

# Search the Past Web with Arquivo.pt

Forum on the archiving of the  
Polish Internet resources

[daniel.gomes@fccn.pt](mailto:daniel.gomes@fccn.pt)



ARQUIVO.PT

# Who are we?



**Free online** service to research the Past Web

Preserves **publicly accessible** information related with:

- Portugal
- **Research** and **Education** (international)

Governmental service provided by  
Foundation for Science and Technology (Portugal)

A digital research infrastructure

# Arquivo.pt begun in 2007

Menu

Tabela

2008

Março

16 Março às 04:52

ARQUIVO.PT

arquivo-web.fccn.pt/portuguese-web-archive?set\_language=en16 Março às 04:52, 2008

Opções

Site MapAccessibilityContact

Search

☐ only in current section

HomeCrawlerTeam

You are here: HomeEnglishPortuguês

## Portuguese Web Archive

### Welcome to the Tomba project: the Portuguese web archive

Publishing tools, such as Blogger, enabled people with limited technical skills to become web publishers. Never before in the history of mankind so much information was published. However, it was never so ephemeral. Web documents such as news, blogs or discussion forums are valuable descriptions of our times, but most of them will not last longer than one year.

If we do not archive the current web contents, the future generations could witness an information gap in our days.

The [Internet Archive](#) collects and stores contents from the world-wide web. However, it is difficult for a single organization to archive the web exhaustively while satisfying all needs, because the web is permanently changing and many contents disappear before they can be archived.

As a result, several countries are creating their own national archives to ensure the preservation of contents of historical relevance to their cultures.

Portugal is now beginning its national web archiving initiative with the Tomba project at [FCCN](#) (National Foundation for Scientific Computing).

Send this — Print this —



Fundação para a Computação Científica Nacional








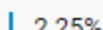

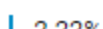



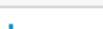

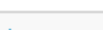

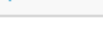

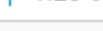

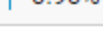
UMIC  
Agência para a Sociedade do Conhecimento

The Plone® CMS — Open Source Content Management System is © 2000-2008 by the Plone Foundation et al.  
Plone® and the Plone logo are registered trademarks of the Plone Foundation. Distributed under the GNU GPL license.

Powered by PloneValid XHTMLValid CSSSection 508WCAG

# Arquivo.pt is used word-wide



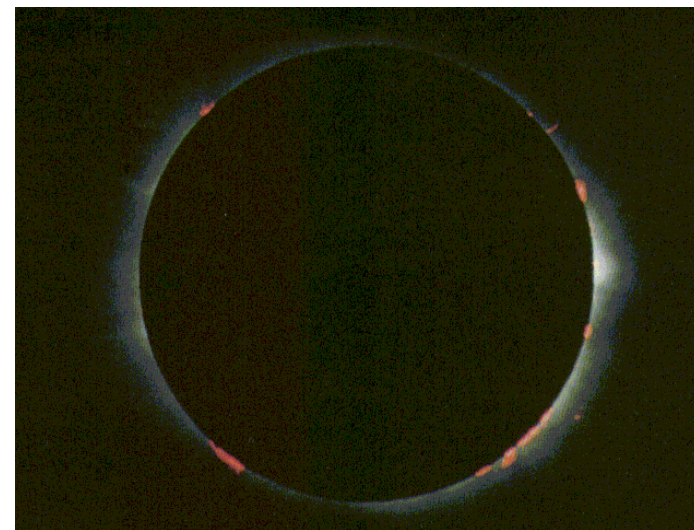
Country	Users	% Users
1.  Portugal	46,891	 46.56%
2.  United States	26,373	 26.19%
3.  Brazil	2,266	 2.25%
4.  Russia	2,234	 2.22%
5.  United Kingdom	2,231	 2.22%
6.  Japan	2,172	 2.16%
7.  Canada	1,237	 1.23%
8.  Mozambique	1,213	 1.20%
9.  India	902	 0.90%
10.  Germany	894	 0.89%

- 53% of users are international
- User Interfaces and documentation also in English
- Combined with Google Translate enables **cross-lingual access** to preserved content

