CAFÉ COM O ARQUIVO.PT

Criação e gestão de sites preserváveis

Recomendações do Arquivo.pt

Introdução

Quando se fala em preservar a Web

- Recolher, armazenar e disponibilizar
- Fragmentação inevitável
- Conservação de recursos únicos para memória futura

Funcionamento do Arquivo.pt

Funcionamento do Arquivo.pt

1

Recolha

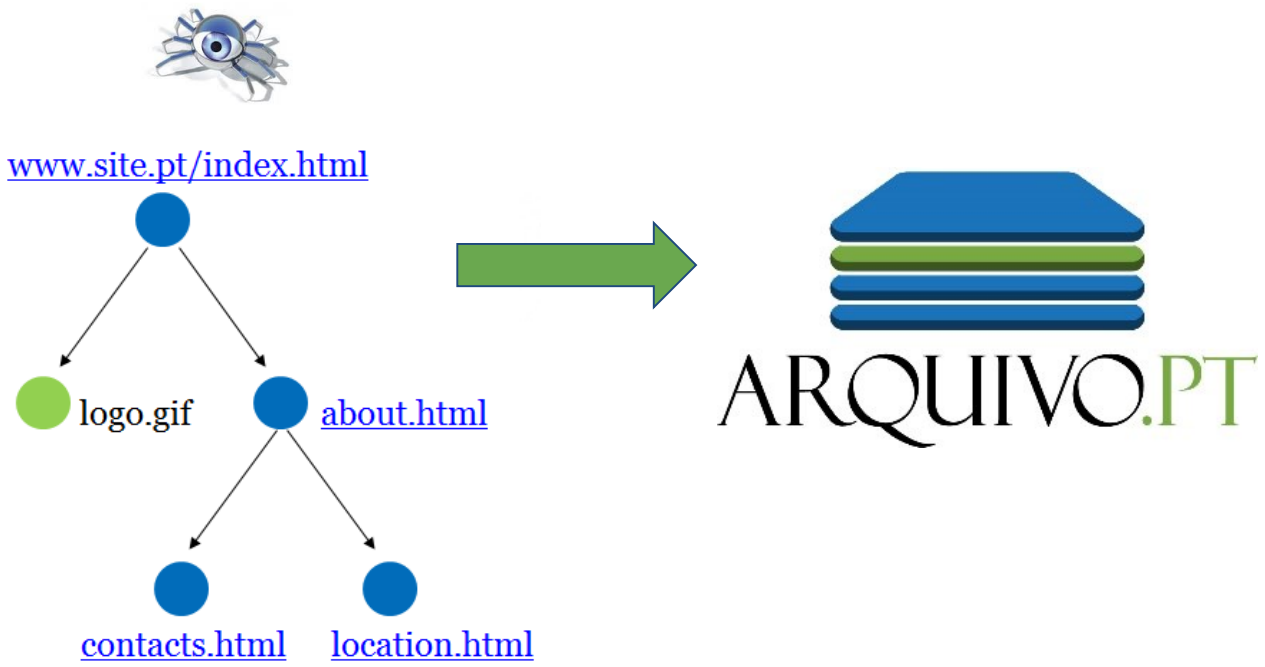
ARQUIVO.PT

2

Reprodução

Funcionamento do Arquivo.pt

Recolha é feita de forma automática por “robots”

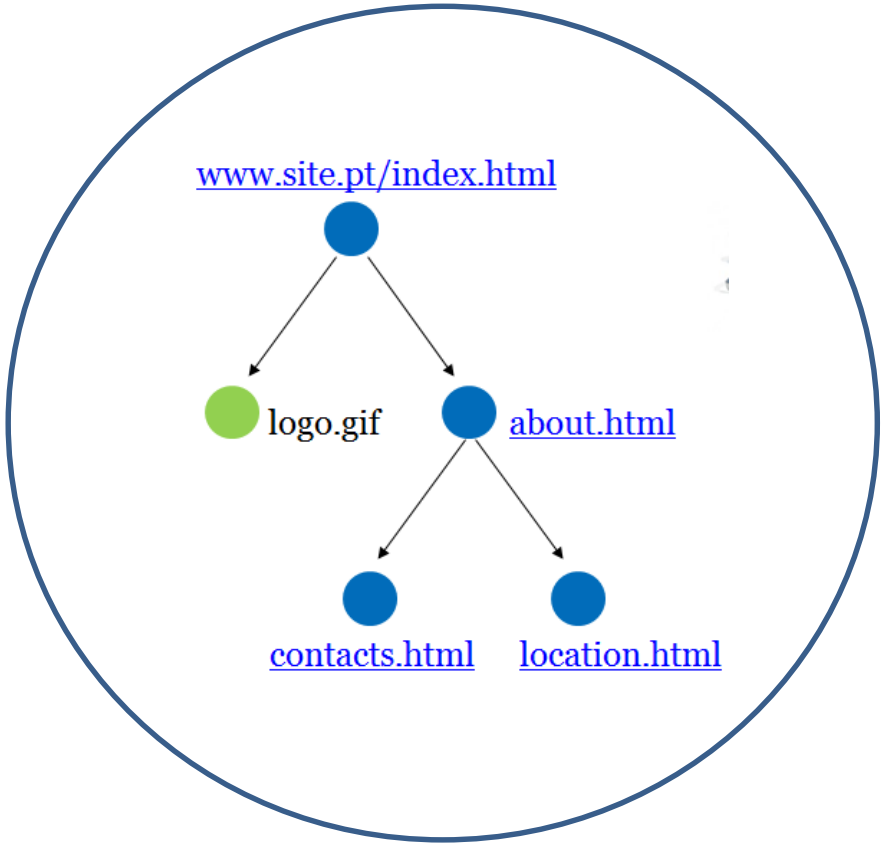


Funcionamento do Arquivo.pt

Recolha é feita dentro de certos limites:

Relacionados com o âmbito de recolha do Arquivo.pt

Relacionados com os websites a recolher

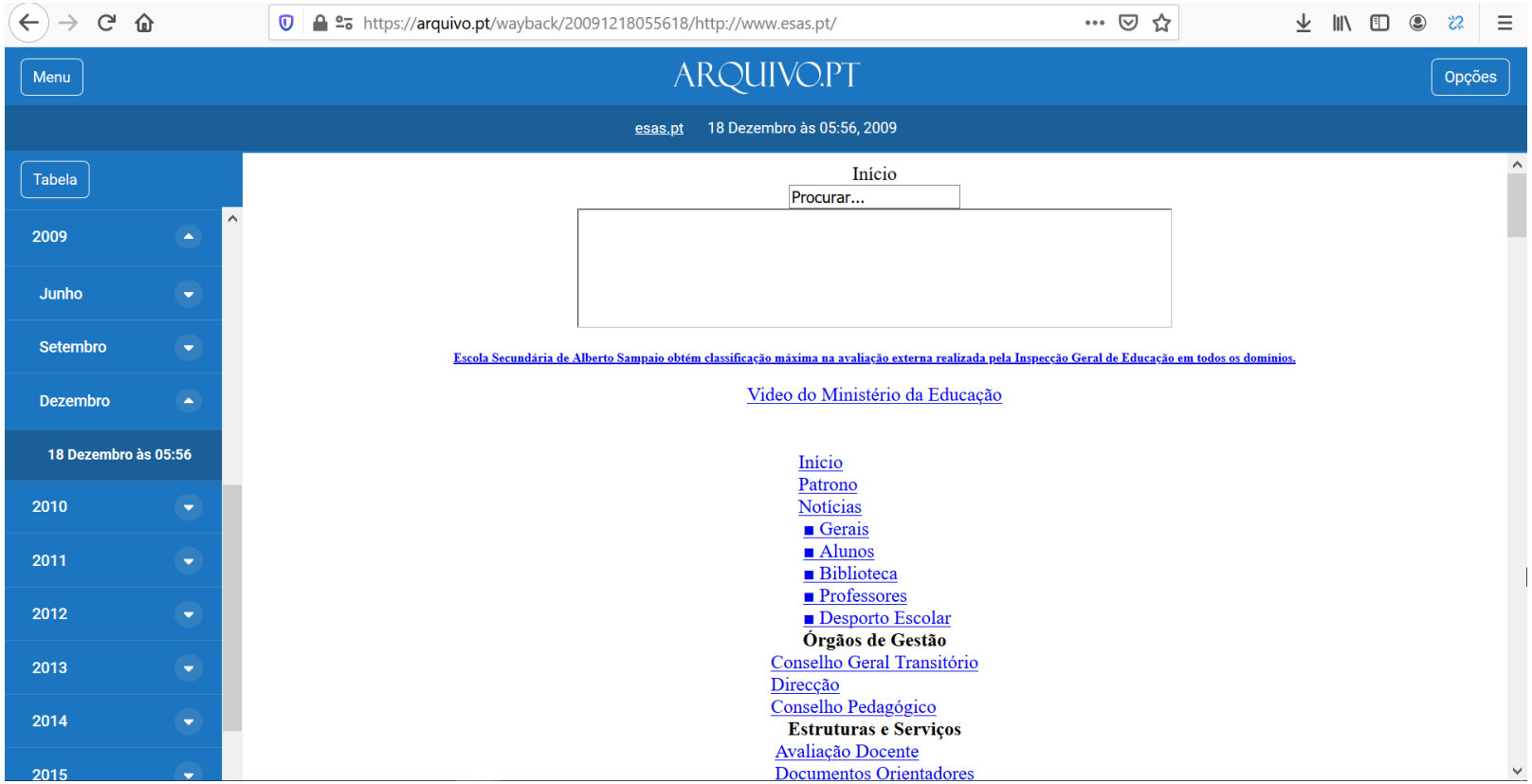


Funcionamento do Arquivo.pt

Reprodução de páginas preservadas

Funcionamento do Arquivo.pt

Nem sempre corre tudo bem!



Recomendações para **Publicar** Informação **Preservável**

1

Identifique corretamente a data de publicação

Identifique corretamente a data de publicação (#1)

Qual a data de publicação?

https://arquivo.pt/wayback/20150108053852/http://observador.pt/2015/01/07/especiali 60%

Menu ARQUIVO.PT Opções

observador.pt/2015/01/07/especialistas-falam-em-ataque-muito-profissional-e-terroristas-com-treino-militar 8 Janeiro às 05:38, 2015

Tabela

2015

Janeiro

8 Janeiro às 05:38

Facebook Twitter Newsletters

OBSERVADOR

INÍCIO POLÍTICA ECONOMIA PAÍS MUNDO DESPORTO CIÊNCIA CULTURA OPINIÃO VAIDADES VÍDEOS EXPLICADORES ESPECIAIS

ATAQUE EM PARIS

Charlie Hebdo: Ataque experiente com precisão "militar"

ATUALIZADO

Vários especialistas em segurança garantem que os terroristas que atacaram a redação do Charlie Hebdo tinham treino militar. Gesto que alude Estado Islâmico está a intrigar analistas.

Terroristas podem estar associados ao Estado Islâmico

Autor

Catarina Falcão

Embora ainda não se saiba nada oficialmente sobre os três homens que mataram

ÚLTIMAS / MUNDO

23:16 O homicídio em nome de Alá que chocou a Holanda

22:11 "Quem teme tempestades passa a vida a rastejar"

19:39 Hackers derrubam sites do governo alemão

19:10 Rushdie: "Religião merece ser objeto de sátira"

19:00 Desenhadores inundam internet com homenagens

18:54 "Terrorismo é um dos resultados da austeridade"

18:45 Xeique Munir: "ficamos chocados com a violência"

FOTOGALERIA EM DESTAQUE

Conheça os dez países mais bonitos do mundo. Portugal é um deles

ATAQUE EM PARIS

Charlie Hebdo: Ataque experiente com precisão “militar”

ATUALIZADO

Vários especialistas em segurança garantem que os terroristas que atacaram a redaç Charlie Hebdo tinham treino militar. Gesto que alude Estado Islâmico está a intrigar analistas.

Identifique corretamente a data de publicação (#1)

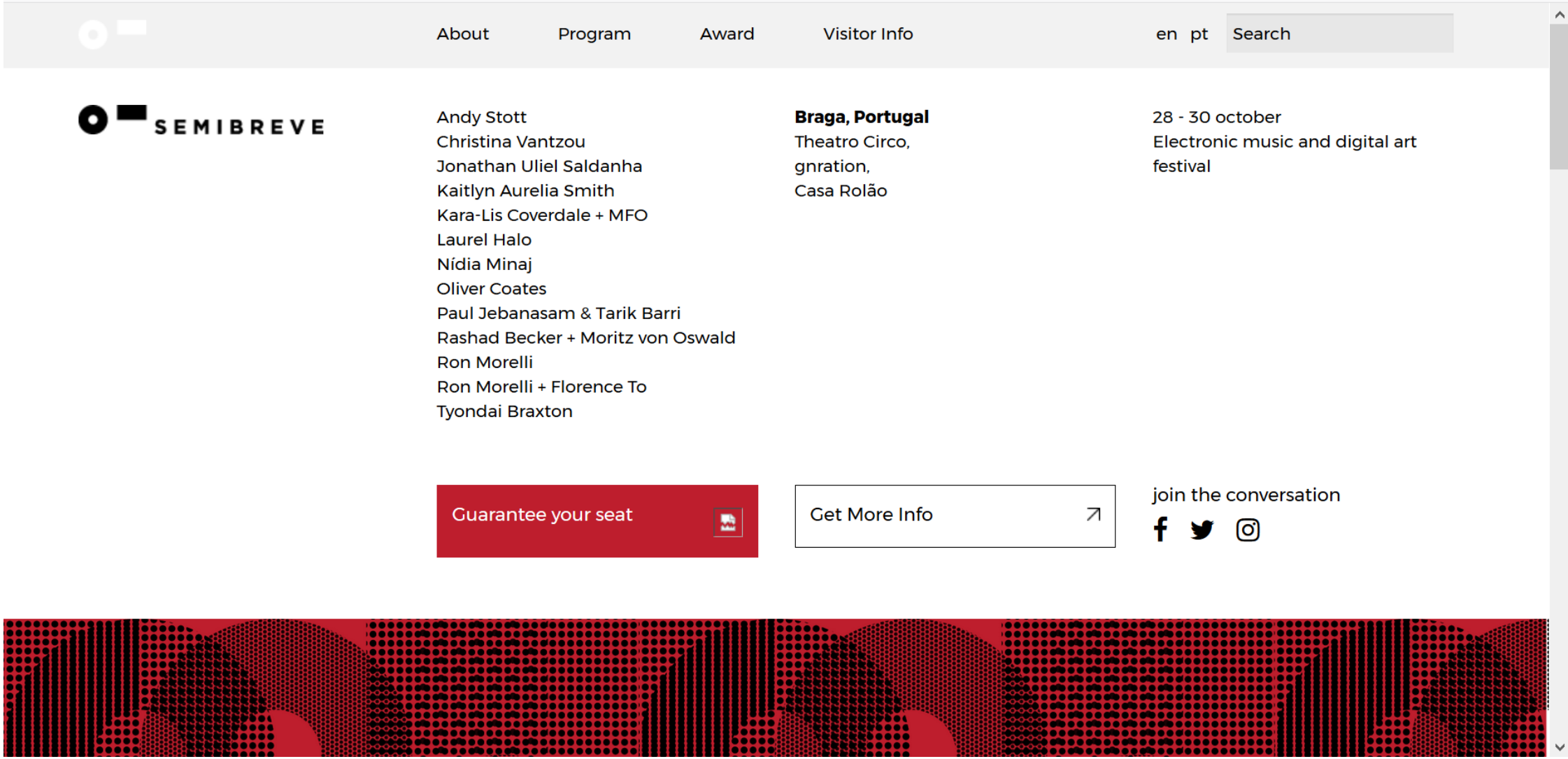
Qual a data de publicação?


