Publish and Preserve

Recommendations to publish preservable information

How to preserve information collected from the Web

Robustify service

How do web archives work?

Web archiving workflow

1

Crawl

INTERNET ARCHIVE

UKWA

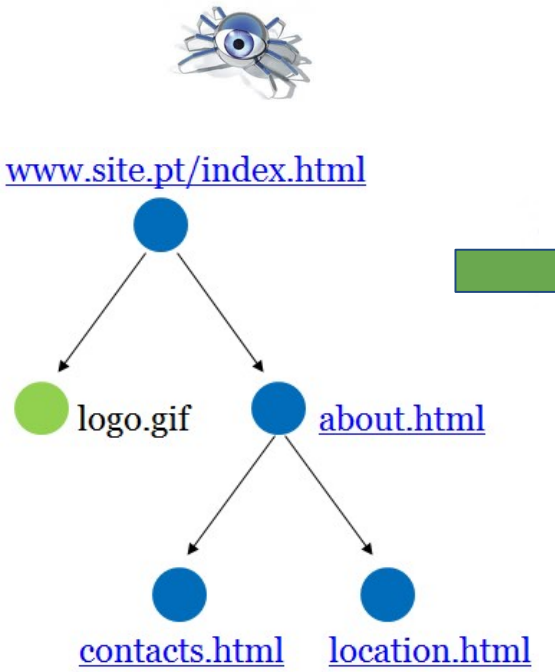
ARQUIVO.PT

Web Archives

2

Replay

Crawls are performed by Robots



