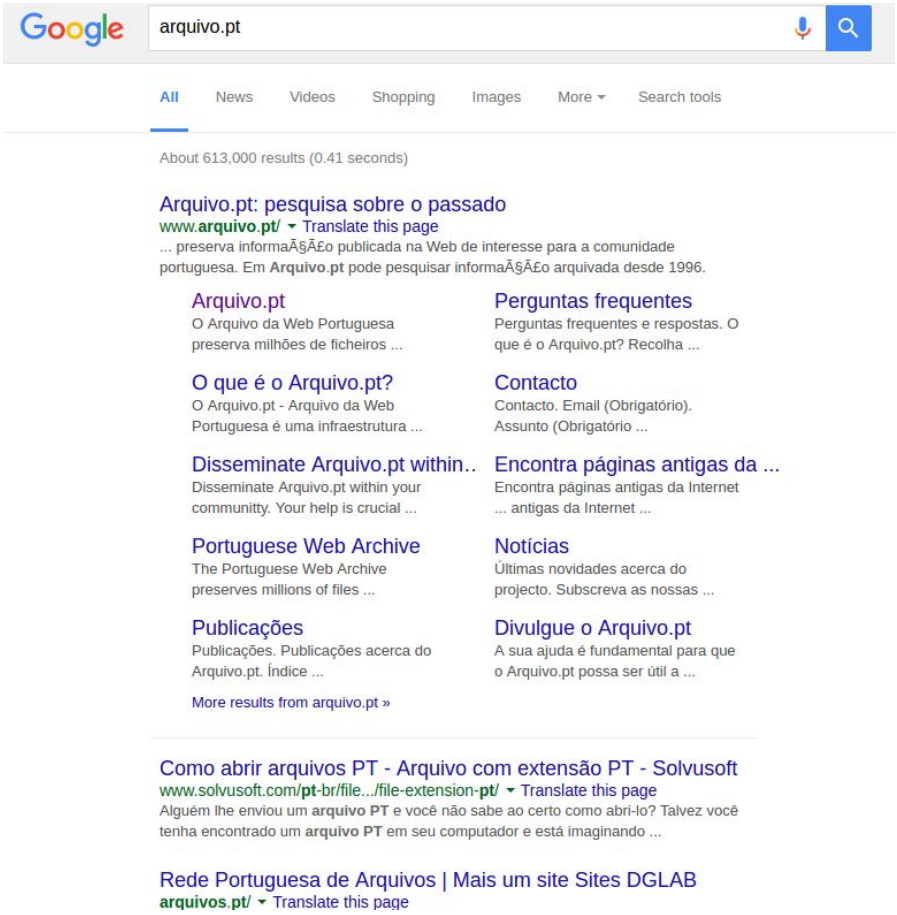
Indexador e ordenador

O indexador extrai as palavras contidas nas páginas armazenadas e constrói índices que irão permitir efectuar pesquisas rápidas.

Apresentador

O apresentador recebe os termos pesquisados pelos utilizadores, acede à informação dos índices e apresenta os resultados da pesquisa na forma de links para as páginas.

