



External Advocacy for Web Archives

DPC / IIPC Training Working Group Workshop

March 6th, 2024

ricardo.basilio@fccn.pt

Web curator



The city of Sines, Portugal

Agenda

- **Do you know** Arquivo.pt? Make your web archive known.
- **Train** people for digital preservation
- Create **challenges**
- Create **services** that people can use

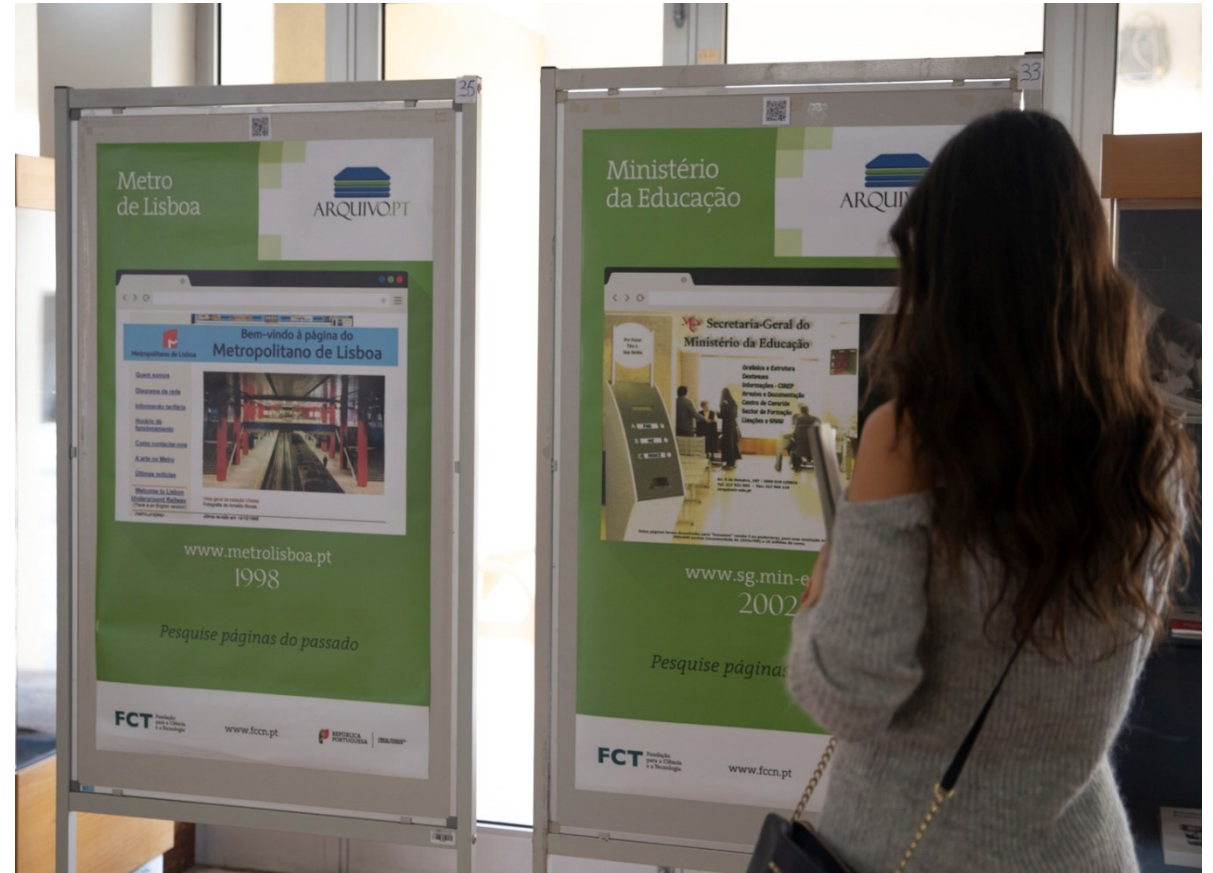
Do you know Arquivo.pt?

- **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
- Available at <https://arquivo.pt/>



Do you know Arquivo.pt?

- Poster exhibitions in universities and other public places
- 14 exhibitions held since 2019



[14 poster exhibitions at universities and other public places held since 2019.](#)

Do you know Arquivo.pt?

- Participation on events and training sessions
- 28 events held on 2023



[Daniel Gomes and Diego Alves presenting at CLEOPATRA final event](#)

Do you know Arquivo.pt?

- Participation on events and training sessions
- 28 events held on 2023



[Portugal Digital Awards 2023 Finalist Pitch: Arquivo.pt](#) | [Fundação para a Ciência e a Tecnologia \(FCT\)](#)

Do you know Arquivo.pt?

- Invite people to visit your headquarters

Welcome to the cloud!



Guided visit to the data center on World Digital Preservation Day

Do you know Arquivo.pt?

Arquivo.pt begun in 2007



The screenshot shows the Arquivo.pt website interface. The header is dark blue with the 'ARQUIVO.PT' logo in the center. To the left is a 'Menu' button, and to the right is an 'Opções' button. Below the header, a breadcrumb trail shows 'Home', 'Crawler', and 'Team'. A search bar is located on the right side of the header. The main content area is titled 'Portuguese Web Archive' and contains a welcome message about the 'Tomba project'. It mentions that publishing tools like Blogger enabled people to become web publishers, but web documents are ephemeral. It also states that the Internet Archive collects and stores contents from the world-wide web, but it is difficult for a single organization to archive the web exhaustively. The text concludes by mentioning that Portugal is now beginning its national web archiving initiative with the Tomba project at FCCN (National Foundation for Scientific Computing). The footer includes logos for FCCN (Fundação para a Computação Científica Nacional) and UMIC (Agência para a Sociedade do Conhecimento), and a copyright notice for the Plone CMS.

Menu

ARQUIVO.PT

Opções

arquivo-web.fccn.pt/portuguese-web-archive?set_language=en 16 Março às 04:52, 2008

Site Map Accessibility Contact

Search

only in current section

Home Crawler Team

You are here: Home English Português

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 —

FCCN
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 Plone Valid XHTML Valid CSS Section 508 WCAG

Do you know Arquivo.pt?

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%

Do you know Arquivo.pt?

