



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 publicar 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

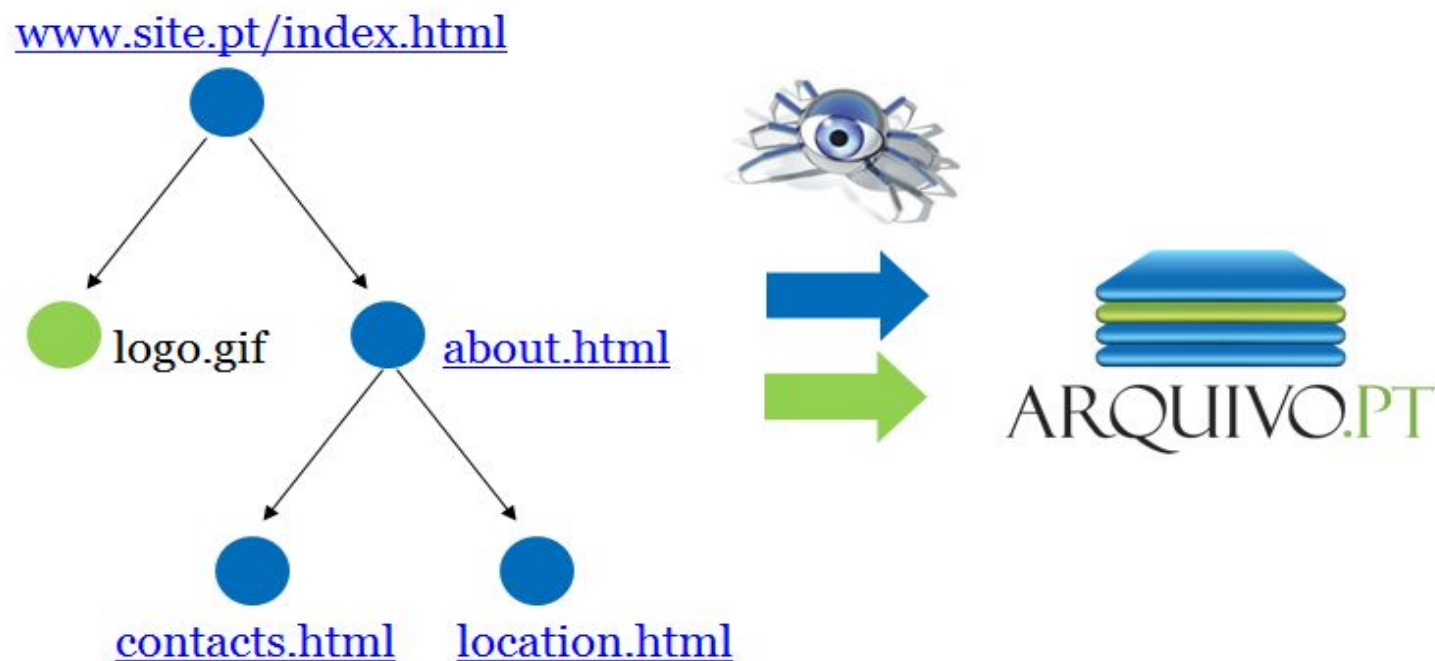
1. Batedor
2. Armazenamento
3. Indexador
4. Ordenador
5. Apresentador

Batedores

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

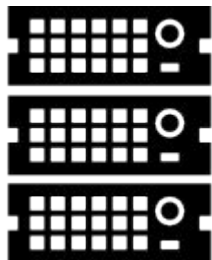


Informação é recolhida automaticamente



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.

