Ciclo de Webinars com o Arquivo.pt

- Webinar 1 - **5 de junho**, das 15 às 16h30 - Arquivo.pt: uma nova ferramenta para pesquisar o passado. **Bem publicar, para bem preservar**
- Webinar 2 - **19 de junho**, das 15 às 16h30 - Acesso e processamento automático de informação preservada da Web através de APIs
- Webinar 3 - **3 de julho**, das 15 às 16h30 - Arquivar a Web: faça-você-mesmo!

POLITÉCNICO DE LISBOA

ARQUIVO.PT

FCCN serviços digitais fct

fct Fundação para a Ciência e a Tecnologia

Bem publicar para bem preservar

Pedro Gomes, responsável pelas recolhas
dos websites no Arquivo.pt

05 de Junho de 2024

pedro.gomes@fccn.pt

Funcionamento do Arquivo.pt

Preservar conteúdos da Web implica **recolher**, **armazenar** e **reproduzir**

1

Recolher

ARQUIVO.PT

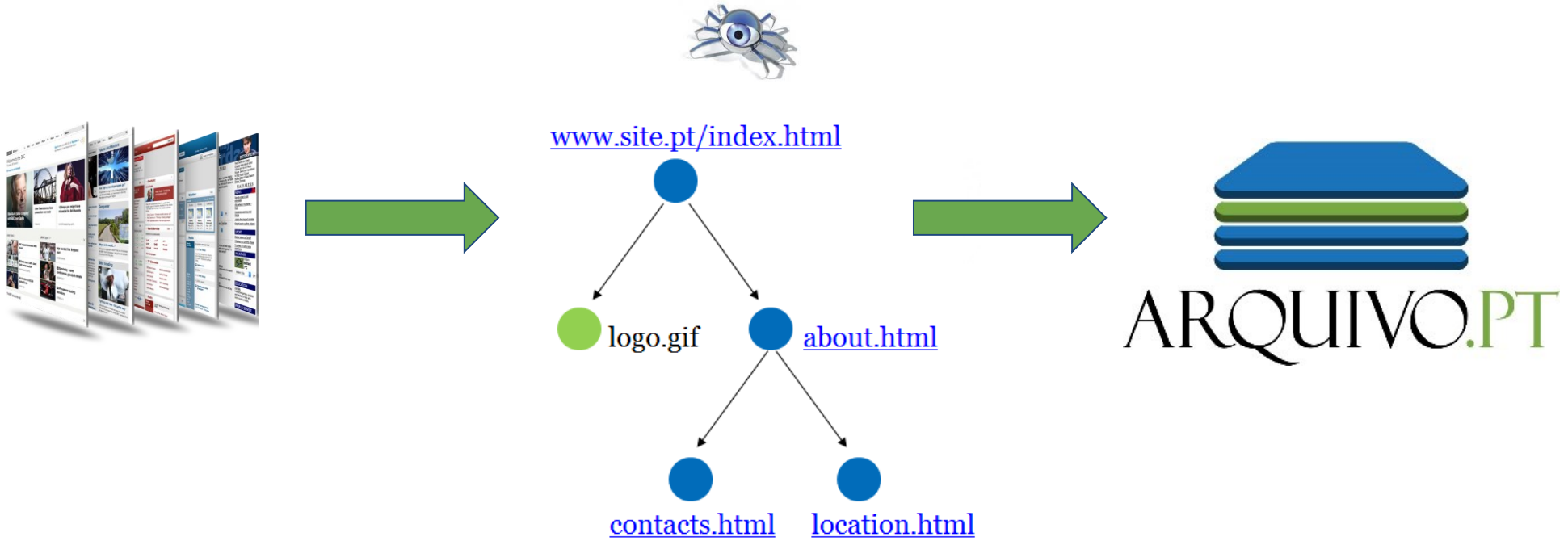
2

Armazenar

3

Reproduzir

Recolha é feita de forma automática por um Robot



Reprodução de páginas preservadas

Nem sempre corre bem ☹

7

recomendações para **publicar**
informação **preservável**

Identifique corretamente a data de publicação **(#1)**

Qual a data de *publicação*?

"ATUALIZADO"?!

Data de *arquivo*

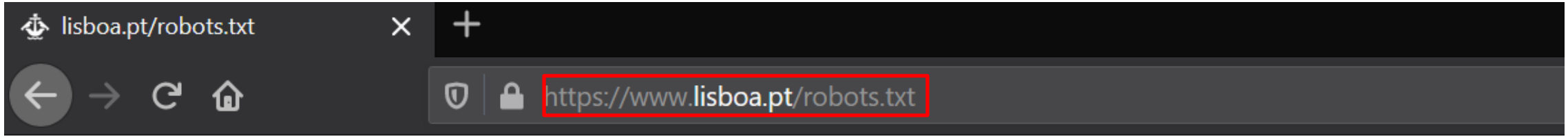
Data da *página*

Ataque ocorreu no dia 7 de janeiro de 2015

Identificar corretamente a **data de publicação** (#1)

Use corretamente o protocolo de exclusão de robots **(#2)**

Protocolo de exclusão de robots




```
User-agent: *
Disallow: /typo3/
Disallow: /typo3_src/
Allow: /typo3/sysext/frontend/Resources/Public/*
Sitemap: https://www.lisboa.pt/?sitemap=pages&type=1533906435&cHash=751dcadd4c80d443b2fd8be5388fa89e
```

