RESAW 2023. Methodological Challenges.
From online Tools to Web Archive?

Time to explore, time to learn from the archived Web

Arquivo.pt training initiative

ricardo.basilio@fccn.pt

Fundação
para a Ciência
e a Tecnologia

Agenda

1. Arquivo.pt

2. ROSSIO Infrastructure for Digital Humanities

3. Arquivo.pt training initiative

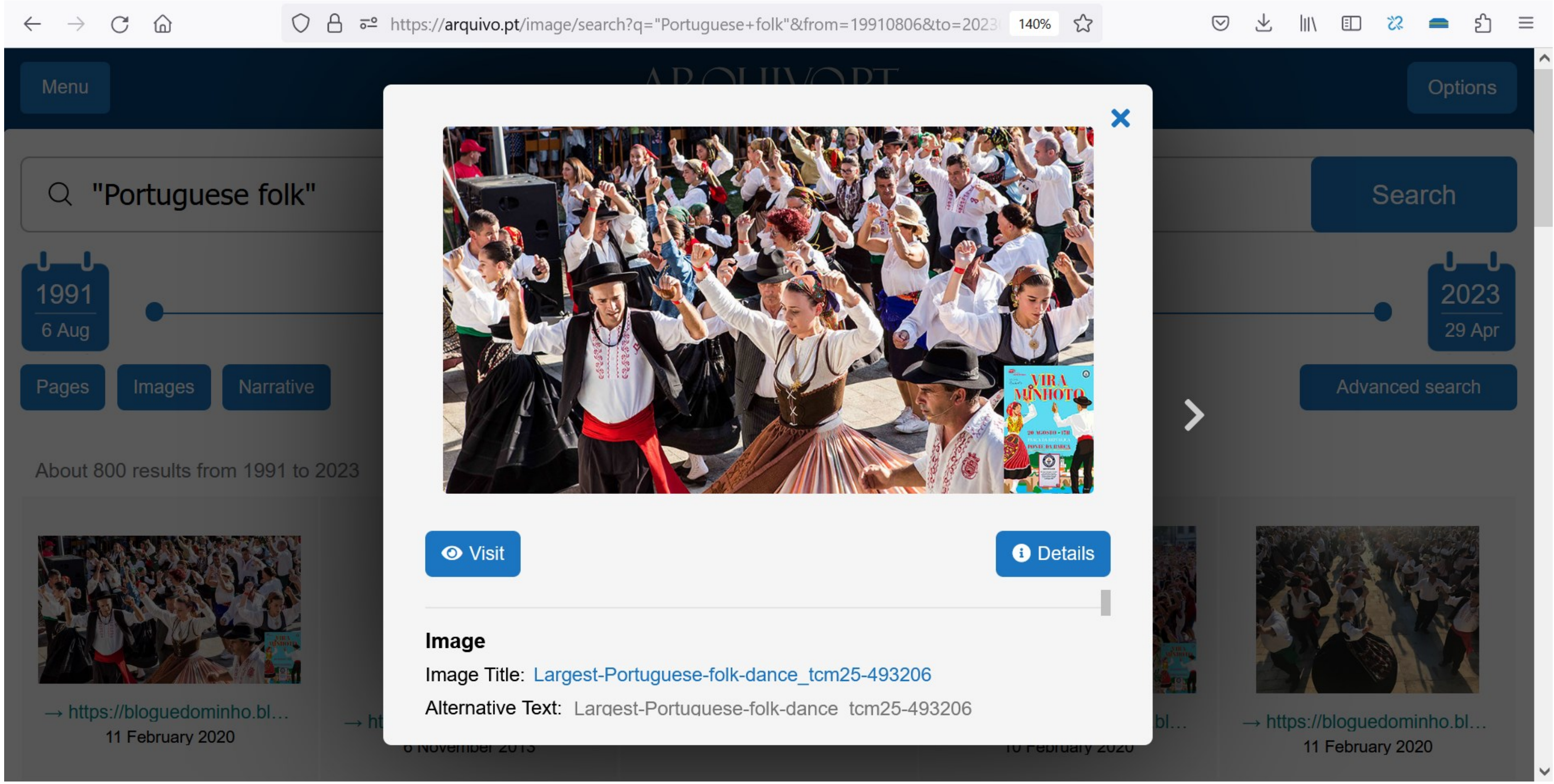
1

Arquivo.pt

Arquivo.pt

- Portuguese public Web preservation service
- Provides a public search interface (by URL, text, image, and API)
- Aimed at academic research, but also at all citizens

