

# How to research governmental web data?

daniel.gomes@fccn.pt





#### Who are we?



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

Preserves **publicly accessible** information.

**Governmental service** provided by the Ministry of Science and Higher Education (Portugal)

A digital **research infrastructure** 

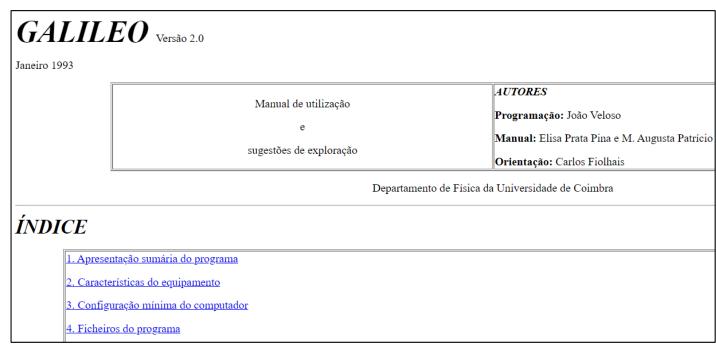
#### Arquivo.pt is used word-wide



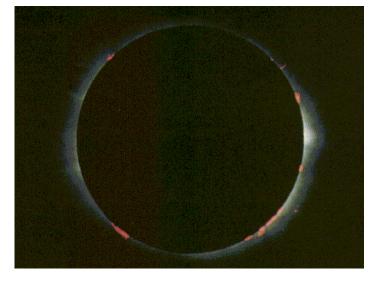
	Country	Users	% Users
1.	Portugal	46,891	46.56%
2.	United States	26,373	26.19%
3.	S Brazil	2,266	2.25%
4.	Russia	2,234	2.22%
5.	United Kingdom	2,231	2.22%
6.	● Japan	2,172	2.16%
7.	<b>○</b> Canada	1,237	1.23%
8.	Mozambique	1,213	1.20%
9.	India	902	0.90%
10.	Germany	894	0.89%

- 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

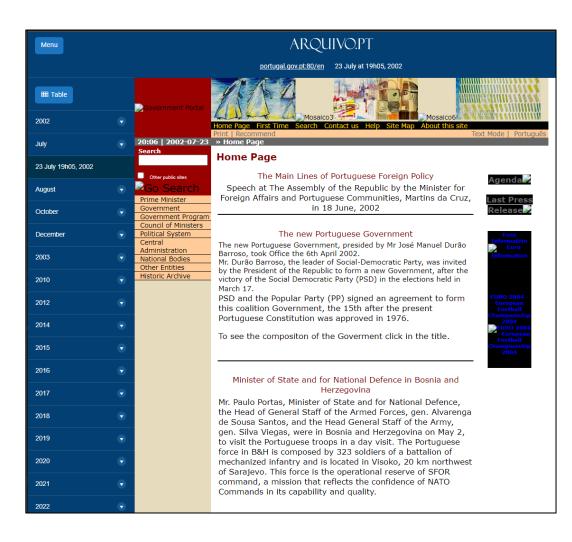


nautilus.fis.uc.pt - 1993 (oldest page)



spacelink.nasa.gov - 1992
(oldest image)

# Special attention to the preservation of governmental web data



Arquivo.pt is responsible for the preservation of Public Administration websites, such as the "Government Portal"

### What is "governmental web data"?

"Governmental web data" is online information **produced by** public sector entities

- Laws, decisions, press-releases, open data sets
- Governments communicate with their citizens through the web (almost exclusively)

"Governmental web data" is online information **published about** public sector entities

- Online news, public discussions or activism platforms
- Scrutinize governmental activities

#### Why is it interesting to research governmental web data?

The web became a crucial resource of information for **contemporary political**, **social or historical research**.

Accurate studies about governments of the XXI century **require the analysis of online information across time**.

The problem is that even governmental data quickly vanishes from the web.