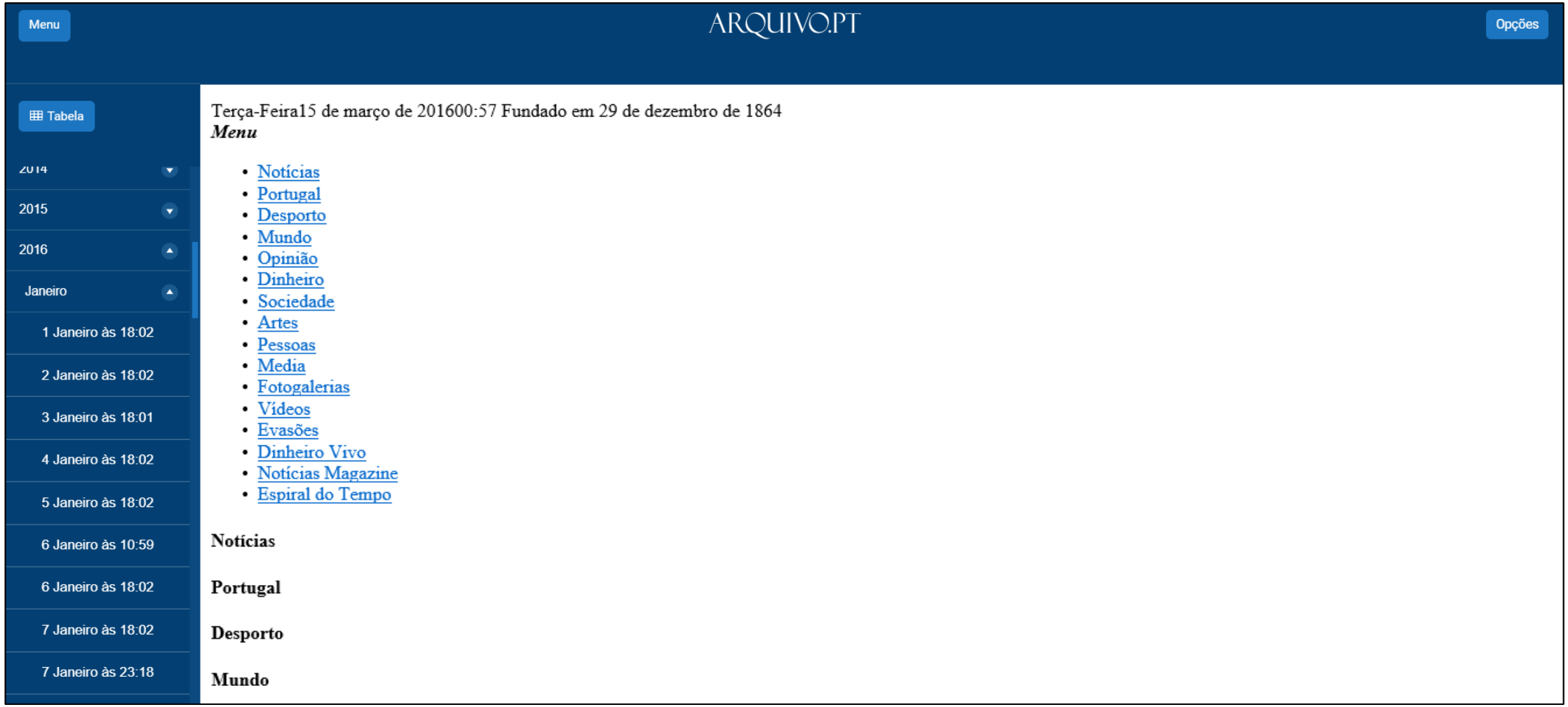
1

2

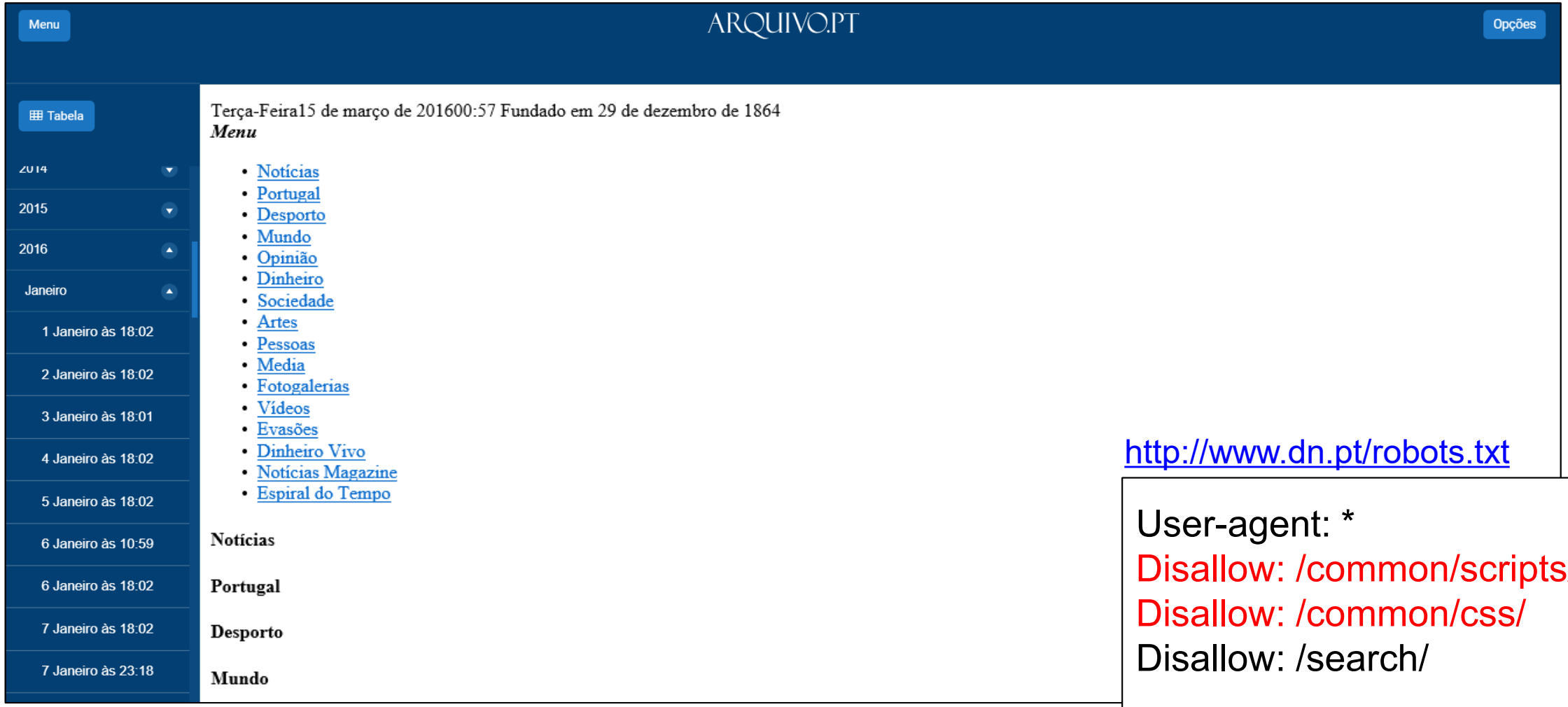
lisboa.pt

dn.pt : como era (2016)

dn.pt : como foi preservado



Protocolo de exclusão de robots
é a origem do problema de preservação



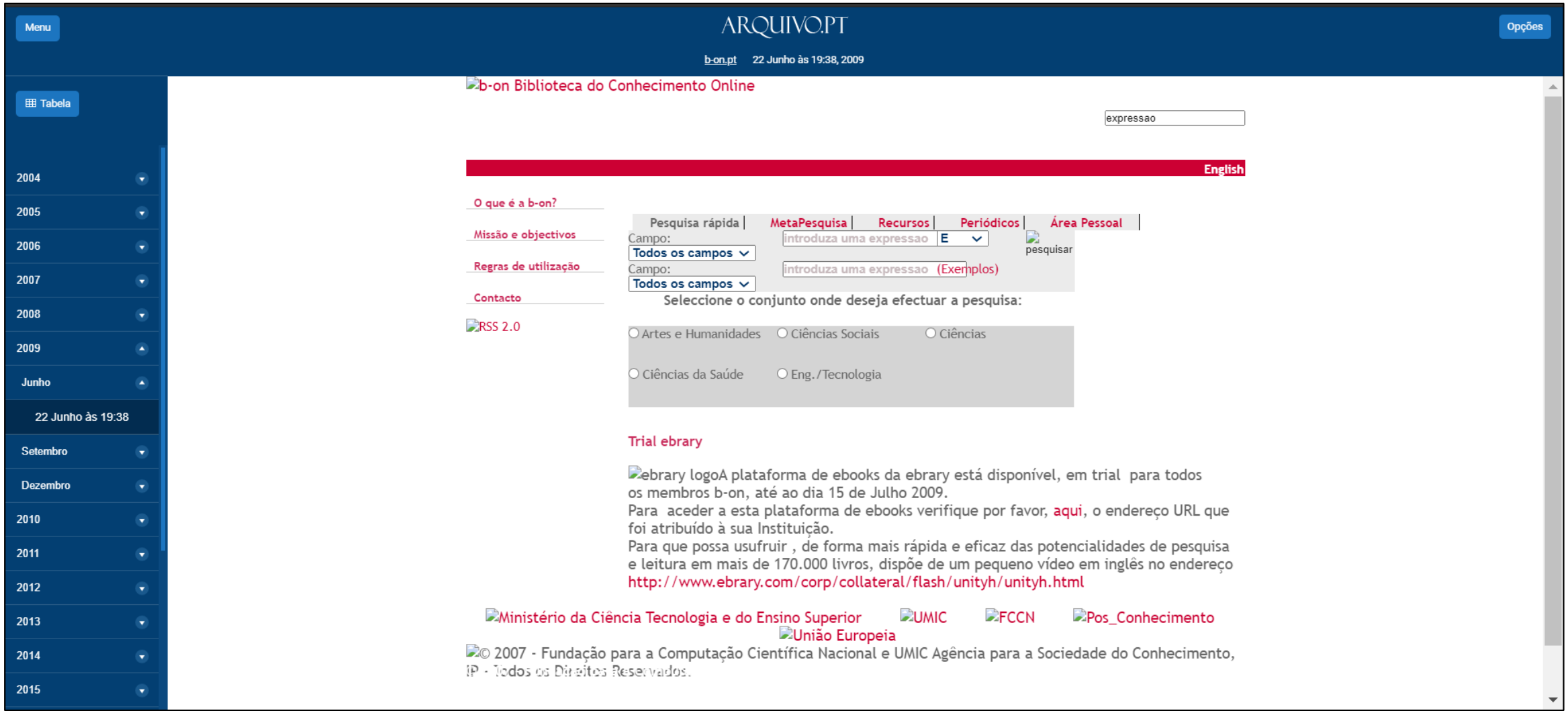
http://www.dn.pt/robots.txt

User-agent: *
Disallow: /common/scripts/
Disallow: /common/css/
Disallow: /search/

“**Não permitir** a recolha dos ficheiros de **javascript** ou **css** através do ficheiro robots.txt **afeta diretamente a renderização e indexação dos conteúdos do site** o que pode resultar num **baixo ranking de pesquisa**.”

https://webmasters.googleblog.com/2014/10/updating-our-technical-webmaster.html?m=1

b-on.pt : como foi preservado



Exclusões pré-definidas pelos Sistemas de Gestão de Conteúdos (CMS) causaram problemas

User-agent: *
Disallow: /administrator/
Disallow: /cache/
Disallow: /components/
Disallow: /editor/
Disallow: /help/
Disallow: /images/
Disallow: /includes/
Disallow: /language/
Disallow: /mambots/
Disallow: /media/
Disallow: /modules/
Disallow: /templates/
Disallow: /installation/
Disallow: /dmdocuments/

b-on.pt : Protocolo de exclusão de robots configurado corretamente

Permitir o acesso ao Arquivo.pt

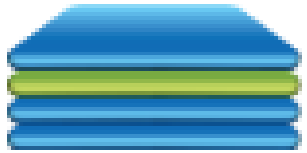
O Sistema de Recolha do Arquivo.pt está devidamente identificado.

```
User-agent: Arquivo-web-crawler
Disallow:
```

Utilize um endereço para cada conteúdo **(#3)**

Conteúdos escondidos atrás de formulários

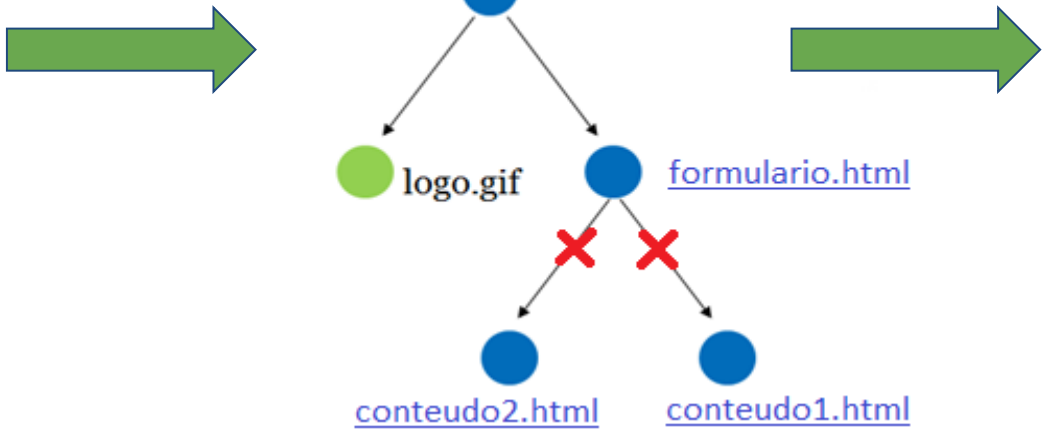
Conteúdos escondidos atrás de formulários