Apresentador do Google



Esquema de motor de busca conventional

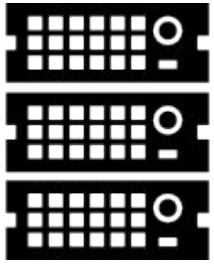
ARQUIVO.PT

Componentes de um arquivo Web com motor de busca

1. Batedor
2. Armazenamento
3. Indexador
4. Ordenador
5. Apresentador
6. Reprodução de conteúdo

Armazenamento

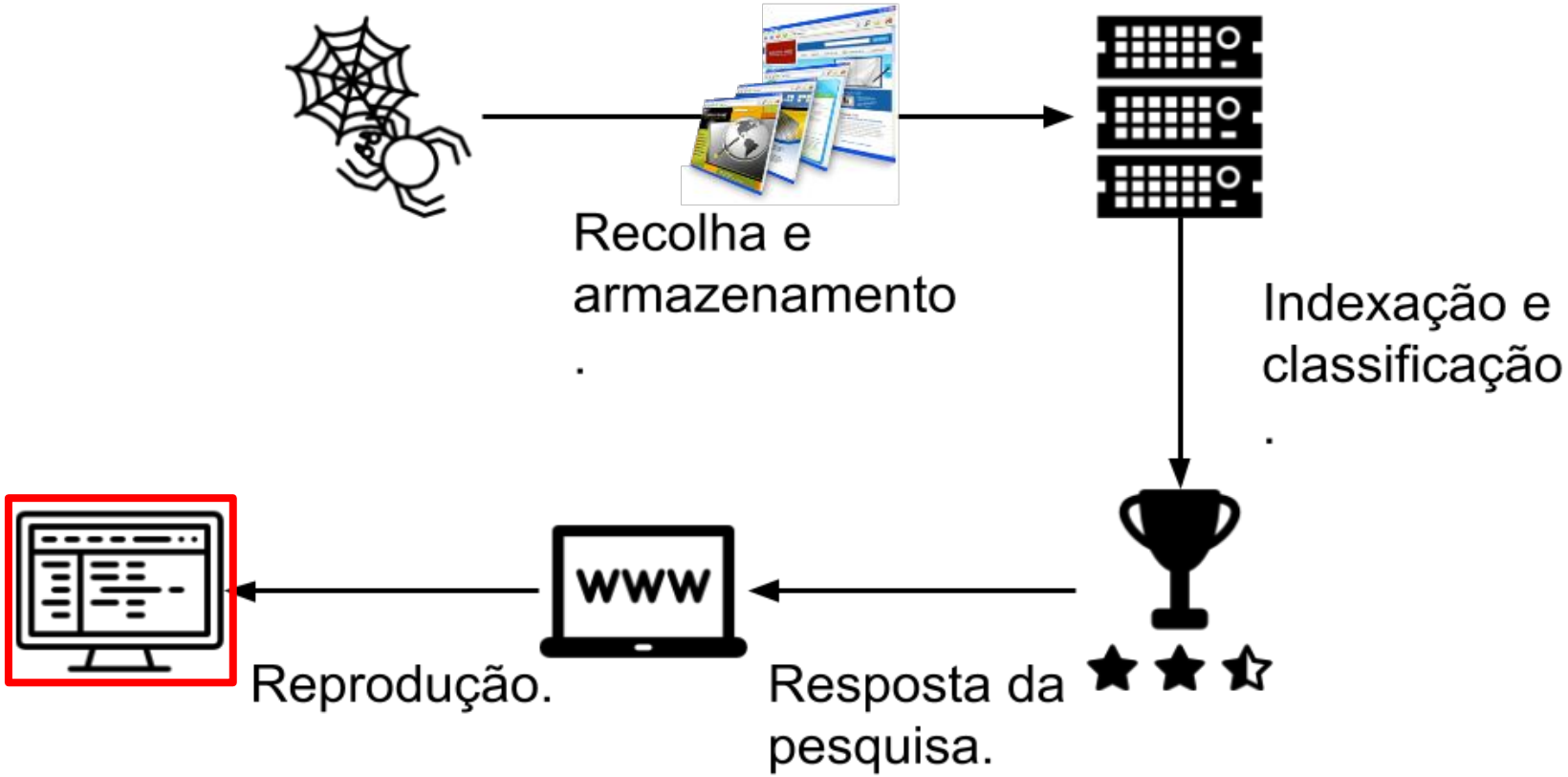
Após a recolha ter terminado, toda a informação recolhida da web fica armazenada no repositório para ser indexada e **reproduzida**.



Reprodução de conteúdo Arquivo.pt (2011)