# How to select governmental web data to be preserved?

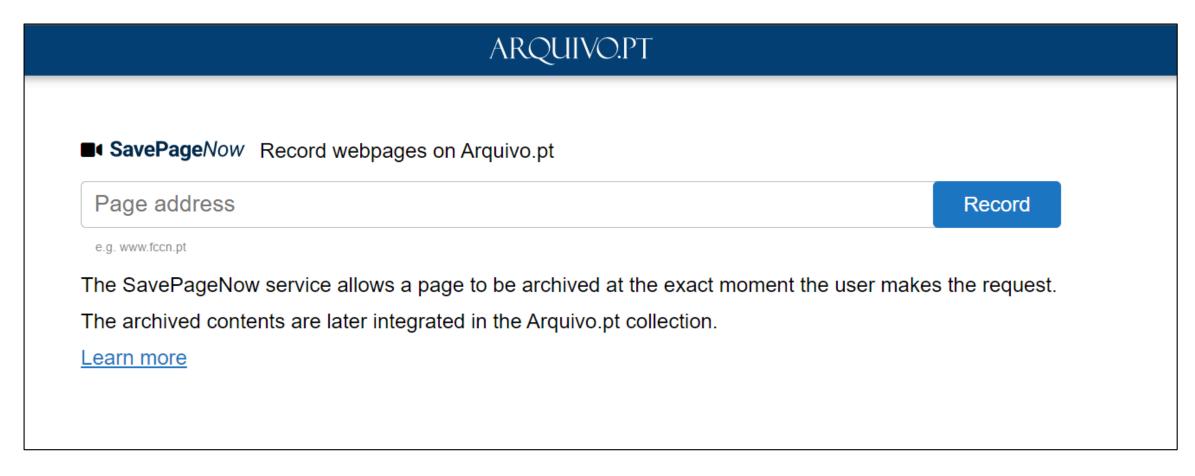
Combining experts with automated methods

#### Receive suggestions of websites to be preserved

- Public form
- Any domain
- Any language
- Submit home page
- > arquivo.pt/suggest

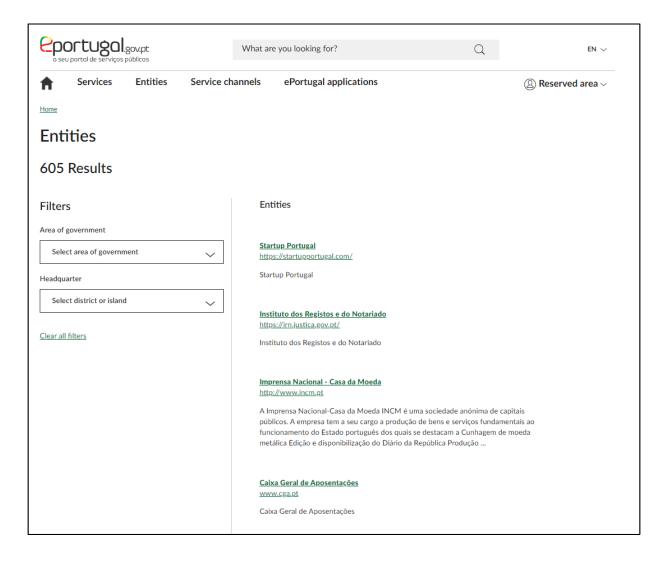
#### Suggest websites to be preserved International websites are welcome! sawfccn@gmail.com Switch accounts $\odot$ \*Required Email \* Your email address Website (1 per line) \* Your answer Description of the Websites Your answer

## SavePageNow: Archive a webpage immediately

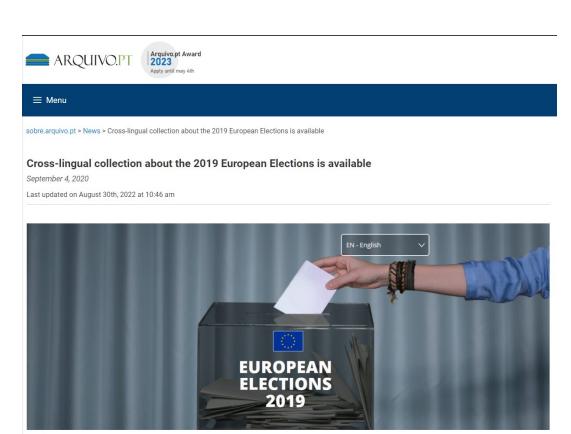


arquivo.pt/savepagenow

#### **Lists** of official public sector websites



#### Selecting online about **events** is harder (eg. elections)



Print from an archived page on Arquivo.pt: https://www.european-elections.eu

The special collection of web pages about the 2019 European Elections is available for search at Arquivo.pt.