https://arquivo.pt/wayback/20120904150219/http://www.diarioviseu.pt/
Menu
ARQUIVO.PT
Opções
diarioviseu.pt 4 Setembro às 15:02, 2012
Tabela
2008
2009
2010
2011
2012
Janeiro
Fevereiro
Março
Julho
Agosto
Setembro
Diário de Viseu
1ª Página da Edição de Hoje
Developed by pictonio
Terça, Setembro 4, 2012 - 15:11
NOTÍCIAS SUPLEMENTOS ASSINATURAS CONTACTOS ESTATUTO EDITORIAL FICHA TÉCNICA
Procurar
Fundador: Adriano Lucas (1883-1950)
Director "In Memoriam": Adriano Lucas (1925-2011)
Director: Adriano Callé Lucas
Há 15 anos a informar...
Siga-nos no Facebook
Diário de...
13000 gostos
Gostar da Página
Sê o primeiro dos teus amigos a gostar disto
Jean Louis Eusebio
Incêndio ameaçou casas em Viseu e não deu descanso aos bombeiros do distrito
Edição de: Terça, Setembro 4, 2012
"Isto não está mau, está péssimo", disse o tem o adjunto do comando dos Bombeiros Voluntários de Viseu, Luís Duarte, que se encontrava a comandar as operações na freguesia de Cepões. Por volta das 18h00 era este o cenário naquela zona do concelho, mais concretamente na Quinta da Sobreira, onde o ...
Ler mais
Empresários do interior pedem fim de metade dos pórticos no novo modelo de cobrança
Tempo em Viseu
Previsão das 19:00-01:00
28°C
0 mm precipitação
Vento de Nordeste, 7.5 km/h
Ver próximos dias
Publicidade

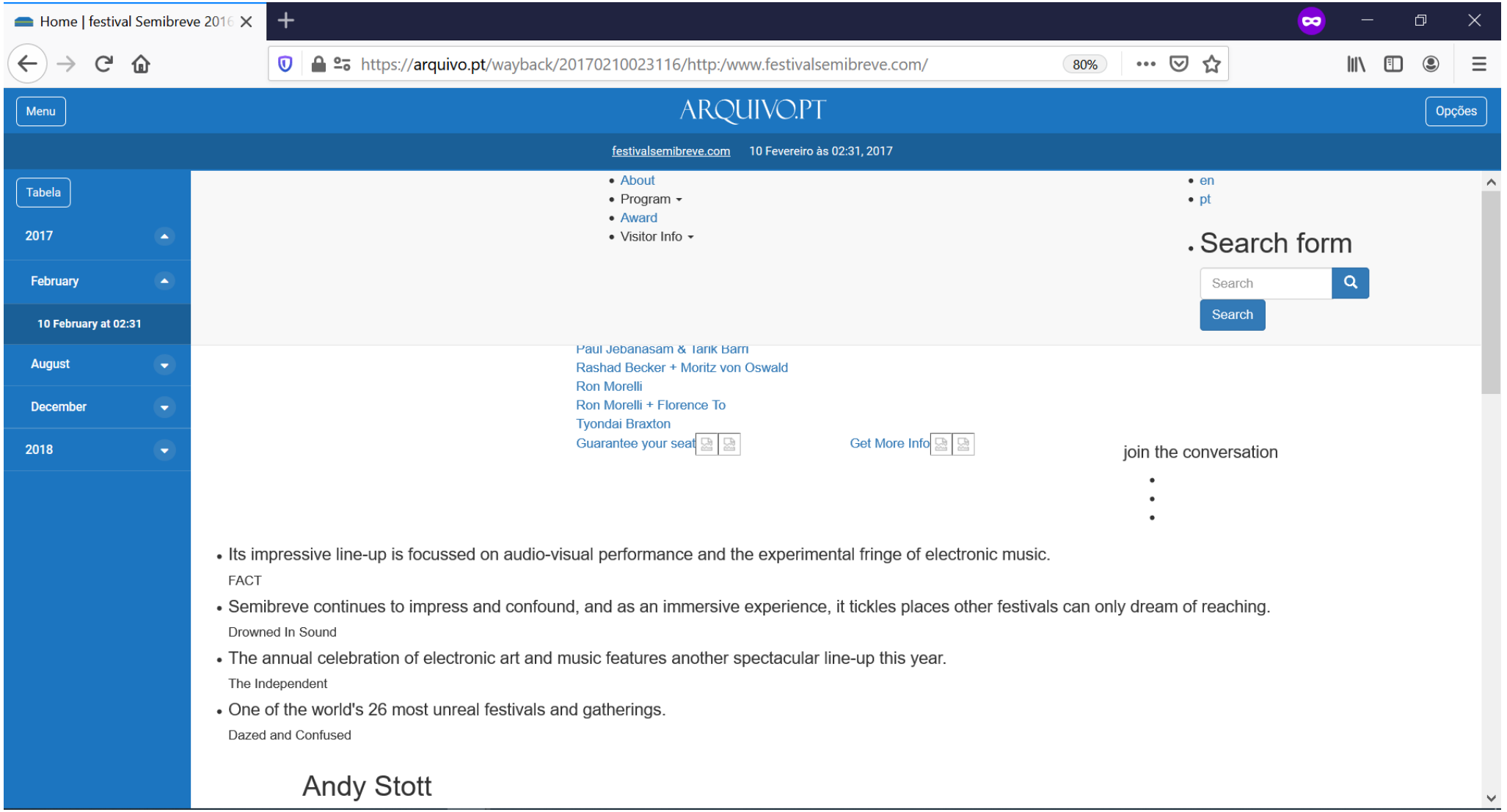
2

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

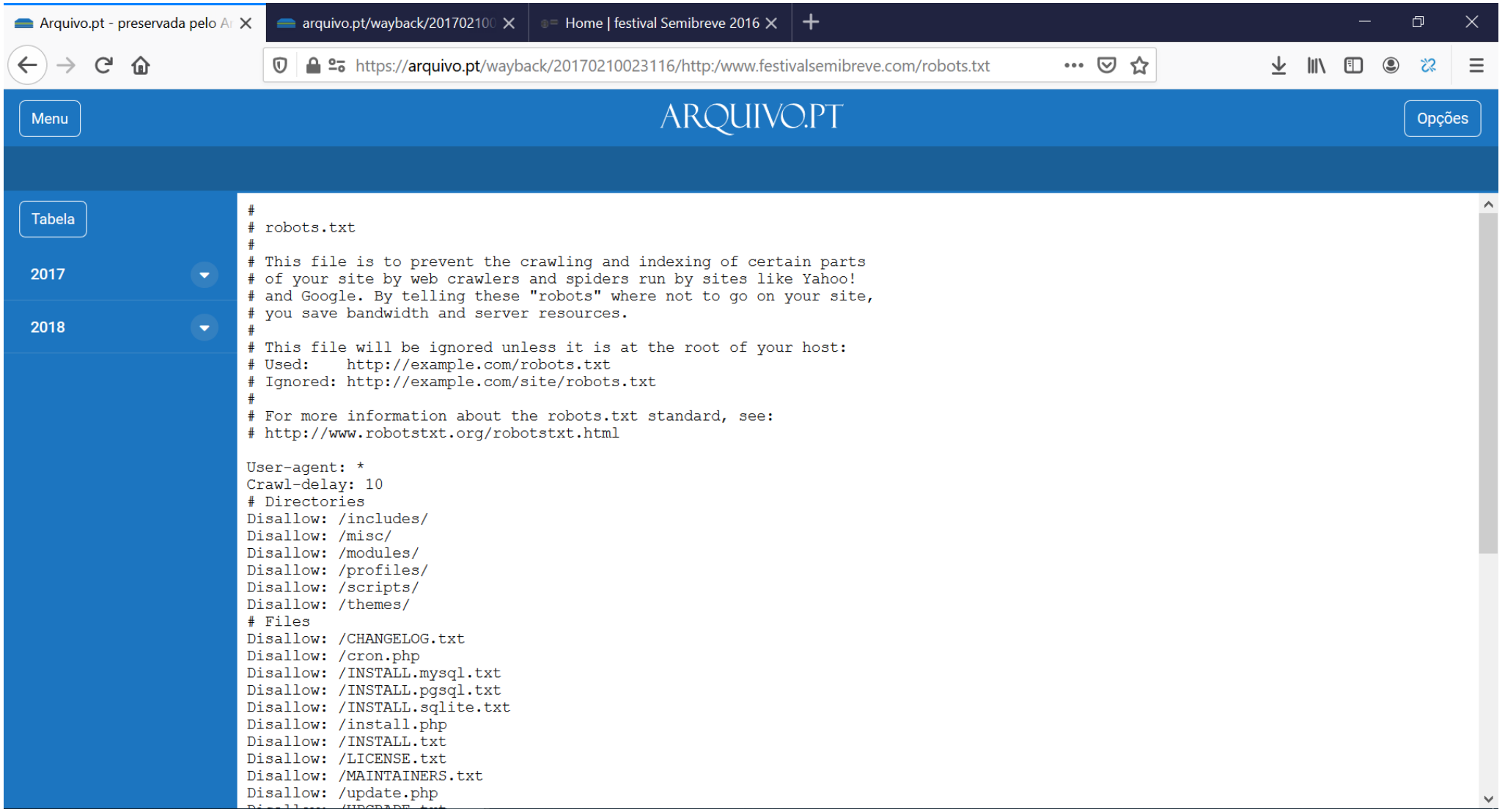
Site do festival Semi-Breve: como era em 2017



Site do festival Semi-Breve: como foi preservado



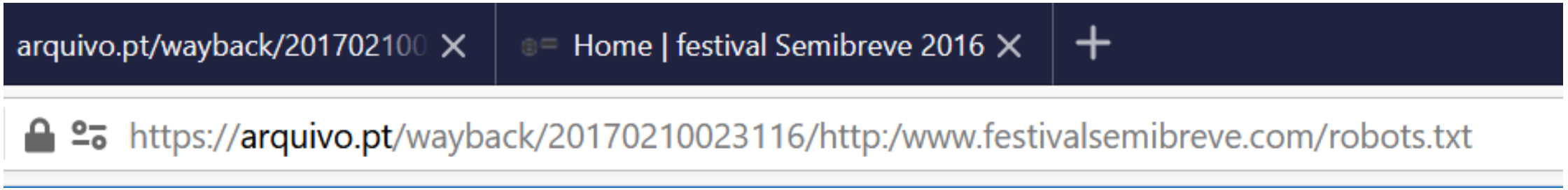
Robots Exclusion Protocol é a origem do problema de preservação



Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

```
#
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# Directories
Disallow: /includes/
Disallow: /misc/
Disallow: /modules/
Disallow: /profiles/
Disallow: /scripts/
Disallow: /themes/
```

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

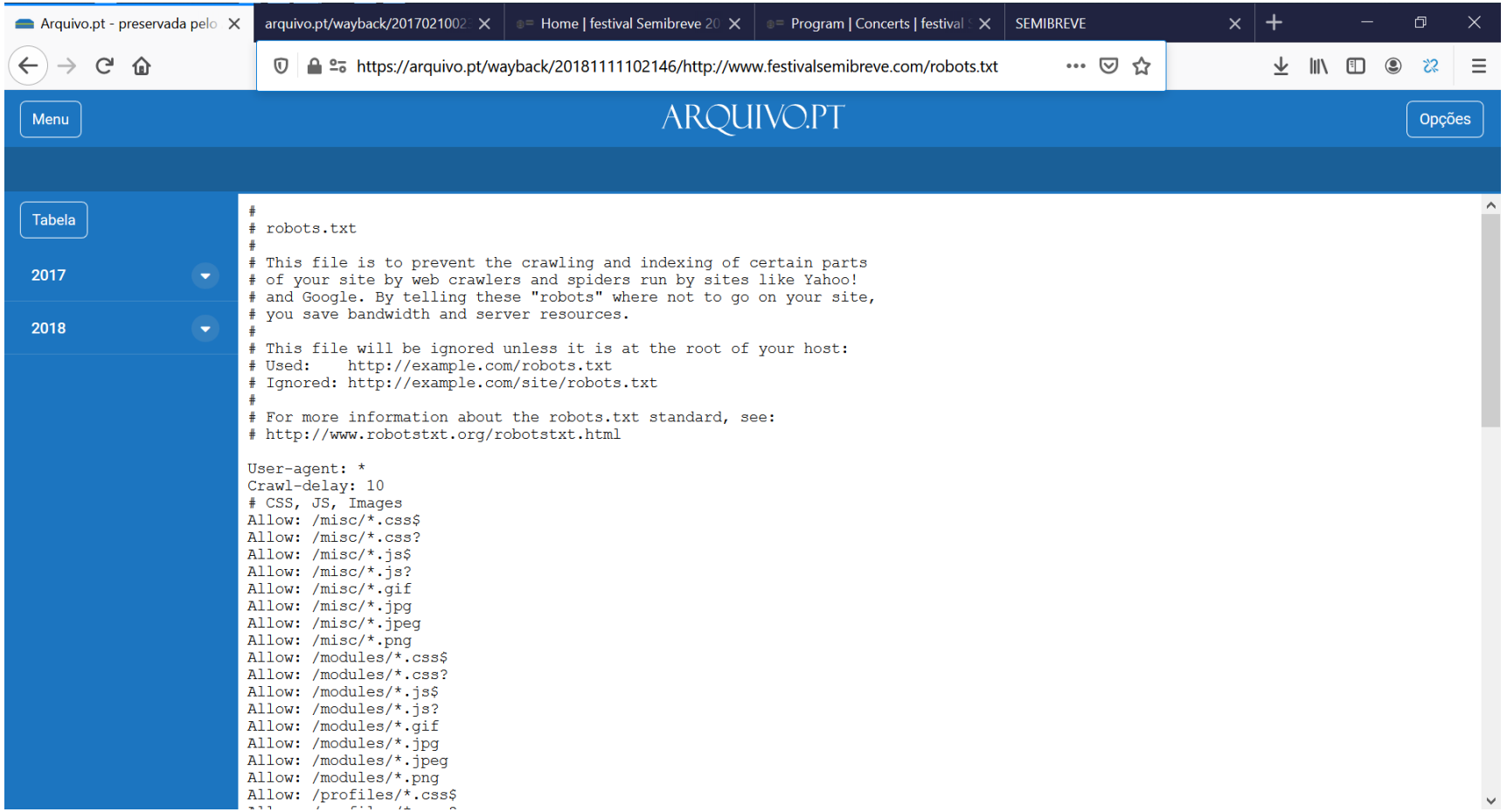


https://arquivo.pt/wayback/20170210023116/http:/www.festivalsemibreve.com/robots.txt