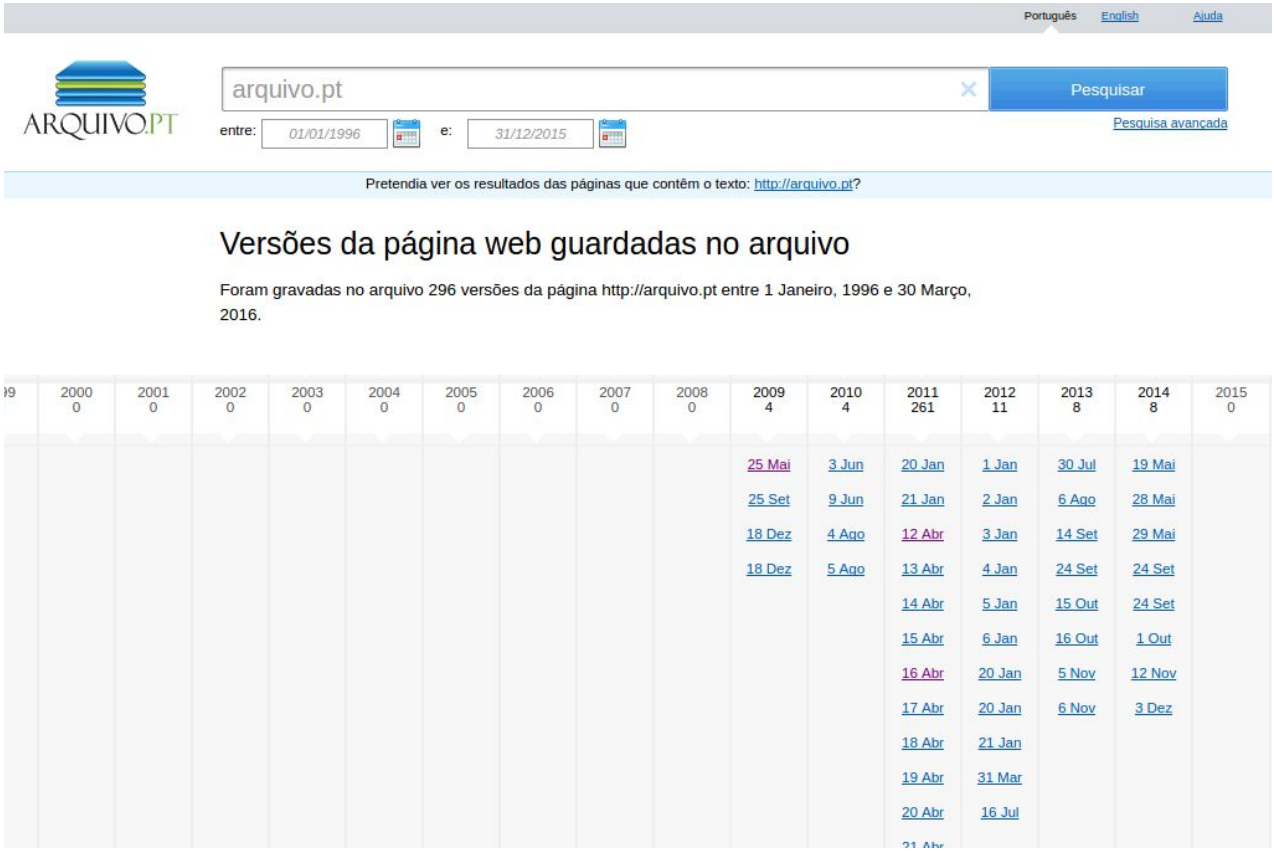
ARQUIVO.PT

Esquema de um motor de busca para arquivos da Web

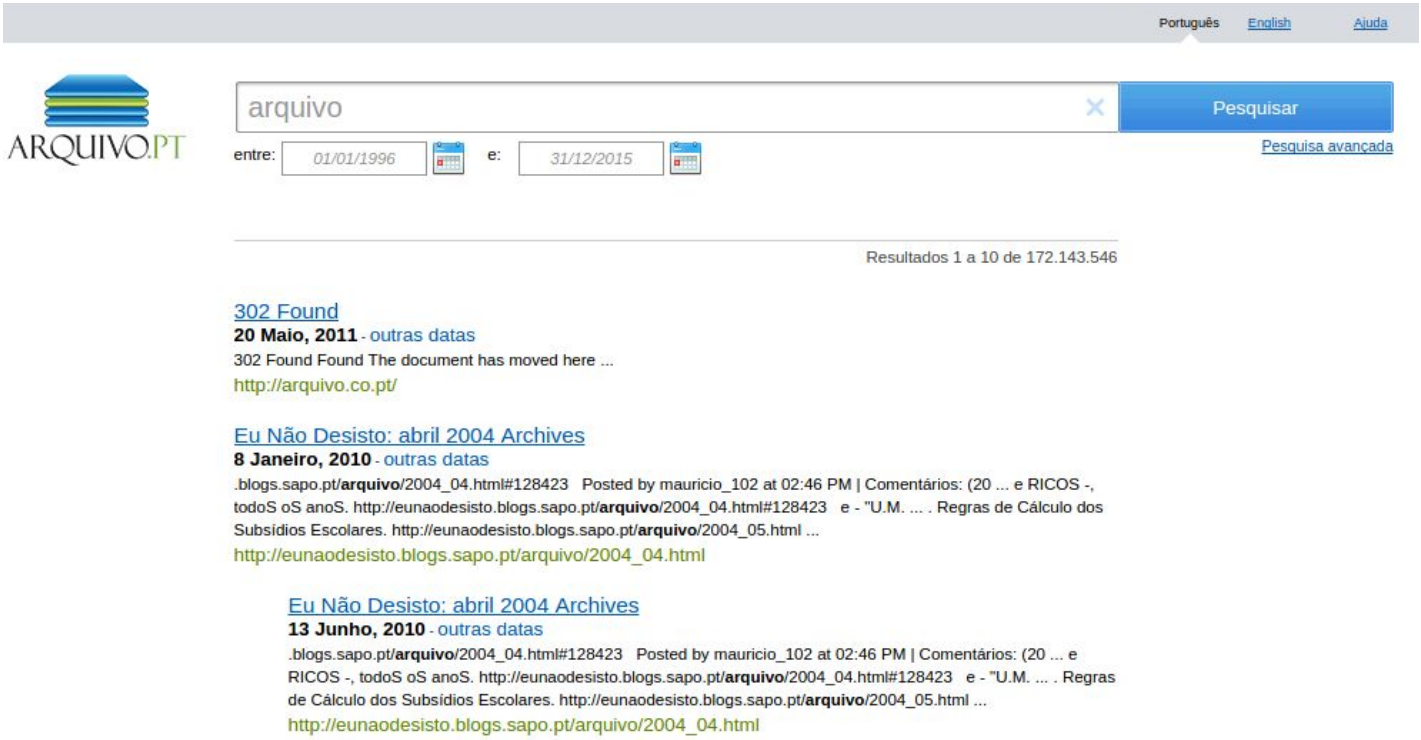


ARQUIVO.PT

Apresentador do Arquivo.pt (pesquisa por URL)



Apresentador do Arquivo.pt (pesquisa por termos)



ARQUIVO.PT

Como públicar conteúdos Web preserváveis?

Uma ligação por conteúdo

http://arquivo.pt/img/logo-home-pt.png

Mapa de navegação para utilizadores

Você está aqui: Entrada → Colabore → Recomendações para autores de sítios Web

Mapa do sítio

Uma vista geral do conteúdo disponível neste sítio. Mantenha o cursor imóvel sobre um item durante alguns segundos para obter a sua descrição.

- Sobre
  - O que é o Arquivo.pt?
  - Objetivos
  - Funcionamento
    - Arquitectura
    - Tecnologia
  - Exemplos de páginas preservadas no Arquivo.pt
  - Serviços e Software em código-aberto
    - Botão Histórico desta página
    - Filtros de rejeição usados nas recolhas
    - Mapa das iniciativas de arquivo da Web a nível mundial - zip

ARQUIVO.PT

Mapa do sitio

https://www.europeia.pt/sitemap.xml

Universidade Europeia XML Sitemap

URL	Priority	Change Frequency	LastChange
https://www.europeia.pt/	NaN%		2016-03-16
https://www.europeia.pt/universidade-europeia	NaN%		2014-09-16
https://www.europeia.pt/universidade-europeia/a-universidade	NaN%		2015-09-16
https://www.europeia.pt/universidade-europeia/reitor	NaN%		2015-10-27
https://www.europeia.pt/universidade-europeia/advisory-board	NaN%		2016-01-20
https://www.europeia.pt/universidade-europeia/lideres-na-universidade	NaN%		2015-01-26
https://www.europeia.pt/universidade-europeia/laureate-international-universities	NaN%		2015-10-14
https://www.europeia.pt/universidade-europeia/student-services	NaN%		2015-08-31
https://www.europeia.pt/universidade-europeia/campus-da-universidade	NaN%		2015-01-26
https://www.europeia.pt/universidade-europeia/localizacao	NaN%		2015-01-26
https://www.europeia.pt/universidade-europeia/institucional	NaN%		2015-01-26
https://www.europeia.pt/universidade-europeia/abre-as-portas-aos-teus-sonhos	NaN%		2015-06-05