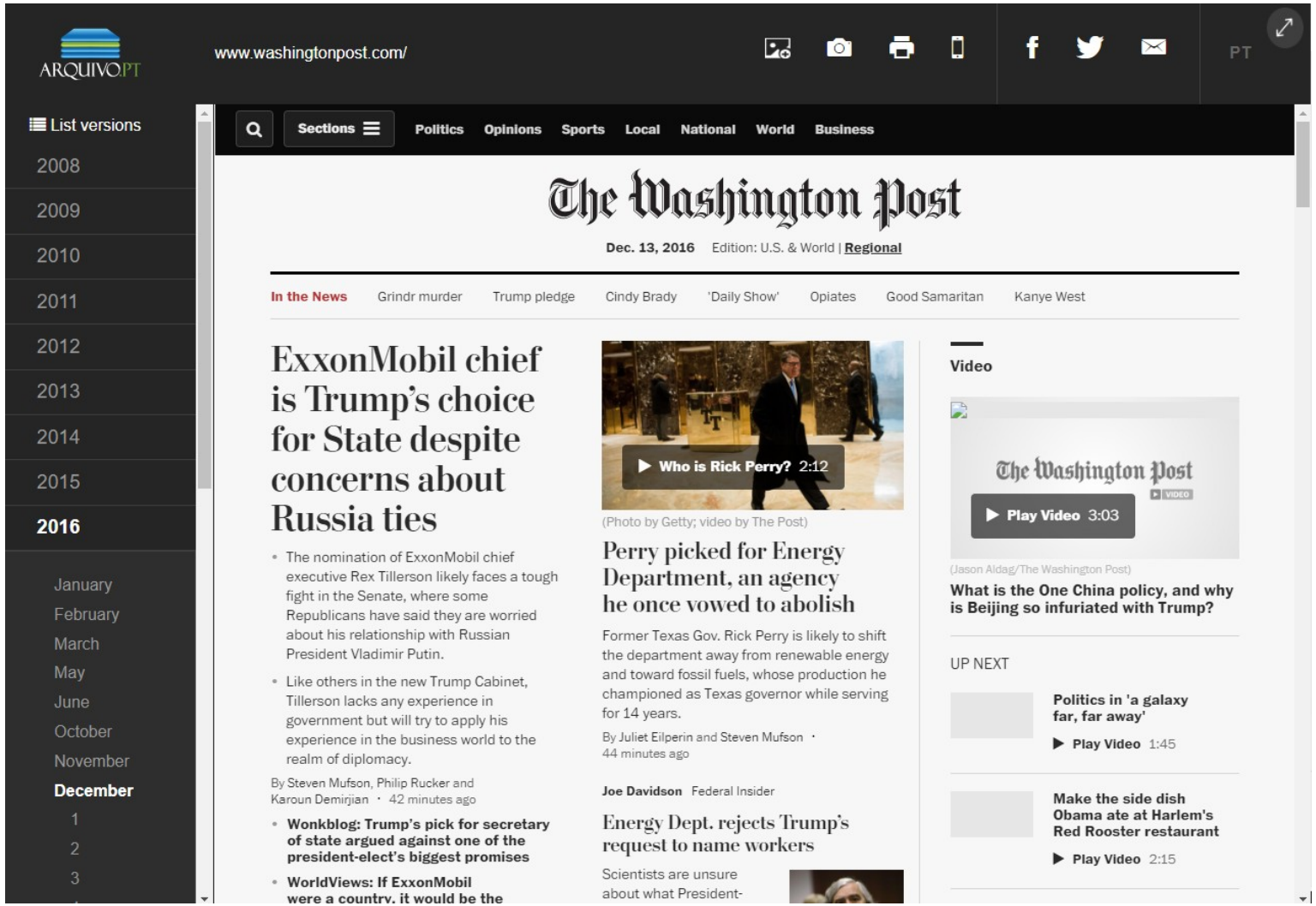
INTERNET ARCHIVE

UKWA

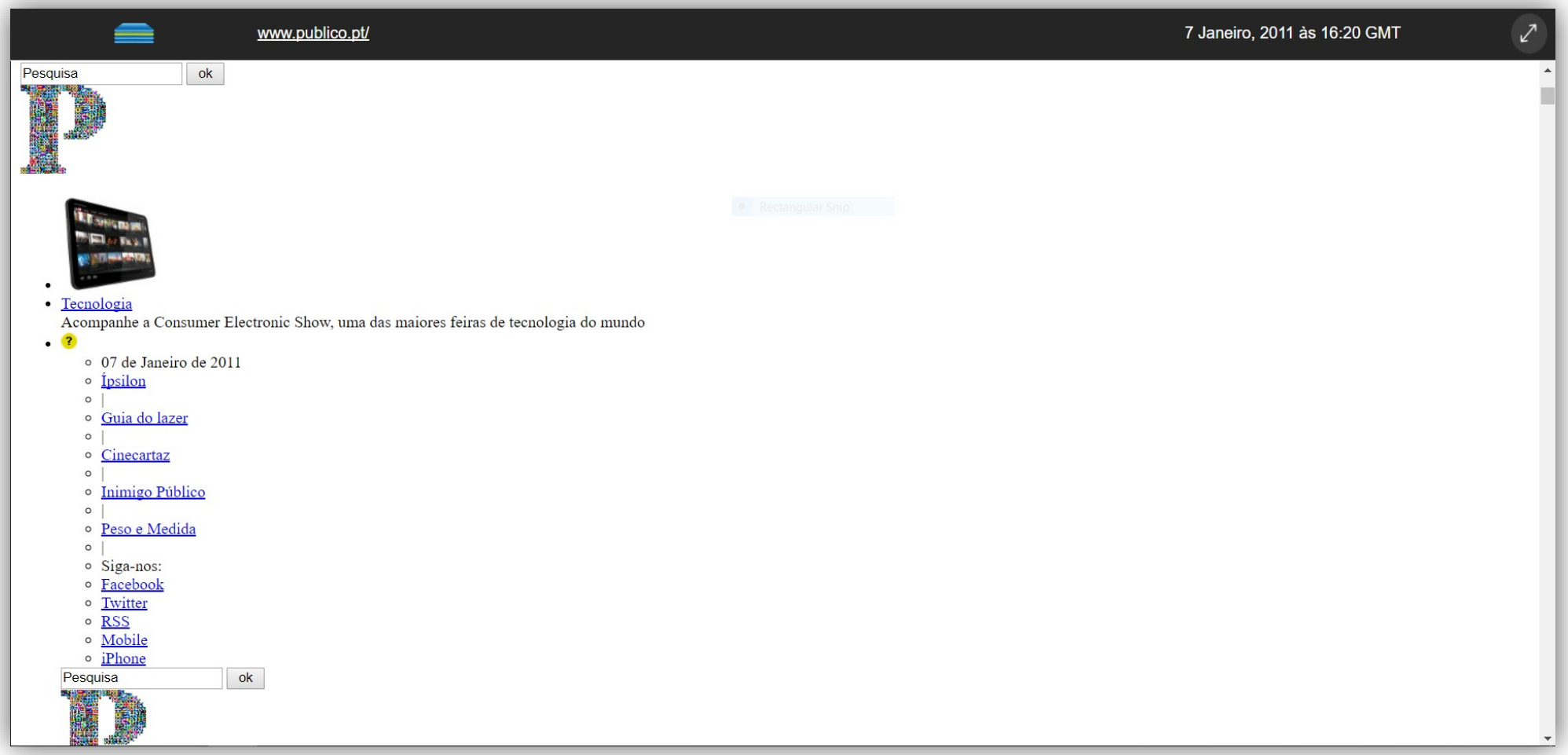
ARQUIVO.PT

Web Archives

Replay of preserved pages



Sometimes, things go wrong. ☹

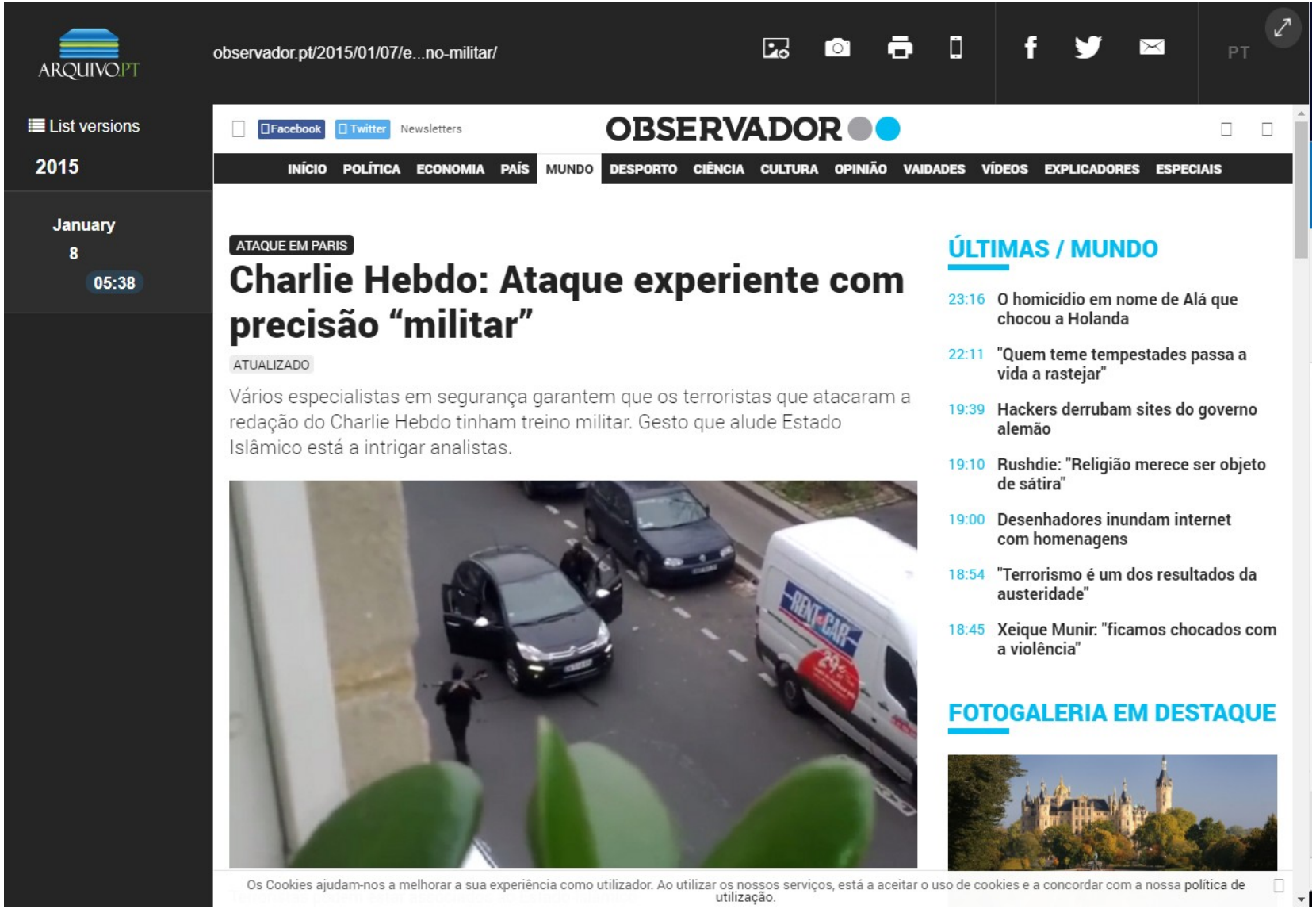


6

Recommendations to **publish preservable** information

Identify the **publication date** (#1)

Which is the publication date?