Não Arquivado

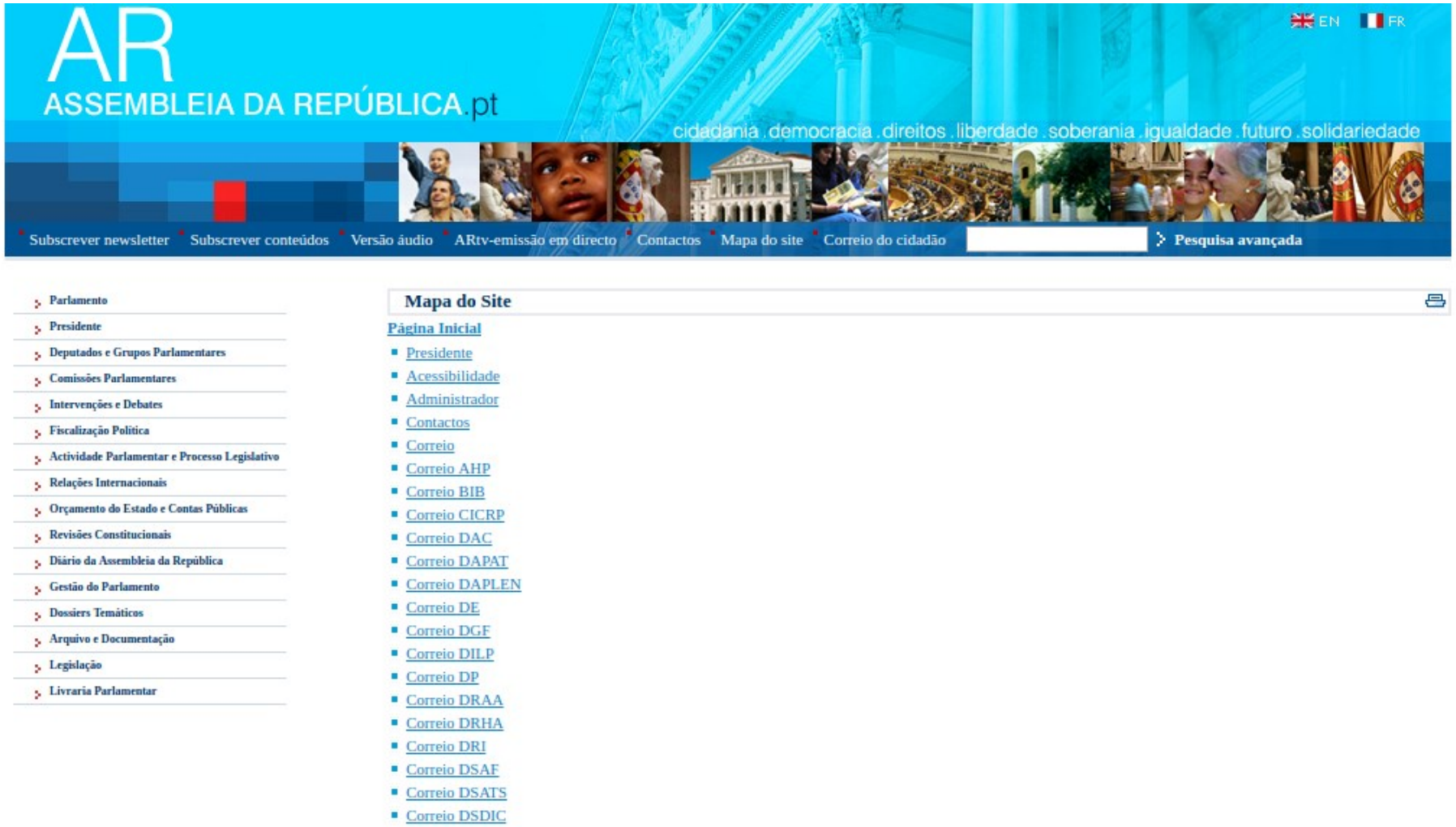
Pesquisar noutros arquivos

Conteúdos escondidos atrás de formulários



ARQUIVO.PT

Mapa do Site facilita acesso a pessoas (usabilidade) e máquinas (SEO)



Mantenha o mesmo endereço ao longo do tempo **(#4)**

Problema: quebra de histórico devido a mudança de endereço do site

Tabela de versões

http://site.pt

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
22 Jan	28 Jan	3 Fev	7 Jan	13 Jan	4 Jan	6 Jan	20 Mai					
25 Mar	2 Fev	21 Mar	12 Jan	2 Fev	9 Jan	18 Jan	20 Mai					

Tabela de versões

http://novo-site.pt

2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
								6 Jun	21 Jan	21 Jan	5 Nov	26 Set
								10 Jun	22 Jan	22 Jan	7 Nov	30 Set
								4 Aug	20 Mai			27 Out
								5 Aug	21 Mai			
									2 Jul			

Problema: quebra de histórico devido a mudança de endereço do site

Cavaco insatisfeito em parte com 2007
Presidente pretende melhores resultados na Economia, Educação e Justiça (SIC)

O Presidente da Republica, Cavaco Silva, afirmou-se insatisfeito com os resultados obtidos em 2007 na Economia, Educação e Justiça e fez um apelo ao diálogo do Governo para "reduzir a conflitualidade e tensões" em 2008.

Cavaco Silva faz avisos e elogios ao Governo

DD | DN | JN | Lusa | SIC | SOL | TSF

Quénia
Comissão eleitoral pressionada a anunciar reeleição de Kibaki - presidente de organismo (Lusa)

Há mais dois turistas portugueses no país, mas estão em segurança

"Isto não é o faroeste"
Autarquia de Gaia reage a morte de jovem vítima de tiro disparado em festejos de Ano Novo (SIC)

Morte em Gaia

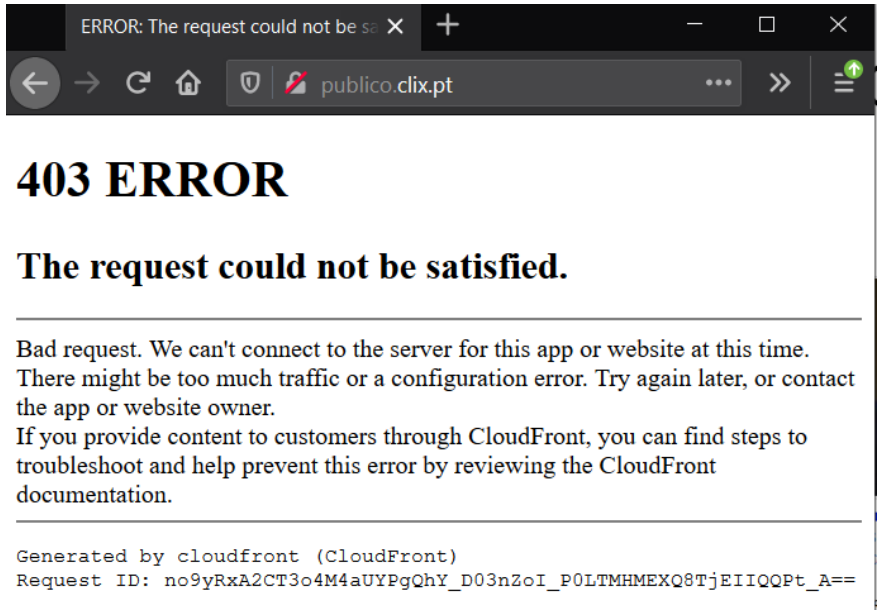
Colômbia
Familiares de Clara fazem testes ADN para confirmar consanguinidade de Emmanuel (Lusa)

Peniche
Veleiro incendiou-se e naufragou, tripulantes estão a salvo (Lusa)