Arquivo.pt



https://arquivo.pt/image/search?q="Portuguese+folk"&from=19910806&to=2023

"Portuguese folk"

1991 6 Aug

2023 29 Apr

About 800 results from 1991 to 2023

→ https://bloguedominho.bl…
11 February 2020

→ https://bloguedominho.bl…
11 February 2020

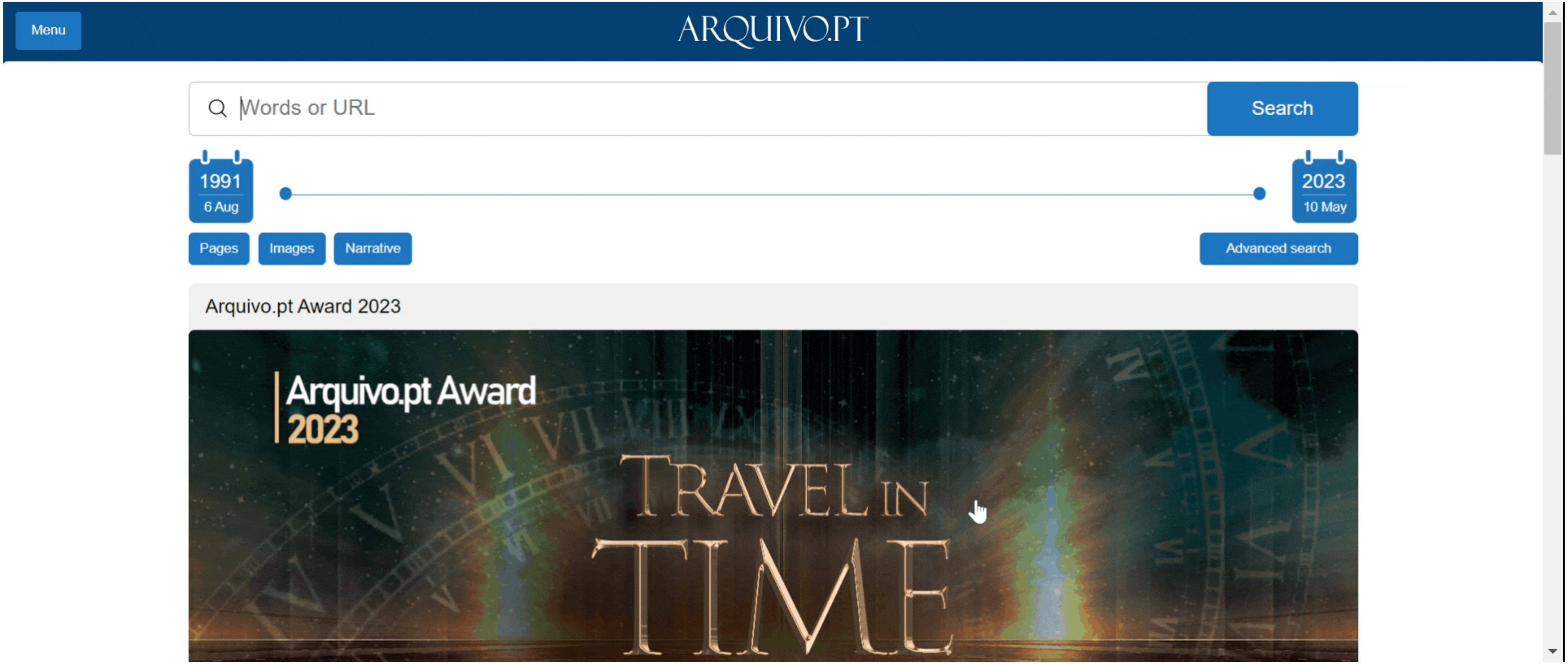
Image

Image Title: Largest-Portuguese-folk-dance_tcm25-493206