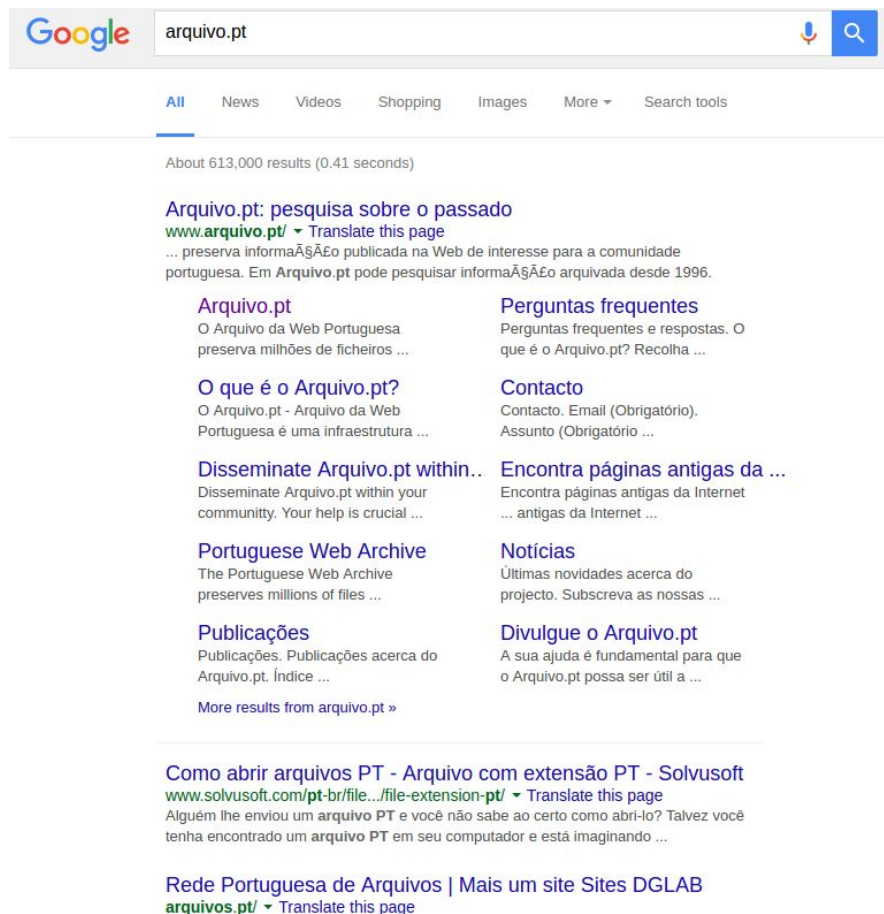


Apresentador

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



Google

[All](#) [News](#) [Videos](#) [Shopping](#) [Images](#) [More](#) [Search tools](#)

About 613,000 results (0.41 seconds)

Arquivo.pt: pesquisa sobre o passado
www.arquivo.pt/ [Translate this page](#)
... preserva informação publicada na Web de interesse para a comunidade portuguesa. Em Arquivo.pt pode pesquisar informação arquivada desde 1996.

Arquivo.pt O Arquivo da Web Portuguesa preserva milhões de ficheiros ...	Perguntas frequentes Perguntas frequentes e respostas. O que é o Arquivo.pt? Recolha ...
O que é o Arquivo.pt? O Arquivo.pt - Arquivo da Web Portuguesa é uma infraestrutura ...	Contacto Contacto. Email (Obrigatório). Assunto (Obrigatório) ...
Disseminate Arquivo.pt within.. Disseminate Arquivo.pt within your community. Your help is crucial ...	Encontra páginas antigas da ... Encontra páginas antigas da Internet ... antigas da Internet ...
Portuguese Web Archive The Portuguese Web Archive preserves millions of files ...	Notícias Últimas novidades acerca do projecto. Subscriba as nossas ...
Publicações Publicações. Publicações acerca do Arquivo.pt. Índice ...	Divulgue o Arquivo.pt A sua ajuda é fundamental para que o Arquivo.pt possa ser útil a ...

[More results from arquivo.pt »](#)

Como abrir arquivos PT - Arquivo com extensão PT - Solvusoft
www.solvusoft.com/pt-br/file.../file-extension-pt/ [Translate this page](#)
Alguém lhe enviou um arquivo PT e você não sabe ao certo como abri-lo? Talvez você tenha encontrado um arquivo PT em seu computador e está imaginando ...

Rede Portuguesa de Arquivos | Mais um site Sites DGLAB
arquivos.pt/ [Translate this page](#)



Esquema de motor de busca convencional

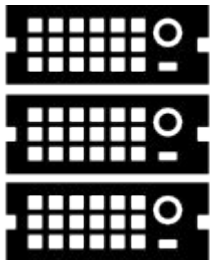


Componentes de um arquivo Web com motor de busca


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 da Web Portuguesa](#) - ligações exteriores, formulários e caixas de pesquisa poderão não funcionar corretamente. URL: <http://www.arquivo.pt/> Data: 15:49:54 12 Abril, 2011 [[esconder](#)]

 entre e
dd/mm/aaaa dd/mm/aaaa
[Pesquisa Avançada](#)
Experimental

Pesquise e aceda a páginas do passado

Veja ou reveja páginas que já desapareceram.