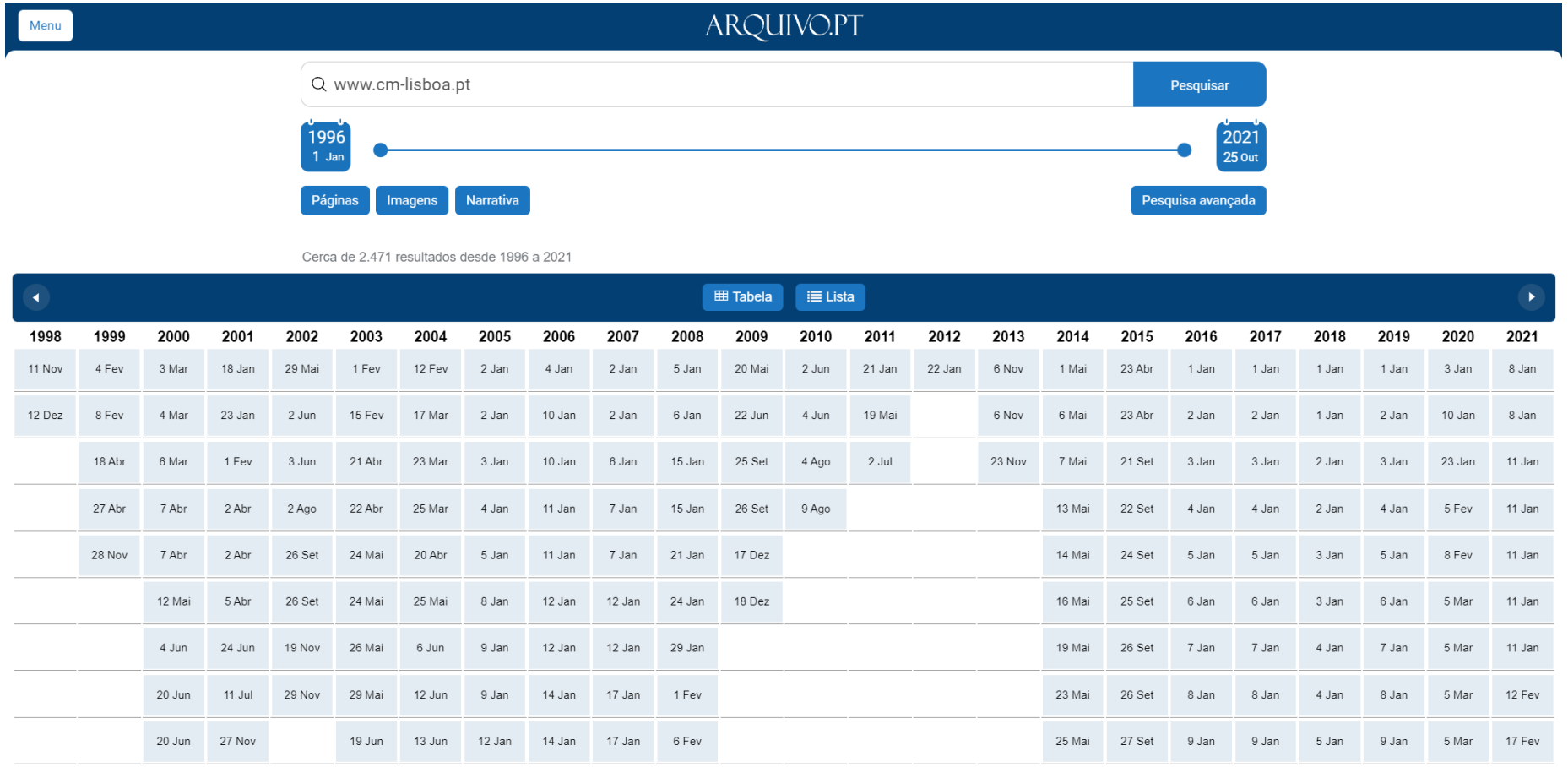
Foto do dia

mundo

publico.clix.pt



Mantenha o histórico redirecionando os endereços antigos para os novos



1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
11 Nov	4 Fev	3 Mar	18 Jan	29 Mai	1 Fev	12 Fev	2 Jan	4 Jan	2 Jan	5 Jan	20 Mai	2 Jun	21 Jan	22 Jan	6 Nov	1 Mai	23 Abr	1 Jan	1 Jan	1 Jan	1 Jan	3 Jan	8 Jan
12 Dez	8 Fev	4 Mar	23 Jan	2 Jun	15 Fev	17 Mar	2 Jan	10 Jan	2 Jan	6 Jan	22 Jun	4 Jun	19 Mai		6 Nov	6 Mai	23 Abr	2 Jan	2 Jan	1 Jan	2 Jan	10 Jan	8 Jan
	18 Abr	6 Mar	1 Fev	3 Jun	21 Abr	23 Mar	3 Jan	10 Jan	6 Jan	15 Jan	25 Set	4 Ago	2 Jul		23 Nov	7 Mai	21 Set	3 Jan	3 Jan	2 Jan	3 Jan	23 Jan	11 Jan
	27 Abr	7 Abr	2 Abr	2 Ago	22 Abr	25 Mar	4 Jan	11 Jan	7 Jan	15 Jan	26 Set	9 Ago				13 Mai	22 Set	4 Jan	4 Jan	2 Jan	4 Jan	5 Fev	11 Jan
	28 Nov	7 Abr	2 Abr	26 Set	24 Mai	20 Abr	5 Jan	11 Jan	7 Jan	21 Jan	17 Dez					14 Mai	24 Set	5 Jan	5 Jan	3 Jan	5 Jan	8 Fev	11 Jan
		12 Mai	5 Abr	26 Set	24 Mai	25 Mai	8 Jan	12 Jan	12 Jan	24 Jan	18 Dez					16 Mai	25 Set	6 Jan	6 Jan	3 Jan	6 Jan	5 Mar	11 Jan
		4 Jun	24 Jun	19 Nov	26 Mai	6 Jun	9 Jan	12 Jan	12 Jan	29 Jan						19 Mai	26 Set	7 Jan	7 Jan	4 Jan	7 Jan	5 Mar	11 Jan
		20 Jun	11 Jul	29 Nov	29 Mai	12 Jun	9 Jan	14 Jan	17 Jan	1 Fev						23 Mai	26 Set	8 Jan	8 Jan	4 Jan	8 Jan	5 Mar	12 Fev
		20 Jun	27 Nov		19 Jun	13 Jun	12 Jan	14 Jan	17 Jan	6 Fev						25 Mai	27 Set	9 Jan	9 Jan	5 Jan	9 Jan	5 Mar	17 Fev