To compile this collection, pages written in 24 European languages were identified through automatic searches on the Bing search engine and suggestions from 17 European countries.

- Information is spread across various websites.
- Quickly appears before the elections.
- Quickly disappears after the elections.
- A challenging problem but a great use case for web archives.

## Developed a low-cost methodology to automatically select and obtain web data for research

- 1. Define research question
- 2. Identify relevant keywords (Google Sheet)
- Search through APIs (Bing) to get seed URLs
- 4. Crawl *seeds* from the live-web (Heritrix)
- 5. Process crawled web data (Hadoop)
- Provide access to research web data through Arquivo.pt services (arquivo.pt/catalog)
- See "Preserving Websites Of Research & Development Projects", Ipres 2016

**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



How to assure the accessibility of historical web data?



### Public list of Arquivo.pt collections

Ħ		collections preserved by Arquiew Insert Format Data Tools		,	•	r by SAW - FCCN Serviços	s e Arquivos Web	
lic	~ <del>6</del> 7	100% ▼   \$ % .0 .0 123 ▼	10 - B	7 S A	<b>♦.</b> ⊞ ₹₹ -   ≡ -	T -  5  - B -  C0	<b>+</b> iii ▼ - ∑ -	
12	→ fX =IMPORTRANGE("https://docs.google.com/spreadsheets/d/1hfLyq9N2ZJEd1hra2V0OPI24jTy42ByR6M667rG-oHo/edit", "Summary!A2:G")							
	А	В	С	D	E	F	G	Н
1	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	ightharpoons
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	$\overline{\mathbf{z}}$
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	$\overline{\mathbf{Z}}$
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	$ \mathbf{Z} $
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	$\mathbf{Z}$
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	$ \mathbf{Z} $
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 web, mainly from the .PT domain, in 2010. The AWP8 crawl is incremental because it was performed using DeDuplicator (http://landsbokasafn.github.io/DeDuplicator/) 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	<b>☑</b>

arquivo.pt/collections

#### Open datasets about thematic collections





arquivo.pt/dadosabertos

## Existing open datasets are complemented with links to URLs and web archives

title	endDate	projectUrl	totalCost	URLsBingSearch	ArchivedProjectURLs
MARine MAnagement and Ecosystem Dynamics under climate change	2019-09-30	https://www.marmaed.uio.no/	4073903,28		https://arquivo.pt/wayback/https://www.marmaed.uio.no/
Services to enhance innovation management capacities of SMEs in England, Northern Ireland and Wales 2015 -2016	2019-09-30	'	4073903,28 2336311,17	https://www.findaphd.com/phds/project/assess-the-safety-and-efficacy-of-renal-regenerative-medicine-therapies-using-advanced-imaging-technologies-renaltoolbox-project/?p101521, https://www.liverpool.ac.uk/mathematical-sciences/staff/rachel-bearon/research/, https://www.liverpool.ac.uk/veterinary-science/staff/lorenzo-ressel/research/, https://www.cyanagen.com/grants/, https://umexpert.um.edu.my/selvee.html, https://iroffice.msu.ac.th/th/index.php/scholarships/128-european-funding-and-fellowship-opportunities, https://www.ukrmp.org.uk/reports_and_papers/imaging-technologies-for-monitoring-the-safety-efficacy-and-mechanisms-of-action-of-cell-based-regenerative-medicine-therapies-in-models-of-kidney-disease/, https://onlinelibrary.wiley.com/doi/full/10.1002/jev2.12033, https://www.autocrat.eu/author/admin/, https://translational-medicine.biomedcentral.com/articles/10.1186/s12967-019-1994-0, https://onlinelibrary.wiley.com/doi/full-xml/10.1002/jev2.12033, https://renaltoolbox.org/imaging-technologies-to-evaluate-safety-and-efficacy/, https://www.findaphd.com/phds/project/developing-mathematical-tools-for-image-analysis-to-evaluate-the-effects-of-regenerative-medicine-therapies-on-kidney-function-renaltoolbox-project/?p101518, https://en.unito.it/research/research-projects-and-initiatives/international-research-projects/marie-sklodowska-curie,	https://arquivo.pt/wayback/http://www.enterprise-europe.co.uk/ https://arquivo.pt/wayback/https://www.findaphd.co m/phds/project/assess-the-safety-and-efficacy-of- renal-regenerative-medicine-therapies-using- advanced-imaging-technologies-renaltoolbox- project/?p101521, https://arquivo.pt/wayback/https://www.liverpool.ac .uk/mathematical-sciences/staff/rachel- bearon/research/, https://arquivo.pt/wayback/https://www.liverpool.ac .uk/veterinary-science/staff/lorenzo- ressel/research/, https://arquivo.pt/wayback/https://www.cyanagen.c om/grants/, https://arquivo.pt/wayback/https://umexpert.um.ed u.my/selvee.html, https://arquivo.pt/wayback/https://iroffice.msu.ac.th /th/index.php/scholarships/128-european-funding- and-fellowship-opportunities,
Developing novel tools and technologies to assess the				https://royalsocietypublishing.org/doi/10.1098/rstb.2017.0229,	of-action-of-cell-based-regenerative-medicine-
safety and efficacy of cell-based regenerative medicine				https://uctranslationalmedicine.org/mtm-final-symposium-2021/,	therapies-in-models-of-kidney-disease/,
therapies, focusing on kidney disease	2022-10-31		4071175,56	https://www.nature.com/articles/s41536-021-00154-y, https://europepmc.org/article/MED/32292894	https://arquivo.pt/wayback/https://onlinelibrary.wile