# Arquivo.pt preserves national and international historical web content

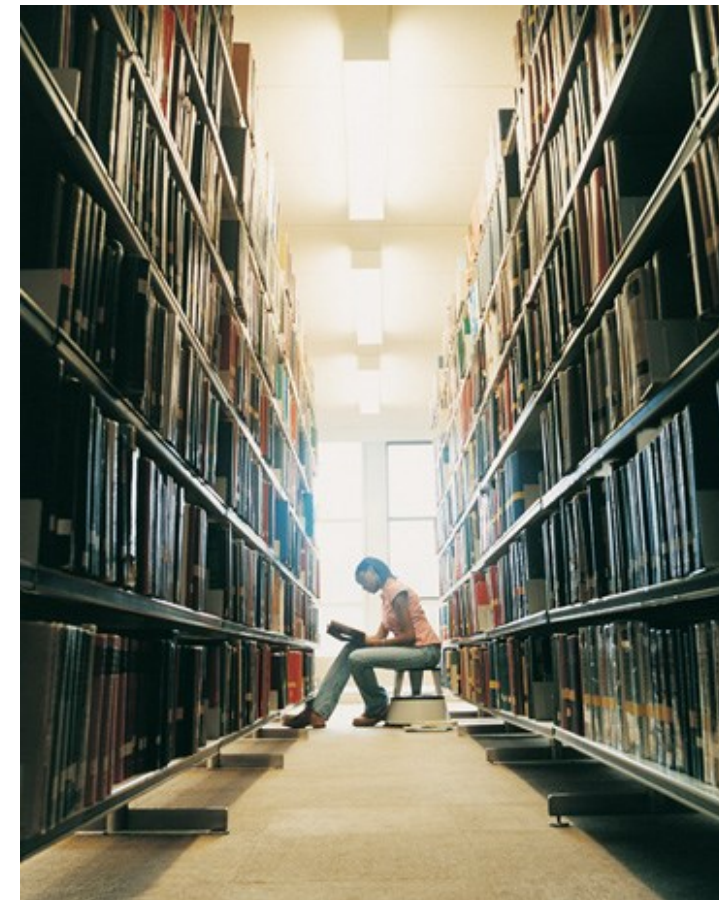
<b><i>GALILEO</i></b> Versão 2.0	
Janeiro 1993	
Manual de utilização e sugestões de exploração	<b><i>AUTORES</i></b> <b>Programação:</b> João Veloso <b>Manual:</b> Elisa Prata Pina e M. Augusta Patrício <b>Orientação:</b> Carlos Fiolhais
Departamento de Física da Universidade de Coimbra	
<b><i>ÍNDICE</i></b>	
<a href="#">1. Apresentação sumária do programa</a> <a href="#">2. Características do equipamento</a> <a href="#">3. Configuração mínima do computador</a> <a href="#">4. Ficheiros do programa</a>	

[nautilus.fis.uc.pt](http://nautilus.fis.uc.pt) - 1993  
(oldest page)



[spacelink.nasa.gov](http://spacelink.nasa.gov) – 1992  
(oldest image)

How to **search** and **access**  
the preserved information?



# Search **texts** from the past in any language

[Menu](#) ARQUIVO.PT [Options](#)

[Search](#)

1996  
1 Jan

2021  
12 Nov

[Pages](#) [Images](#) [Narrative](#) [Advanced search](#)

About 23,322 results since 1996 until 2021

→ [berlin.msz.gov.pl/pl](http://berlin.msz.gov.pl/pl)

**Ambasada RP w Berlinie**

30 March 2019

Ambasada RP w Berlinie Masz wyłączoną obsługę języka JavaScript Polski Deutsch RSS Wersja tekstowa Być wiernym Ojczyźnie mej, Rzeczypospolitej Polskiej Aktualności **Ambasada** Współpraca dwustronna ... się zwrócić? Zasady realizacji projektów polonijnych Praktyki **Ambasada** **Ambasada** **Ambasada RP** w ...

→ [dublin.msz.gov.pl/pl](http://dublin.msz.gov.pl/pl)

**Ambasada RP w Dublinie**

10 June 2017

Ambasada RP w Dublinie Masz wyłączoną obsługę języka JavaScript English Polski RSS Wersja tekstowa Polsce – służyć, Europę – tworzyć, Świat – rozumieć Aktualności **Ambasada** Współpraca dwustronna ... WYDZIAŁEM KONSULARNYM **Ambasada** **Ambasada RP** w Dublinie 5 Ailesbury Road Ballsbridge D04 W221 Dublin ...



# Search the **history** of a web address

Menu ARQUIVO.PT

Q bbc.co.uk Search

1996 1 Jan 2021 12 Nov

Pages Images Narrative Advanced search

Cerca de 2,538 resultados desde 1996 a 2021

Table List

2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
15 Feb	4 May	29 May	4 Jan	21 Jan	10 Jan	5 Sep	1 Jan	1 Jan	20 Jan	1 Jan	1 Jan	1 Jan	3 Jan
14 Mar	20 May	30 May	6 Jan	27 Jan	11 Jan	7 Sep	2 Jan	2 Jan	9 Feb	1 Jan	2 Jan	1 Jan	3 Jan
22 Oct	20 May	4 Aug	12 Jan	31 Jan	12 Jan	23 Sep	3 Jan	4 Jan	10 Feb	2 Jan	3 Jan	2 Jan	3 Jan
22 Oct	11 Jun		12 Jan	11 Mar	15 Jan	24 Sep	4 Jan	8 Jan	3 Apr	2 Jan	4 Jan	3 Jan	4 Jan
23 Oct	22 Jun		14 Jan	12 Mar	17 Jan	26 Sep	5 Jan	8 Jan	4 Apr	2 Jan	5 Jan	4 Jan	4 Jan
	6 Aug		21 Jan	13 Mar	18 Jan	27 Sep	6 Jan	17 Jan	5 Apr	3 Jan	6 Jan	5 Jan	4 Jan
	6 Aug		21 May	14 Mar	19 Jan	28 Sep	8 Jan	17 Jan	6 Apr	3 Jan	7 Jan	6 Jan	6 Jan
	3 Sep		1 Jul	15 Mar	22 Jan	30 Sep	10 Jan	21 Jan	7 Apr	3 Jan	8 Jan	7 Jan	6 Jan
	3 Sep		1 Dez	16 Mar	23 Jan	4 Oct	10 Jan	24 Jan	8 Apr	3 Jan	9 Jan	8 Jan	6 Jan
	4 Sep		2 Dez	17 Mar	26 Jan	5 Oct	12 Jan	28 Jan	9 Apr	3 Jan	10 Jan	9 Jan	6 Jan
	8 Sep			18 Mar	29 Jan	6 Oct	17 Jan	30 Jan	10 Apr	3 Jan	11 Jan	10 Jan	8 Jan