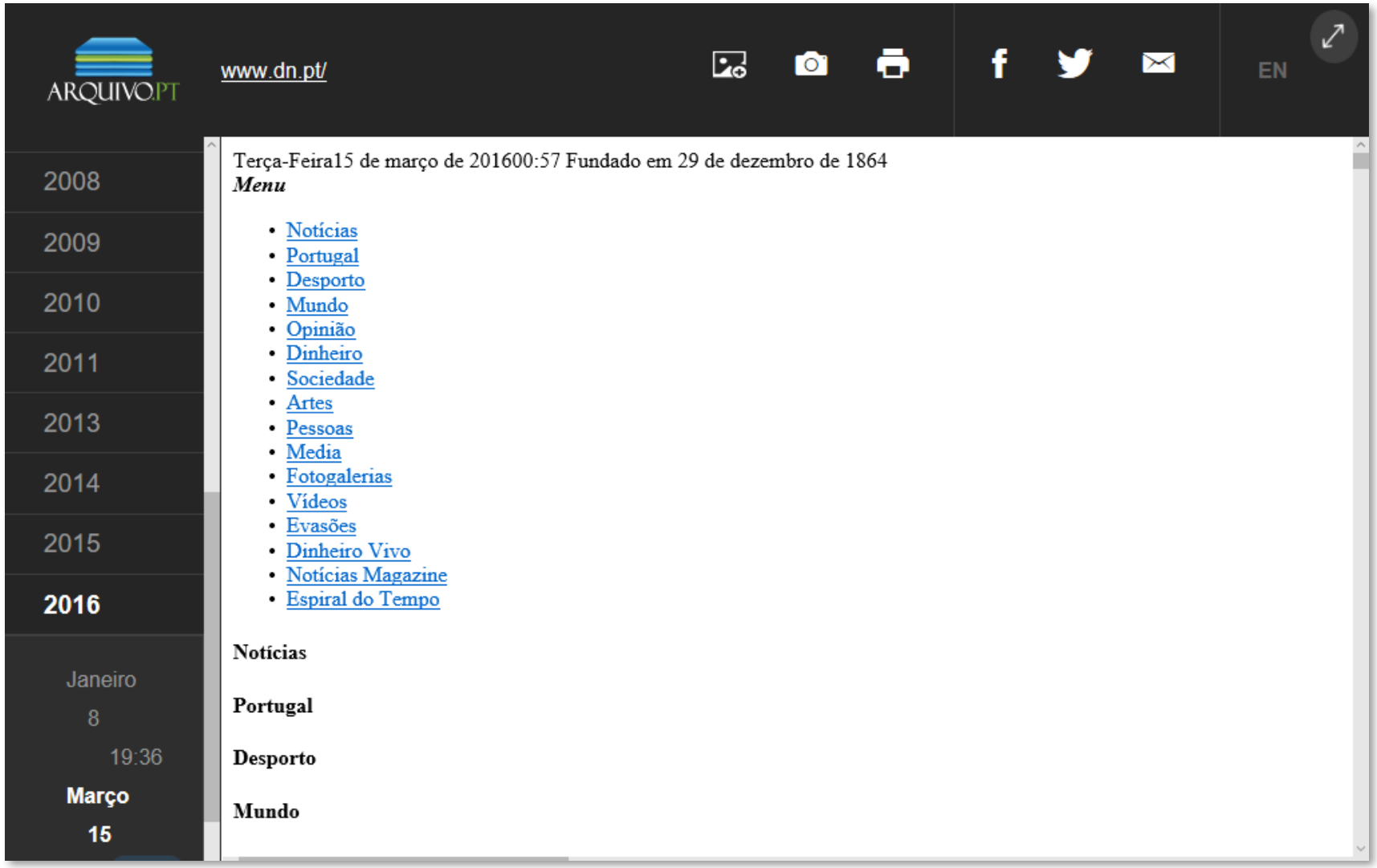


Identify the publication date (#1)

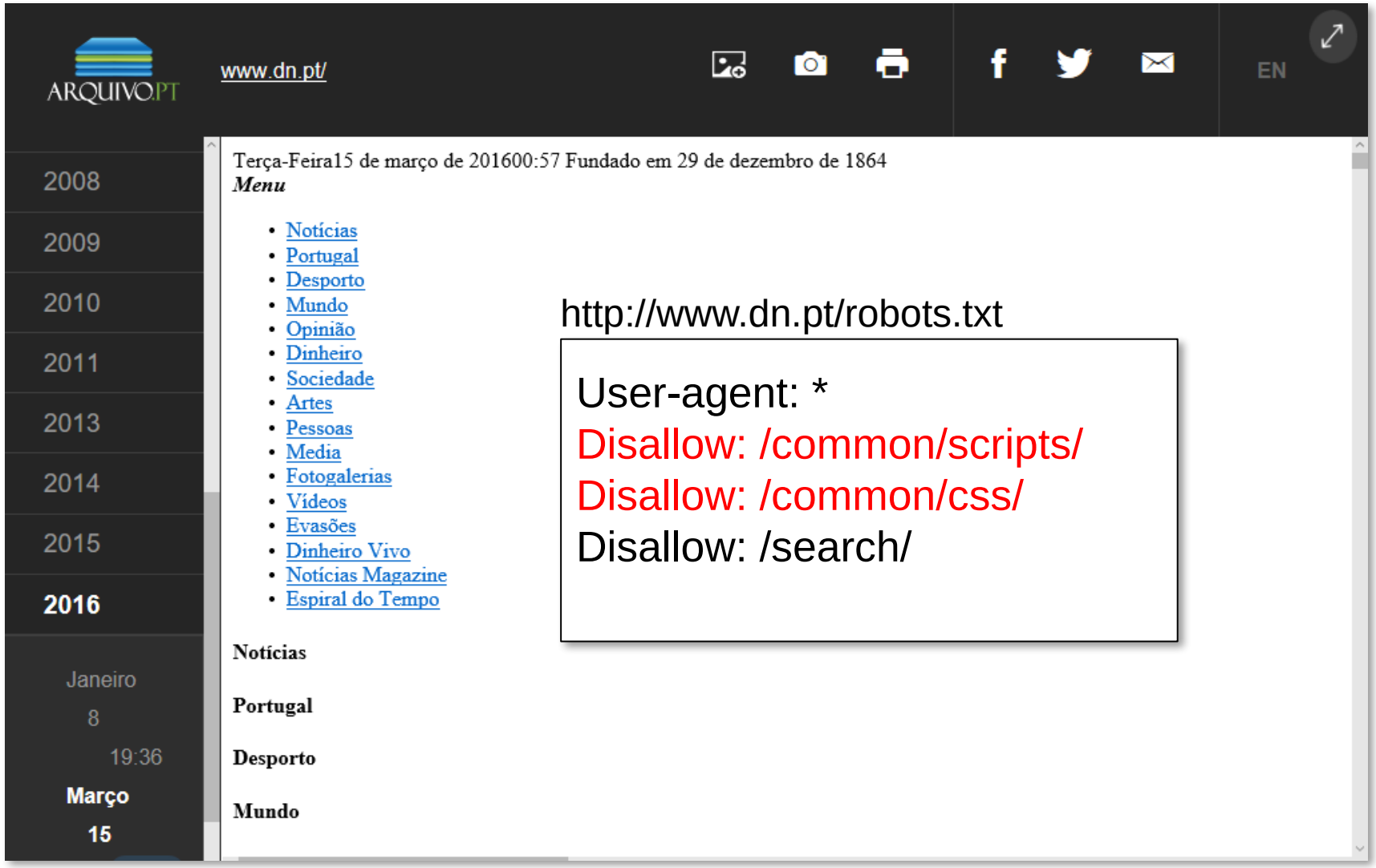
Allow web archive robots (*robots.txt)* #2