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

```
User-agent: *

Disallow:
```

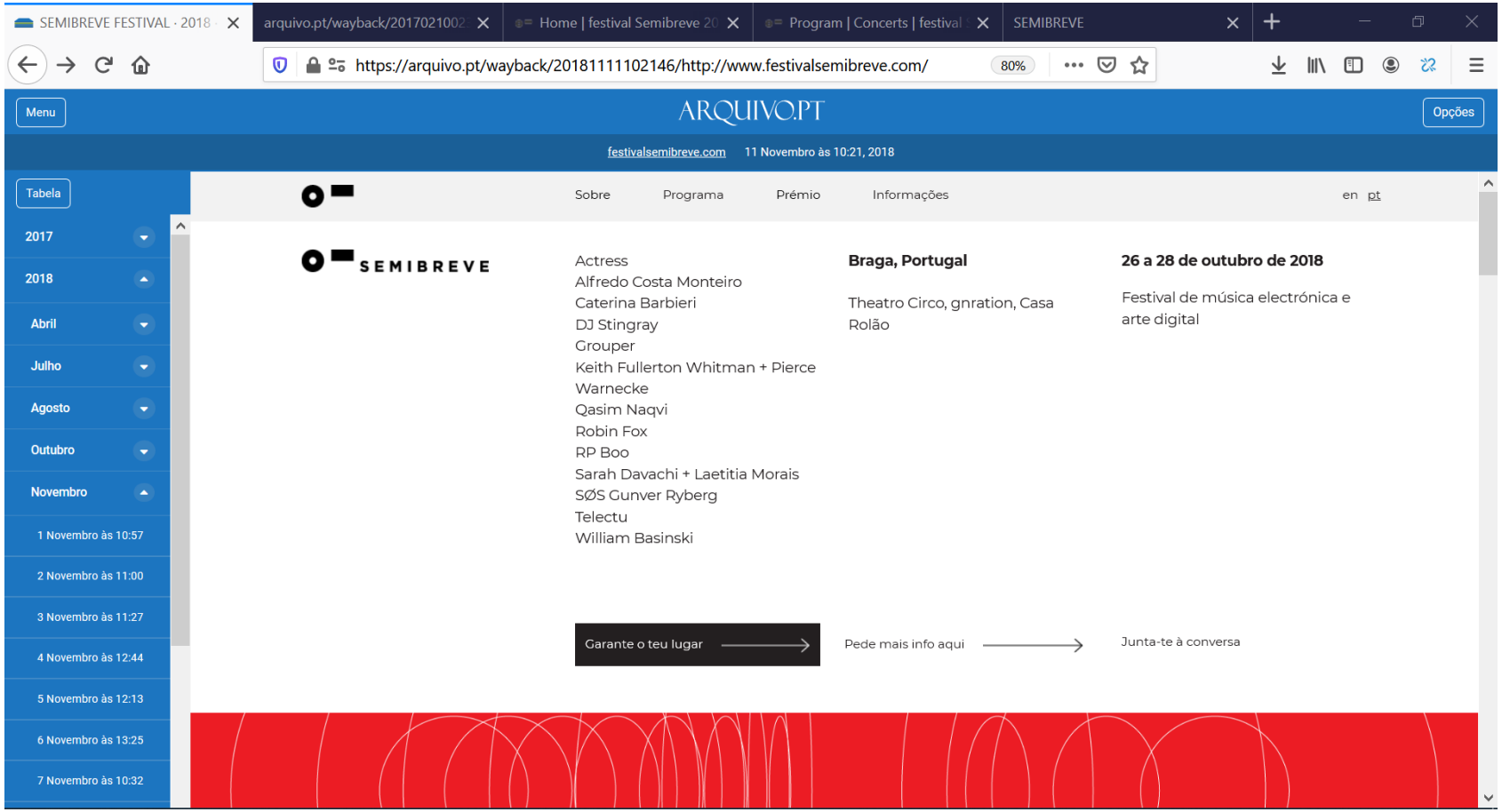
Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*



Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

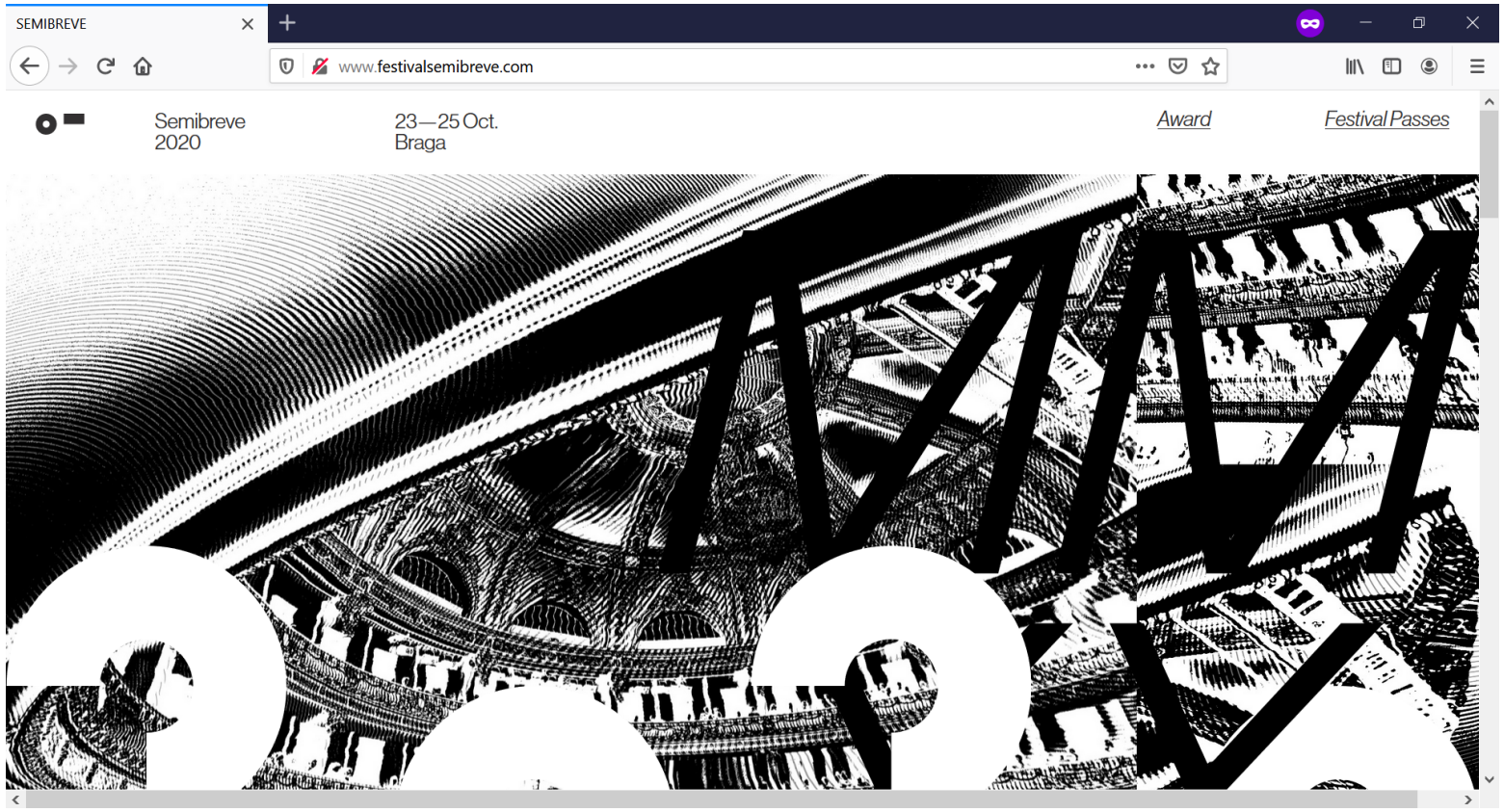
```
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html

User-agent: *
Crawl-delay: 10
# CSS, JS, Images
Allow: /misc/*.css$
Allow: /misc/*.css?
Allow: /misc/*.js$
Allow: /misc/*.js?
Allow: /misc/*.gif
Allow: /misc/*.jpg
```

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*



Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

Em alternativa: utilize de um mapa do site



Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

Em alternativa: utilize de um mapa do site


```
−<urlset>
  −<url>
      <loc>http://festivalsemibreve.com/index.html</loc>
      <lastmod>2020-06-04</lastmod>
      <changefreq>weekly</changefreq>
      <priority>0.5</priority>
    −<image:image>
      −<image:loc>
          http://festivalsemibreve.com/images/pasted-svg-84244x36.svg
        </image:loc>
      </image:image>
    −<image:image>
      −<image:loc>
          http://festivalsemibreve.com/images/moshed-2019-11-7-18-56-211024x683.jpg
        </image:loc>
      </image:image>
    −<image:image>
      −<image:loc>
          http://festivalsemibreve.com/images/moshed-2019-11-7-18-56-21_2x.jpg
        </image:loc>
```


http://festivalsemibreve.com/sitemap.xml

Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

Porque continua a haver bloqueios no Robots.txt ?

"**Disallowing** crawling of **Javascript** or **CSS** files in your site's robots.txt directly **harms** how well our algorithms render and index **your content** and can result in suboptimal rankings."

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

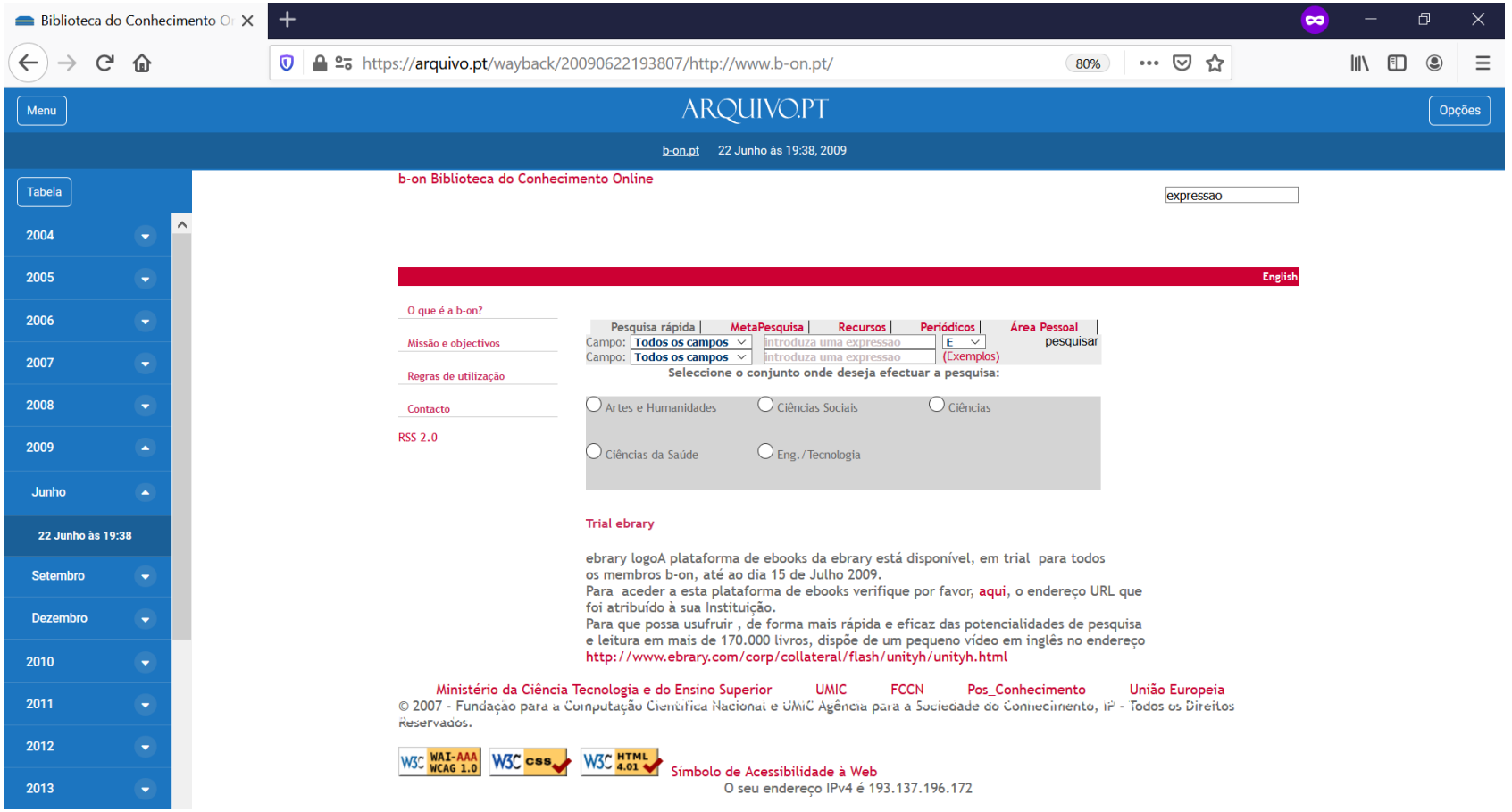
...