cm-lisboa.pt → lisboa.pt

Redireciona

Mantenha os antigos domínios

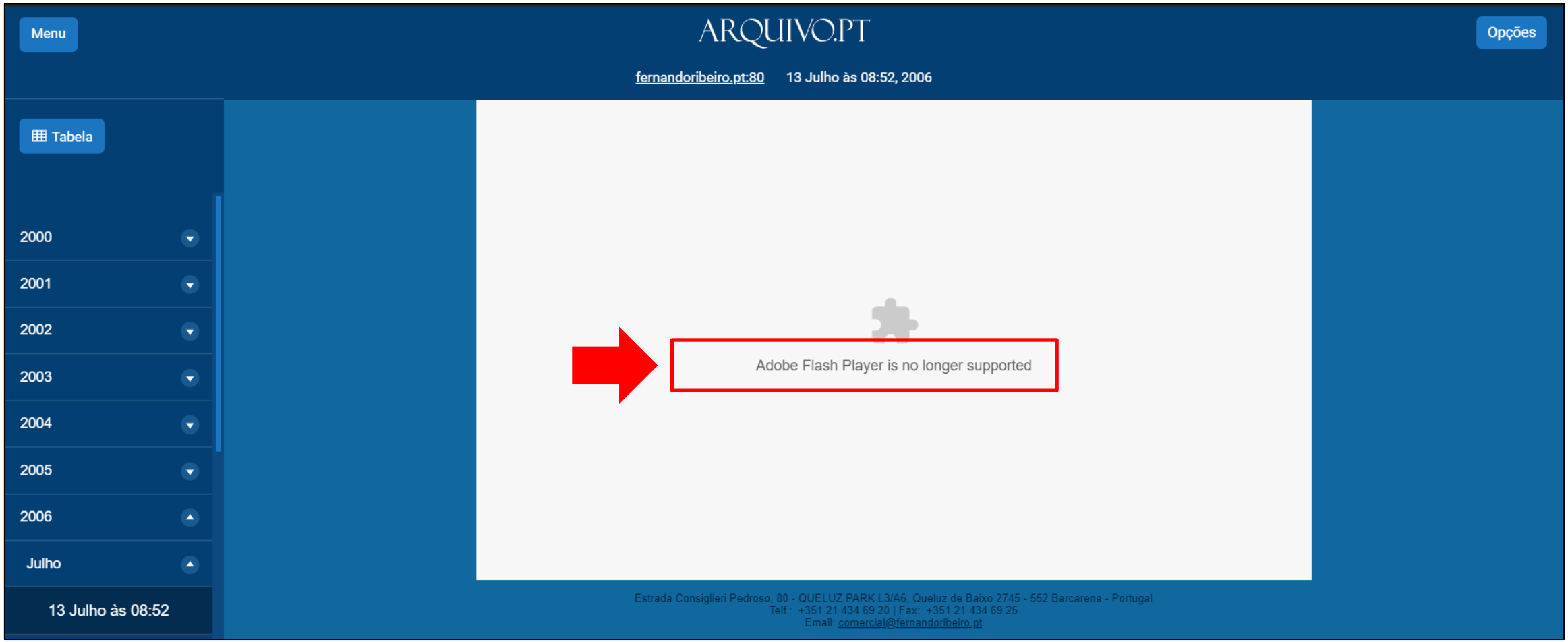
2005@Arquivo.pt

vivapraia.com

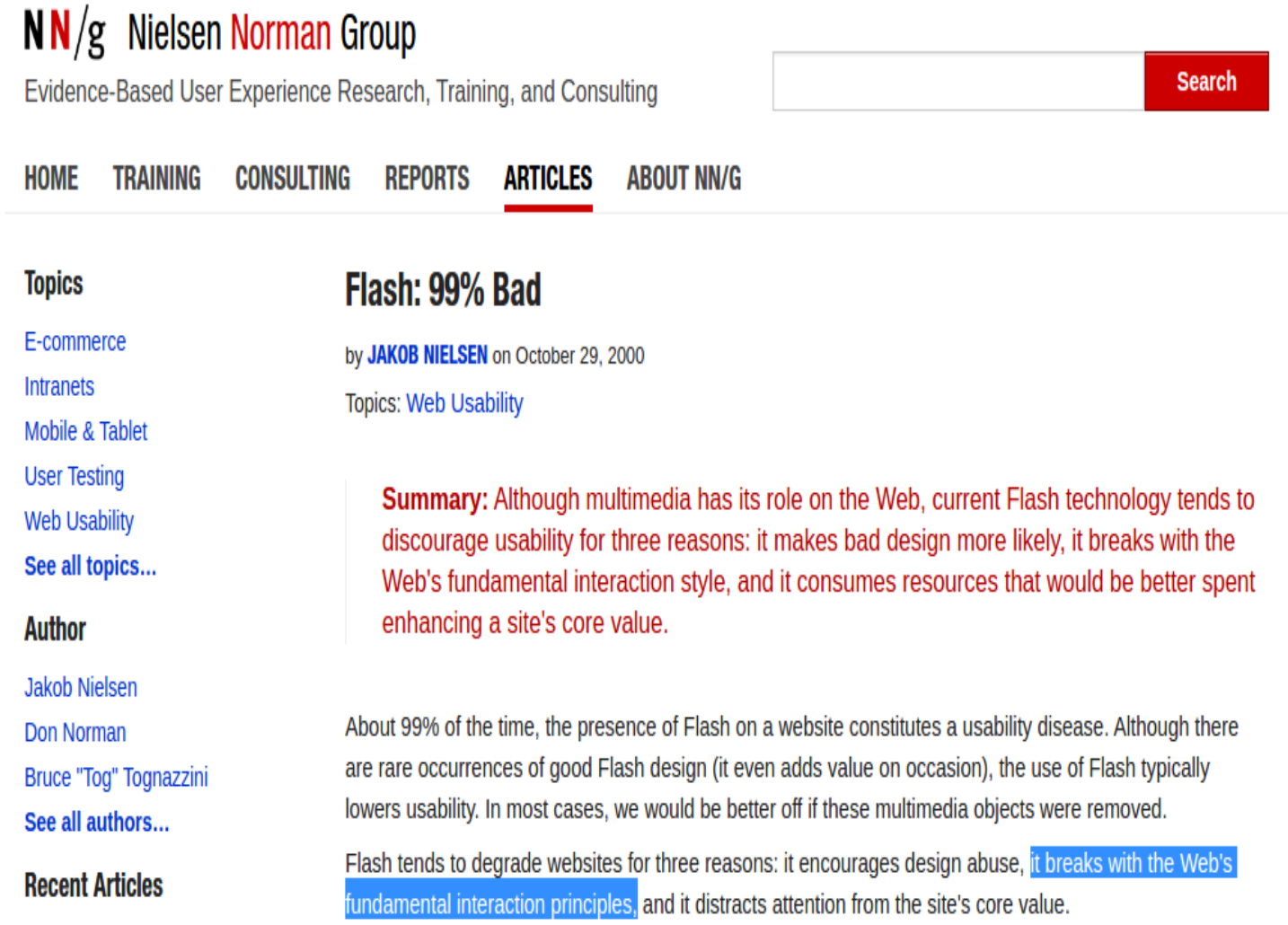
2021@Arquivo.pt

Utilize formatos adequados para preservação **(#5)**

Problema: Utilização de Flash



“It **breaks** with the Web’s **fundamental** interaction principles”



NN/g Nielsen Norman Group

Evidence-Based User Experience Research, Training, and Consulting

Search

HOME TRAINING CONSULTING REPORTS ARTICLES ABOUT NN/G

Topics

E-commerce
Intranets
Mobile & Tablet
User Testing
Web Usability
See all topics...

Author

Jakob Nielsen
Don Norman
Bruce "Tog" Tognazzini
See all authors...

Recent Articles

Flash: 99% Bad

by JAKOB NIELSEN on October 29, 2000

Topics: Web Usability

Summary: Although multimedia has its role on the Web, current Flash technology tends to discourage usability for three reasons: it makes bad design more likely, it breaks with the Web's fundamental interaction style, and it consumes resources that would be better spent enhancing a site's core value.

About 99% of the time, the presence of Flash on a website constitutes a usability disease. Although there are rare occurrences of good Flash design (it even adds value on occasion), the use of Flash typically lowers usability. In most cases, we would be better off if these multimedia objects were removed.

Flash tends to degrade websites for three reasons: it encourages design abuse, it breaks with the Web's fundamental interaction principles, and it distracts attention from the site's core value.

Escolha formatos com:

- Condições de **licenciamento** que permitam a sua utilização.
- **Normas** emitidas por um organismo oficial (W3C).
- Documentados abertamente através de uma **especificação pública.**
- Lidos e escritos por múltiplas plataformas de software, incluindo código-aberto.
- Amplamente usados.

Formatos **adequados** para preservação

Texto
HTML, XHTML ou XML
Open Document Text (.odt)
PDF/A-1 segundo a norma ISO 19005-1 (.pdf)

Imagem
PNG (.png)
JPEG2000

Video
AVI sem compressão (.avi)
MP4

Formatos **não adequados** para preservação

Texto:
Microsoft Word (.doc)

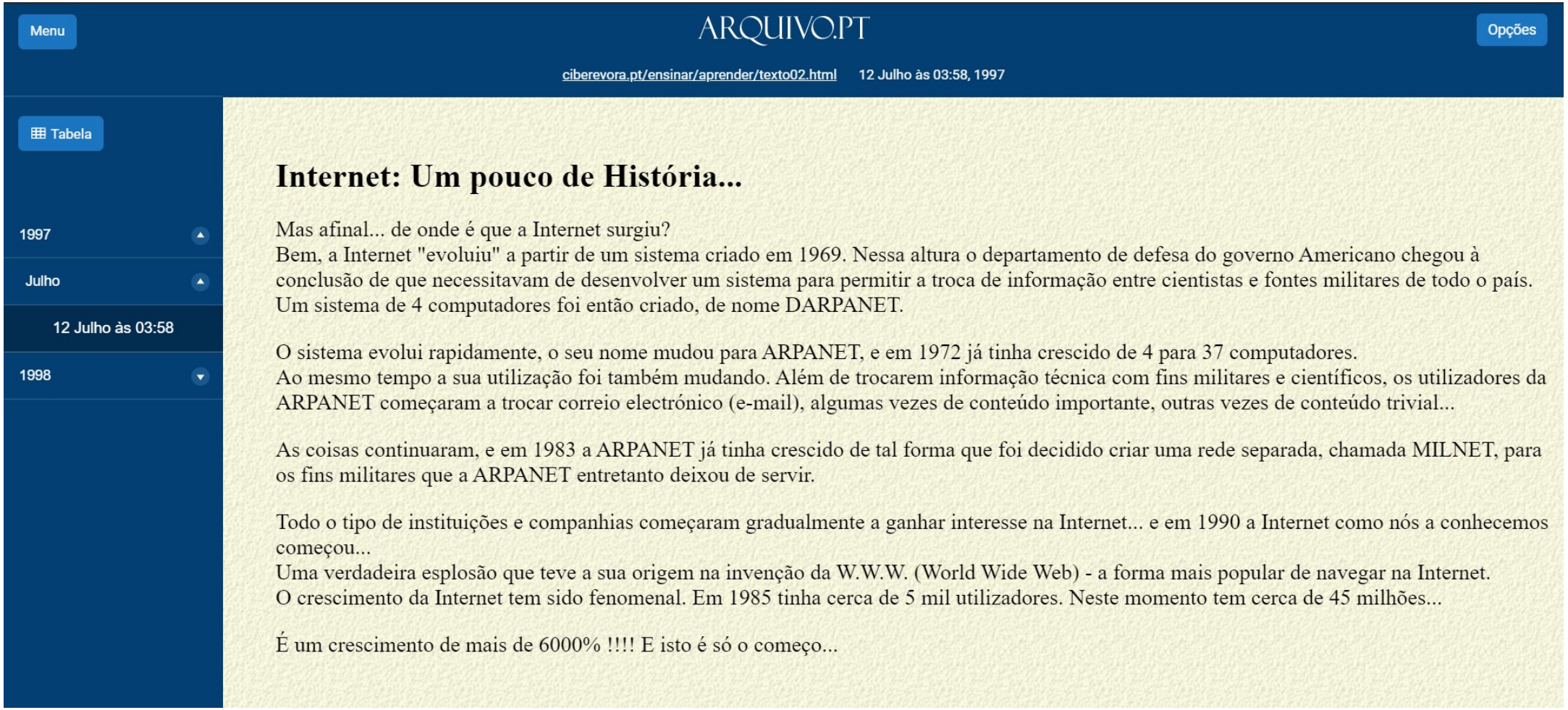
Imagem:
Macromedia Flash (*.swf)
PhotoShop (.psd)

Vídeo:
Windows Media Video (.wmv)

Utilize meta-dados para descrever os conteúdos **(#6)**

Quem é o *autor* da página?