# Browse [bbc.co.uk](http://bbc.co.uk) 10 years ago

Menu

Table

2008

February

15 February at 02:28

March

October

2009

2010

2011

January

May

July

1 July at 07:43

December

2012

2013

2014

2015

2016

ARQUIVO.PT

bbc.co.uk 1 July at 07:43, 2011

BBC Mobile


News Sport Weather Travel TV Radio More

Search the BBC


**Top News story**

**'Doubts' over Strauss-Kahn case**


The sex assault case against Dominique Strauss-Kahn appears to be in trouble amid reported concerns about the alleged victim's credibility.



**News**

**Hu warns communists on corruption**

President Hu Jintao warns China's communists that corruption could cost them the support of the people, as the party marks its 90th anniversary.

**Chavez tells of cancerous tumour**

Venezuelan President Hugo Chavez says he has had a second operation to remove a cancerous tumour, in his first televised speech since flying to Cuba for treatment.

**Moroccans vote on king's reforms**

**US in CIA detainee death inquiry**


Poland to take over EU presidency

Crowds welcome royals in Canada

Japan manufacturers 'pessimistic'

**Spotlight**

**One in a Billion**



**China's haves and have-nots**

It is officially a Communist country - but economic data points to a stark gap between rich and poor.

Building dreams: Dai Zhikang

From hawker to billionaire: Zhou Xiaoguang

China's top art collector: Liu Yiqian

**Most Popular in News**

Shared Read Watched/Listened

Meet the world's most powerful computer

Monaco Prince Albert to wed Olympic star

China celebrates longest sea bridge


# Search **images** from the past

Menu
ARQUIVO.PT
Options


1996 1 Jan
2021 12 Nov

Pages Images Narrative
Advanced search


About 4,550,586 results since 1996 until 2021




→ dentons.com/en/fin...  
5 September 2019 at 11:06




→ epsmaps.com/search...  
30 May 2010 at 05:50




→ warsawguide.com...  
18 May 2016 at 05:17




→ timeshighereducatio...  
7 June 2019 at 02:28




→ epsmaps.com/Wars...  
24 December 2009 at 17:32




→ timeshighereducatio...  
9 December 2019 at 21:19




→ ilconservation.org/s...  
13 May 2016 at 15:21




→ realpoland.eu/packa...  
14 June 2017 at 00:31




→ ilconservation.org/s...  
13 May 2016 at 15:22




→ vidiani.com/large-ro...  
3 December 2019 at 11:49




→ nationsonline.org/o...  
1 December 2019 at 00:35




→ vidiani.com/detailed...  
3 December 2019 at 11:44




→ vidiani.com/large-de...  
3 December 2019 at 11:45



→ lawyerpoland.eu/ho...  
31 March 2019 at 03:22




→ 9flats.com/radmieci...  
22 December 2019 at 11:11



→ cdn.neurope.eu/nati...  
28 November 2014 at 01:46

ARQUIVO.PT



Visit
Details

**Image**

**Title:**  
Warsaw city centre

**Alt text:** Warsaw city centre

**Caption:** Poland's education system has seen some extraordinary changes in recent years. Following a huge drive in the nation's Katowice #6 1001+ Wrocław University of Science and Technology Wrocław Read mor...

**URL:**  
timeshighereducation.com/sites/default/files/styles/the\_breaking\_news\_image\_style/public/warsaw\_city\_c...  
entre.jpg?tok=NYFWGvNj

**Resolution:** jpeg 620 x 413

**Capture date:** 7 June 2019 at 02:28

**Page**

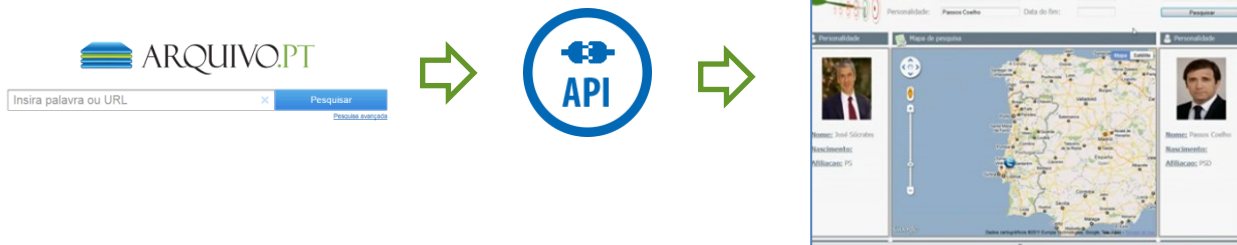
**Title:**  
Best universities in Poland | Times Higher Education (THE)