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/

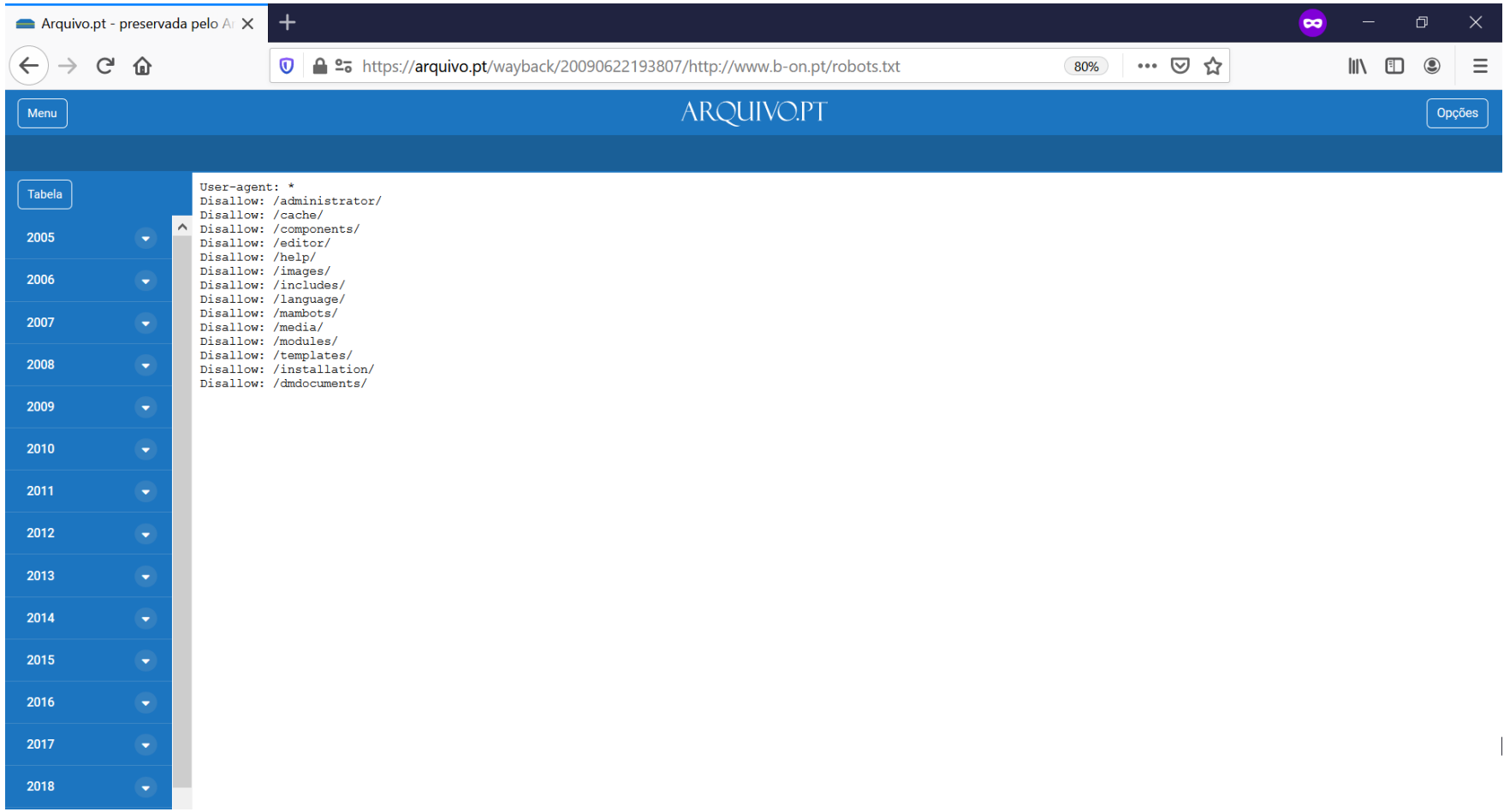
...

B-on.pt: como foi preservado



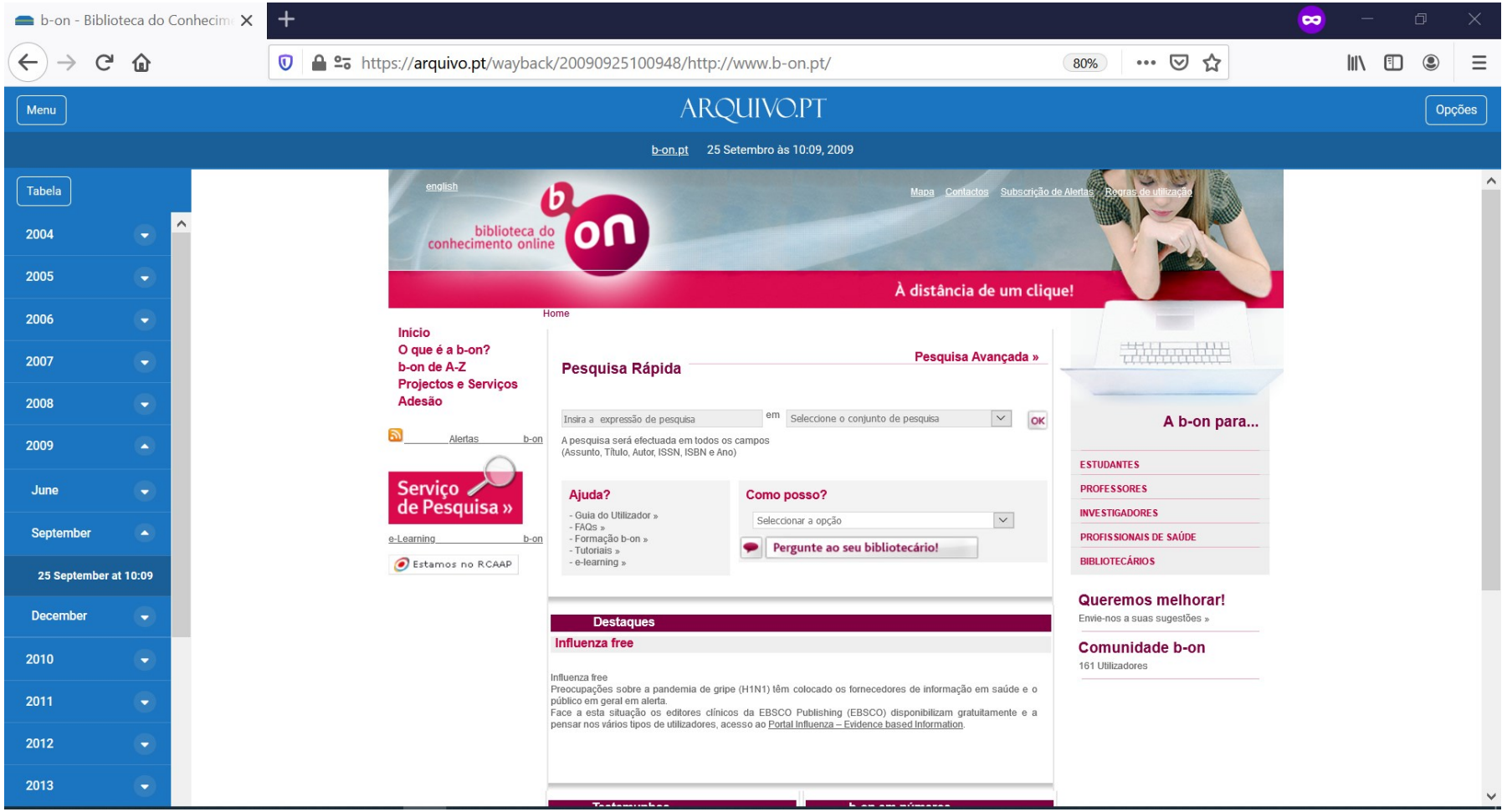
...

B-on.pt: como foi preservado



...

B-on.pt: como foi preservado



Autorize a recolha de conteúdos importantes através do *Robots Exclusion Protocol*

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

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

https://sobre.arquivo.pt/pt/recomendacoes/robots-exclusion-protocol-autorizando-a-recolha-de-conteudos-importantes/