DN.pt: how it was (2016)

DN.pt: how it was preserved



Robots Exclusion Protocol was the source of the problem



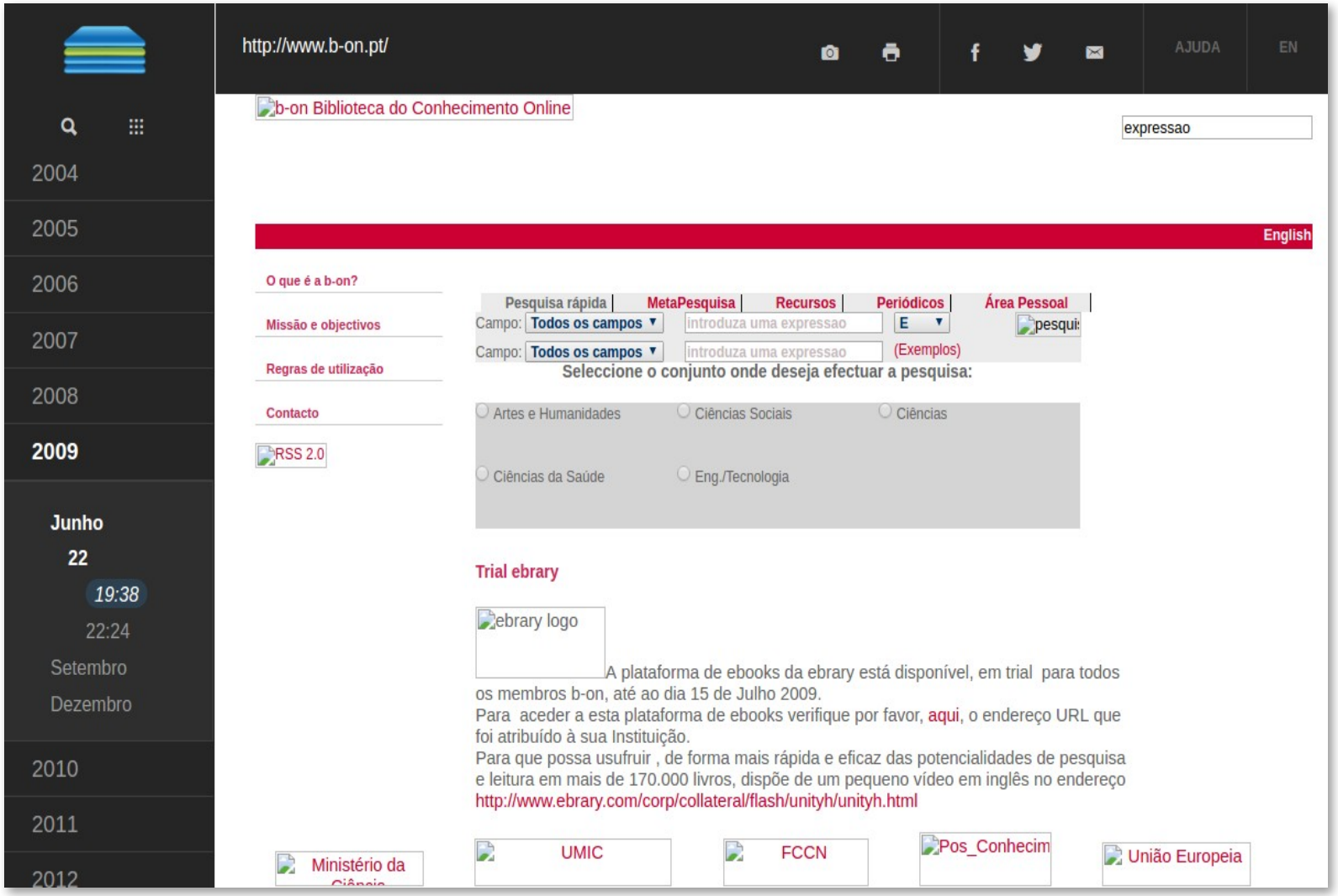
http://www.dn.pt/robots.txt

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

"**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."

source: **Google** Webmaster Central Blog, Updating our technical Webmaster Guidelines