#### Cordis-h2020organizationsComplementedByArquivoPT.xlsx

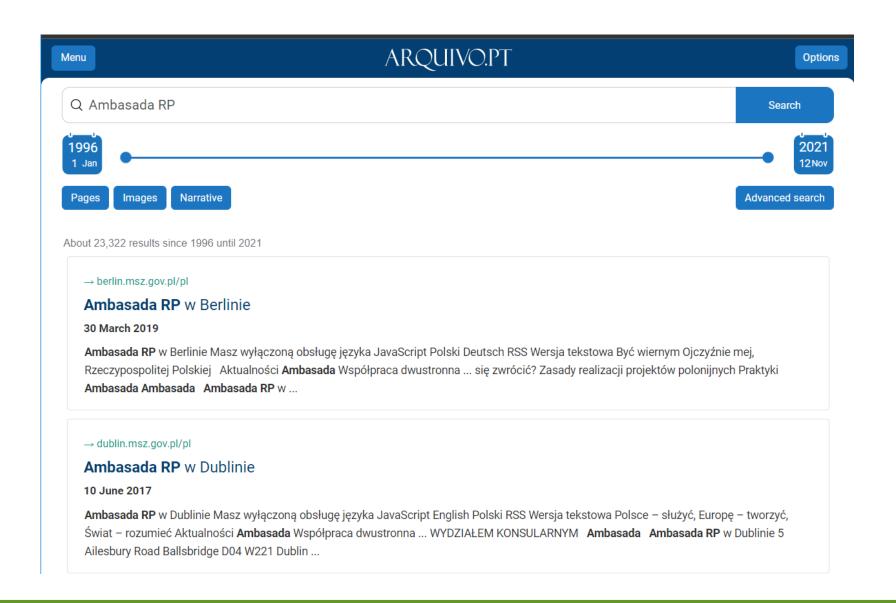
Dataset contendo a listagem das organizações dos projetos financiados pelo programa H2020 (Horizon 2020 - European Commission