3

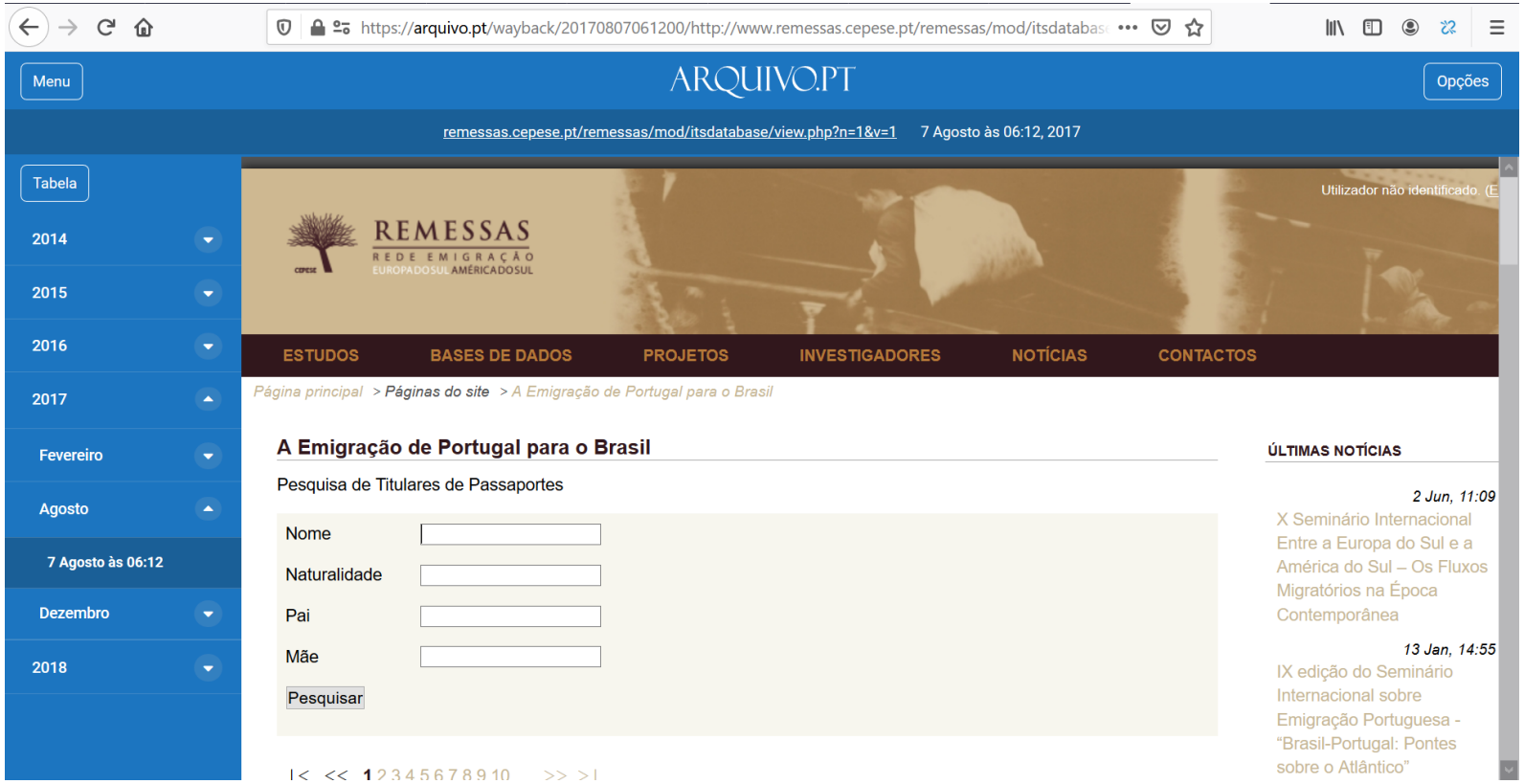
Utilize um endereço para cada conteúdo

Utilize um endereço para cada conteúdo

Conteúdos escondidos atrás de formulários escapam às recolhas

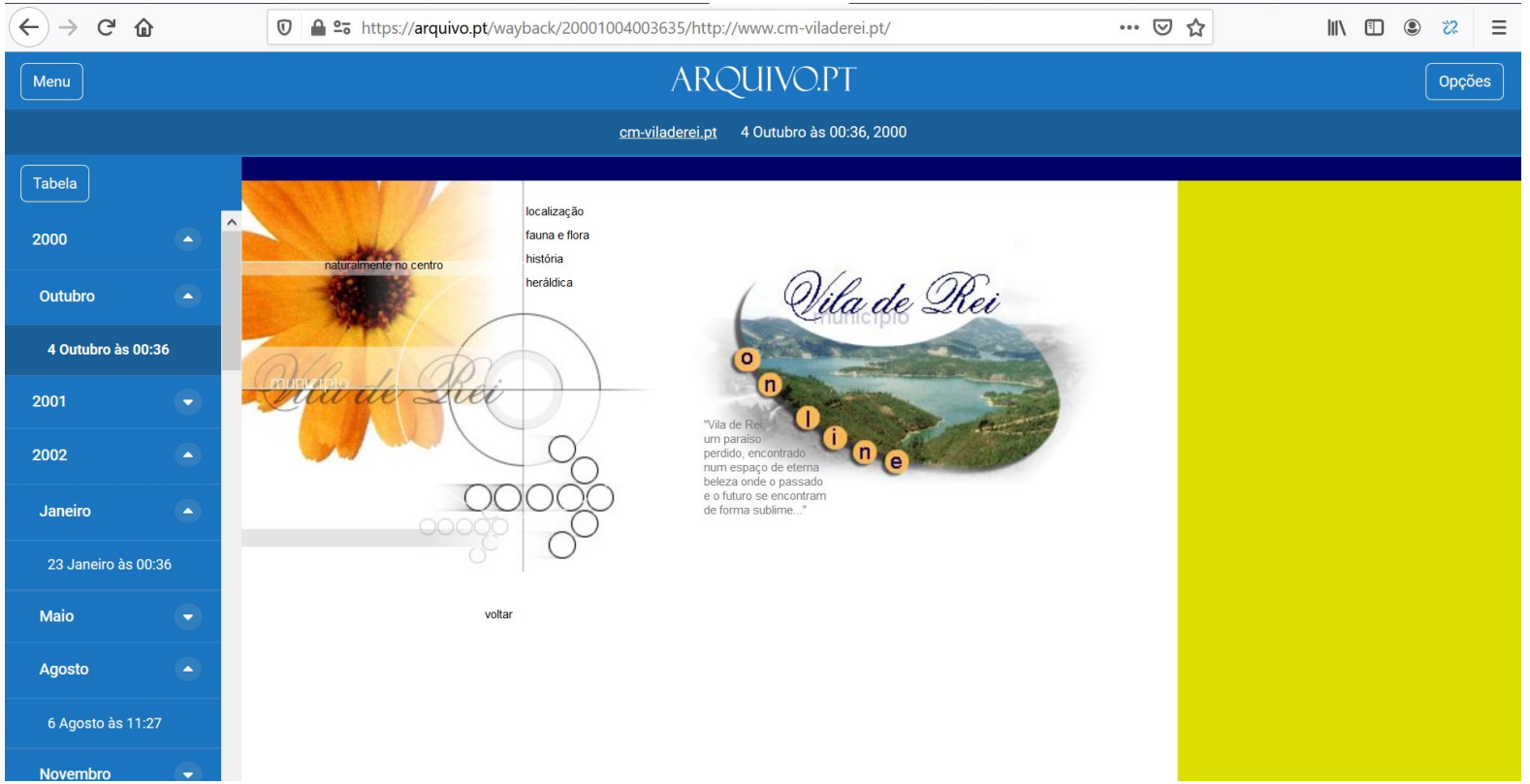
Utilize um endereço para cada conteúdo

Conteúdos escondidos atrás de formulários escapam às recolhas



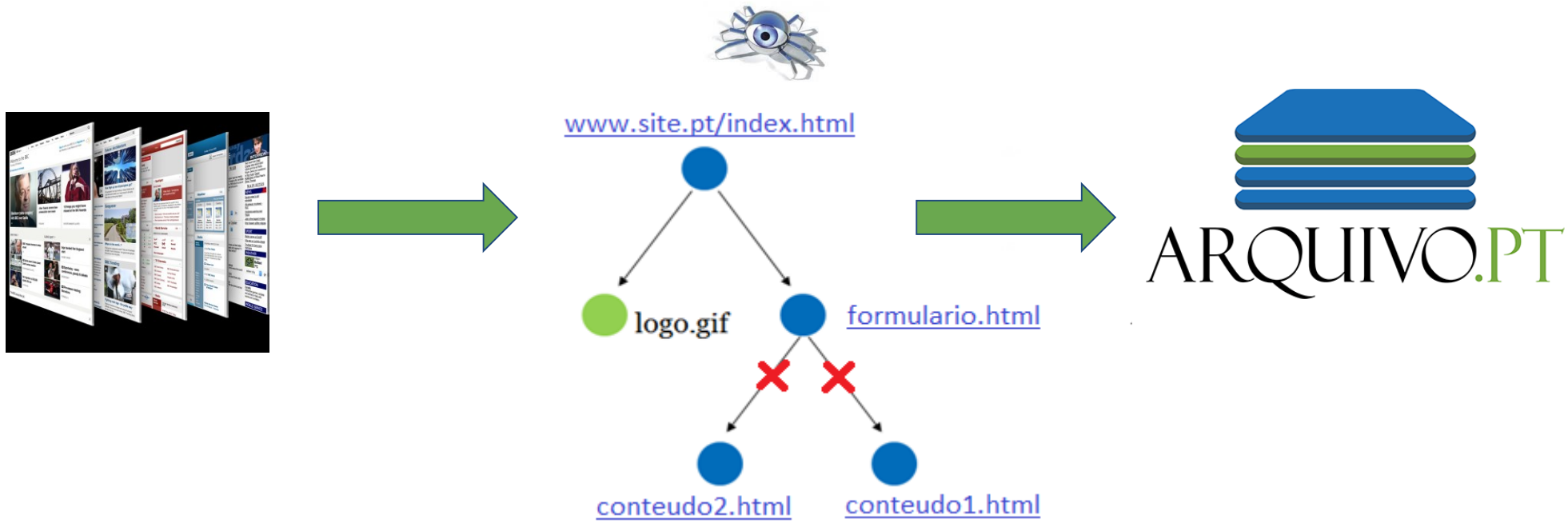
Utilize um endereço para cada conteúdo

Conteúdos não identificados por um endereço único escapam às recolhas



Utilize um endereço para cada conteúdo

Conteúdos escondidos atrás de formulários escapam às recolhas



Utilize um endereço para cada conteúdo

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

Utilize um endereço para cada conteúdo

Conteúdos escondidos atrás de formulários escapam às recolhas

Alternativa para recuperar conteúdo perdido

4

Mantenha o mesmo endereço ao longo do tempo

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

Tabela de versões

460 versões de iscte.pt

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

24 versões de iscte-iul.pt

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			

Mantenha o histórico redirecionando os endereços antigos para os novos

Tabela de versões

460 versões de iscte.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	31 Mai	21 Jan	21 Jan	5 Nov	26 Set
25 Mar	2 Fev	21 Mar	12 Jan	2 Fev	9 Jan	18 Jan	20 Mai	6 Jun	20 Mai	22 Jan	7 Nov	30 Set
25 Mai	18 Fev	21 Abr	13 Jan	6 Fev	13 Jan	26 Jan	21 Mai	5 Aug	22 Mai			
27 Mai	24 Mar	5 Jun	1 Fev	7 Fev	19 Jan	28 Jan	23 Jun					
5 Jun	27 Mar	11 Jun	4 Fev	2 Mar	21 Jan	6 Fev	24 Set					
24 Set	9 Abr	11 Jun	4 Fev	2 Abr	24 Jan	15 Fev	26 Set					
27 Set	10 Abr	12 Jun	8 Fev	25 Abr	27 Jan	11 Mar	17 Dez					
13 Nov	10 Abr	14 Jun	8 Fev	27 Abr	5 Fev	14 Mar	18 Dez					
26 Nov	31 Mai	23 Jun	9 Fev	2 Jun	20 Fev	21 Out						

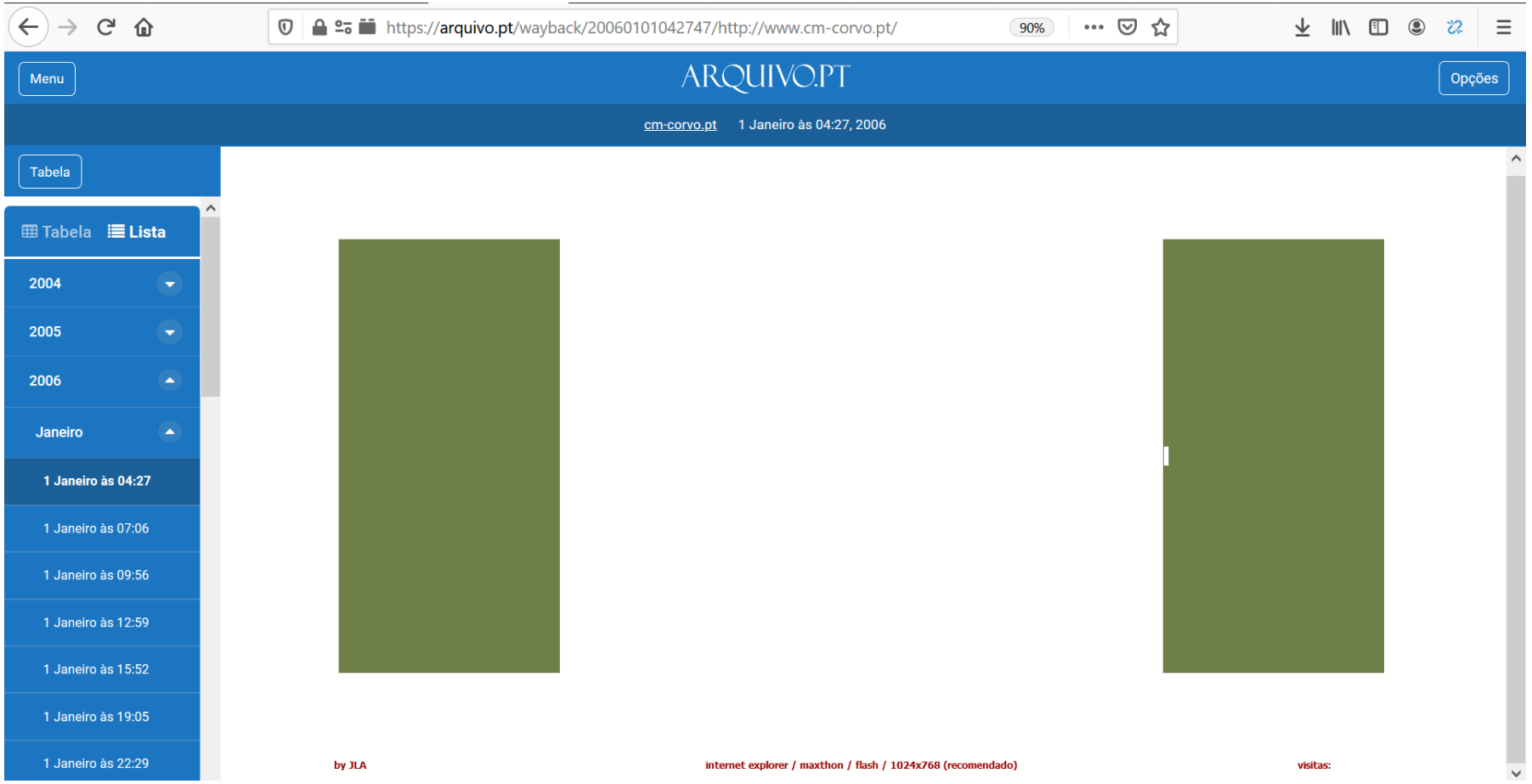
http://iscte.pt → http://iscte-iul.pt

Redireciona

5

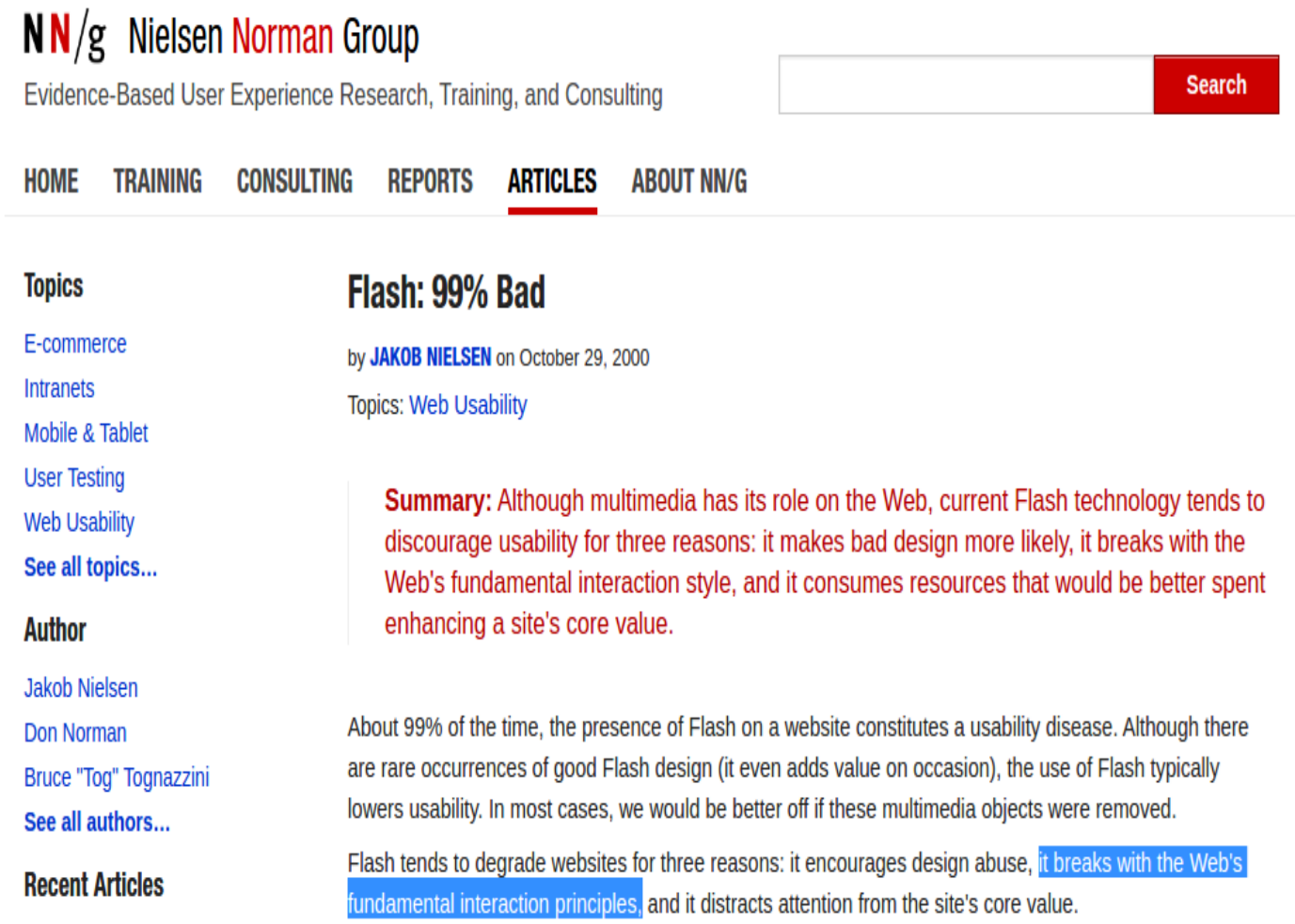
Utilize formatos adequados para preservação

Problema: Utilização de Flash



https://arquivo.pt/wayback/20060101042747/http://www.cm-corvo.pt/

“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.

http://www.occupyflash.org/

Escolha formatos adequados:

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.