Alternative Text: Largest-Portuguese-folk-dance_tcm25-493206

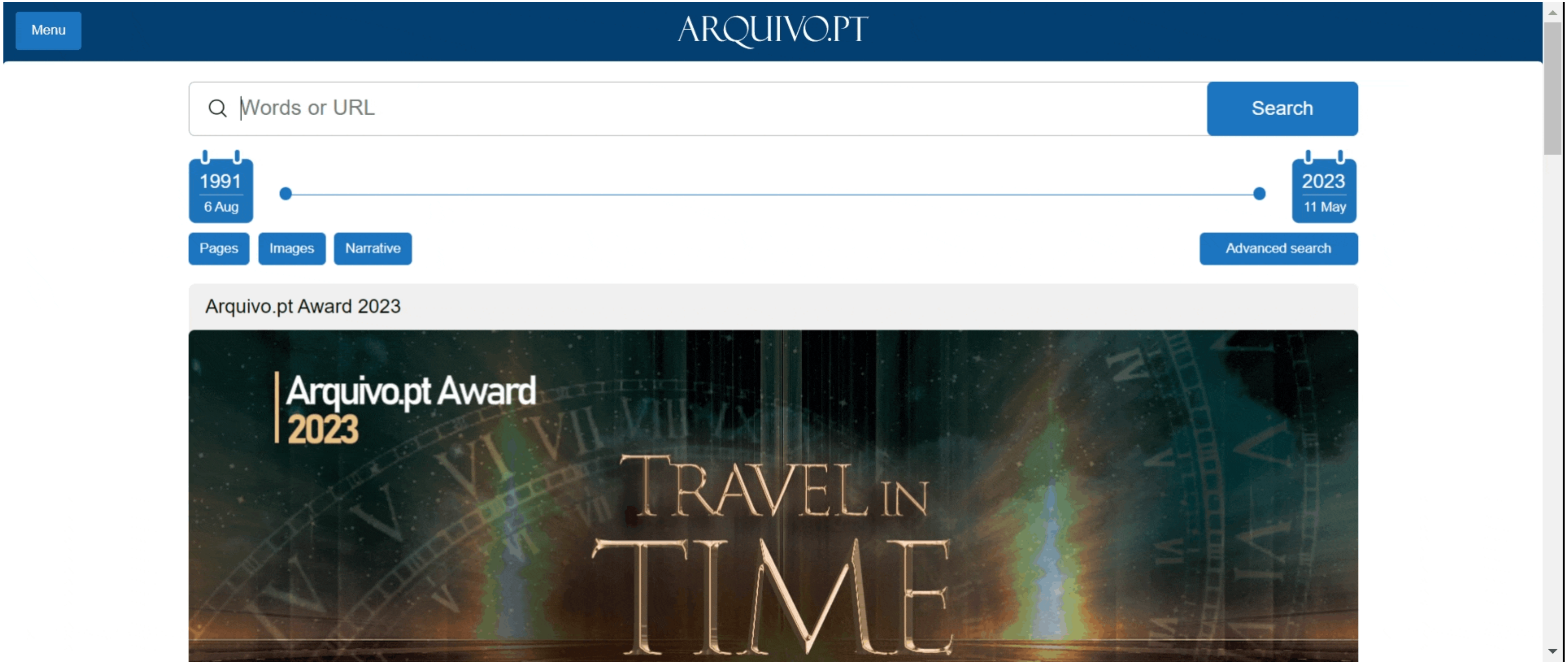
Example: a search for "Portuguese folk" on Arquivo.pt, the Portuguese Web Archive

Arquivo.pt



Arquivo.pt public interface

Arquivo.pt



Arquivo.pt public interface and the use of the narrative button that uses the external service *Tell me stories/Conta-me histórias*