**URL:** timeshighereducation.com/student/best-universities/best-universities-poland

**Capture date:** 7 June 2019 at 02:28

# Application Programming Interfaces to develop new applications

Your new app



## Arquivo.pt APIs

### APIs

Daniel Gomes edited this page 11 days ago · 33 revisions

APIs specific to Arquivo.pt that enable the full exploration of its functions

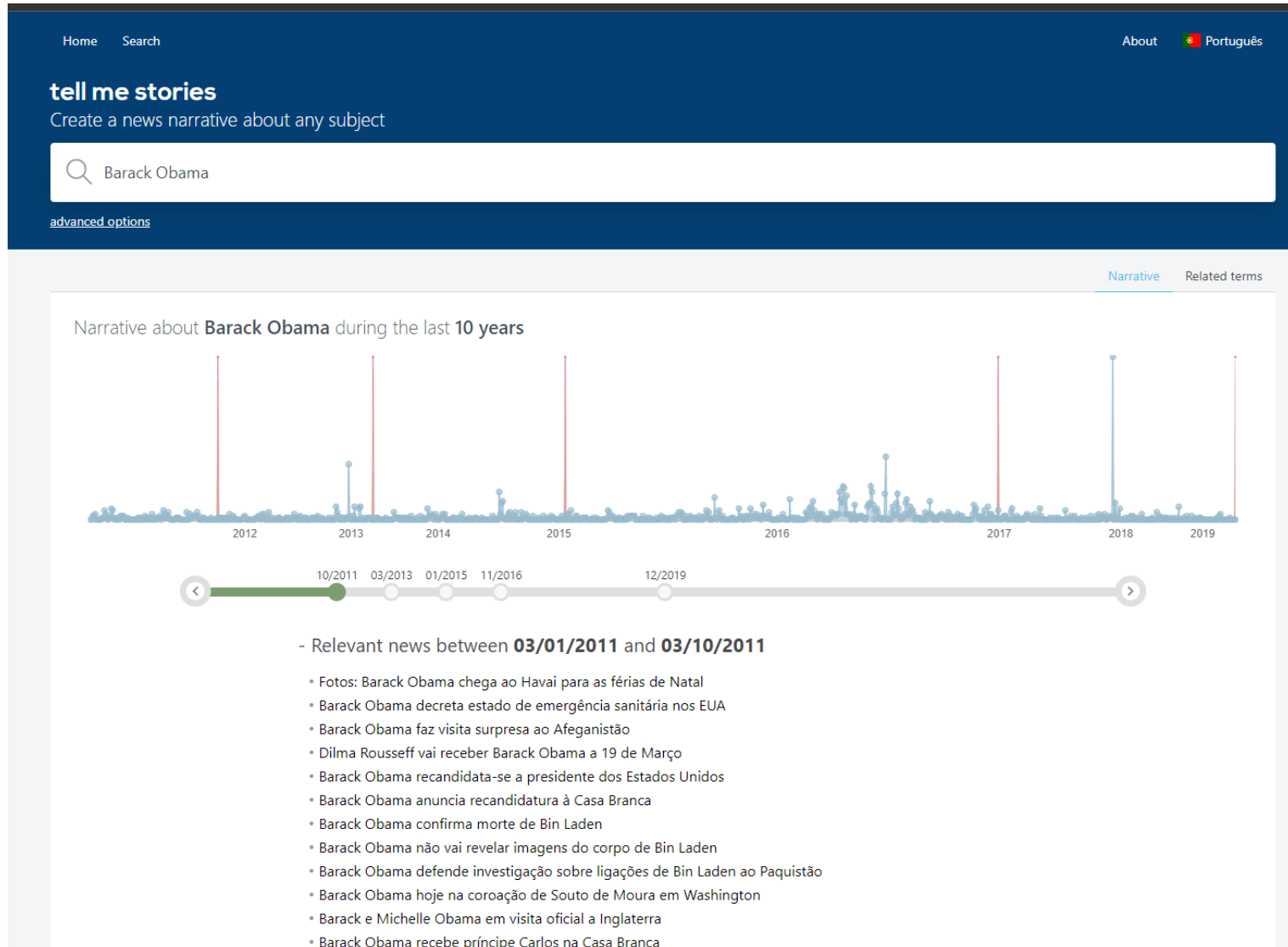
- [Arquivo.pt API \(Full-text & URL search\)](#)
- [Image Search API v1.1 \(beta version\)](#)

APIs based on international standards to enable interoperability among web archives and code reuse

- [CDX-server API \(URL search\)](#)
- [Memento API \(URL search\)](#)

[arquivo.pt/api](https://arquivo.pt/api)

# Arquivo.pt API in action: **TellMeStories.pt**



**Automatically** generates **narratives** about any subject based on online news from the Past.

Winners of the Arquivo.pt Award 2018.