São mais de 130 milhões de páginas, arquivadas entre [1996](#) e [2010](#), que estão ao seu dispor.

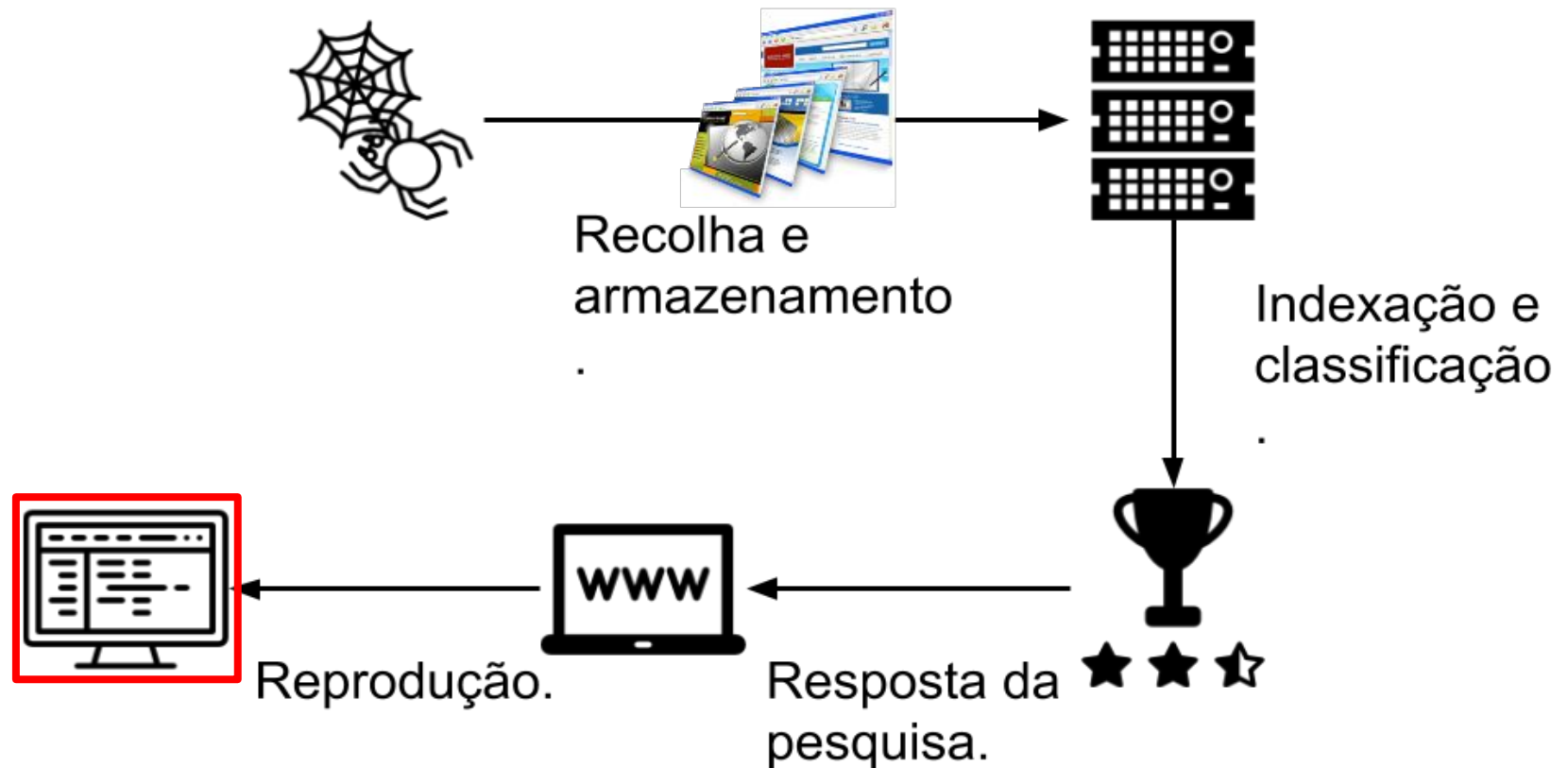
Exemplos de páginas arquivadas:

- [Expo 98](#)
- [Euro 2004](#)
- [Sapo \(1996\)](#)
- [Jornal "Público" \(1996\)](#)
- [Figo eleito melhor jogador do mundo \(2001\)](#)
- [Tim Berners-Lee: página pessoal \(1996\)](#)
- [Resultados de eleições \(2001\)](#)
- [José Saramago: página pessoal \(2000\)](#)

[Sobre o Arquivo](#) | [Termos & Condições](#) | [Tecnologias](#)




Esquema de um motor de busca para arquivos da Web



Apresentador do Arquivo.pt (pesquisa por URL)

PortuguêsEnglishAjuda



arquivo.pt

entre: 01/01/1996 e: 31/12/2015

Pesquisar

Pesquisa avançada

Pretendia ver os resultados das páginas que contêm o texto: <http://arquivo.pt?>


Versões da página web guardadas no arquivo

Foram gravadas no arquivo 296 versões da página <http://arquivo.pt> entre 1 Janeiro, 1996 e 30 Março, 2016.

1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
										25 Mai	3 Jun	20 Jan	1 Jan	30 Jul	19 Mai	
										25 Set	9 Jun	21 Jan	2 Jan	6 Ago	28 Mai	
										18 Dez	4 Ago	12 Abr	3 Jan	14 Set	29 Mai	
										18 Dez	5 Ago	13 Abr	4 Jan	24 Set	24 Set	
												14 Abr	5 Jan	15 Out	24 Set	
												15 Abr	6 Jan	16 Out	1 Out	
												16 Abr	20 Jan	5 Nov	12 Nov	
												17 Abr	20 Jan	6 Nov	3 Dez	
												18 Abr	21 Jan			
												19 Abr	31 Mar			
												20 Abr	16 Jul			
												21 Abr				

Apresentador do Arquivo.pt (pesquisa por termos)

PortuguêsEnglishAjuda



arquivo

x

Pesquisar

entre: 01/01/1996

e: 31/12/2015

Pesquisa avançada

Resultados 1 a 10 de 172.143.546

[302 Found](#)

[20 Maio, 2011](#) - outras datas

302 Found Found The document has moved here ...

<http://arquivo.co.pt/>

[Eu Não Desisto: abril 2004 Archives](#)

[8 Janeiro, 2010](#) - outras datas

.blogs.sapo.pt/arquivo/2004_04.html#128423 Posted by mauricio_102 at 02:46 PM | Comentários: (20 ... e RICOS -, todoS oS anoS. http://eunaodesisto.blogs.sapo.pt/arquivo/2004_04.html#128423 e - "U.M. ... Regras de Cálculo dos Subsídios Escolares. http://eunaodesisto.blogs.sapo.pt/arquivo/2004_05.html ...

http://eunaodesisto.blogs.sapo.pt/arquivo/2004_04.html

[Eu Não Desisto: abril 2004 Archives](#)

[13 Junho, 2010](#) - outras datas

.blogs.sapo.pt/arquivo/2004_04.html#128423 Posted by mauricio_102 at 02:46 PM | Comentários: (20 ... e RICOS -, todoS oS anoS. http://eunaodesisto.blogs.sapo.pt/arquivo/2004_04.html#128423 e - "U.M. ... Regras de Cálculo dos Subsídios Escolares. http://eunaodesisto.blogs.sapo.pt/arquivo/2004_05.html ...

http://eunaodesisto.blogs.sapo.pt/arquivo/2004_04.html



Como publicar 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](#)

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

País > Lisboa > Lisboa

Taxa para entradas no aeroporto de Lisboa limitada a voos internacionais

11.11.2014 - 23:17

A taxa turística para entradas no aeroporto de Lisboa será limitada aos voos internacionais, afirmou esta terça-feira o vice-presidente da autarquia, Fernando Medina.



PUB

Recolhida em 12/11/2014

Fazemos Bem SETEMBRO A DEZEMBRO SETOR TERCIÁRIO

BT Edições Multimédia

Ocasão/Zaask - Destaque 300x100 JN (Cartão)

JN Descontos Natal 300x100

Volta ao Mundo Novembro 300x100

Brasil goleia Turquia em Istambul

Empresa investigada por surto de legionela garante cumprir a lei

Manter o mesmo endereço ao longo do tempo

PSD2011.com (2011)



PSD2011.com (2014)



Outras recomendações: *Robots.txt*

Robots.txt ?! Para que serve?!