2

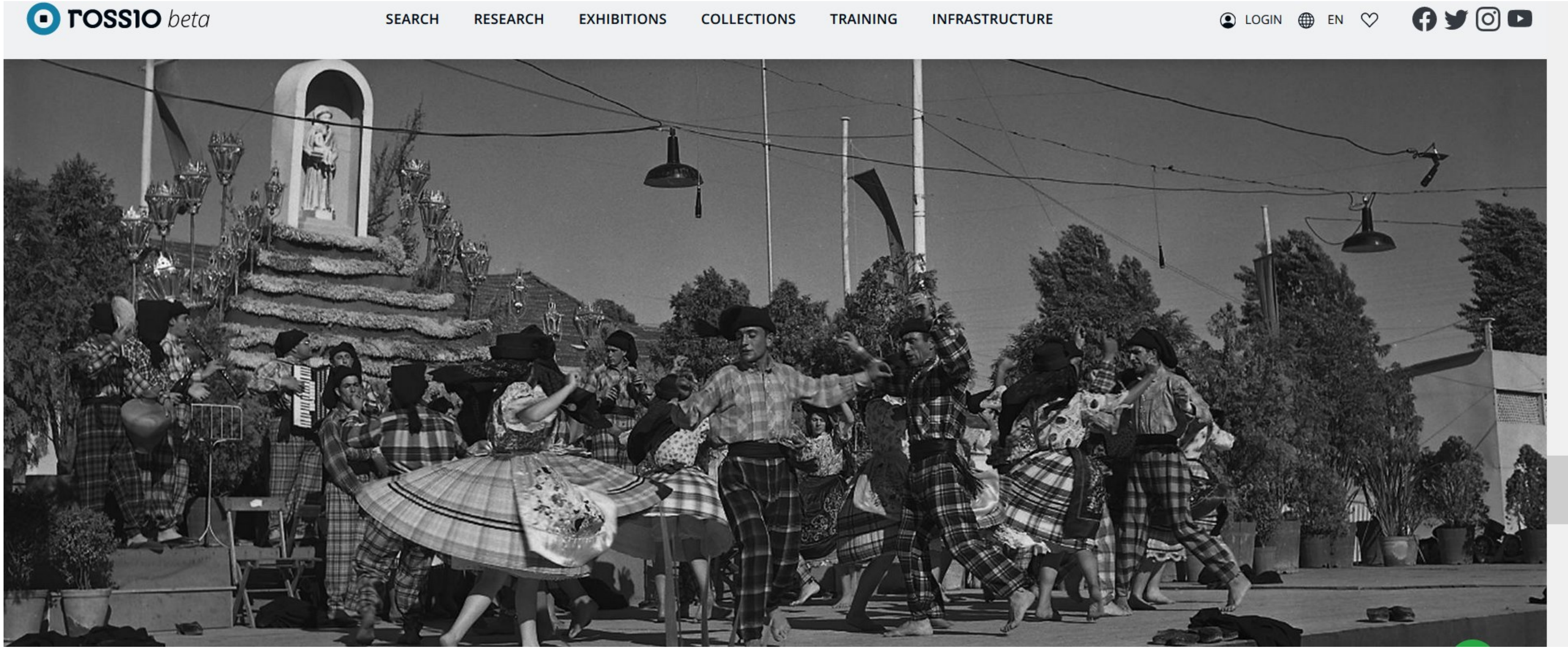
ROSSIO

ROSSIO

- Consortium of cultural and educational institutions
- Coordinated by the NOVA School of Social Sciences and Humanities of Lisbon
- Aggregates, contextualizes, enriches and disseminates digital content