② Pontual

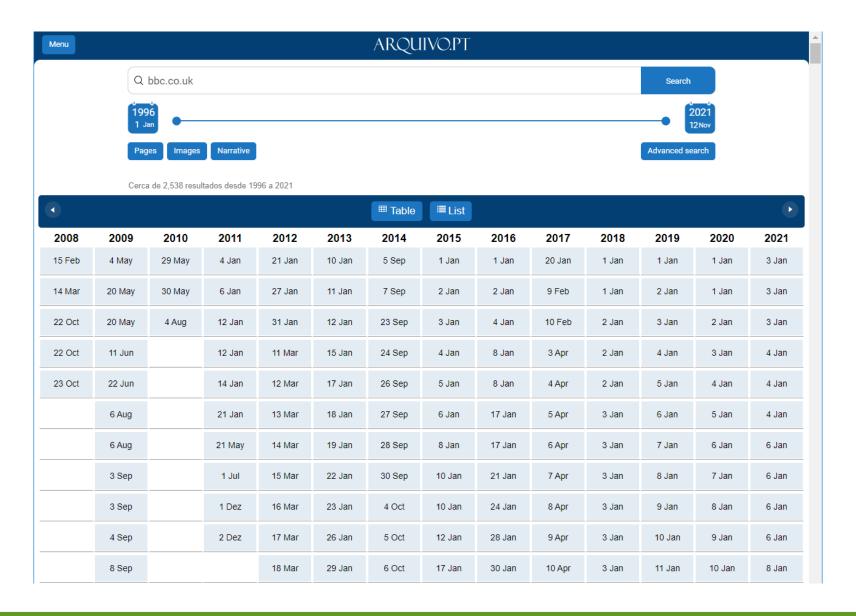
**13** 0



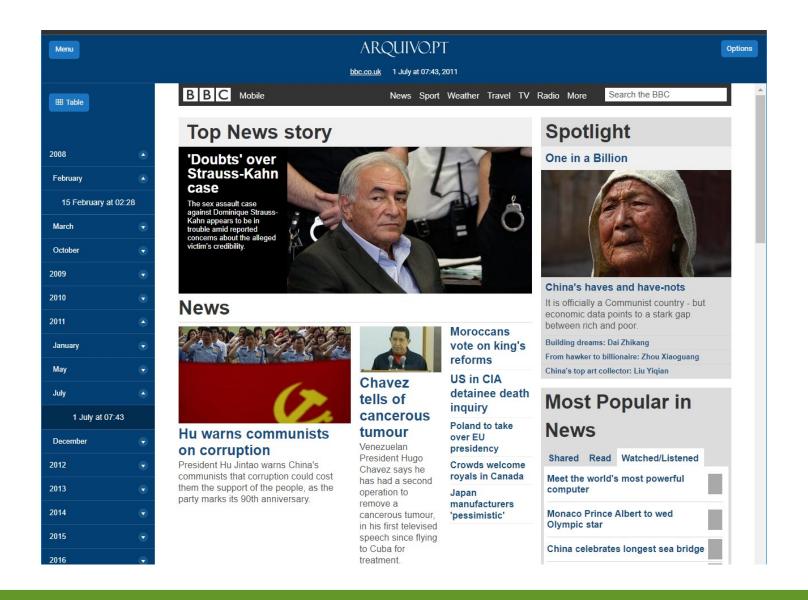
## Search texts from the past in any language



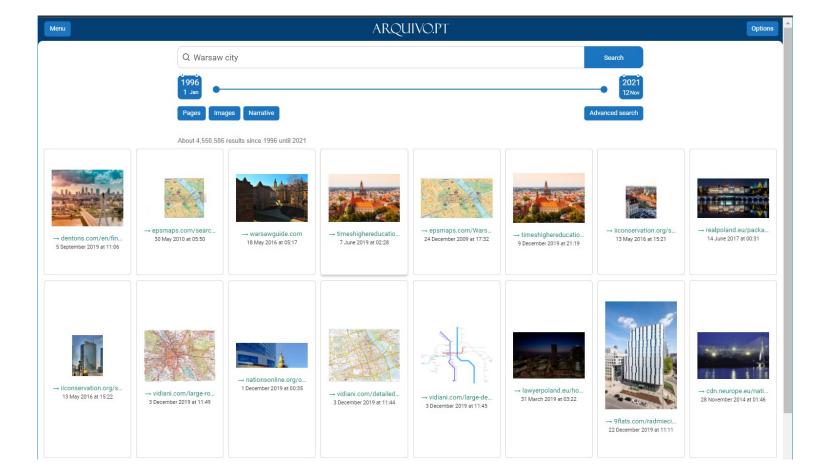
#### Search the **history** of a web address



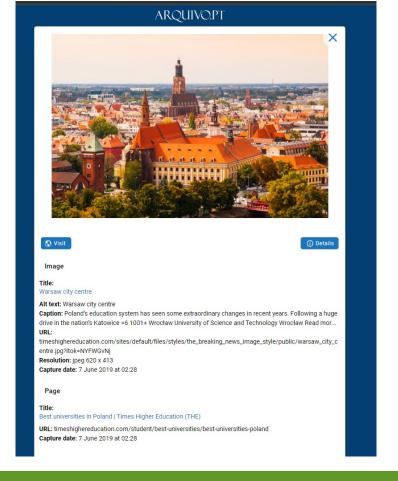
### Browse bbc.co.uk 10 years ago













How to provide access to support research over large-scale volumes of historical web data?