Escolha formatos adequados:

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)

Escolha formatos adequados

Evite formatos não adequados para preservação

Texto:

Microsoft Word (.doc)

Imagem:

Macromedia Flash (*.swf)

PhotoShop (.psd)

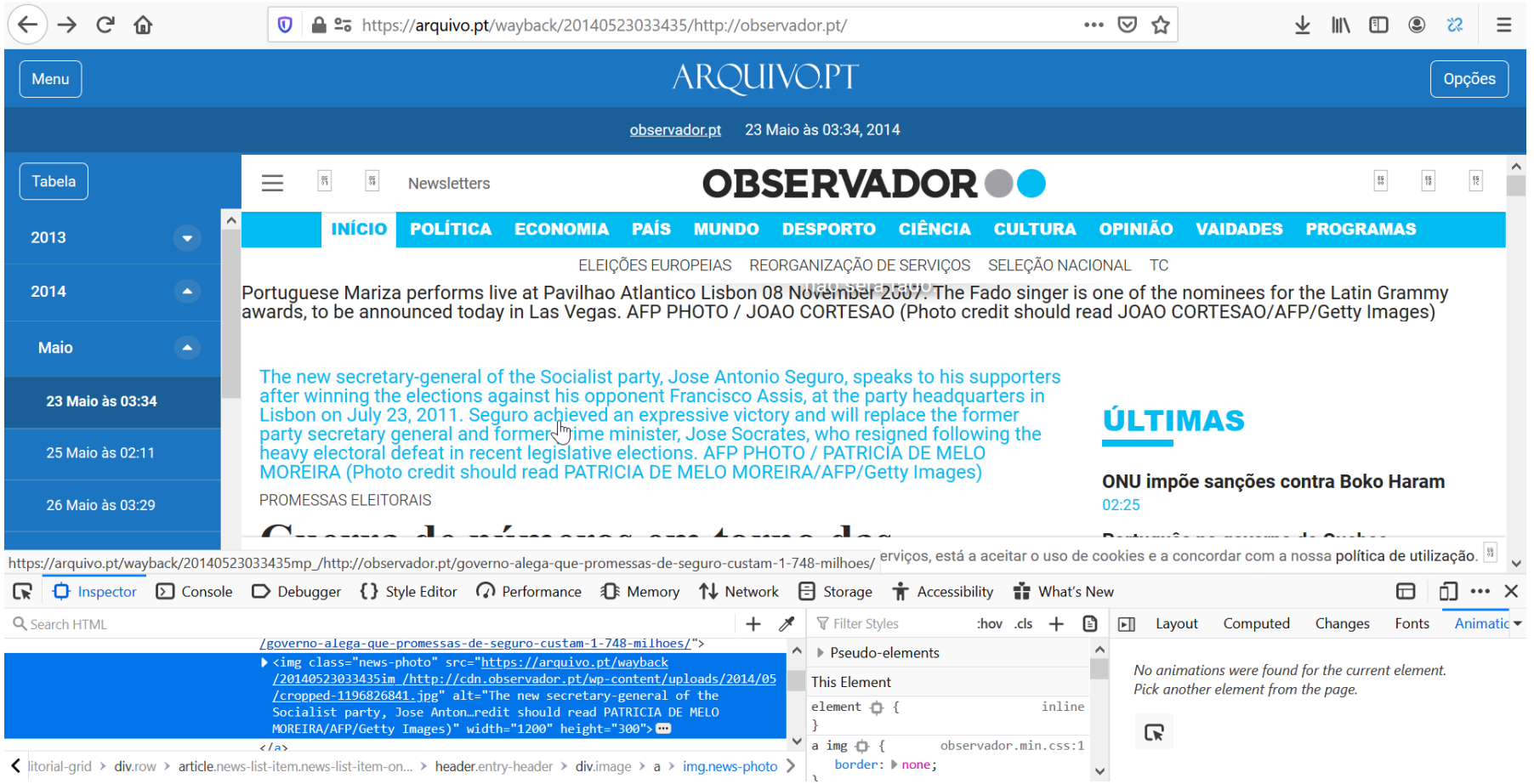
Vídeo:

Windows Media Video (.wmv)

Novas formas de inserir os conteúdos nas páginas

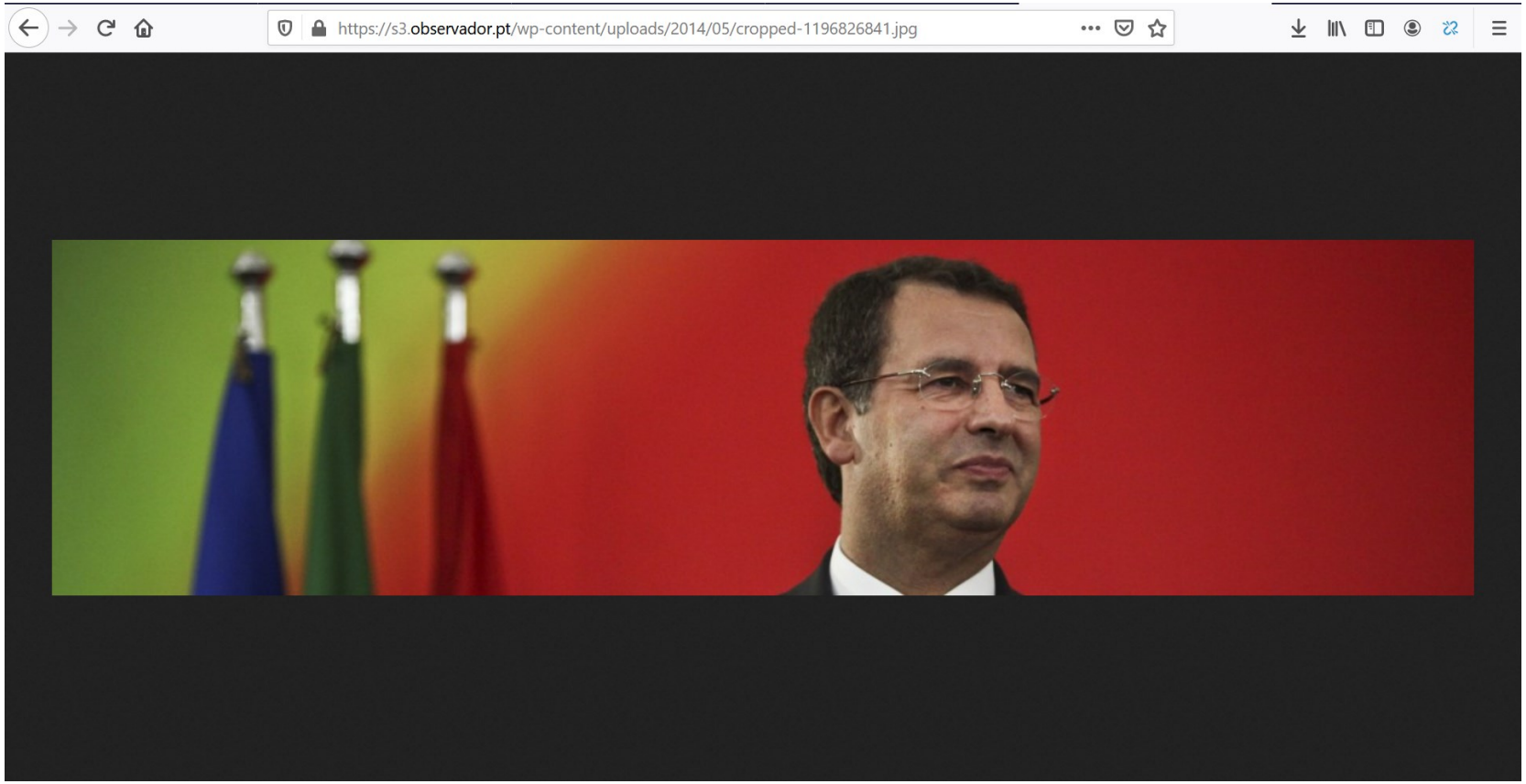
imagem

Novas formas de inserir os conteúdos nas páginas



http://cdn.observador.pt/wp-content/uploads/2014/05/cropped-1196826841.jpg

Novas formas de inserir os conteúdos nas páginas



http://s3.observador.pt/wp-content/uploads/2014/05/cropped-1196826841.jpg

6

Utilize metadados para descrever os conteúdos

Quem é o autor da página?

Utilize metadados (Dublin Core)