Qual a *data da publicação*?

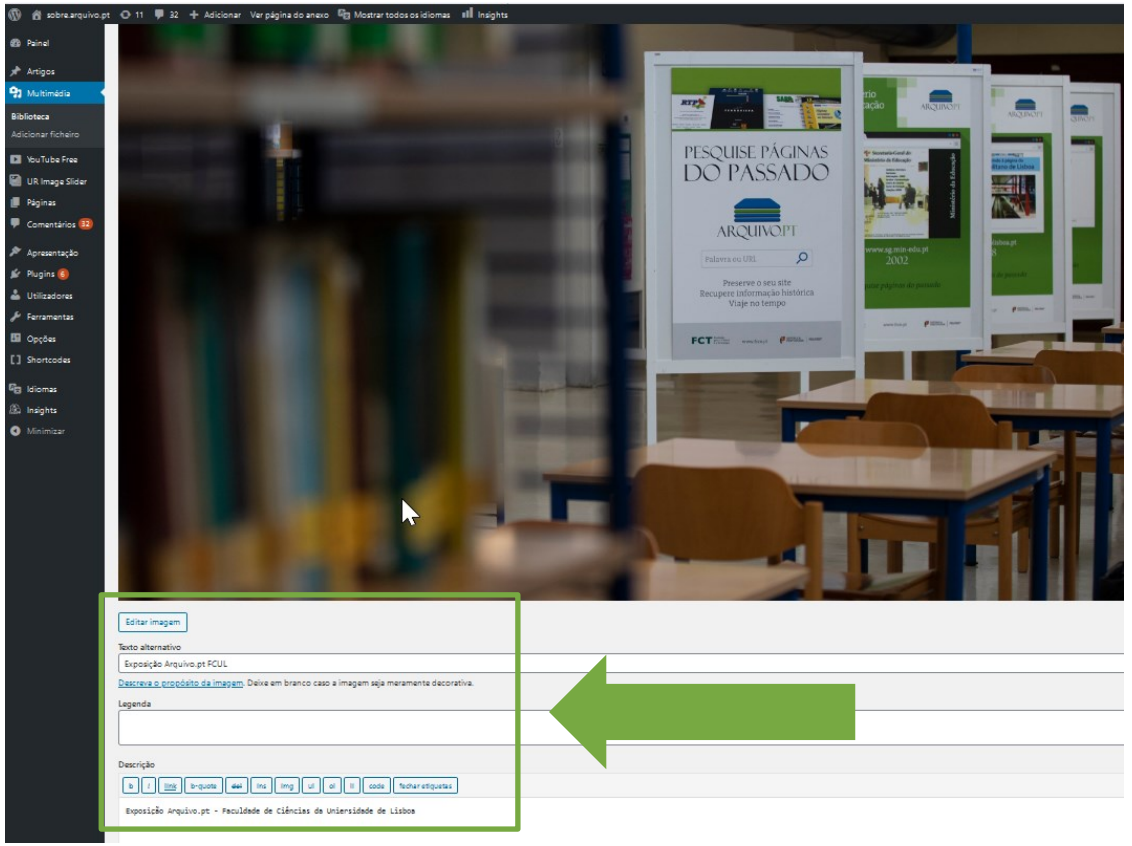


Utilizar meta-dados (**Dublin Core**)

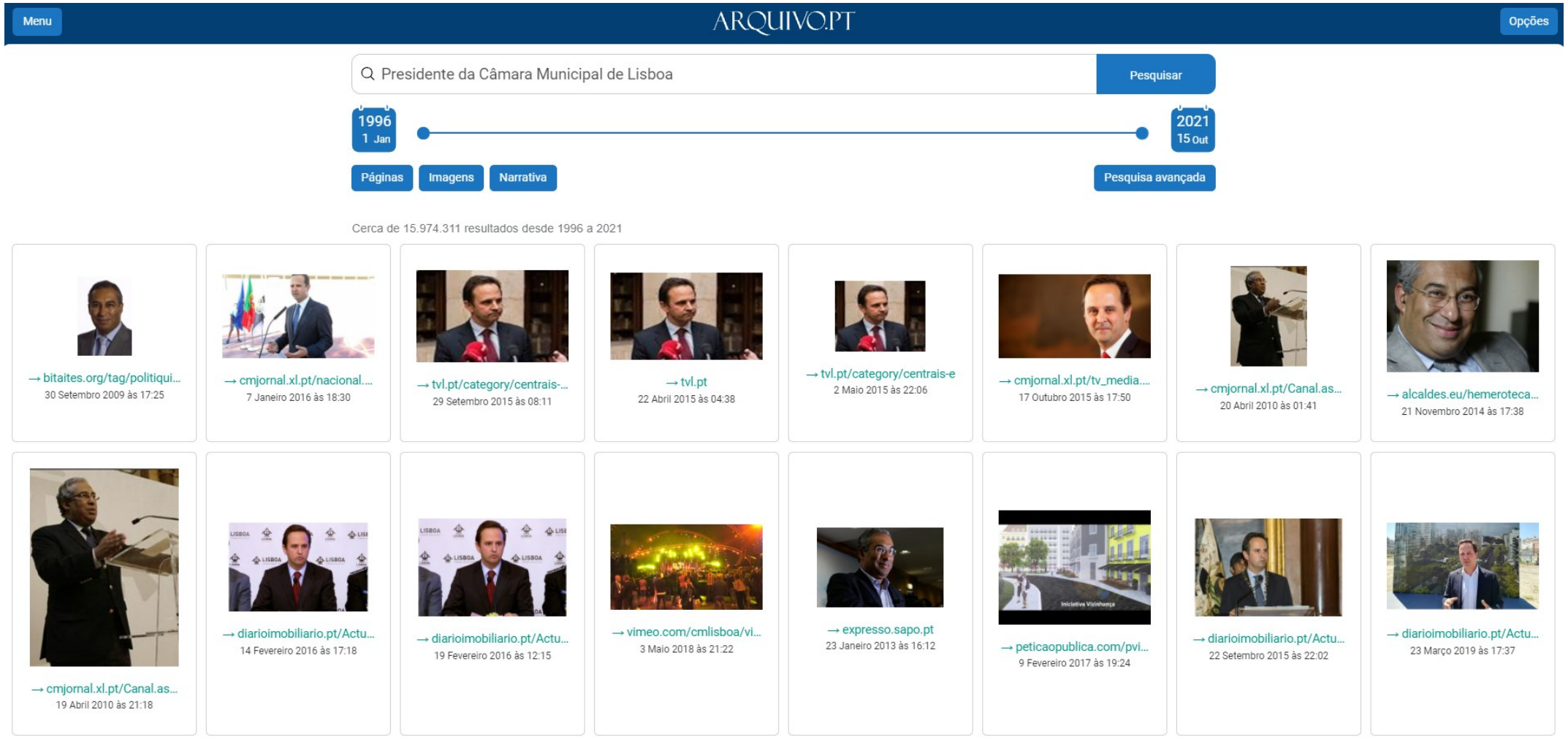
```
<meta name="DC.Type" content="Text" />
<meta name="DC.Creator" content="Daniel Gomes" />
<meta name="DC.Date.Created" content="2009-08-21" />
<meta name="DC.Date.Modified" content="2009-11-10" />
```

Esta informação resume, **enriquece** ou complementa os conteúdos, produzindo assim um potencial incremento de informação.