Data de publicação correctamente identificada

ARQUIVO.PT

Manter o mesmo endereço ao longo do tempo

PSD2011.com (2011)

ARQUIVO.PT

PSD2011.com (2014)

Outras recomendações: *Robots.txt*

Robots.txt ?! Para que serve?!

ARQUIVO.PT

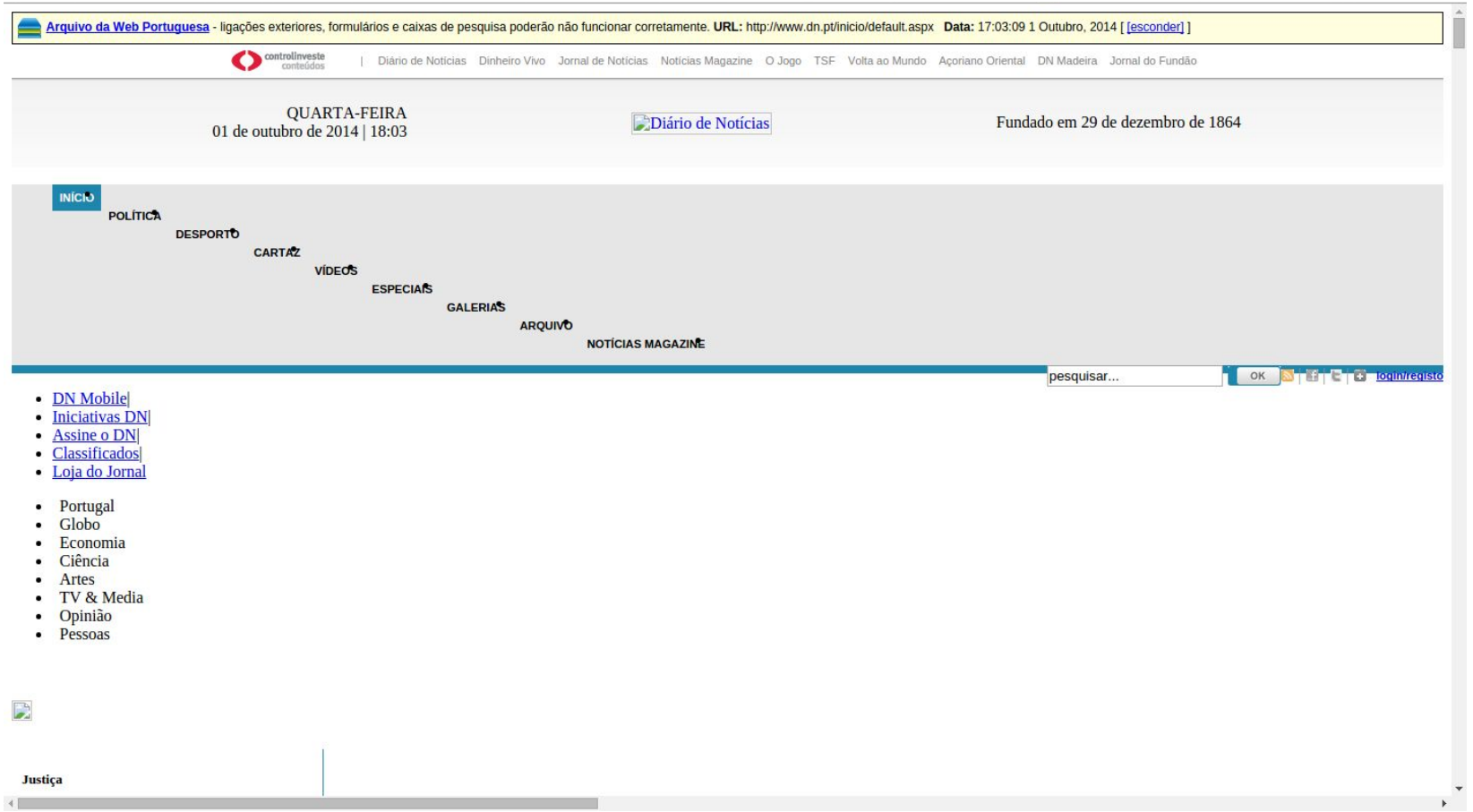
Protocolo de exclusão de Robôs: Robots.txt