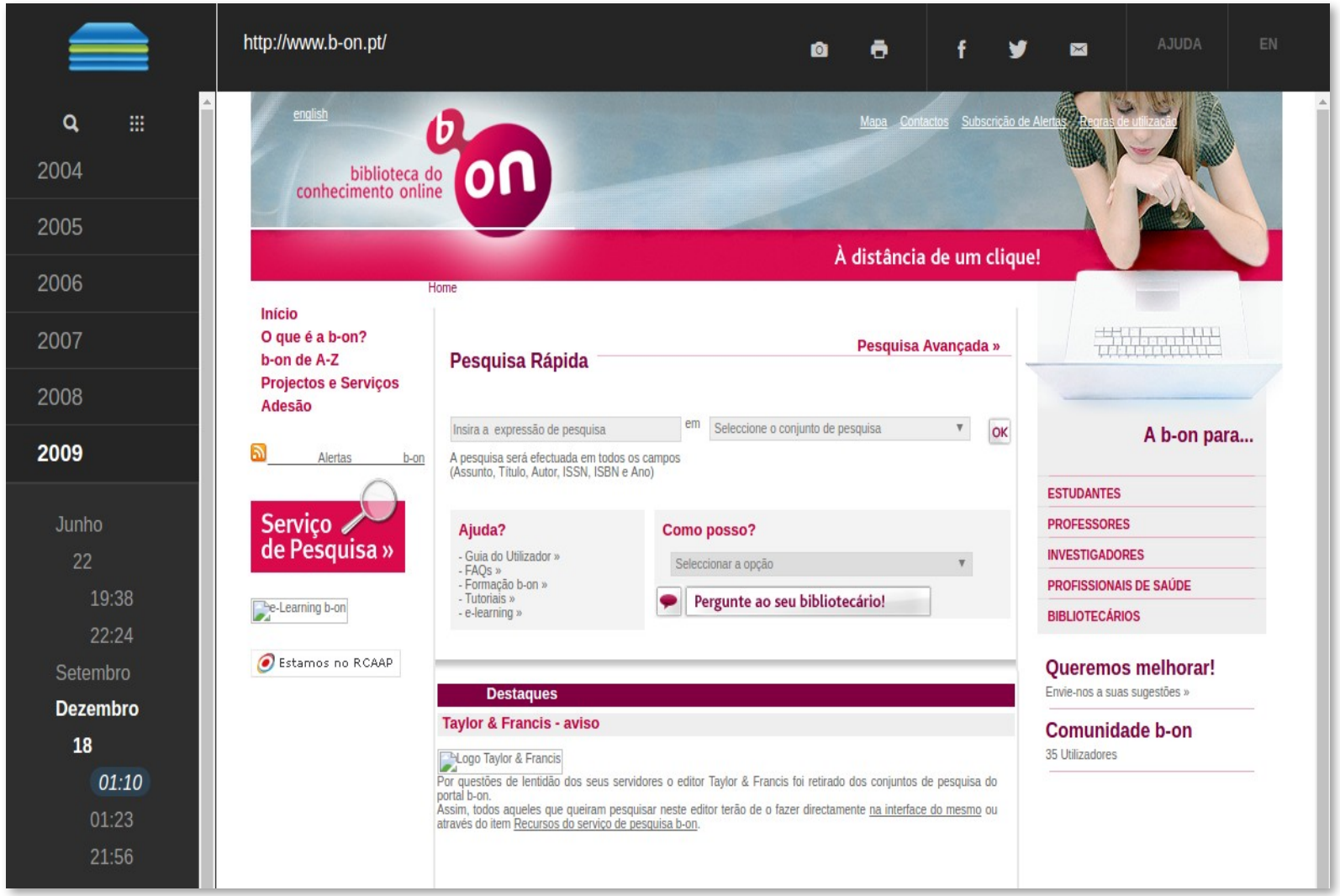
B-on.pt: how it was preserved



Default exclusions by the Content Management Systems (CMS) are the source of the problem

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: *robots.txt* correctly configured



Allow access to Web Archives

Arquivo.pt robot is correctly **identified**.

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

An **U**niversal **R**esource **L**ocator for each **resource** (#3)

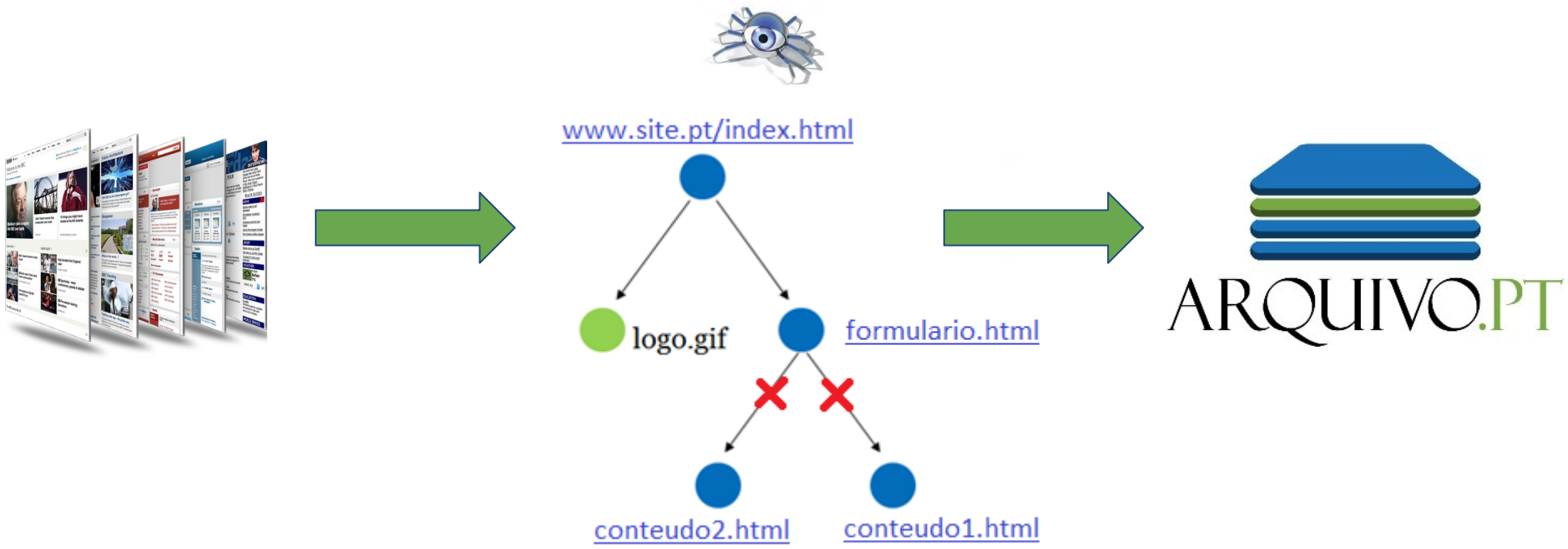
Resources hidden behind forms

Resources hidden behind forms

Não Arquivado

Pesquisar noutros arquivos

Resources hidden behind forms



Site Maps help people (usability) and machines (SEO) to access the information

Maintain the same URL along time (#4)

Problem: broken timeline

Versions list

463 versions of iscte.pt

http://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					
25 Mar	2 Fev	21 Mar	12 Jan	2 Fev	9 Jan	18 Jan	20 Mai					