Arquivo.pt preserves national and international historical web content

GALILEO
Versão 2.0

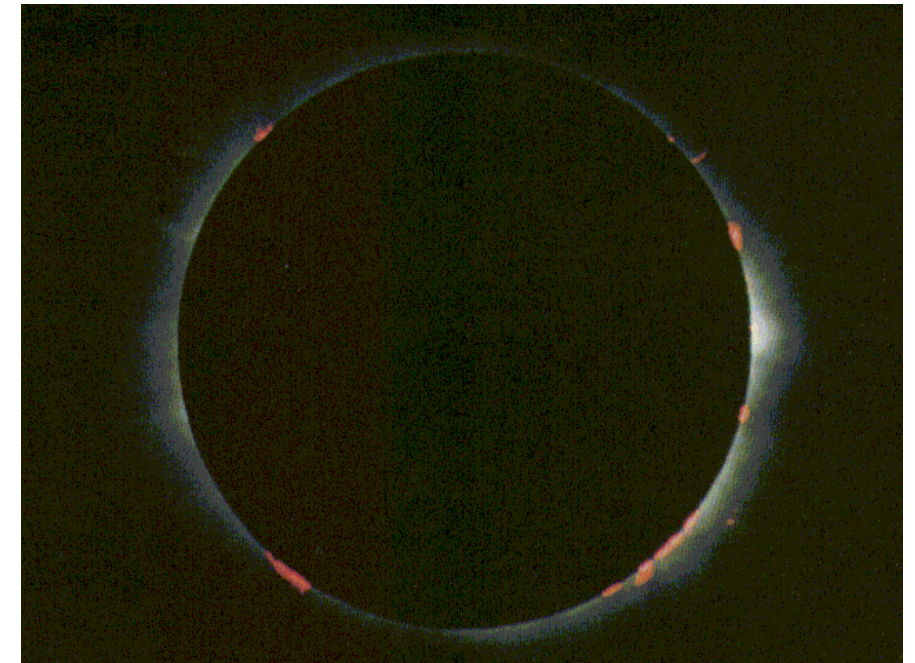
Janeiro 1993

Manual de utilização e sugestões de exploração	AUTORES Programação: João Veloso Manual: Elisa Prata Pina e M. Augusta Patricio Orientação: Carlos Fiolhais
--	--

Departamento de Física da Universidade de Coimbra

ÍNDICE

1. Apresentação sumária do programa 2. Características do equipamento 3. Configuração mínima do computador 4. Ficheiros do programa
--

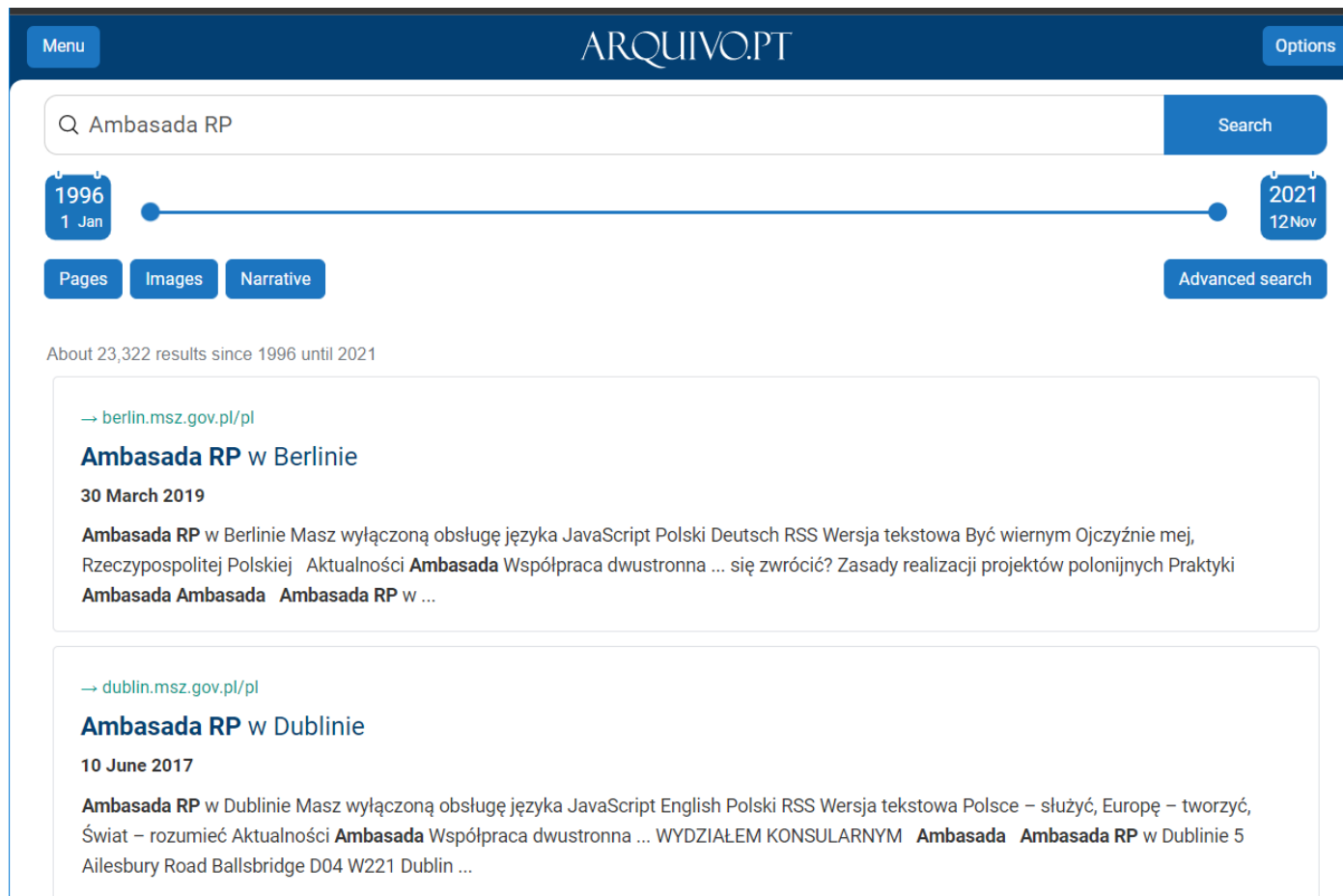


nautilus.fis.uc.pt- 1993
(oldest page)

spacelink.nasa.gov – 1992
(oldest image)

Do you know Arquivo.pt?

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

A screenshot of the Arquivo.pt website's search interface. The header is dark blue with "Menu" and "Options" buttons on the left and right, and "ARQUIVO.PT" in the center. Below the header is a search bar containing "Ambasada RP" and a blue "Search" button. Under the search bar is a timeline slider with a blue line and dots, showing dates "1996 1 Jan" and "2021 12 Nov". Below the slider are three buttons: "Pages", "Images", and "Narrative", and an "Advanced search" button on the right. The main content area shows search results. The first result is for "Ambasada RP w Berlinie" dated "30 March 2019", with a link to "berlin.msz.gov.pl/pl". The second result is for "Ambasada RP w Dublinie" dated "10 June 2017", with a link to "dublin.msz.gov.pl/pl".