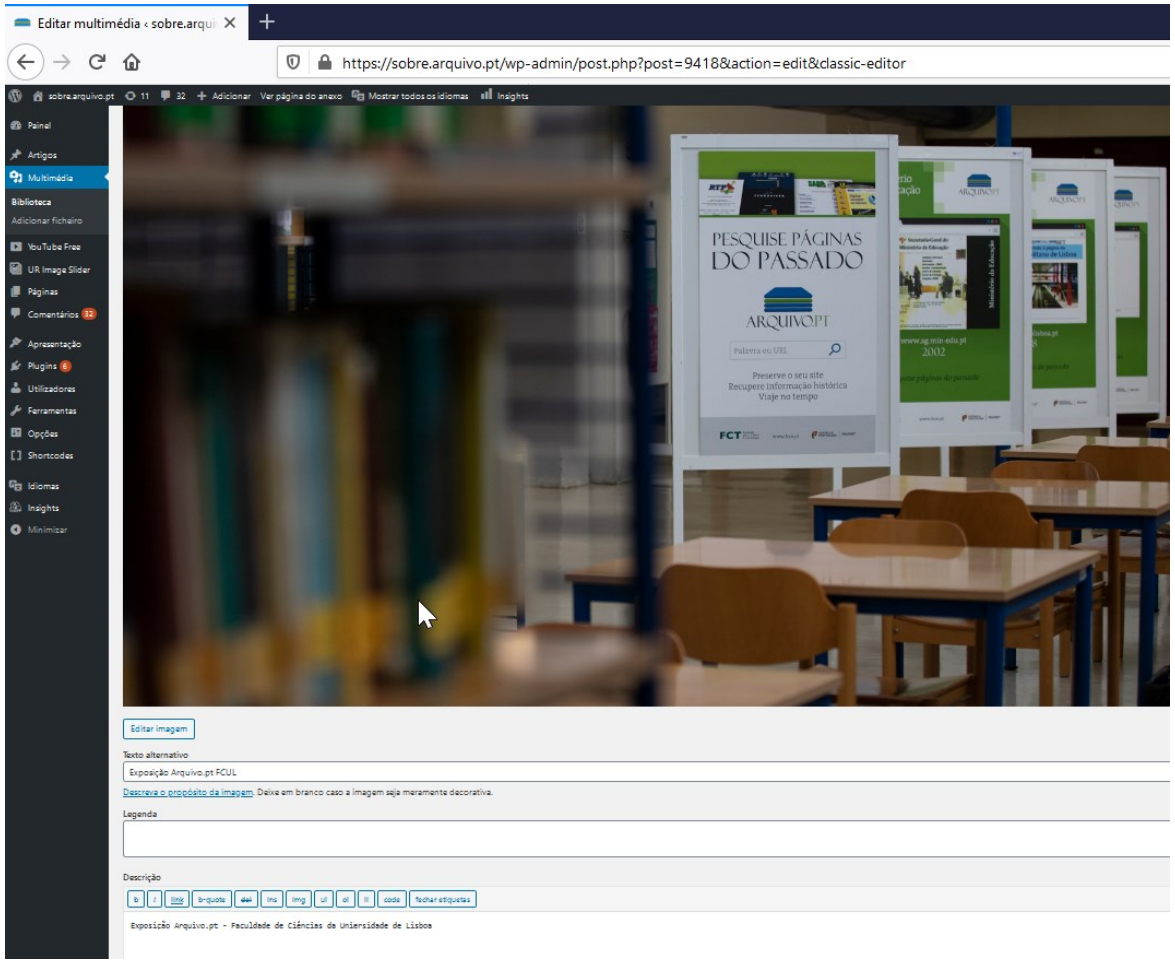
Exemplo de campos de descrição do Wordpress image title

```
<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.

Utilize metadados



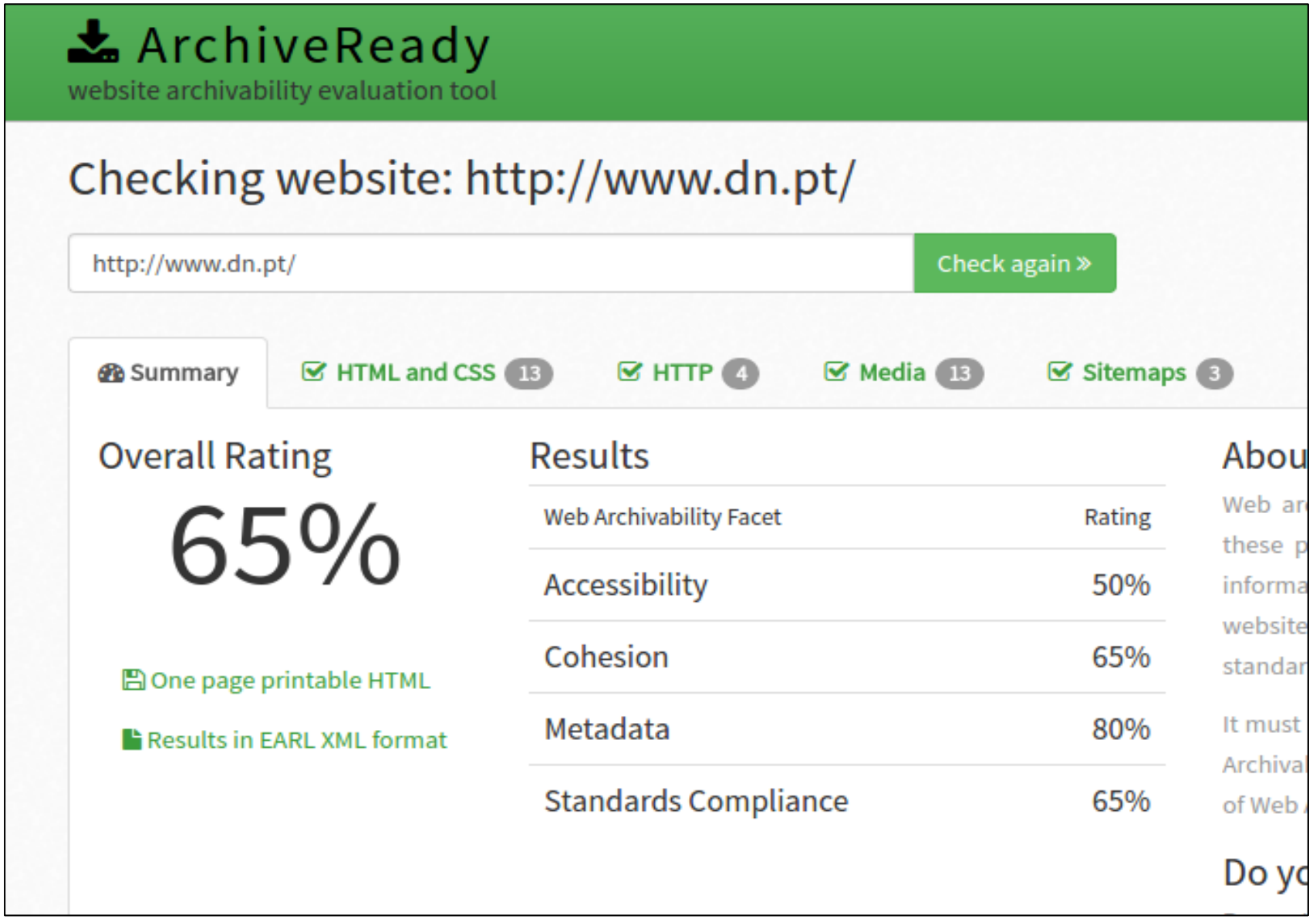
7

Torne-se curador dos seus websites

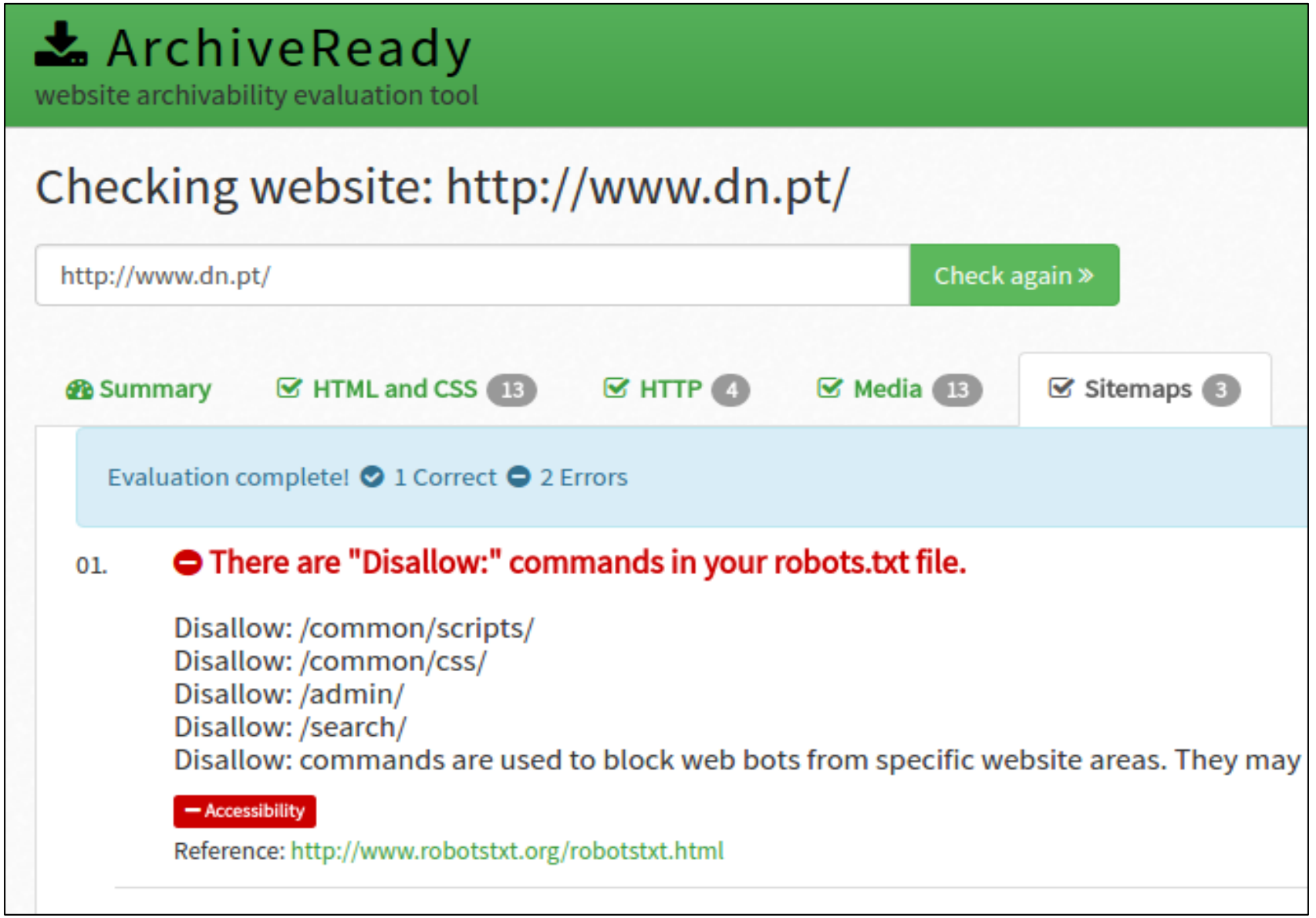
Use ferramentas para avaliar se uma página é preservável

Archive Ready

http://archivereadyready.com



http://archiveready.com



Recupere conteúdos perdidos

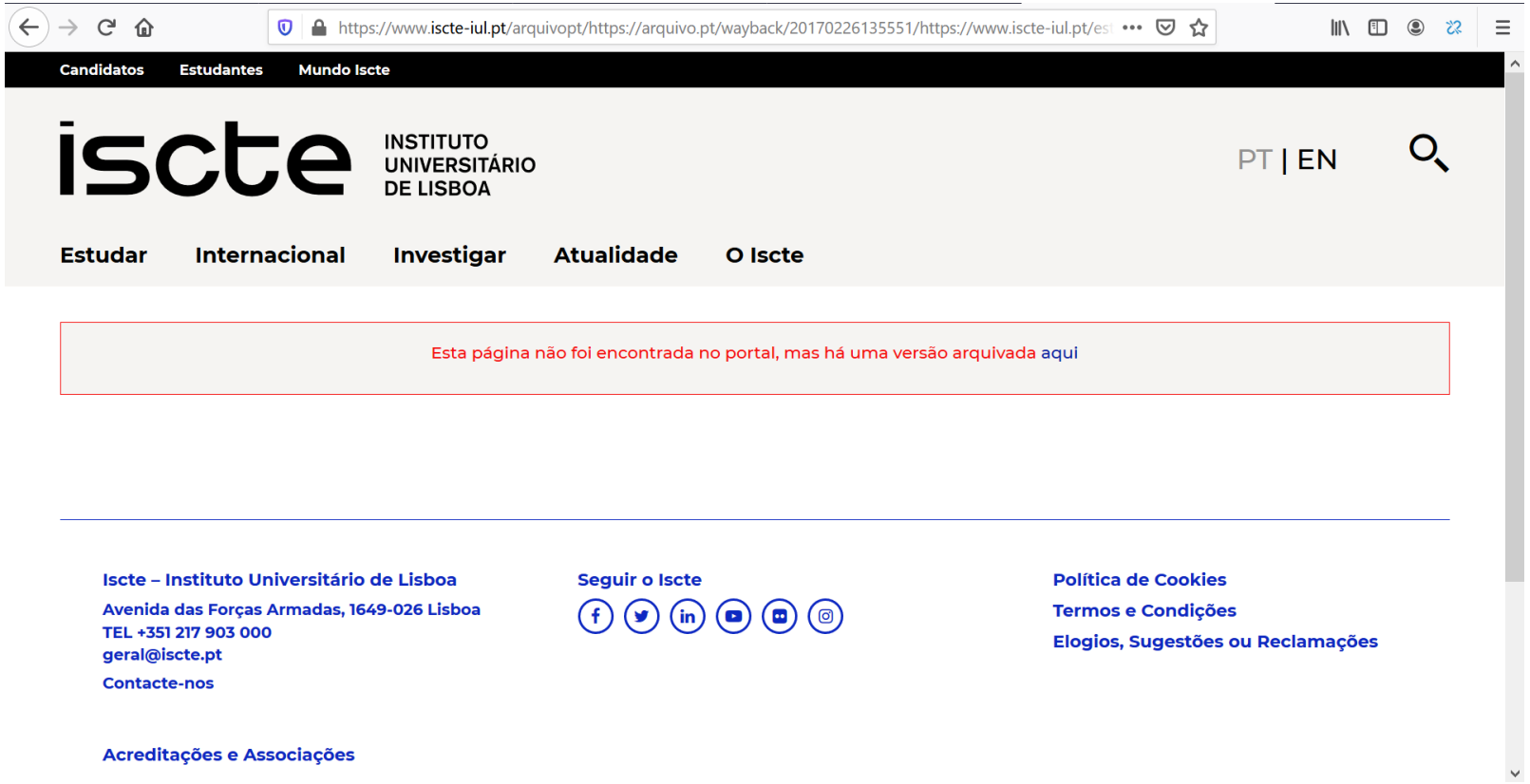
Soft404

Páginas Web com ligações quebradas

Utilizadores seguem a ligação para uma página preservada no Arquivo.pt

Soft404

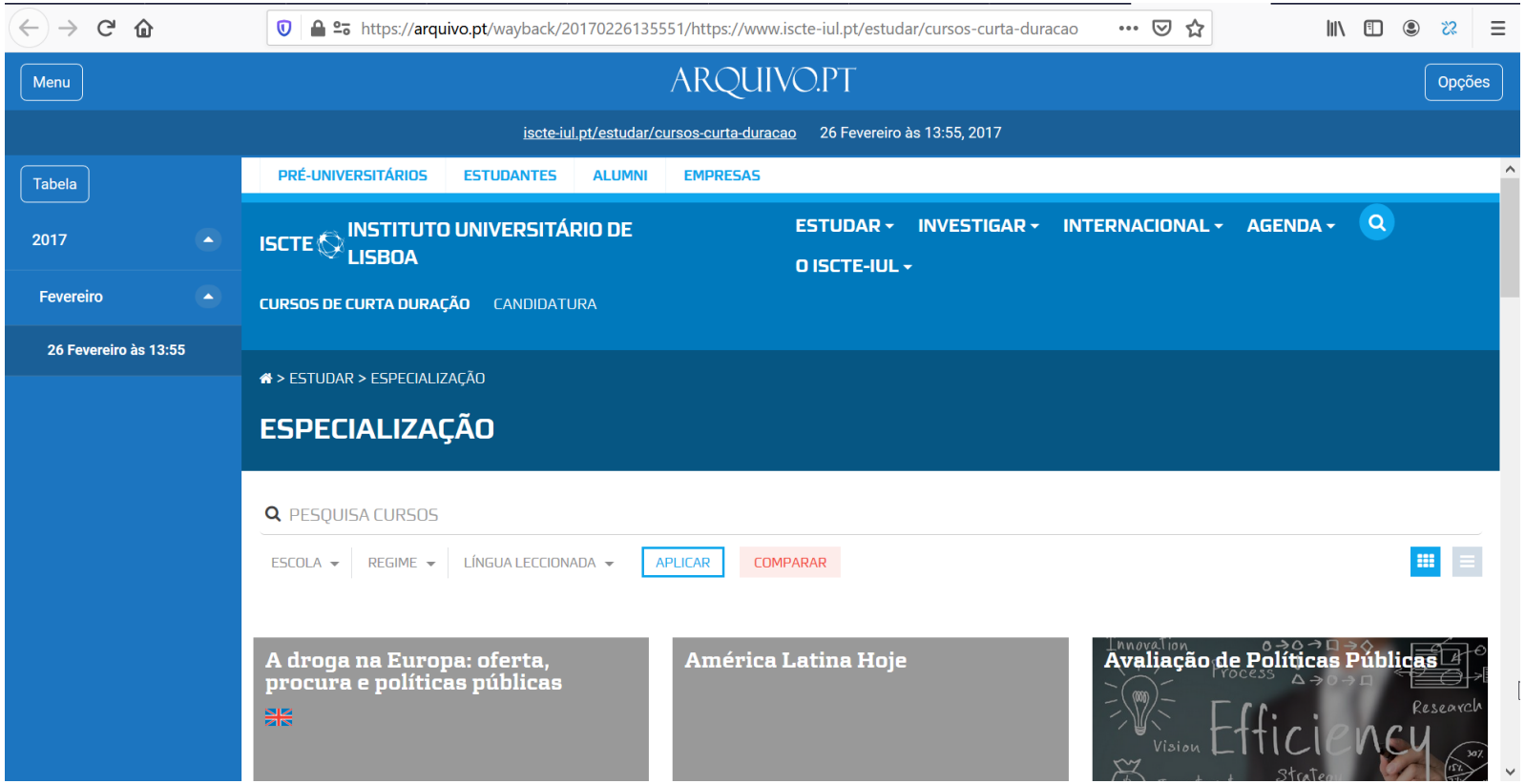
https://www.iscte-iul.pt/estudar/cursos-curta-duracao



https://github.com/arquivo/example-cdx-api

Soft404

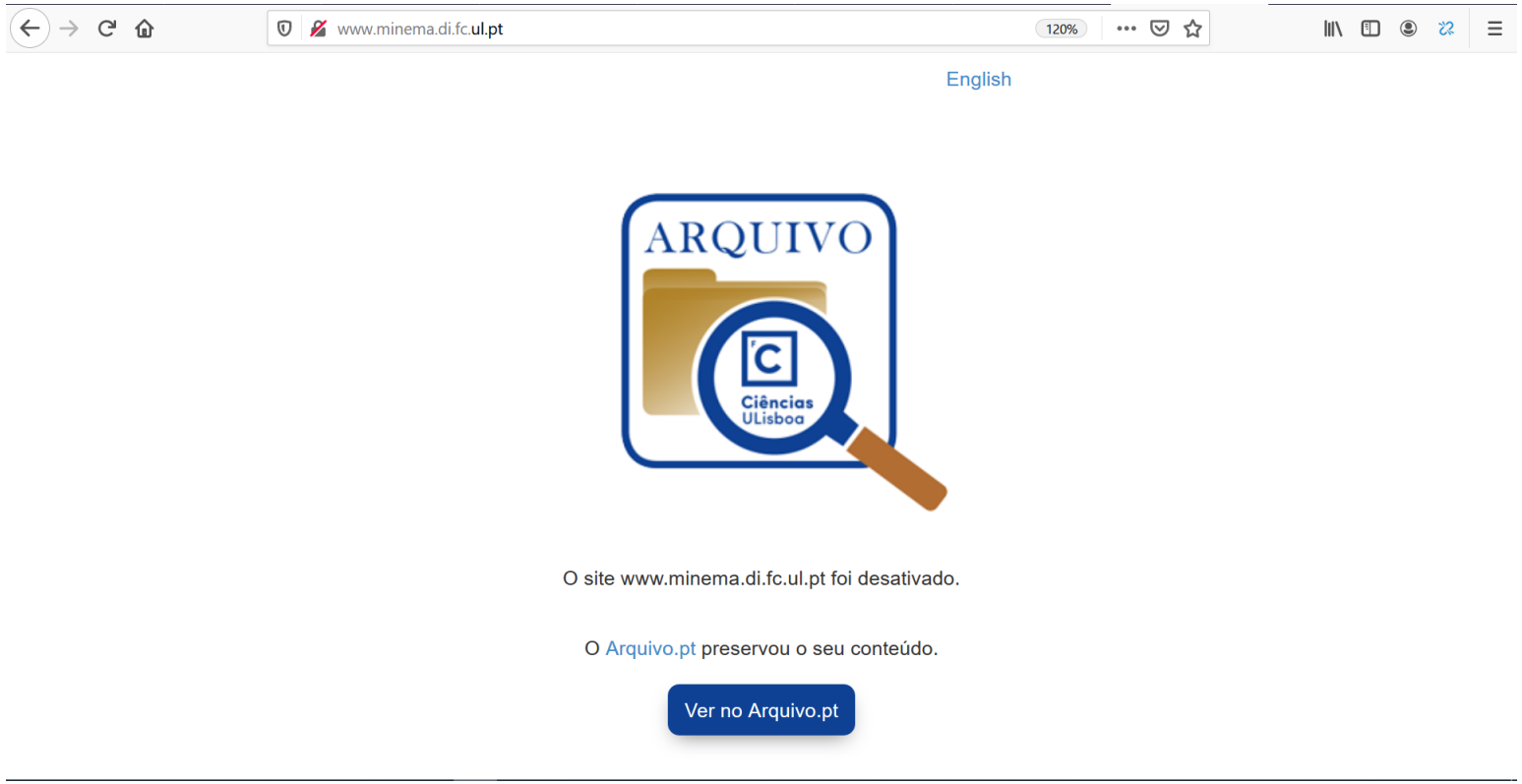
https://www.iscte-iul.pt/estudar/cursos-curta-duracao



https://github.com/arquivo/example-cdx-api

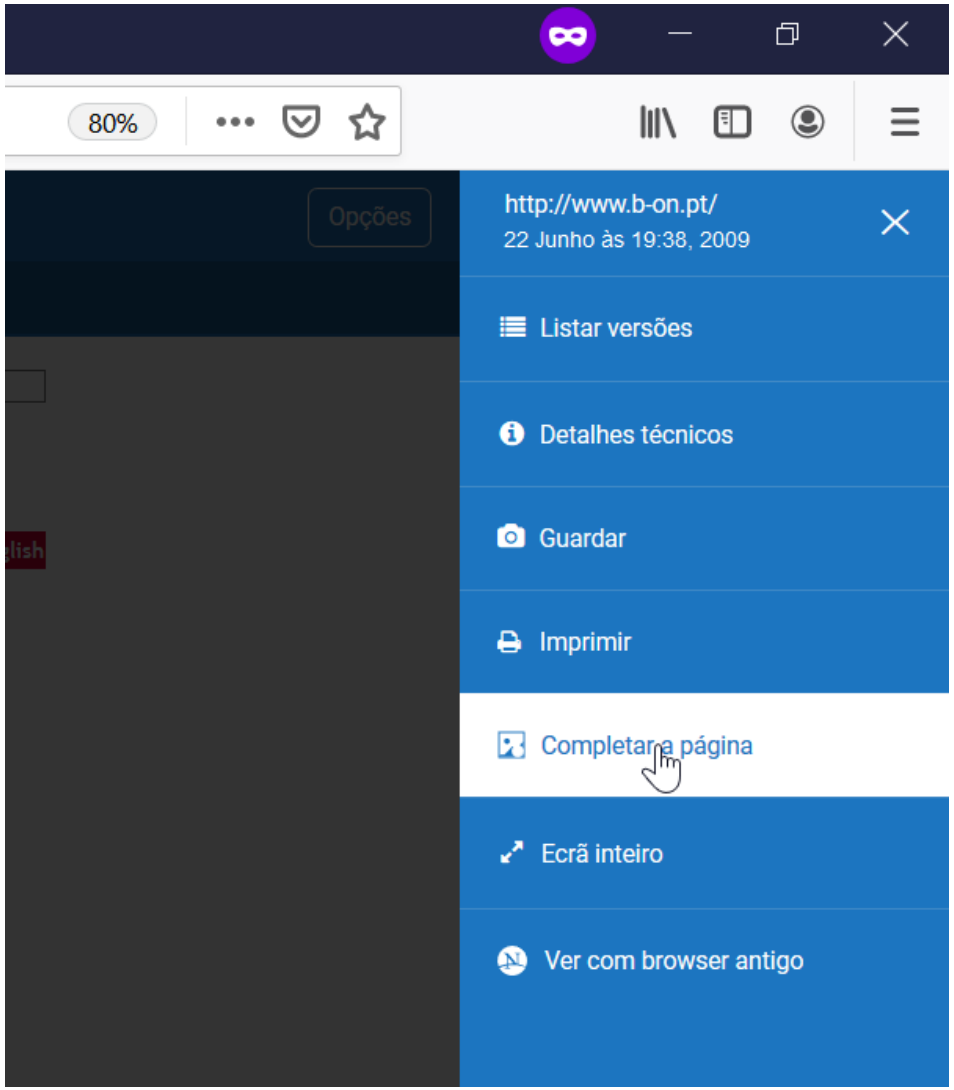
Memorial do Arquivo.pt

Exemplo, site do projeto Minema da Faculdade de Ciências da Universidade de Lisboa
http://www.minema.di.fc.ul.pt/



https://arquivo.pt/memorial

Completar Página



WARC (Web ARChive)

ISO 28500:2017

1

Selecionar

Webrecorder

2

Capturar

WARC

WARC

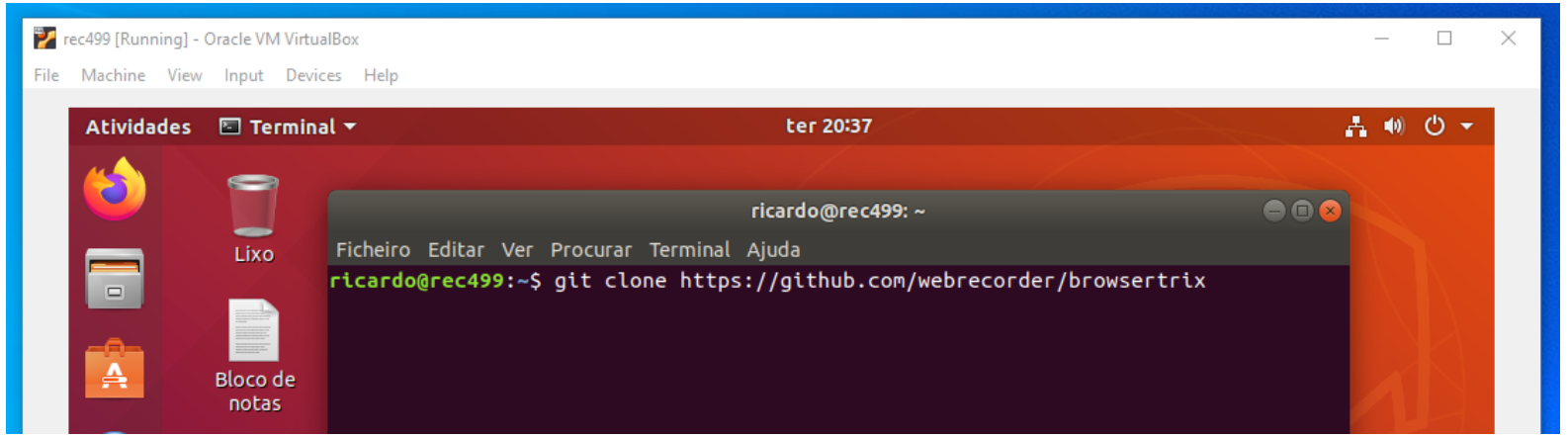
3