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/>

[Diário de Notícias](#)[Jornal de Notícias](#)[TSF](#)[O Jogo](#)[Dinheiro Vivo](#)[Volta ao Mundo](#)[Delas](#)[Classificados](#)[ASSINAR](#)[LOGIN QUIOSQUE](#)[▶ OUVIR RÁDIO](#)

Terça-Feira | 15 de março de 2016 | 11:28 | Fundado em 29 de dezembro de 1864

☰ Diário de Notícias



◉ ÚLTIMA HORA EXPLOÇÃO DE CARRO EM BERLIM FAZ UM MORTO

Esquerda quer menos alunos por turma para reduzir indisciplina

Num grupo de 4200 alunos houve mais de 9 mil expulsões no ano passado. PS, BE e PCP de acordo: solução é ter menos crianças por...

- ↳ Manuais vão ser dados aos alunos. Mas é só um empréstimo por...
- ↳ Orçamento (real) da Educação ultrapassa os 6019 milhões

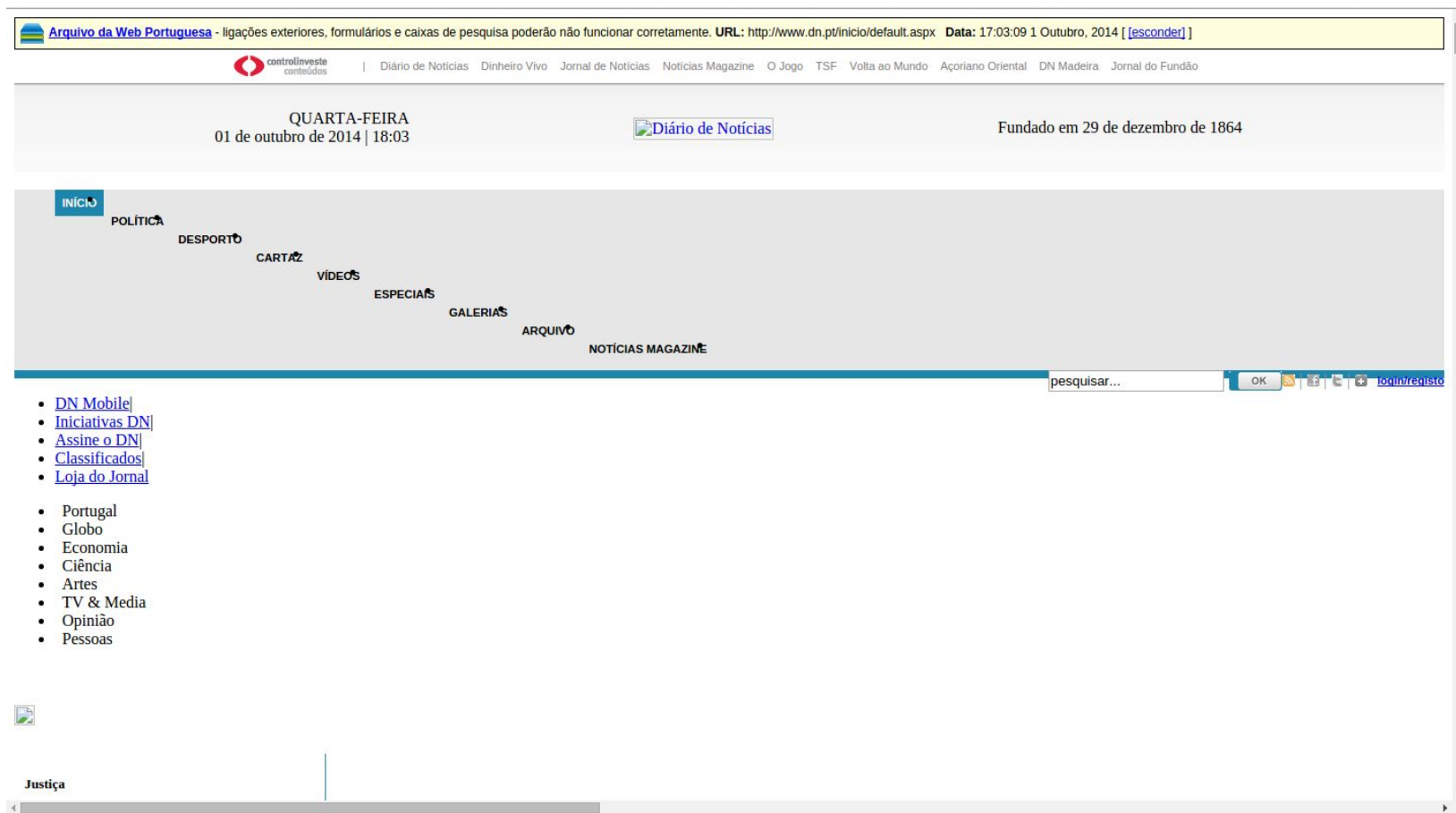


PUB

ASSINAR DN

<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

INTERNET ARCHIVE
Wayback Machine

1.181 captures
1 jul 11 - 29 mar 16

http://www.dn.pt/inicio/default.aspx Go

ABR MAI JUN
26
2013 2014 2015

Close X
Help ?

controlinveste conteúdos

Diário de Notícias Dinheiro Vivo Jornal de Notícias Notícias Magazine O Jogo TSF Volta ao Mundo Açoriano Oriental DN Madeira Jornal do Fundão

SEGUNDA-FEIRA
26 de maio de 2014 | 02:57

Diário de Notícias
Fundado em 29 de dezembro de 1864

AUMENTO DE CAPITAL BANIF 2014

NÃO DEIXE PASSAR O MOMENTO.

AUMENTO DE CAPITAL BANIF 2014
NÃO DEIXE PASSAR O MOMENTO
www.banif.pt
808 200 200
Dias úteis das 7h às 1h
Sábados das 10h às 20h
Não dispensa a consulta do prospecto de oferta disponível em www.cmvm.pt e www.banif.pt
BANIF
A força de acreditar

CDU

PS 31,45%	A. Portugal 27,71%	CDU 12,68%	MPT 7,15%	BE 4,56%	Outros Partidos 8,98%	Abstenção 66,09%
7 Mandatos 1032143 Votos	6 Mandatos 609263 Votos	2 Mandatos 416102 Votos	1 Mandato 234516 Votos	1 Mandato 149546 Votos	0 Mandatos 294538 Votos	6396510

ACEDA À PÁGINA ESPECIAL DAS ELEIÇÕES