Menu ARQUIVO.PT Options

Q Ambasada RP 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

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

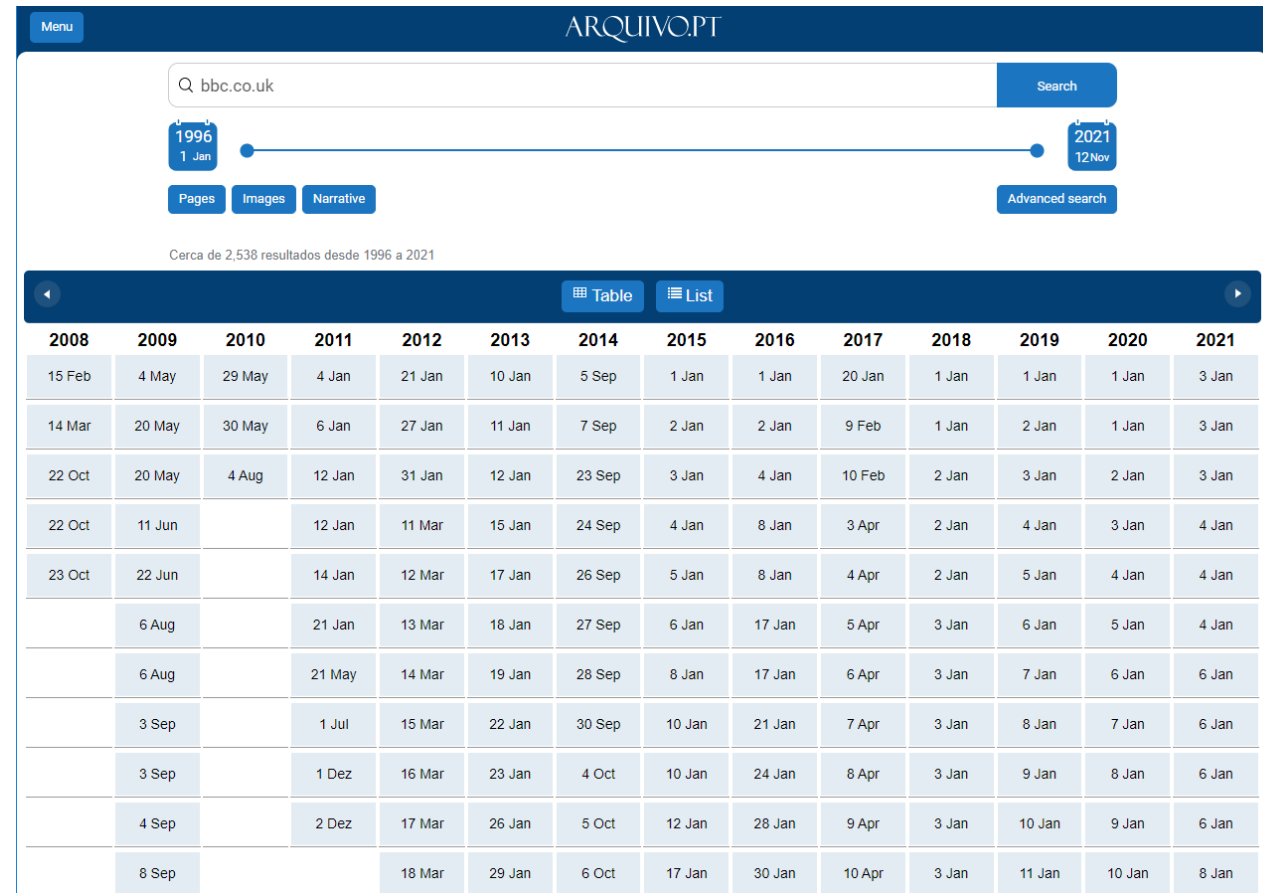
Ambasada RP w Dublinie

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 ...

Do you know Arquivo.pt?

Search the **history** of a web address

The screenshot shows the Arquivo.pt search interface. At the top, there's a search bar with "bbc.co.uk" entered and a "Search" button. Below the search bar is a timeline slider from 1996 to 2021. There are tabs for "Pages", "Images", and "Narrative", and an "Advanced search" button. Below the search bar, it says "Cerca de 2.538 resultados desde 1996 a 2021". At the bottom, there's a table with columns for years from 2008 to 2021 and rows of dates. The table is currently in "Table" view, with a "List" view option also available.

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

Do you know Arquivo.pt?

Search **images** from the past

















Menu ARQUIVO.PT Options

Q Warsaw city Search


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/searc... 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	 → liconservation.org/s... 13 May 2016 at 15:21	 → realpoland.eu/packa... 14 June 2017 at 00:31
 → liconservation.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+ Wroclaw University of Science and Technology Wroclaw Read mor...

URL:
timeshighereducation.com/sites/default/files/styles/the_breaking_news_image_style/public/warsaw_city_centre.jpg?tok=NYFWGvNj

Resolution: jpeg 620 x 413

Capture date: 7 June 2019 at 02:28

Page

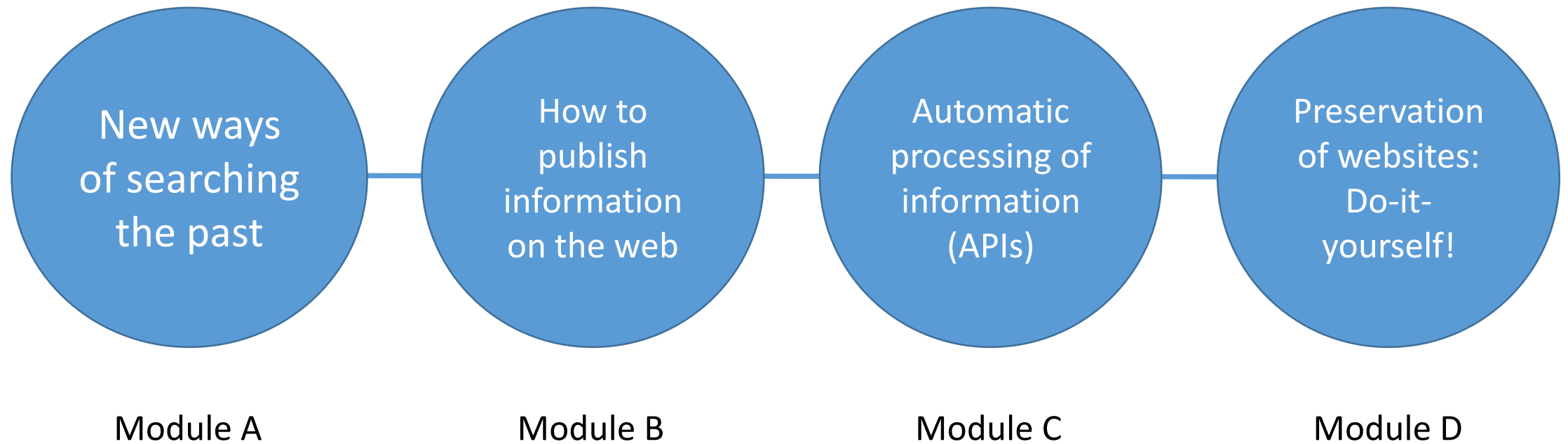
Title:
Best universities in Poland | Times Higher Education (THE)

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

Capture date: 7 June 2019 at 02:28

Train people for digital preservation

Arquivo.pt training programme initiative



arquivo.pt/training

Train people for digital preservation

Arquivo.pt training programme initiative



Module A



Module B



Module C




Module D

arquivo.pt/training



Arquivo.pt training initiative



New ways
of searching
the past


Module A

Training contents

- The problem of 80% of Web content disappearing
- Search for historical contents in Arquivo.pt
- **Everyday use cases**