Versions list

34 versions of http://iscte-iul.pt

http://iscte-iul.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			

Single unbroken timeline

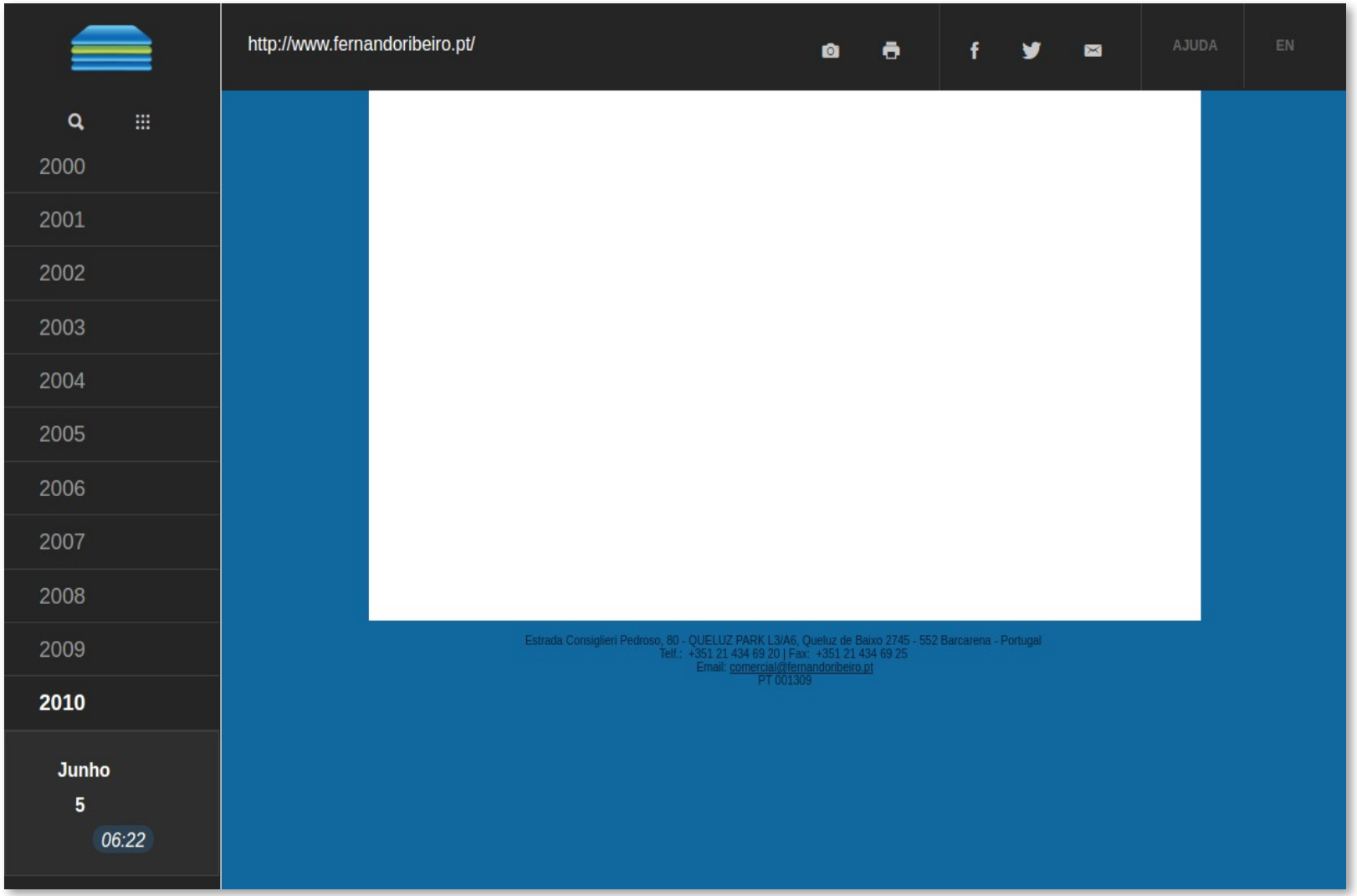
Versions list

463 versions of 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						

Use **preservable** data formats (#5)

Problem: Flash usage



“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/

Choose a media format with:

Freely available, without legal rights that restrict its use.

A standard issue by an official organization (e.g. W3C).

Openly documented through a public and free specification.
Widely used.

Read and written by several software platforms, including open-source technologies.

Not compressed or compressed without information loss.

Preservable media formats

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

Image
PNG (.png)
JPEG2000

Video
AVI without compression (.avi)

Not preservable media formats

Text
Microsoft Word (.doc)

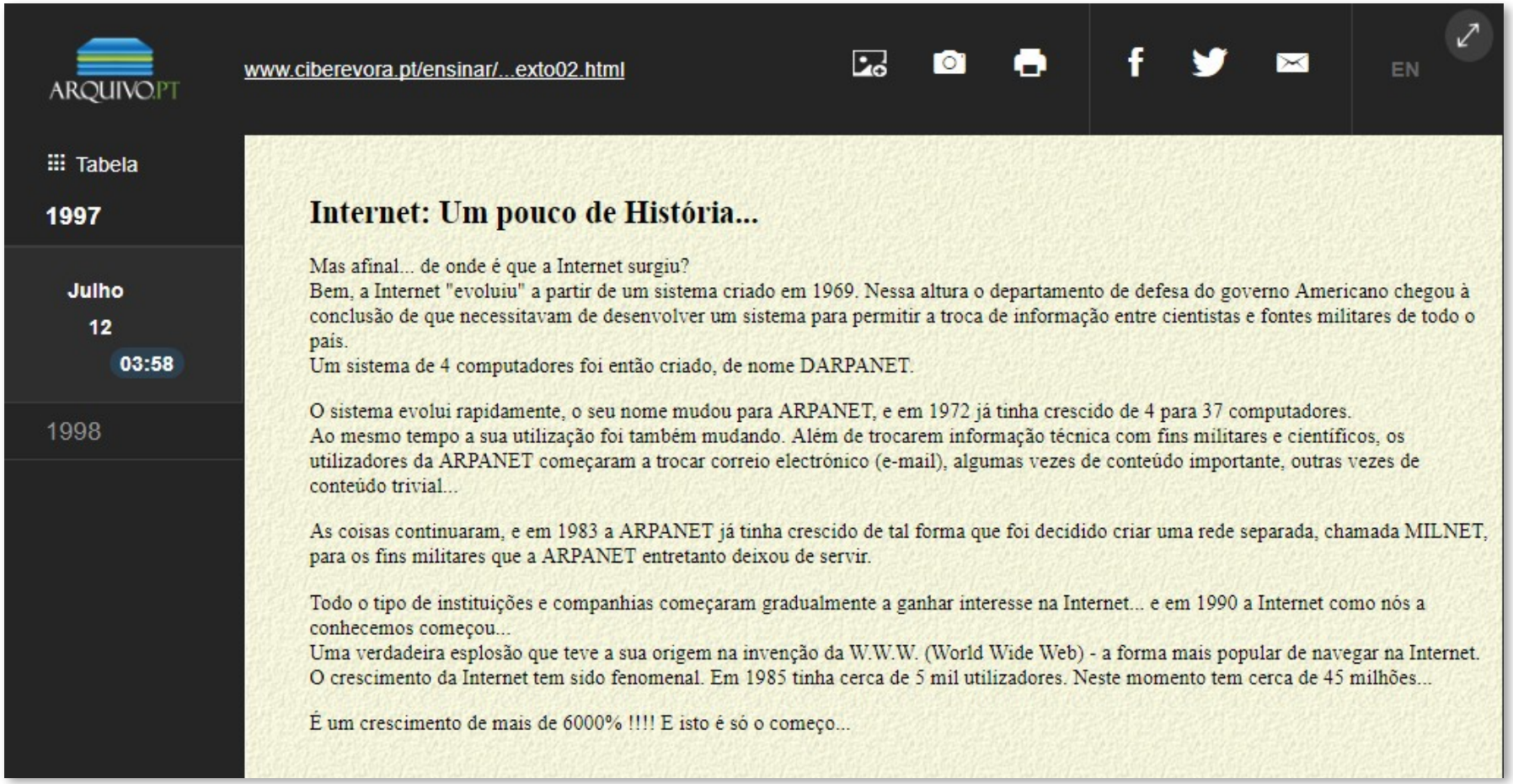
Images
Macromedia Flash (*.swf)
PhotoShop (.psd)

Video
Windows Media Video (.wmv)

Metadata to describe content (#6)

Who is the author?