Trata-se de um arquivo que, apesar da imponência do nome, *não é robô* e, na maioria das vezes, é de uma simplicidade impressionante.

http://www.dn.pt/

http://www.dn.pt

arquivada pelo Arquivo.pt



http://www.dn.pt/robots.txt

```
User-agent: *
Disallow: /common/scripts/
Disallow: /common/css/
Disallow: /admin/
Disallow: /search/
Sitemap: http://www.dn.pt/google_news.ashx
```

http://www.dn.pt

arquivada pelo Internet Archive

ARQUIVO.PT

Respeito pelos direitos de autor

ARQUIVO.PT

Para que serve Robots.txt

- Páginas protegidas por login;
- Páginas protegidas por formulários;
- Conteúdo repetidos;
- Informação privada.

Protocolo de exclusão de Robôs

- É importante que os autores autorizem a recolha de conteúdos importantes (para evitar problemas como o do http://www.dn.pt)
- robots.txt deverá estar na raiz do sítio web (ex. http://arquivo.pt/robots.txt).

Dicas para criar o seu Robots.txt

Permitir o arquivo pelo Arquivo.pt

User-agent: Arquivo-web-crawler

Disallow:

Controlar acessos consecutivos

User-agent: *

Disallow:

Crawl-delay: 100 # exige 100 segundos entre acessos

Proibir acesso a diretoria usando o robots.txt

User-agent: Arquivo-web-crawler

Disallow: /calendar/

Proibir a recolha e indexação usando a meta tag ROBOTS

<meta name="ROBOTS" content="NOINDEX, NOFOLLOW" />

```
<html>
<head>
<title></title>
<META NAME=”ROBOTS CONTENT=”NOINDEX,NOFOLLOW”>
</head>
```