Arquivo.pt training initiative



How to
publish
information
on the web

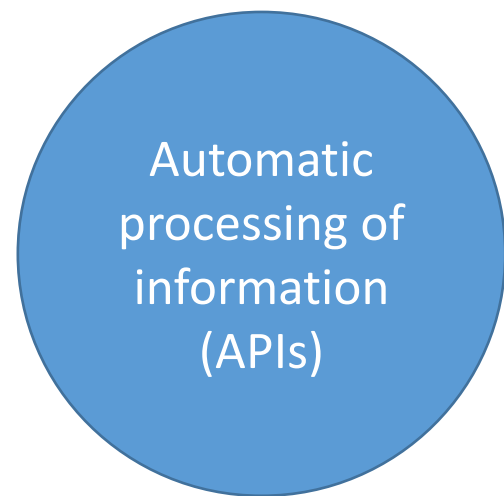
Module B

Training contents

- Recommendations for publication
- Robots.txt configuration
- Examples of poorly preservable Web publishing technologies



Arquivo.pt training initiative



Module C

Training contents

- Arquivo.pt APIs
- CDXJ Indexes (arquivo.pt/api)
- Use cases



Arquivo.pt training initiative

Preservation
of websites:
Do-it-
yourself!

Module D

Training contents

- Record Web contents locally in a standard format
- ArchiveWeb.page tutorial
- Practical exercises (websites, social media like Facebook, Twitter)

Train people for digital preservation

Training programme –
Every day use cases

Management: documentation/audit of finished project



“Everything was on the website”

Linguateca	Publications on the computational processing of portuguese
Structure Team Overview Access to resources Evaluation contents Catalogue of resources Catalogue of tools Catalogue of actors Catalogue of publications <ul style="list-style-type: none"> Publication search Publication advice Sitemap Interesting information FAQ Glossary Racetracks Links  Comments and suggestions <small>Please use the appropriate options in the left sidebar. Please inform us if the data is not correct or if you have any suggestions (tel: +351 21 361 10 00)</small>	Linguateca <p>Our goal is to gather a collection of bibliographic references on the computational Processing of portuguese in this pages. We appreciate any contribution that you as an author or reader can give us, by providing us with additional information or to direct us to pages with relevant content.</p> <p>To look for publications we suggest the use of our search interface.</p> <p>To contribute with new references you can use this form.</p> <p>To use SUPORB as your publications manager see here.</p> <p>We currently list 2375 references under the following categories:</p> <ul style="list-style-type: none"> Publications in journals (184 ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Books (1 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Book chapters (159 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Book (Author) (37 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Book (Compilation) (33 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Series (112 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Articles in international conferences (1124 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Reports and monographs (192 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Other academical work (3 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Publications only available on the web (85 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) Presentations (297 publications, ordered alphabetically, by date, in BibTeX: ordered alphabetically, by date) <p>Publications inserted last month</p> <p>Last update 30 November 2015.</p> <p>Comments & suggestions</p>

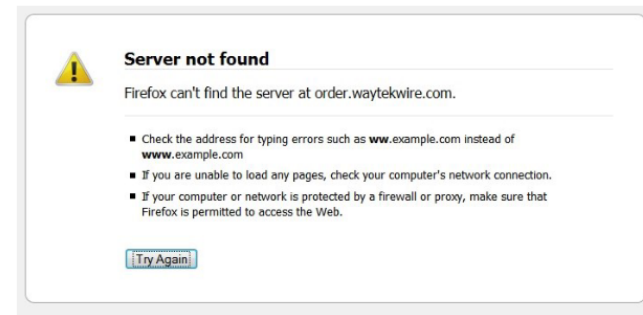
Train people for digital preservation

Training programme –
Every day use cases

Maintain Portfolio/CV

“My best work was a website that no longer exists...”

60% of the websites disappear just after 2 years.



Train people for digital preservation

- Emphasizing the preservation of scientific research results (e.g. [H2020 European Funded Projects](#))
- Emphasizing citations of Web content in knowledge bases (e.g. [Colaboration with Portuguese Wikipedia](#))



Example of broken links on the Wikipedia [page about web archiving initiatives](#)

Train people for digital preservation

Training programme – Publish well to preserve better



User-agent: Arquivo-web-crawler
Disallow:



Train people for digital preservation

Training programme –
Arquivo.pt APIs

arquivo.pt/api

Text search - API



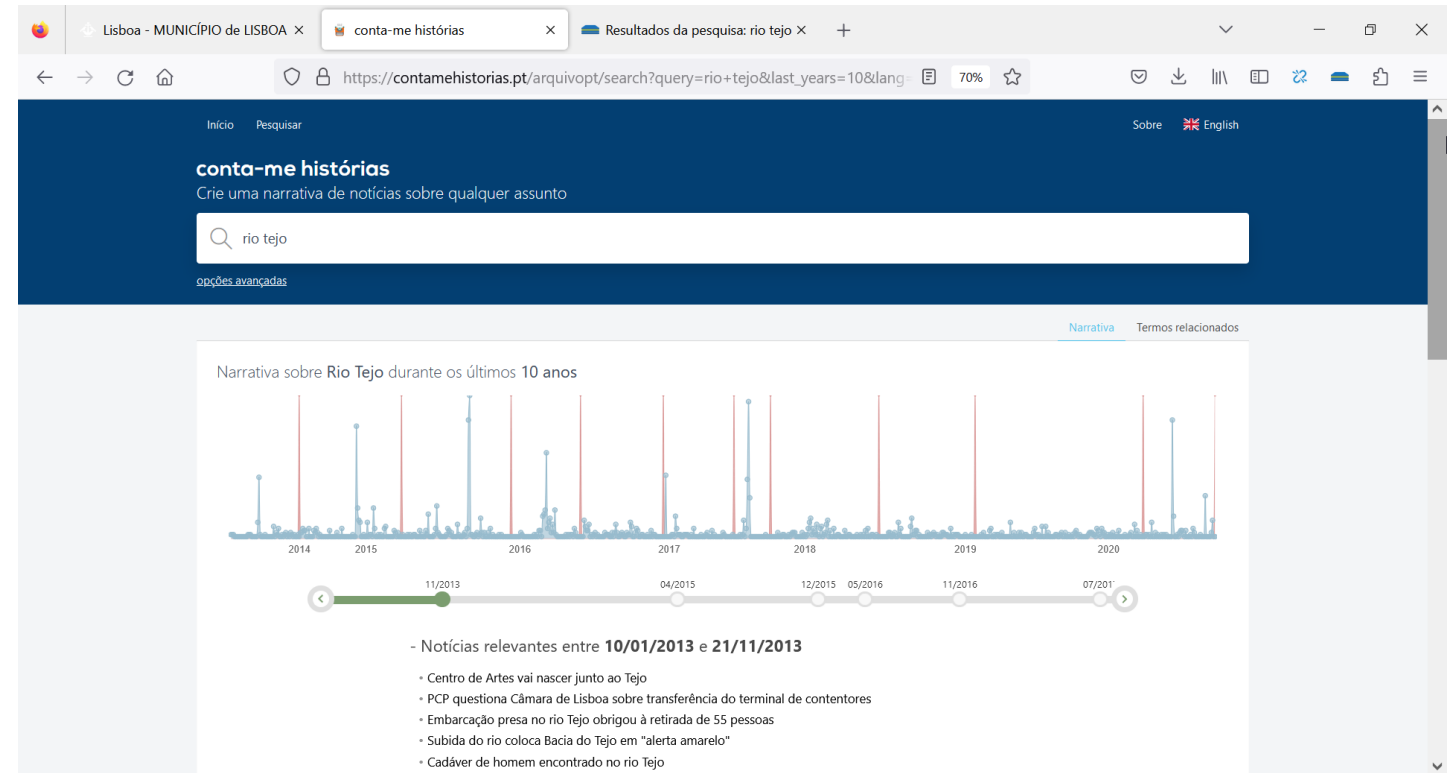
```
{
  "serviceName": "Arquivo.pt - the Portuguese web-archive",
  "linkToService": "https://arquivo.pt",
  "next_page": "https://arquivo.pt/textsearch?q=universidade%20de%20aveiro&offset=50",
  "estimated_nr_results": 32195375,
  "request_parameters": {
    "offset": 0,
    "dedupValue": 2,
    "dedupField": "site",
    "q": "universidade de aveiro",
    "maxItems": 50
  },
  "response_items": [
    {
      "title": "Universidade de Aveiro > Página inicial",
      "originalURL": "http://www.ua.pt/",
      "linkToArchive": "https://arquivo.pt/wayback/20180720085256/http://www.ua.pt/",
      "tstamp": "20180720085256",
    }
  ]
}
```