When was it published?



Internet: Um pouco de História...

Mas afinal... de onde é que a Internet surgiu?
Bem, a Internet "evoluiu" a partir de um sistema criado em 1969. Nessa altura o departamento de defesa do governo Americano chegou à conclusão de que necessitavam de desenvolver um sistema para permitir a troca de informação entre cientistas e fontes militares de todo o país.
Um sistema de 4 computadores foi então criado, de nome DARPANET.

O sistema evolui rapidamente, o seu nome mudou para ARPANET, e em 1972 já tinha crescido de 4 para 37 computadores.
Ao mesmo tempo a sua utilização foi também mudando. Além de trocarem informação técnica com fins militares e científicos, os utilizadores da ARPANET começaram a trocar correio electrónico (e-mail), algumas vezes de conteúdo importante, outras vezes de conteúdo trivial...

As coisas continuaram, e em 1983 a ARPANET já tinha crescido de tal forma que foi decidido criar uma rede separada, chamada MILNET, para os fins militares que a ARPANET entretanto deixou de servir.

Todo o tipo de instituições e companhias começaram gradualmente a ganhar interesse na Internet... e em 1990 a Internet como nós a conhecemos começou...
Uma verdadeira esplosão que teve a sua origem na invenção da W.W.W. (World Wide Web) - a forma mais popular de navegar na Internet.
O crescimento da Internet tem sido fenomenal. Em 1985 tinha cerca de 5 mil utilizadores. Neste momento tem cerca de 45 milhões...

É um crescimento de mais de 6000% !!!! E isto é só o começo...