Reprodução

ARQUIVO.PT

Browsertrix

[Instalar no próprio computador - Guia de instalação]



Short link: https://tinyurl.com/instalar-browsertrix

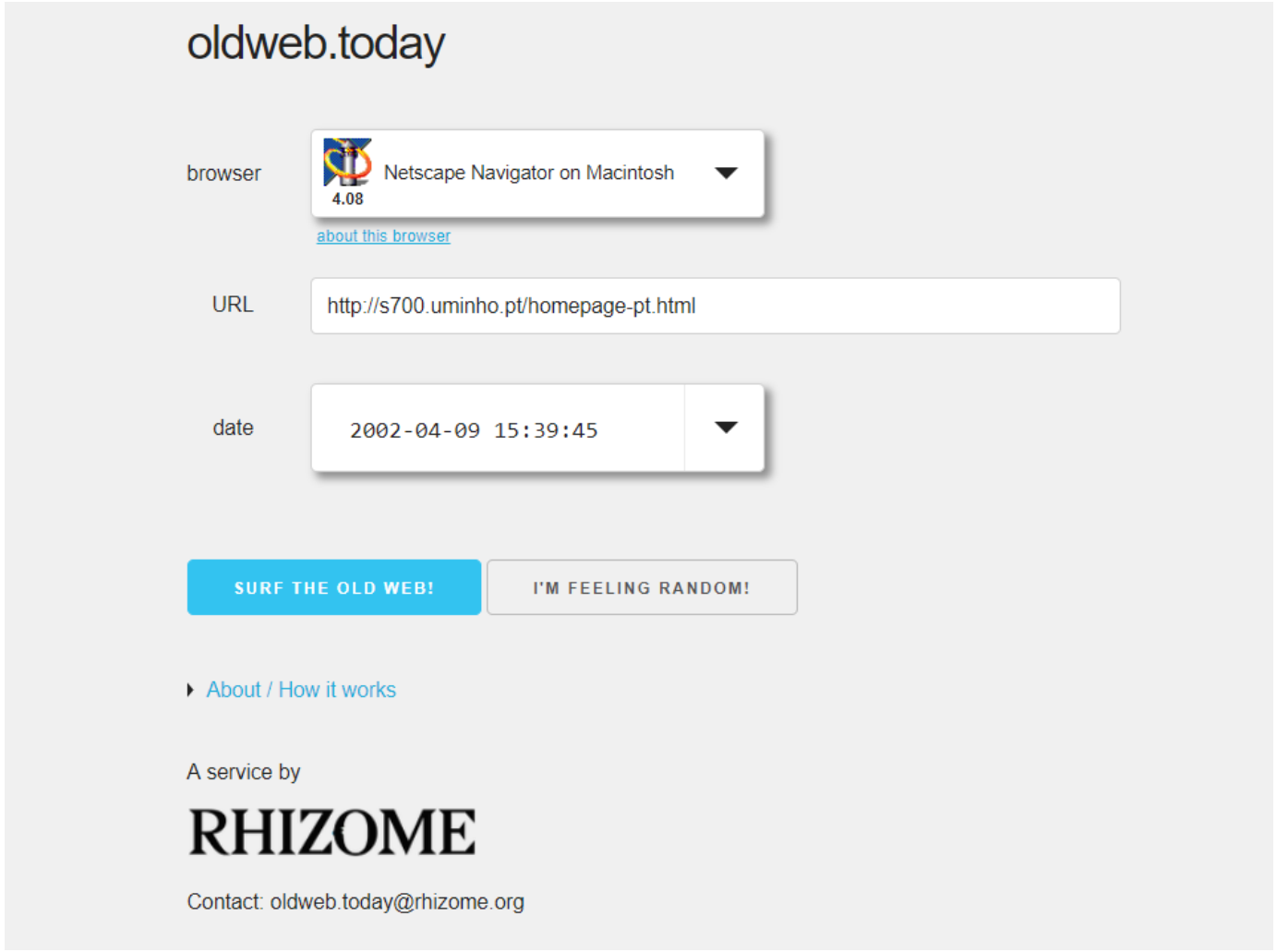
Torne-se curador dos seus websites

Recolhas locais feitas pela instituição ou pessoa proprietária do site :

- Recolhe páginas escolhidas
- Determina a periodicidade que acha adequada
- É objeto de uma verificação humana
- Utiliza ferramentas de recolha de alta qualidade – ex. Webrecorder e Browsertrix
- Gera compromisso com o desenvolvedor do site
- Envolve outros intervenientes - comunicação e imagem, arquivo e a própria gestão

Oldweb.today

Look and Feel do passado



http://oldweb.today

oldweb.today

Time Left
08:27

NCSA Mosaic on Linux 2.2

about this browser

Current Page Archived On:

2002-02-04 12:23:05

Requested Date/Time:

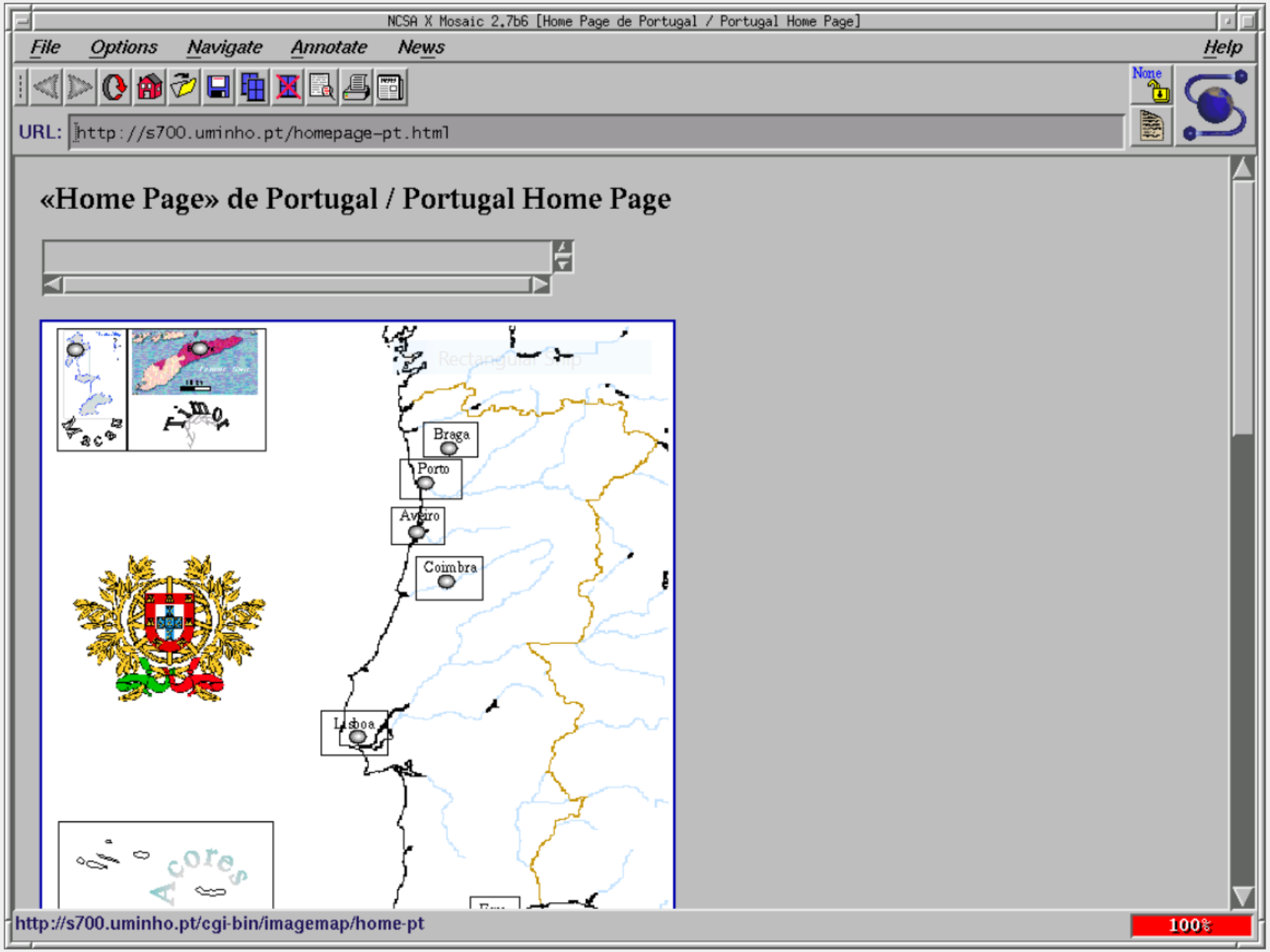
2002-04-09 15:32:31

Loaded **11** resources, spanning

2001-08-05 10:27:41 to 2013-08-02 16:22:08

from public web archives:

- Internet Archive
- Portuguese Web Archive



Source code on github!

Contact: oldweb.today@rhizome.org

Conclusão

Recomendações criar e gerir websites preserváveis

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. Torne-se curador dos seus websites

Recomendações:

arquivo.pt/recomenda