# Collections and automatic selection

# Public list of Arquivo.pt **collections**

List of the collections preserved by Arquivo.pt (publicly available) ☆ 📄

File Edit View Insert Format Data Tools Add-ons Help Last edit was made on 10 September by SAW - FCCN Serviços e Arquivos Web

	A	B	C	D	E	F	G	H
	Collection ID	Description	Collection Start date	Collection End date	Total files (Number of URLs collected)	Total seeds (Preserved sites)	Total volume of WARCS files (TB)	Collection in Production? (URL, Page and Image indexed)
2	AWP1	1st Complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-02-12	2008-03-06	56,046,288	154,787	1.60	✓
3	AWP2	2nd Complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-03-11	2008-05-30	48,718,404	-	1.60	✓
4	AWP3	3rd Complete crawl of the Portuguese web, mainly from the .PT domain, in 2008.	2008-10-21	2008-12-10	51,863,006	193,294	2.00	✓
5	AWP4	4th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-05-01	2009-05-31	68,776,707	366,880	2.50	✓
6	AWP5	5th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-10-01	2009-10-31	119,135,566	373,323	3.80	✓
7	AWP6	6th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2009.	2009-12-01	2009-12-31	118,810,364	340,018	3.50	✓
8	AWP7	7th Complete crawl of the Portuguese web, mainly from the .PT domain, in 2010.	2010-05-01	2010-05-31	87,988,812	389,957	2.90	✓
9	AWP8	Incremental crawl of the Portuguese web, mainly from the .PT domain, in 2010. The AWP8 crawl is incremental because it was performed using DeDuplicator ( <a href="http://landsbokasafn.github.io/DeDuplicator/">http://landsbokasafn.github.io/DeDuplicator/</a> ) taking the content of AWP7 as baseline. Thus, the files that remained unchanged from the AWP7 complete crawl were not archived (duplicated) on the AWP8 incremental crawl.	2010-08-01	2010-08-31	75,771,317	411,562	1.90	✓

[arquivo.pt/collections](http://arquivo.pt/collections)



Developed a low-cost methodology to automatically **select** and preserve online information about given topics

### Preserving Websites Of Research & Development Projects

Daniel Bicho  
Foundation for Science and Technology:  
Arquivo.pt  
Av. do Brasil, 101  
1700-066 Lisboa, Portugal  
daniel.bicho@fccn.pt

Daniel Gomes  
Foundation for Science and Technology:  
Arquivo.pt  
Av. do Brasil, 101  
1700-066 Lisboa, Portugal  
daniel.gomes@fccn.pt

International Conference on Digital Preservation 2016, available at [arquivo.pt/publications](http://arquivo.pt/publications)



# Experiments to preserve European Union web content

 ARQUIVO.PT

Arquivo.pt Award  
**2021**  
Meet the winners

Menu

[sobre.arquivo.pt](#) > [News](#) > Arquivo.pt preserved websites about Research & Development projects funded by the EU

**Arquivo.pt preserved websites about Research & Development projects funded by the EU**

*April 10, 2017*

 ARQUIVO.PT

Arquivo.pt Award  
**2021**  
Meet the winners

Menu

[sobre.arquivo.pt](#) > [News](#) > Cross-lingual collection about the 2019 European Elections is available

**Cross-lingual collection about the 2019 European Elections is available**

*September 4, 2020*



Horizon 2020  
European Union funding  
for Research & Innovation


# Tools to support web preservation

# Suggest websites to be preserved

- Public form
  - Any domain
  - Any language
  - Submit home page
- [arquivo.pt/suggest](https://arquivo.pt/suggest)

## Suggest websites to be preserved

International websites are welcome!

sawfccn@gmail.com [Switch accounts](#) 