Metadata usage (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" />
```

This information summarizes, enrich or completes the contents adding more value to the information.

Computers can use this information.

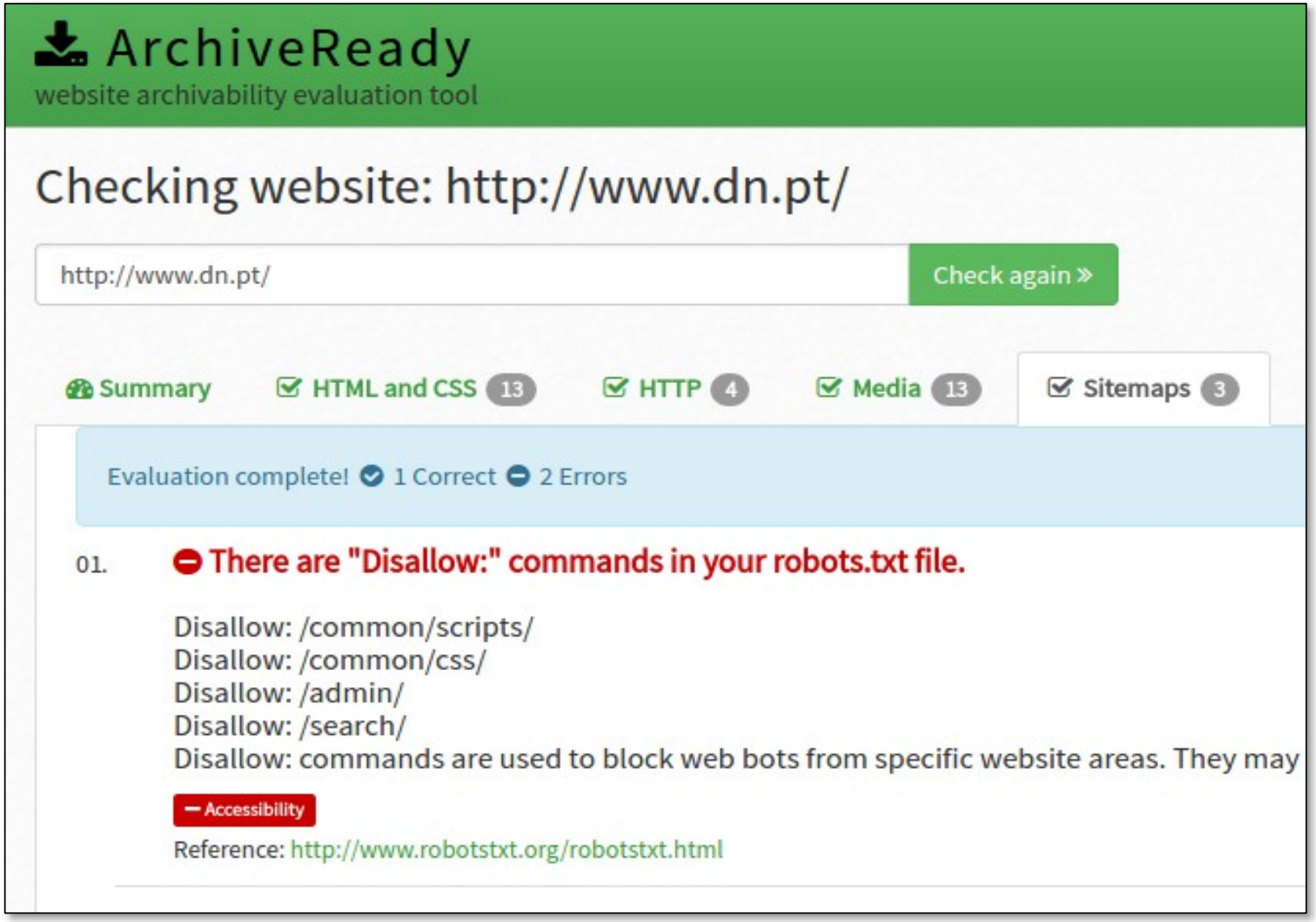
Recommendations to avoid preservation problems

1. Identify the **publication date**
2. **Allow** web archive robots
3. An **U**niversal **R**esource **L**ocator for each **resource**
4. Maintain the **same URL along time**
5. Use **preservable** data formats
6. Metadata to **describe** content

How preservable is a website?

http://archiveready.com

http://archiveready.com



Create your own Web Archive

Web Recorder

https://webrecorder.io

A Project By
RHIZOME

Webrecorder / temp-ZKEA7XN5 / Temporary Collection / Recording

Submit a bug Temporary Collection Sign Up Login

Capturing 2.58 MB Current Browser http://www.tpdl.eu/tpdl2018/

REGISTRATION PROGRAM + WORKSHOPS CONFERENCE INFORMATION + TRAVEL INFORMATION + CONTACT

TPDL 2018

TPDL 2018

Porto
Portugal
Sep. 10-13 2018

22nd International Conference on Theory and Practice of Digital Libraries

FACULTY OF ENGINEERING, UNIVERSITY OF PORTO

Digital Libraries for Open Knowledge

When

The TPDL 2018 Conference will take place from Monday,

How

Interested in participating in the TPDL 2018 Conference?

Who

Want to know more about the people behind the TPDL

Where

The TPDL 2018 Conference will be held at the Faculty of

Twitter

Tweets by @tpdl2018

1

Select

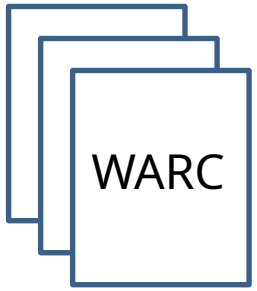
Webrecorder

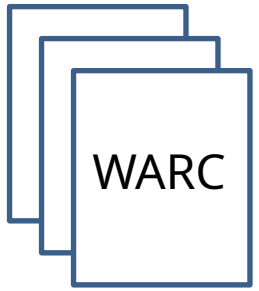
2

Capture

WARC (Web ARChive)

ISO 28500:2017





3

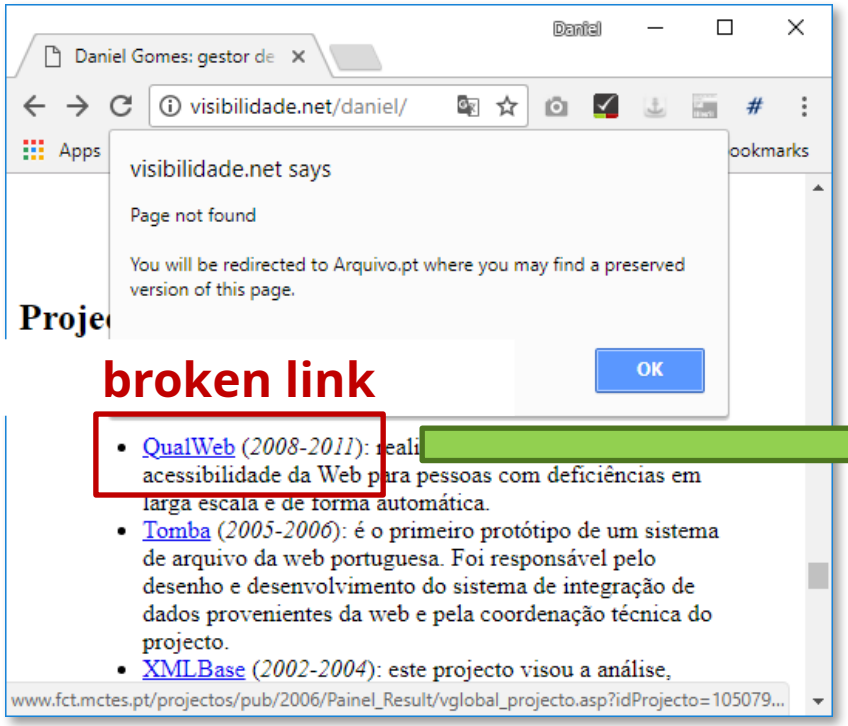
Replay

ARQUIVO.PT

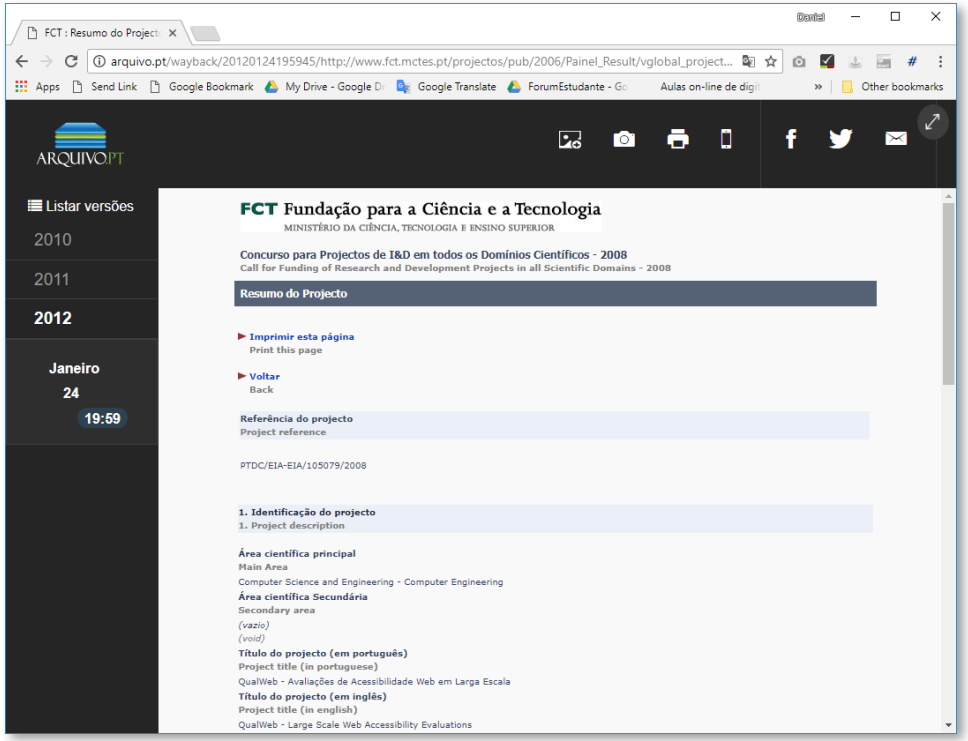
Robustify

Minimizes errors due to broken links, redirecting to a web-archived resources

robustify.arquivo.pt: broken links redirect to preserved pages @Arquivo.pt



Live-web page with broken link



Users follow link to preserved page at Arquivo.pt

How to use robustify in your website

Insert this snippet of code in your website:

```
<script src="http://robustify.arquivo.pt/robustifyArquivoPT.js"></script>
<script> robustify({}); </script>
```

Recommendations:
arquivo.pt/recommendations

daniel.bicho@fccn.pt

TPDL 2018