Marcelo fala em "sova monumental" e "derrota histórica" dos partidos da coligação

EUROPEIAS

António Costa diz que vitória do PS "soube a pouco"

26 maio 2014

O dirigente socialista António Costa considerou esta noite que o PS obteve nas

Resultados Nacionais Provisórios Fonte: DGAJ

Inscritos	9677954	
Votantes	3281444	33,91%
Abstenções	6396510	66,09%
Branco	144832	4,41%
Nulos	100484	3,06%
Freguesias Apuradas	3092	
Freguesias por Apurar	0	
Consulados Apurados	54	
Consulados por Apurar	17	

Ver resultados em detalhe

Subscreva o Aumento de Capital Banif 2014

- > Montante total: até €138.504.779,57
- > Período de subscrição: 16 a 30 de Maio 2014
- > Preço de subscrição: €0,01
- > Destinado ao público em geral, com alocação prioritária a Accionistas

Não dispensa a consulta do prospecto de oferta disponível em www.cmvm.pt e www.banif.pt
BANIF
A força de acreditar

Waiting for web.archive.org...

Respeito pelos direitos de autor



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

sobre.arquivo.pt

Ajuda

Teste ao ficheiro robots.txt

Edite o seu ficheiro robots.txt e verifique se existem erros. [Saiba mais.](#)

Última versão vista em 03/03/16, 02:30 OK (200) 549 bytes

[Consultar ficheiro robots.txt publicado](#)

```
3
4
5
6 # By default we allow robots to access all areas of our site
7 # already accessible to anonymous users
8
9 User-agent: *
10 Disallow:
11
12
13
14 # Add Googlebot-specific syntax extension to exclude forms
15 # that are repeated for each piece of content in the site
16 # the wildcard is only supported by Googlebot
17 # http://www.google.com/support/webmasters/bin/answer.py?answer=40367&ctx=sibling
18
```

0 erros 0 avisos

Enviar

http://sobre.arquivo.pt

Introduza um URL para testar se está bloqueado

Googlebot

TESTAR

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.



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