## Four Application Programming Interfaces to automatically search and access web-archived content

#### 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

#### Bulk download of web-archived resources

If you need to download a large amount of web-archived resources, such as all the URLs archived from a large website along time, we suggest the following methodology:

- 1. Analyse the Arquivo.pt collections so that you may choose those which may contain the most interesting web-archived data for your use case. If you have any doubt, contact us.
- 2. Download the CDXJ index files, (what is CDXJ?) of the Arquivo.pt collections you selected to process. For this purpose, analyse the "column A: Collection ID" and the corresponding CDXJ index files on "column H: Collection CDXJ File");
- 3. Create a list of selected URLs to be downloaded, extracted from the CDXJ index files (e.g. using Linux grep command);
- 4. Download the web-archived resources for the list of selected URLs from Arquivo.pt by using the above APIs or, by building links to directly access the web-archived resources. These links are available on the Technical details of the Options top-right menu when accessing a web-archived page. For instance, for the URL http://publico.pt/ with timestamp 20120201160355 extracted from the CDXJ index file, build the following links to download the:
- original file of the web-archived page (loses replay quality because the original internal links are not rewritten to reference web-archived images or stylesheets): https://arquivo.pt/noFrame/replay/20120201160355id\_/http://publico.pt/
- web-archived page without the Arquivo.pt UI frame (internal links are rewritten to reference web-archived resources): https://arquivo.pt/noFrame/replay/20120201160355/http://publico.pt/

#### arquivo.pt/api#bulk



# Examples of works performed over governmental web data

## Arquivo.pt **Annual** Awards





Support of the President of the Republic of Portugal



#### Arquivo-parlamento.pt: Winner 2023 "Parliament Archive"

Web application for researching news and opinion articles preserved by Arquivo.pt about the Parliament's activities across time

Combined web-archived information with Parliament's official open datasets.



#### Politiquices.pt

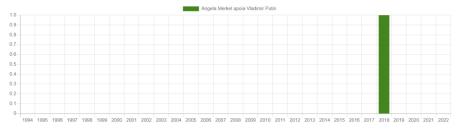
Web application for researching support or opposition relations between political personalities and parties expressed in news headlines across time.

Web-archived news that document the relation are presented.





#### Frequência de Notícias



#### **Notícias**

	Pesquisa	Exportar dados
Data	Título	Artigo
2018-03-20	Merkel vai felicitar Vladimir Putin e abordar divergências	ANQUIVO (1



## Learn more

## 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!

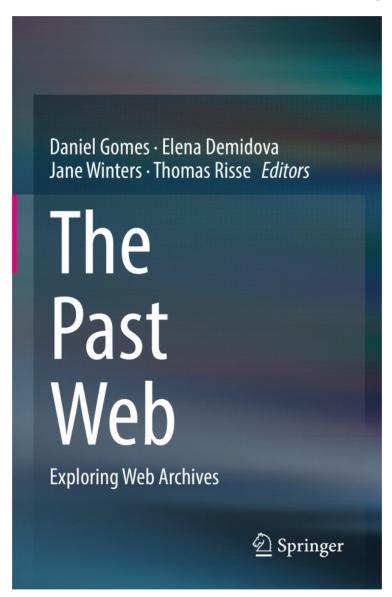


arquivo.pt/training

#### Suggested publications

- Web archives as research infrastructure for digital societies: the case study of Arquivo.pt (2022), Daniel Gomes
- The Anatomy of a Web Archive Image Search Engine (2022), André Mourão
- Information Search in Web Archives (2014), Miguel Costa
- Learning Temporal-Dependent Ranking Models (2014), Costa et al.
- arquivo.pt/publications

#### Read this book ©



Preprint version available in open access at:

arquivo.pt/book

## Keep in touch!

#### **Subscribe our channels:**

- > <u>arquivo.pt/subscribe</u>: mailing list in English
- <u>arquivo.pt/news</u>: social networks and videos