\*Required

Email \*

Your email address

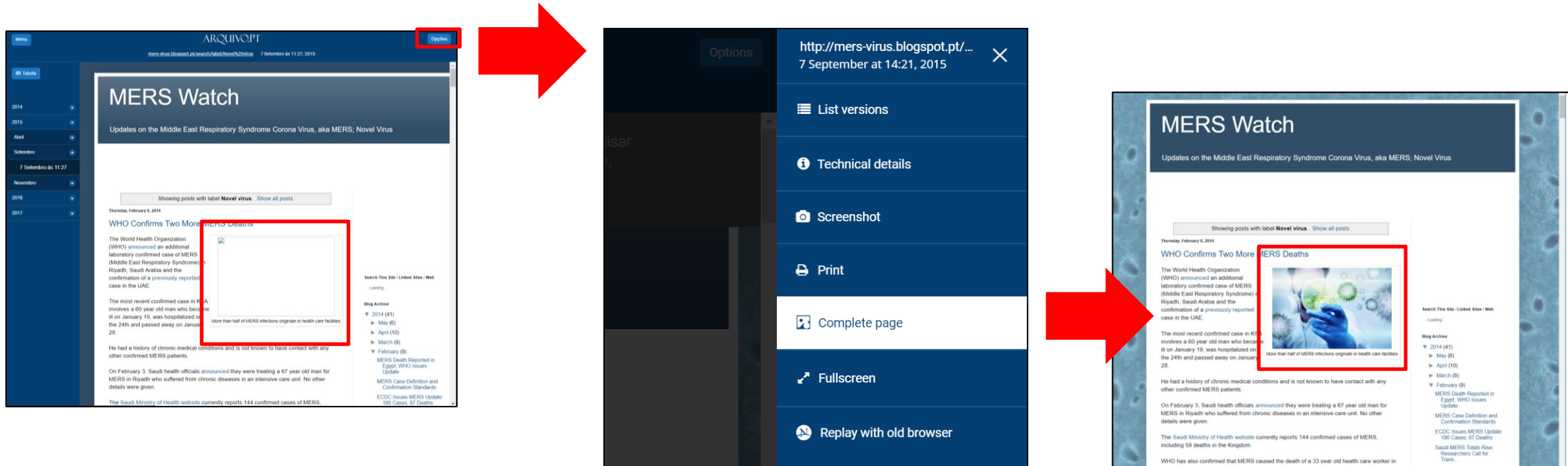
Website (1 per line) \*

Your answer

Description of the Websites

Your answer

# Missing content? Complete the page!

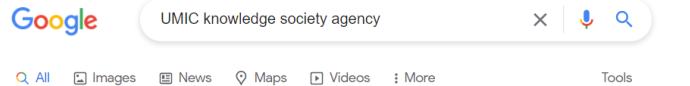


Missing files obtained from the live-web/archives are then integrated to improve the quality of the archived page

➤ Collaborative curation: **Give it a try!**

# Don't kill your historical website!

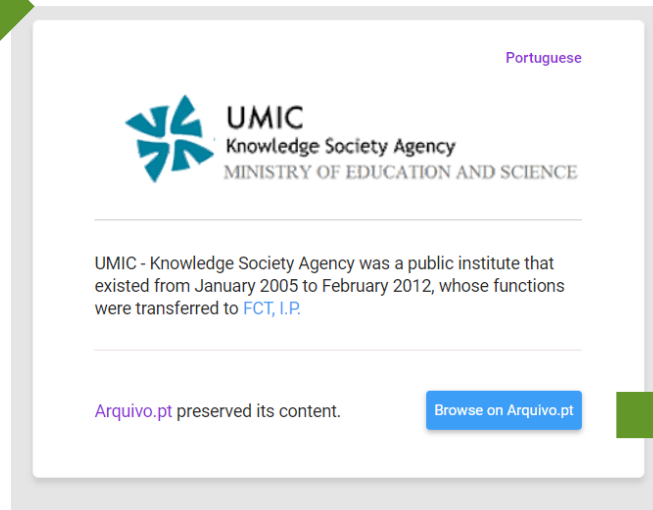
## Preserve it in the **Arquivo.pt Memorial**: [arquivo.pt/memorial](https://arquivo.pt/memorial)



https://www.english.unic.pt · Translate this page

**Umic - Home**

UMIC - Knowledge Society Agency was a public institute that existed from January 2005 to February 2012, whose functions were transferred to FCT, I.P..



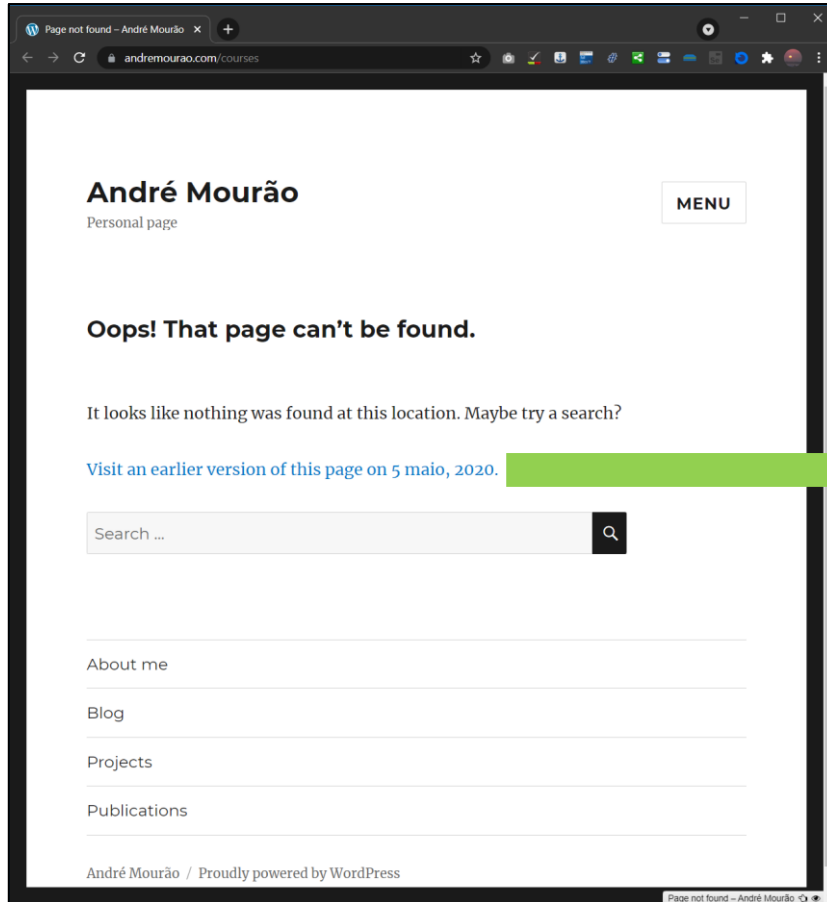
Arquivo.pt preserved its content.