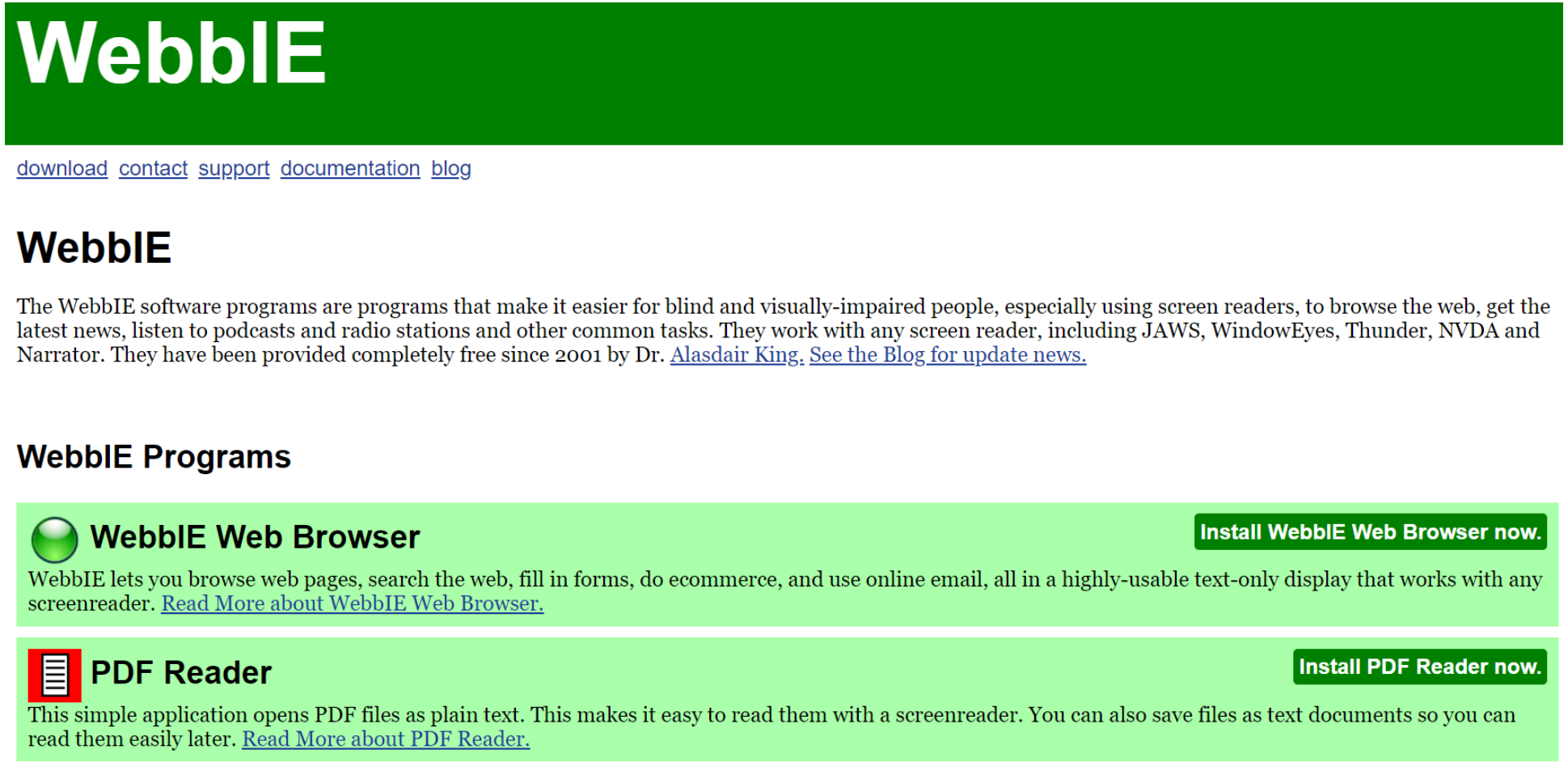
Computadores conseguem utilizar esta informação.



Pesquisa de imagens



Ferramentas de leitura de sites



WebbIE

download contact support documentation blog

WebbIE

The WebbIE software programs are programs that make it easier for blind and visually-impaired people, especially using screen readers, to browse the web, get the latest news, listen to podcasts and radio stations and other common tasks. They work with any screen reader, including JAWS, WindowEyes, Thunder, NVDA and Narrator. They have been provided completely free since 2001 by Dr. Alasdair King. See the Blog for update news.

WebbIE Programs

WebbIE Web Browser — Install WebbIE Web Browser now.

WebbIE lets you browse web pages, search the web, fill in forms, do ecommerce, and use online email, all in a highly-usable text-only display that works with any screenreader. Read More about WebbIE Web Browser.

PDF Reader — Install PDF Reader now.

This simple application opens PDF files as plain text. This makes it easy to read them with a screenreader. You can also save files as text documents so you can read them easily later. Read More about PDF Reader.

Sugira o seu site para ser preservado **(#7)**

Sugira site para ser preservado quando é **criado** ou **descoberto**

- arquivo.pt/sugerir
- Formulário simples e público
- Basta a página de entrada

Sugerir websites para serem preservados

Última atualização em 16 de Outubro de 2020 às 11:22

Partilhe esta página: arquivo.pt/sugerir

*Required

Email address *

Your email address

Endereços dos sites (1 por linha) *

Your answer

Descrição dos sites

Your answer

Site preservado antes de **ataque informático**

Media

Sites do Expresso e da SIC alvo de ataque informático

André Veríssimo
2 Janeiro 2022

Os sites do jornal Expresso e da SIC estão em baixo. No caso da página da SIC é exibida uma mensagem com um pedido de resgate de um grupo de nome Lapsus$, que fez vários ataques em dezembro no Brasil.

Follow Like

Getty images

PAÍS

Sistemas informáticos do Continente "estão a ser progressivamente repostos" após ciberataque

MC Sonae é a mais recente empresa a ser alvo de um ciberataque em Portugal.

my Moeller/Getty Images

PAÍS

Vodafone alvo de ciberataque

"Ato criminoso" estará na origem dos problemas técnicos registados. Até ao momento não há "quaisquer indícios de que os dados de clientes tenham sido acedidos ou comprometidos".

Ataque ao ipleiria.pt

QUINTA-FEIRA, 18 ABRIL 2024 | LEITORES | CONTACTO

JL 40 anos JORNAL DE LEIRIA

ABERTURA ENTREVISTA SOCIEDADE SAÚDE ECONOMIA DESPORTO VIVER OPINIÃO CLASSIFICADOS PODCA

SOCIEDADE

Ataque informático deixa Politécnico de Leiria sem acesso a plataforma e-learning

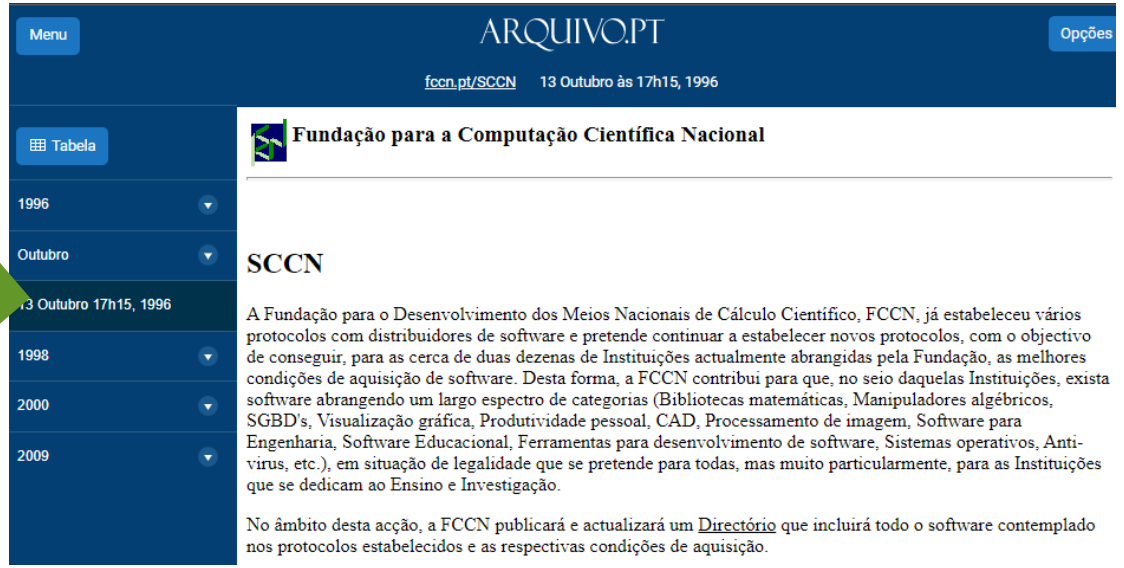
2 MAI 2023 17:10

Instituição de ensino superior desconhece ainda danos provocados pelo ciberataque

Não mate o seu site histórico!
Preserve-o no Memorial do Arquivo.pt: arquivo.pt/memorial

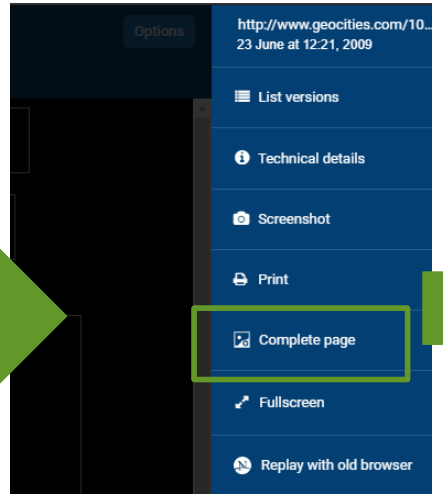
- Antes que o seu site "abandonado" seja atacado, envie-o para o Memorial e evite problemas de segurança, custos e perdas de informação!

arquivo.pt/arquivo404
acabe com as ligações quebradas no seu site!



Arquive você mesmo **(#7)**

Completar a Página

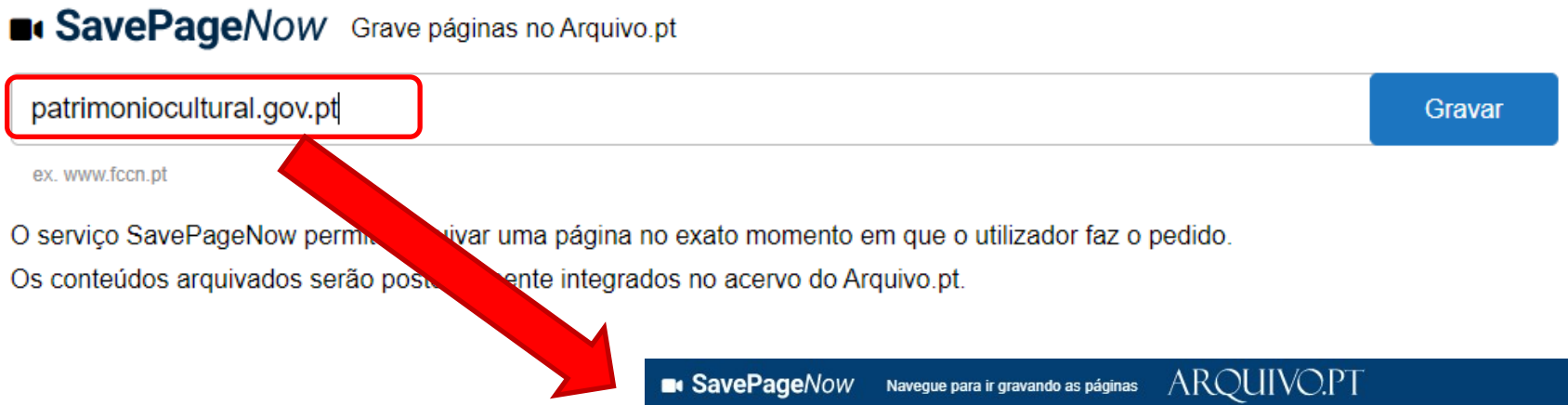


Os ficheiros obtidos estão integrados no Arquivo.pt
Curadoria colaborativa!