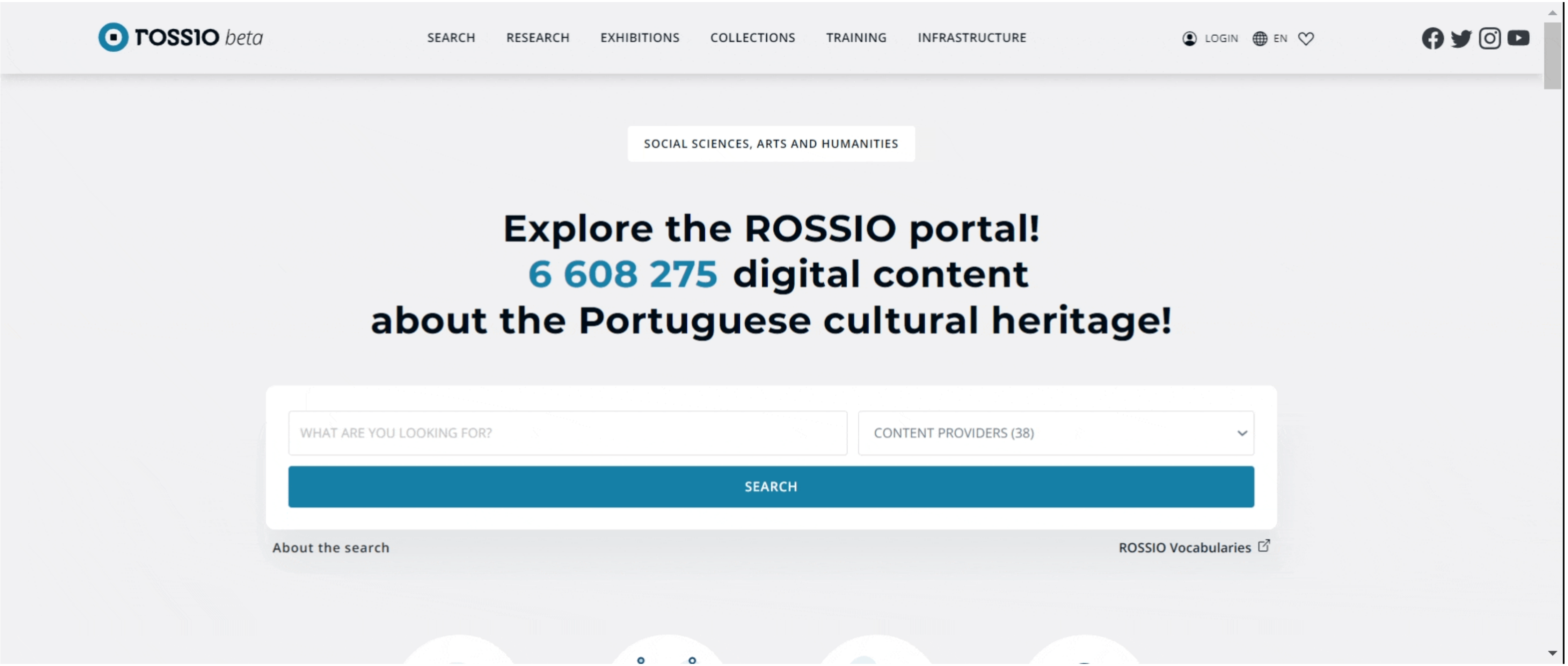
Portuguese node of the DARIAH.EU

ROSSIO



Web portal of the ROSSIO Infrastructure: https://rossio.pt

ROSSIO



SOCIAL SCIENCES, ARTS AND HUMANITIES

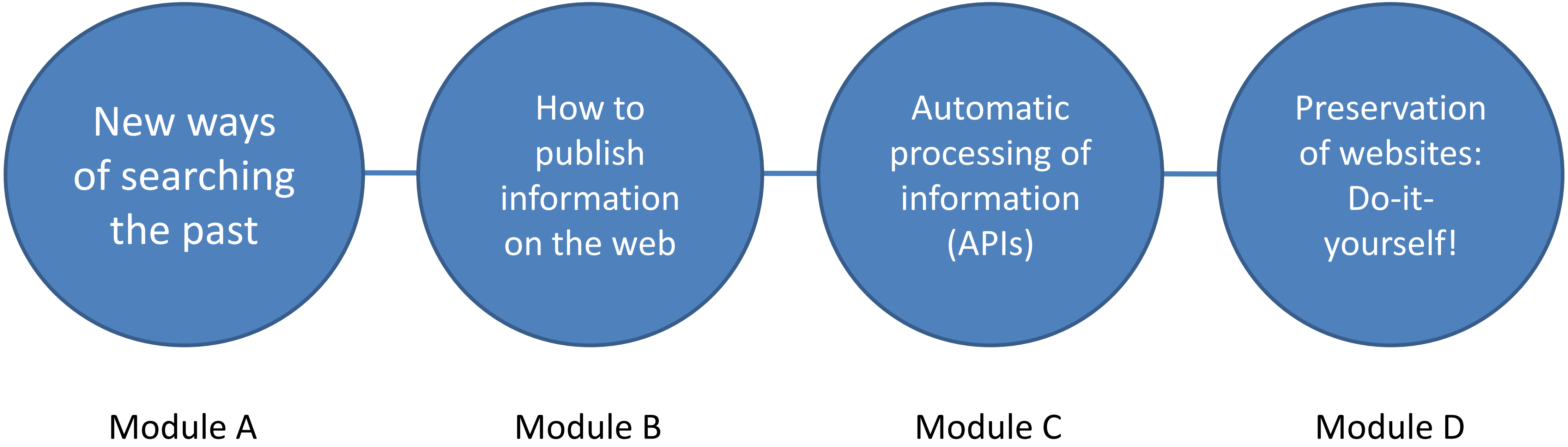
Explore the ROSSIO portal!
6 608 275 digital content
about the Portuguese cultural heritage!