Browse on Arquivo.pt



[arquivo.pt/arquivo404](http://arquivo.pt/arquivo404)

**fix** the **broken links** in your website!



Page not found at live website



Page available at Arquivo.pt

# Explore the Past Web

# Meet the Winners of the Arquivo.pt **Annual** Awards

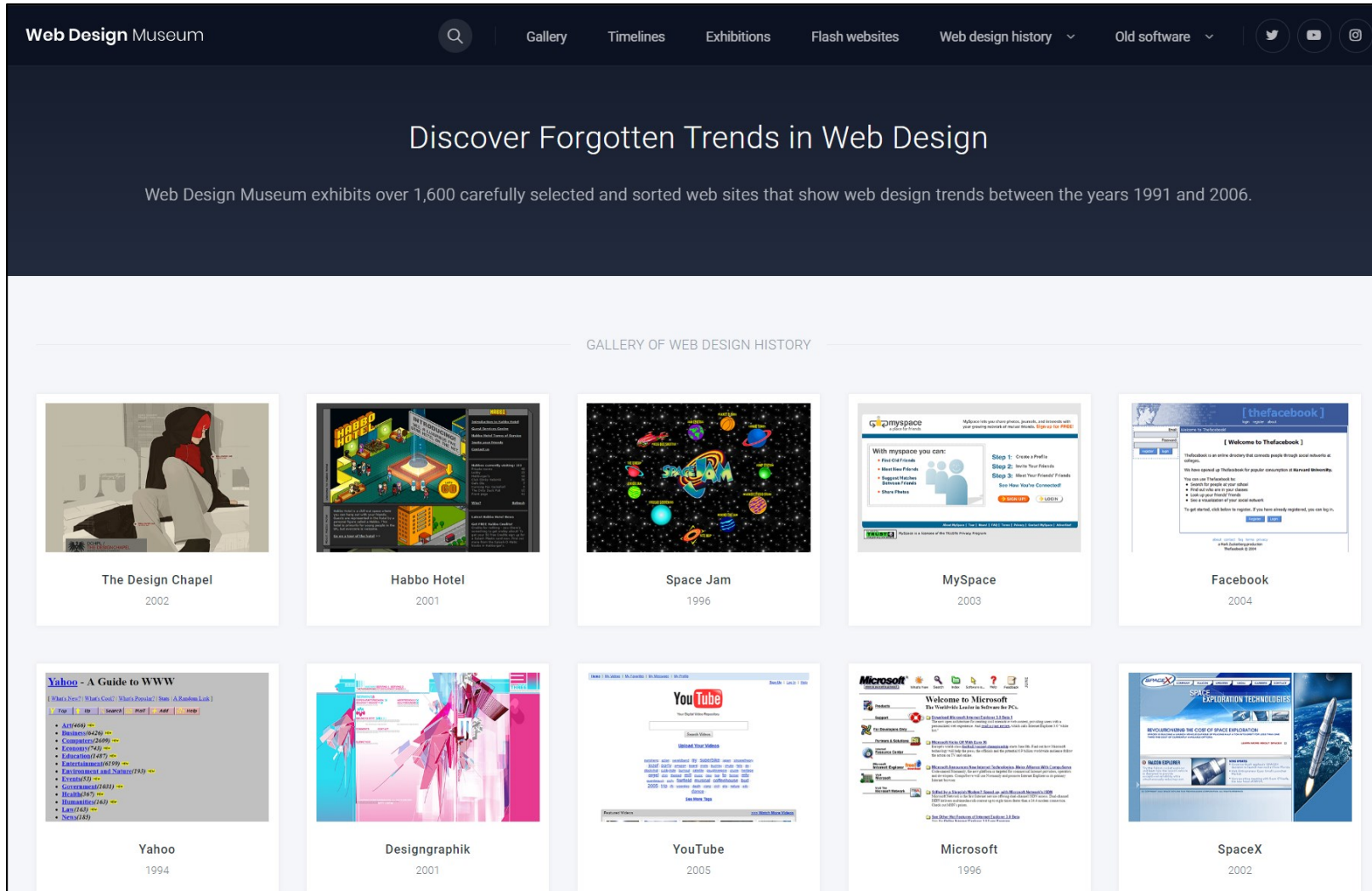


Support of the President of the Republic  
of Portugal

➤ [arquivo.pt/awards](https://arquivo.pt/awards)



# Create Web Design museums



# Attend Training courses about web preservation and research

## **New ways of searching the past** (module A)

- Any Internet user

## **Publish preservable information on the web** (B)

- Web authors

## **Automatic processing of information preserved from the Web** (C)

- Developers, Researchers

## **Do-it-yourself web archiving** (D)

- Researchers, Librarians, Archivists

## **Videos in English also available!**



[arquivo.pt/training](http://arquivo.pt/training)

# Challenges and recommendations to start a web archive

# Challenges

## **Hire and maintain skilled human resources**

- Web archiving requires specialists but is not taught at any course
- Web Archiving is Big Data, must compete with the Internet giants (e.g. Google, Facebook) to hire

## **Costs of dissemination**

- Difficult “to sell” something for free in the attention economy

## **Lack of awareness about the value of online information**

- Information that rules our current World is published exclusively online
- “Information is power”, “Data is the new oil”, “Look into the past to see the future”, ...
- However, societies passively lose most of their online information

# **Lack of awareness about the value of online information**

Let's measure the value of preserving online information or

the waste of not doing it.