Cuidado com os Robots.txt dos CMS's

Robots.txt do Wordpress por omissão

User-agent: *

Disallow: /wp-admin/

Disallow: /wp-includes/

Robots.txt do Joomla por omissão

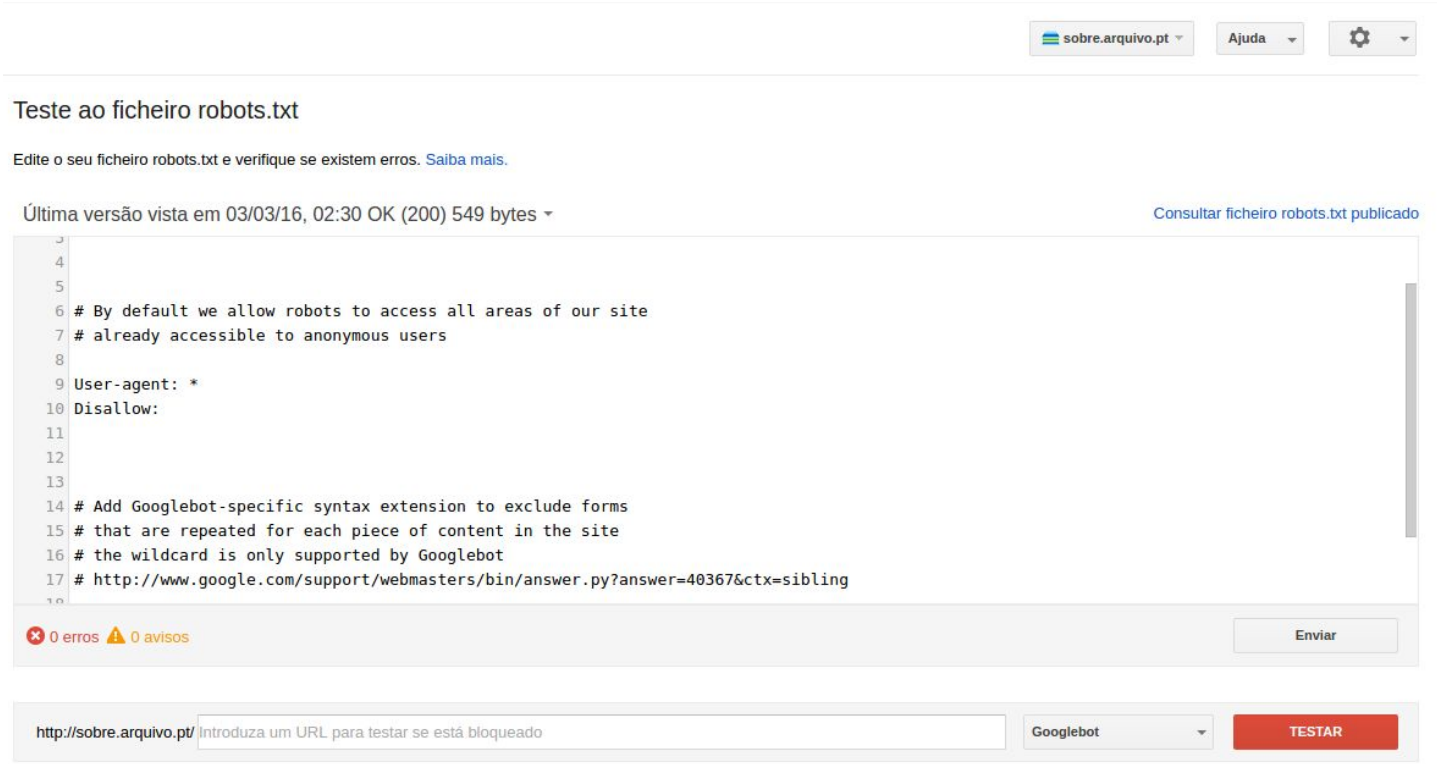
User-agent: *
Disallow: /administrator/
Disallow: /bin/
Disallow: /cache/
Disallow: /cli/
Disallow: /components/
Disallow: /includes/

Como testar o Robots.txt

https://www.google.com/webmasters/tools/robots-testing-tool

Google
robots.txt

com/webmasters/tools/robots-testing-tool



Exemplos de Robots.txt

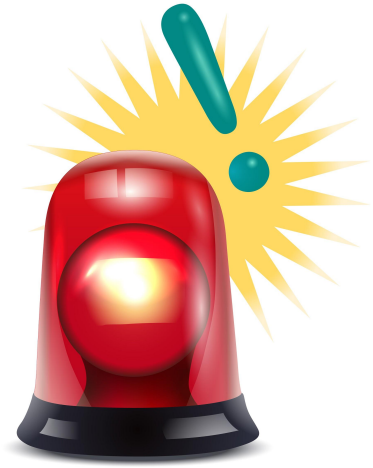
https://fccn.pt/robots.txt

User-agent: *
Allow: /

https://arquivo.pt/robots.txt

User-agent: *
Disallow: /nutchwax/search
Disallow: /search
Disallow: /wayback/
Disallow: /wayback/wayback/

Desafío: Verificar o Robots.txt



Não tenho o Robots.txt, *e agora?!*

O Arquivo.pt recolhe o seu conteúdo.

ARQUIVO.PT

Limitações impostas pelo Arquivo.pt

- tamanho máximo dos conteúdos descarregados da Web
- número de conteúdos por sítio
- número de ligações que o batedor percorre desde um endereço inicial até chegar a um conteúdo

Caso pretenda saber mais:

- http://sobre.arquivo.pt/colabore/recomendacoes-para-autores-de-sitios-web
- http://sobre.arquivo.pt/colabore/recomendacoes-para-autores-de-sitios-web/contacto

ARQUIVO.PT

Obrigado!

hugo.viana@fccn.pt