Save Page Now – Grava páginas no Arquivo.pt

arquivo.pt/savepagenow

Navegue para ir gravando as páginas

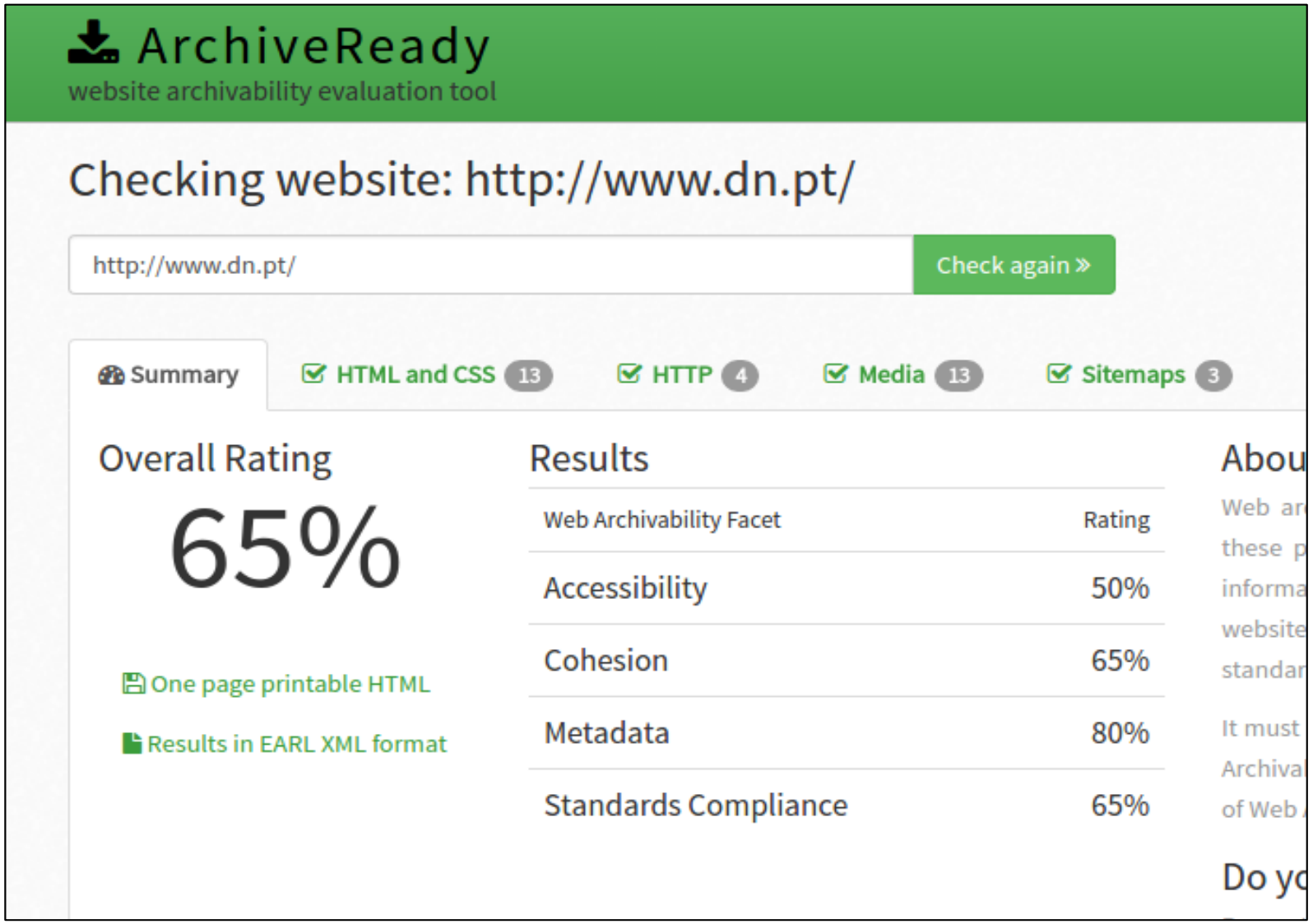


Recomendações para evitar problemas de preservação

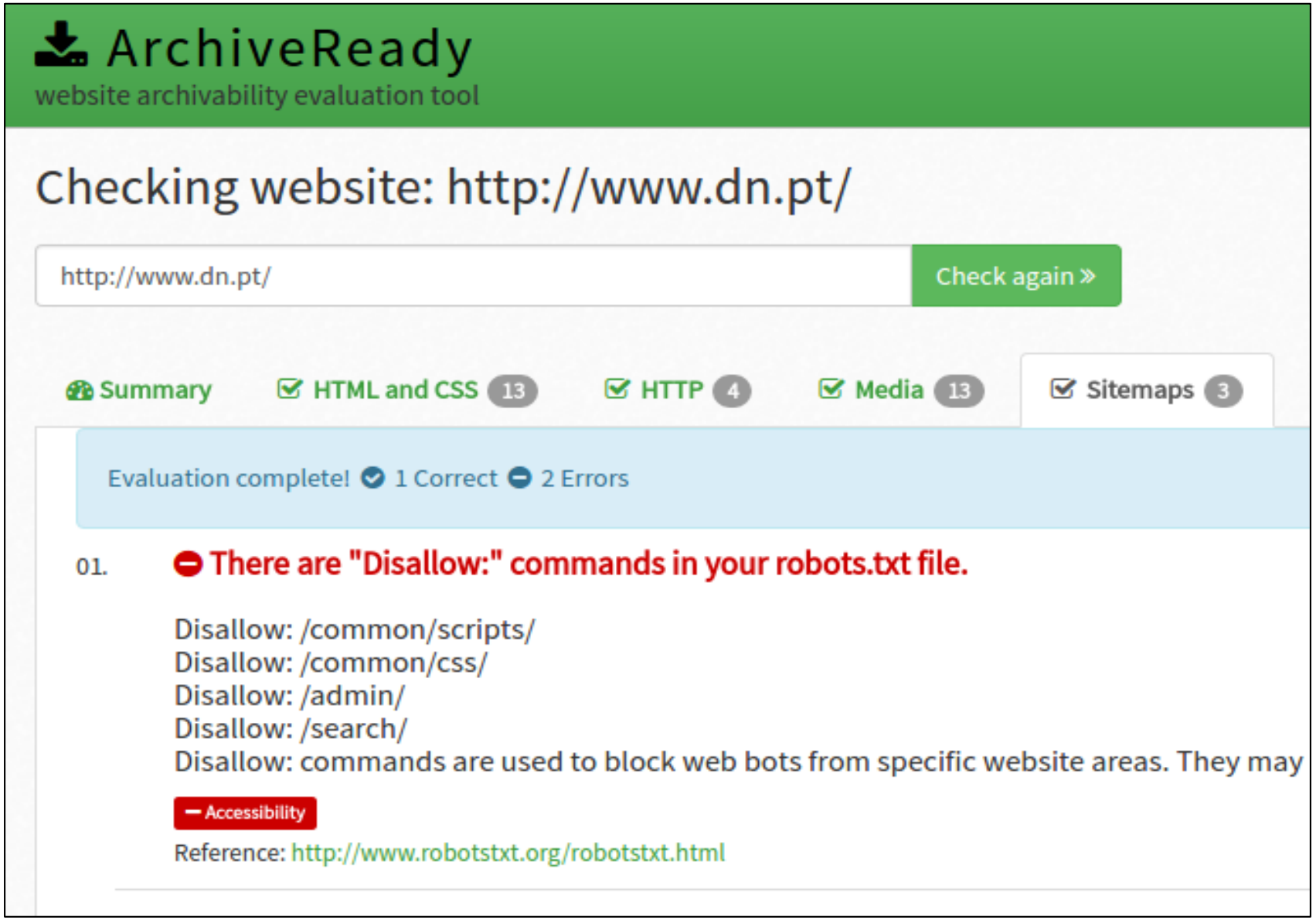
1. Identifique corretamente a **data de publicação**
2. Use corretamente o **protocolo de exclusão de robots**
3. Use um **endereço para cada conteúdo**
4. Mantenha **endereços ao longo do tempo**
5. Utilize **formatos adequados** para preservação
6. Publique **metadados** para enriquecer os conteúdos
7. **Preserve a informação**

Ferramenta para avaliar se uma página é preservável

archivereready.com



Problemas identificados pelo archiveready.com



Memória digital é Património!

A informação publicada online é **valiosa**.

Preservar a memória digital é responsabilidade de todos.