<https://arquivo.pt/textsearch?q=universidade%20de%20aveiro>

Train people for digital preservation

Training programme –
Arquivo.pt APIs

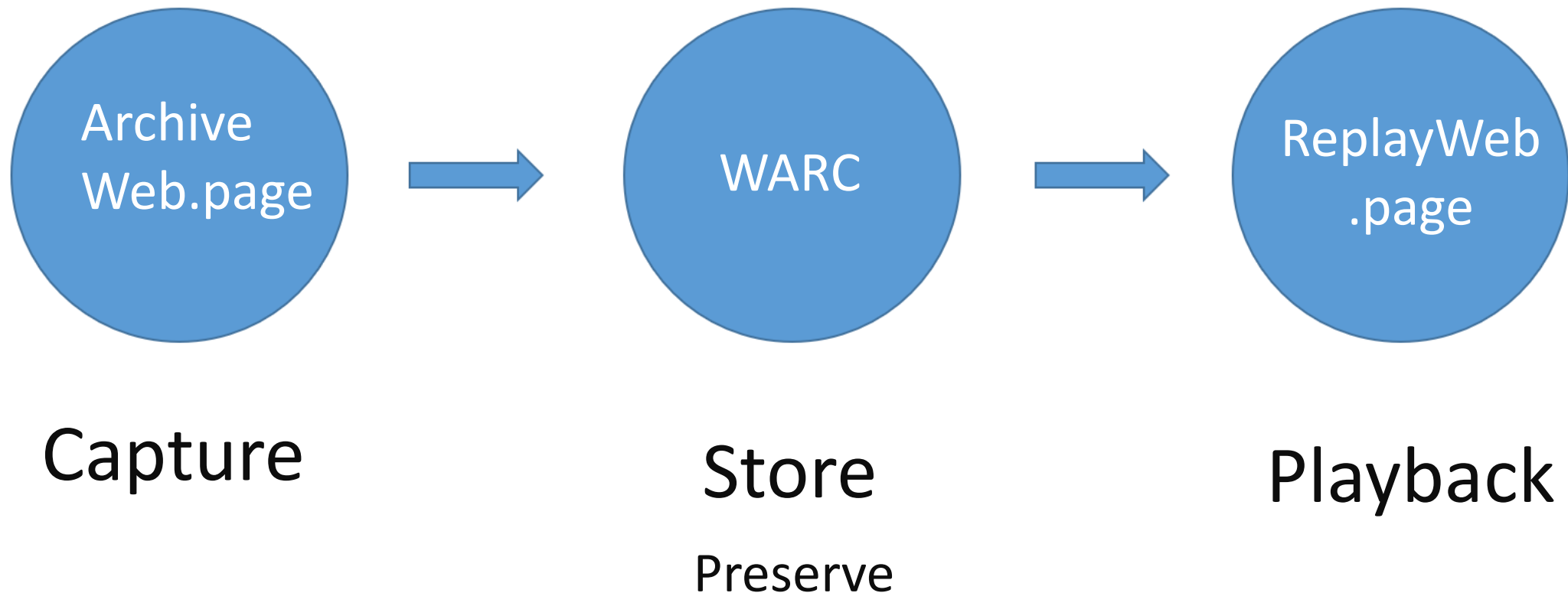
arquivo.pt/api



Example of temporal summarization through the external service, [“Tell me stories”](https://contamehistorias.pt). Search for [“Tejo river”](https://arquivo.pt). Arquivo.pt has the data. The community can build applications and new layers or services.

Train people for digital preservation

Training programme – Do it yourself!



Create challenges

Arquivo.pt Award

- 167 candidates since 2018
- 10.000 Euros for the winner
- Promoted in the media

More information:

arquivo.pt/award



Create challenges



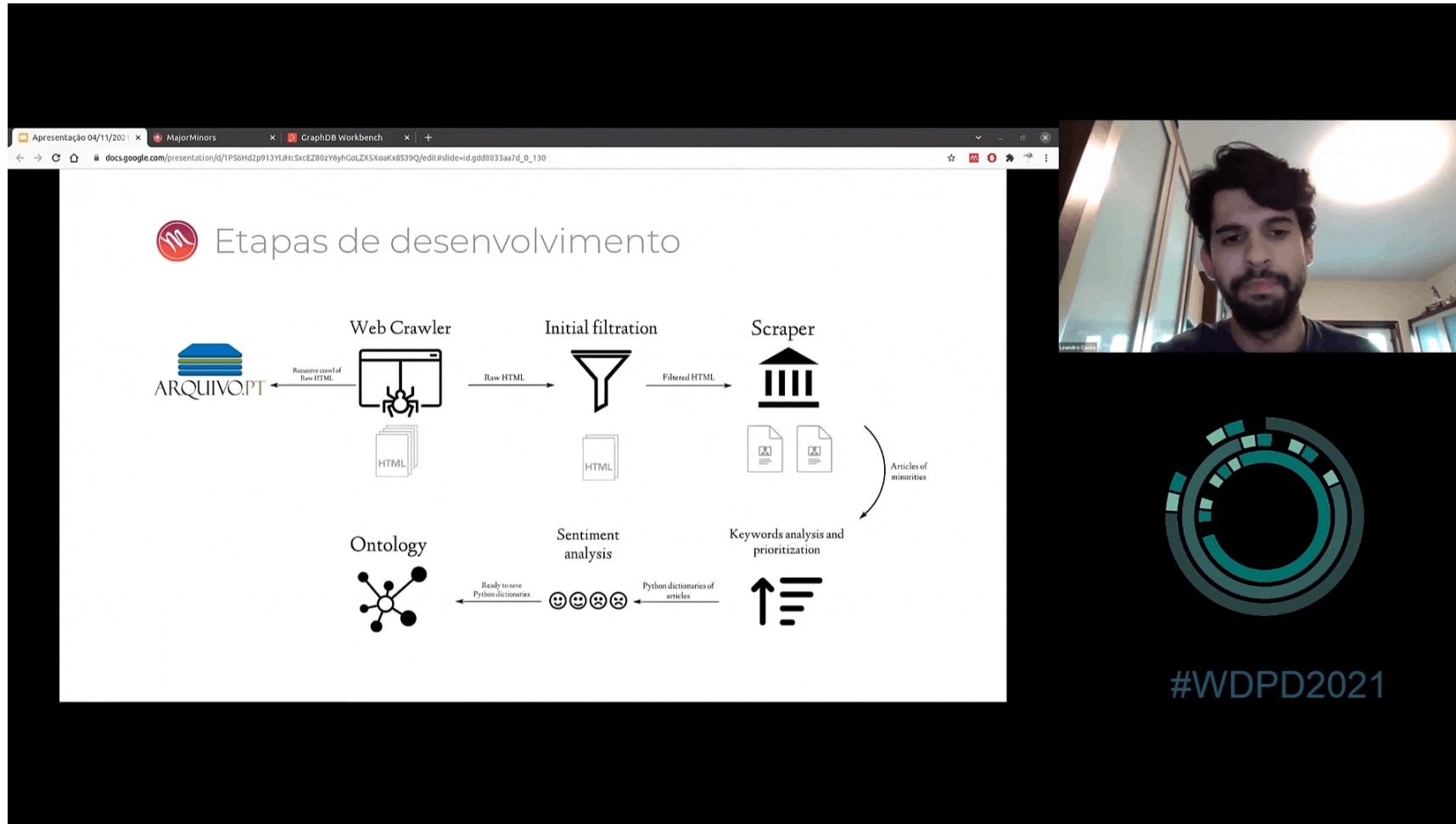
arquivo.pt/winners2023

Create challenges



arquivo.pt/winners2020

Create challenges



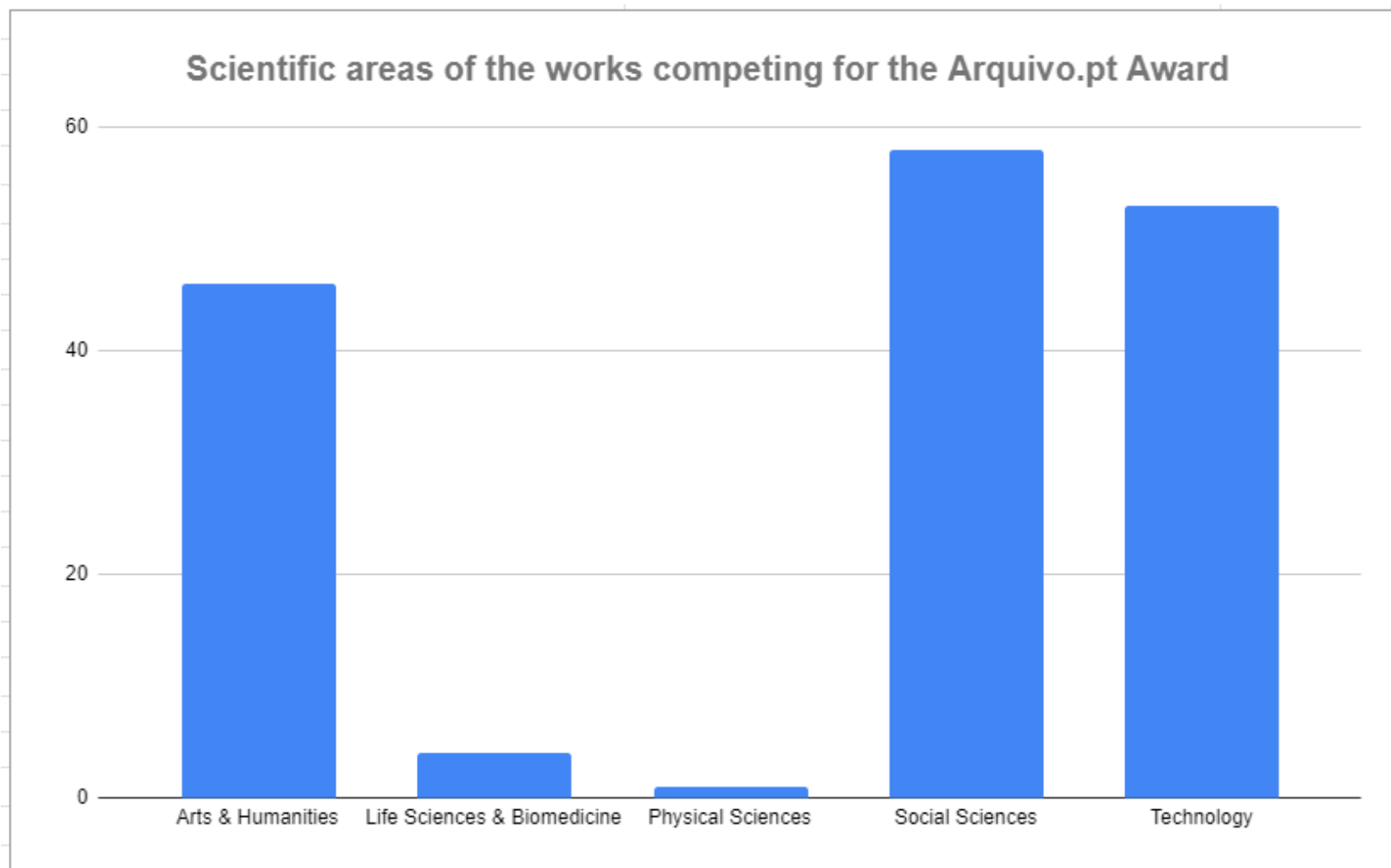
The screenshot shows a Google Docs presentation titled "Etapas de desenvolvimento" (Development Stages). The workflow is as follows:

- Web Crawler**: Represented by a spider icon and a browser window. It performs a "Recursive crawl of Raw HTML" from **ARQUIVO.PT** (represented by a blue and green logo) and outputs **Raw HTML** (represented by a document icon).
- Initial filtration**: Represented by a funnel icon. It takes **Raw HTML** and outputs **Filtered HTML** (represented by a document icon).
- Scraper**: Represented by a building icon. It takes **Filtered HTML** and outputs **Articles of minorities** (represented by two document icons).
- Keywords analysis and prioritization**: Represented by an upward arrow and a list icon. It takes **Articles of minorities** and outputs **Python dictionaries of articles** (represented by a document icon).
- Sentiment analysis**: Represented by a row of four smiley face icons (two happy, two sad). It takes **Python dictionaries of articles** and outputs **Ready to save Python dictionaries** (represented by a document icon).
- Ontology**: Represented by a network graph icon. It takes **Ready to save Python dictionaries** and outputs **Ontology** (represented by a document icon).