Investment made on creating the online information  
preserved in Arquivo.pt

Design/development: 27 Msites x 5 600€/website = 151 000 M€

Content: 8 000 Mpages x 68€/page = 544 000 M€

Value of the preserved heritage =  
**695 000 M€**

Gross Domestic Product of Portugal in 2020 =  
**202 000 M€**





Can we afford to waste all this investment?



# Recommendations to start a web archive

## **No part-time web archivists**

- Web archiving is complex, requires full dedication
- Start with a small but autonomous web archiving team so that it does not have to permanently compete for resources

## **Use existing tools and services**

- Webrecorder.net, Archive-it

## **Listen to your users**

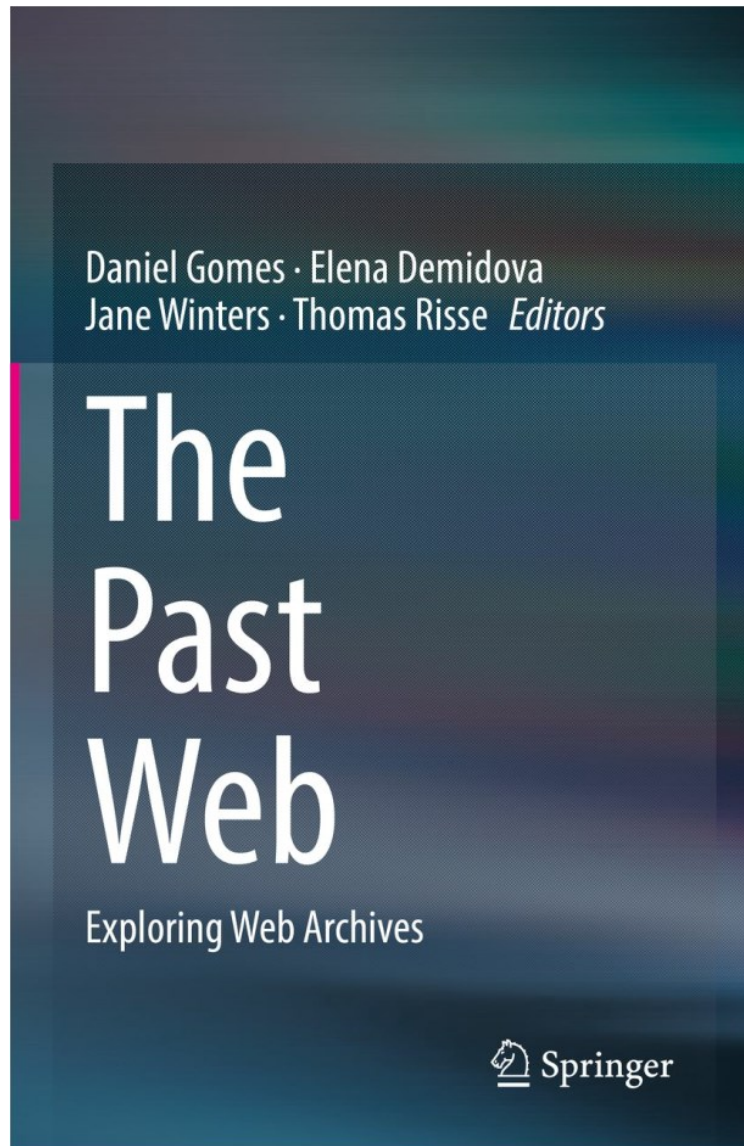
- Don't guess what they need, hire User Experience researchers

## **Start small but perform the full workflow**

- Short list of websites and perform the full preservation cycle: collect, store, provide access and disseminate service
- Doing a little is infinitely more than to do nothing



Read this book 😊



Preprint version available in  
open access at:  
[arquivo.pt/book](https://arquivo.pt/book)

# Part 6:

## The future of web archiving

### A Look into the Future

Front Matter  
Pages 283-283

#### The Past Web: A Look into the Future

Julien Masanès, Daniela Major, Daniel Gomes  
Pages 285-291

✓ The Past Web: A Look into the Future

- 1 Evolution of the Web
- 2 Operation of Web Archives
- 3 Sustainability of Web Archives
- 4 The Future of the Past Web
- 5 Final Thoughts



Julien Masanès: the “father” of web archiving in Europe

# Keep in touch!

## **Subscribe our channels:**

- [arquivo.pt/subscribe](https://arquivo.pt/subscribe): mailing list in English
- [arquivo.pt/news](https://arquivo.pt/news): social networks and videos