In the top right corner, there is a video feed of a man with dark hair and a beard, wearing a dark shirt, speaking. Below the video feed, there is a circular graphic with teal and grey segments, and the hashtag **#WDPD2021**.

2021 Arquivo.pt Award winner presents its work in [Café with Arquivo.pt](#), in the WDPD.

Create challenges




Areas of study of the 167 works competing for the Arquivo.pt Award, between 2018 and 2023

Create services that people can use

SavePageNow: Archive a webpage immediately

ARQUIVO.PT

 **SavePageNow** Record webpages on Arquivo.pt

Page address

e.g. www.fccn.pt

Record

The SavePageNow service allows a page to be archived at the exact moment the user makes the request.

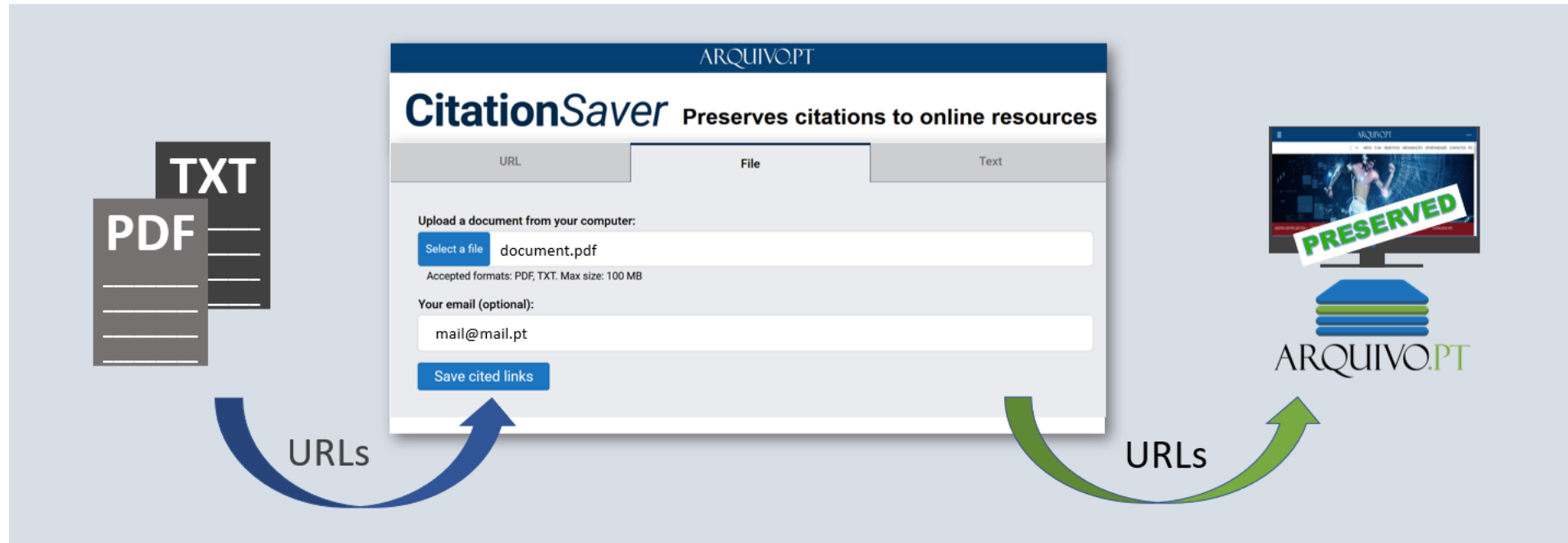
The archived contents are later integrated in the Arquivo.pt collection.

[Learn more](#)

arquivo.pt/savepagenow

Create services that people can use

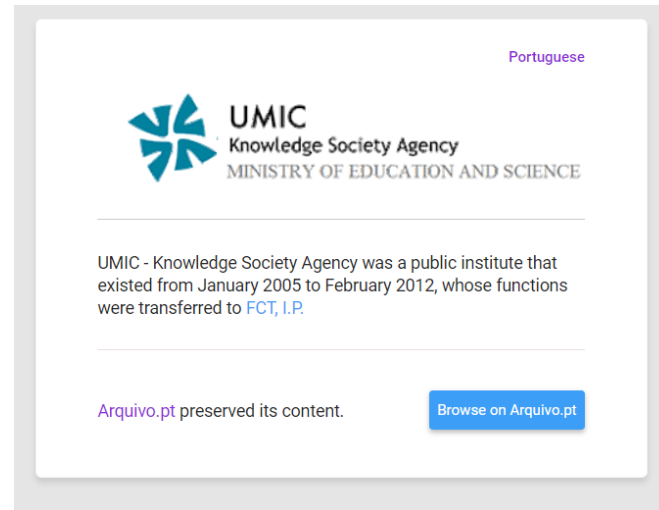
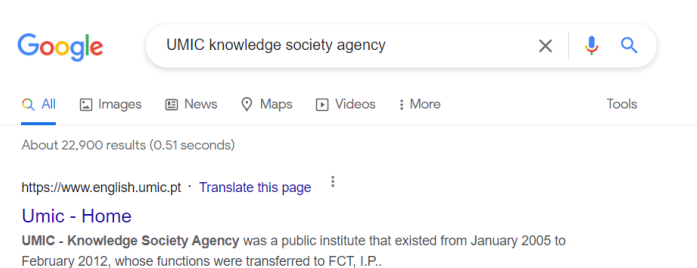
Arquivo.pt CitationSaver



[Use CitationSaver to preserve the integrity of your documents](#)

Create services that people can use

Arquivo.pt Memorial

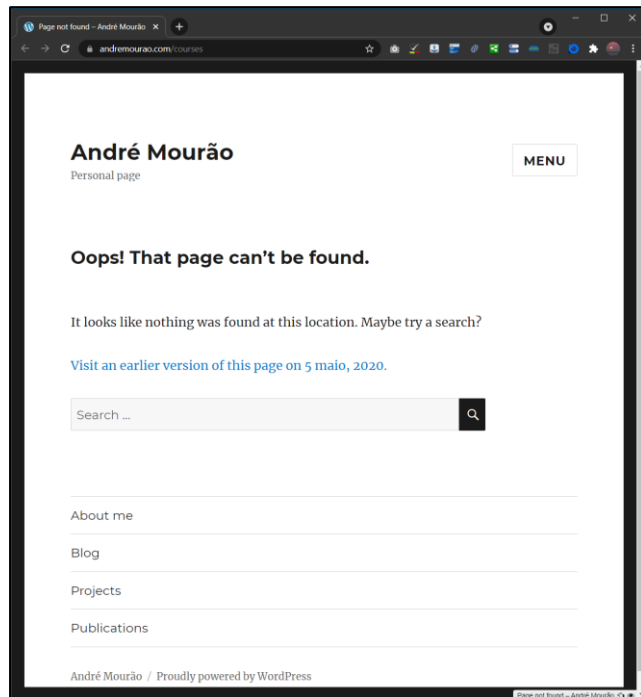


Don't kill your historical website!

Preserve it in the **Arquivo.pt Memorial**: arquivo.pt/memorial

Create services that people can use

Arquivo.pt Arquivo404 arquivo.pt/arquivo404 offer historical content in your website!



Page not found at live website



Page available at Arquivo.pt

Create services that people can use

High Quality Collections (pt: RAQs)

Azores Government website - HighQualityCollection

Actions ▾

Replay

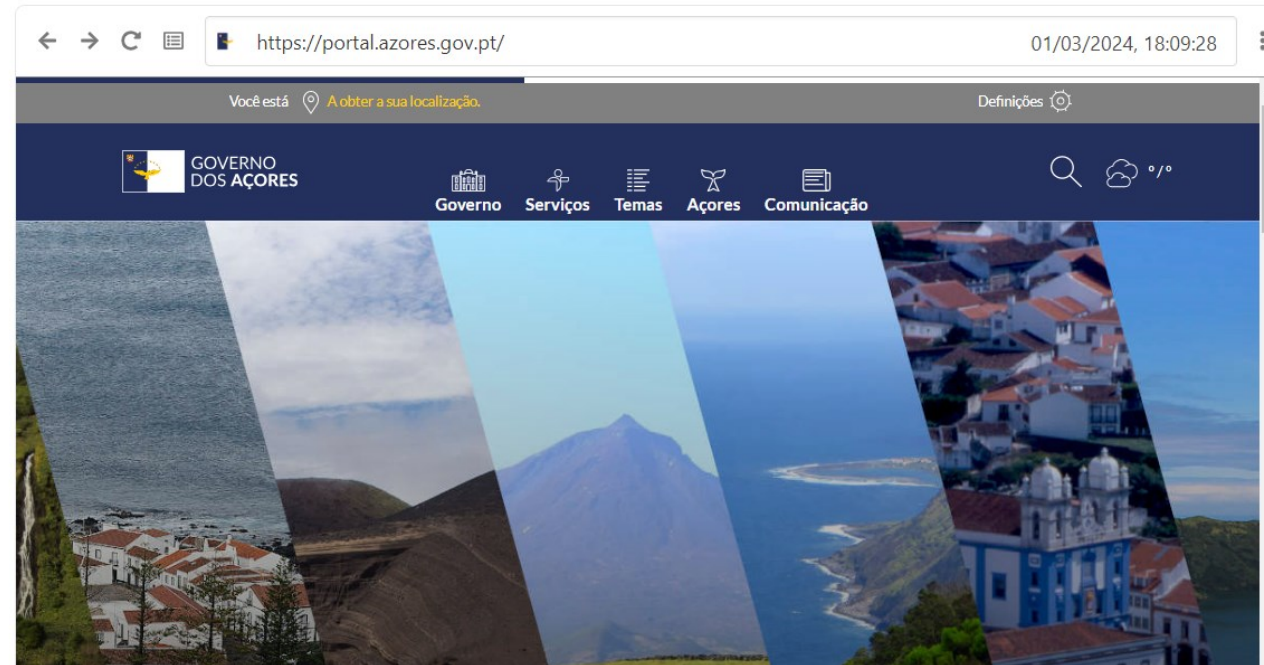
Overview

Replay

Files

Error Logs

Crawl Settings



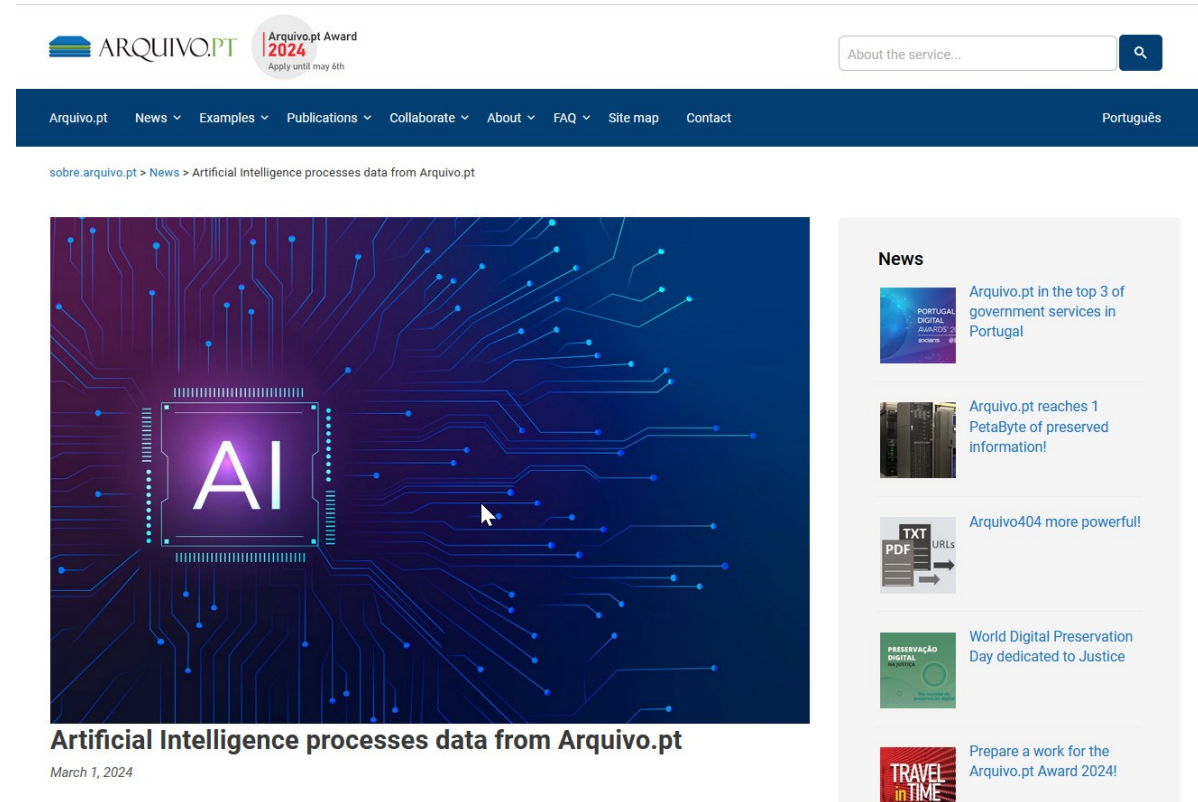
[Example of a high quality collection on demand of Azores Government Website using Browsertrix cloud \(already put on Arquivo.pt\)](#)

Create services that people can use

Emphasize external projects
that used your web archive.

Example:

[GlorIA – A Generative and
Open Large Language
Model for Portuguese](#)



Conclusions

- Make your web archive known
- **Train** people for digital preservation
- Create **challenges**
- Create **services** that people can use

Thank you!

contacto@arquivo.pt

arquivo.pt/subscribe