3

Arquivo.pt training initiative

Café with Arquivo.pt
Webinars Cycles

Arquivo.pt training initiative



New ways of searching the past

Module A

How to publish information on the web

Module B

Automatic processing of information (APIs)

Module C

Preservation of websites: Do-it-yourself!

Module D

arquivo.pt/training

Arquivo.pt training 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

Automatic processing of information (APIs)

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)

Arquivo.pt training initiative

Café with Arquivo.pt

Arquivo.pt training initiative

2019

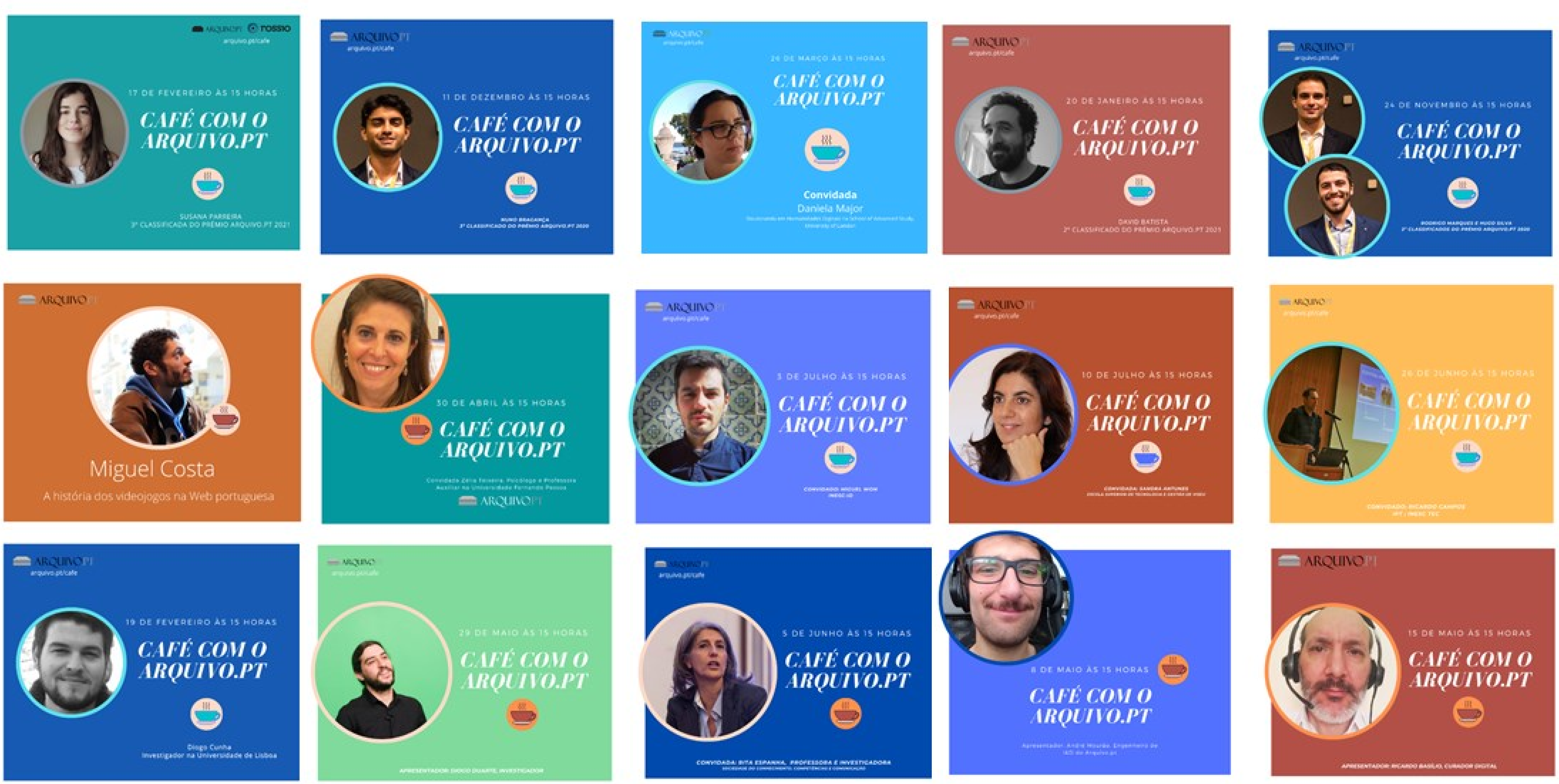
2020

Arquivo.pt training initiative - Café with Arquivo.pt

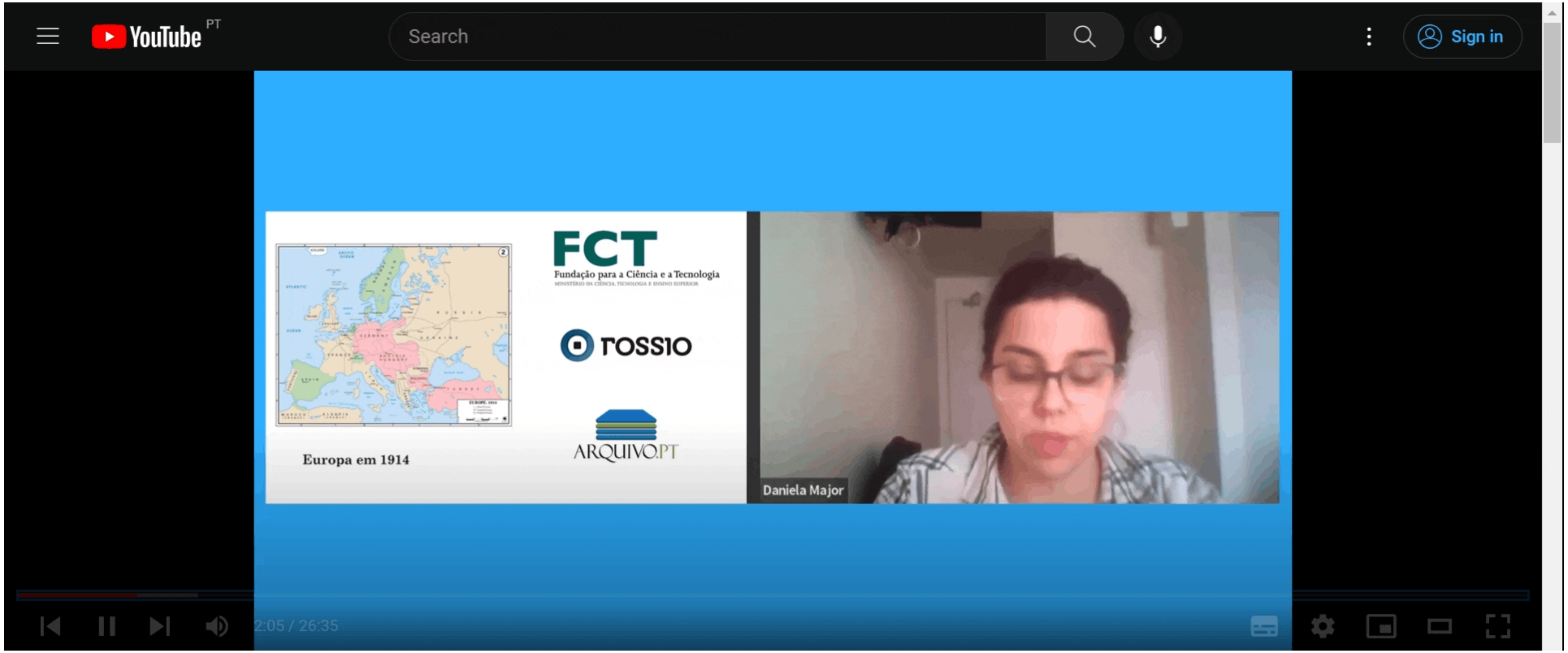
- 680 participants
- 23 sessions
- 30 participants/session
- 90% satisfaction

https://arquivo.pt/onlinecafe

Arquivo.pt training initiative - Café with Arquivo.pt



Arquivo.pt training initiative - Café with Arquivo.pt

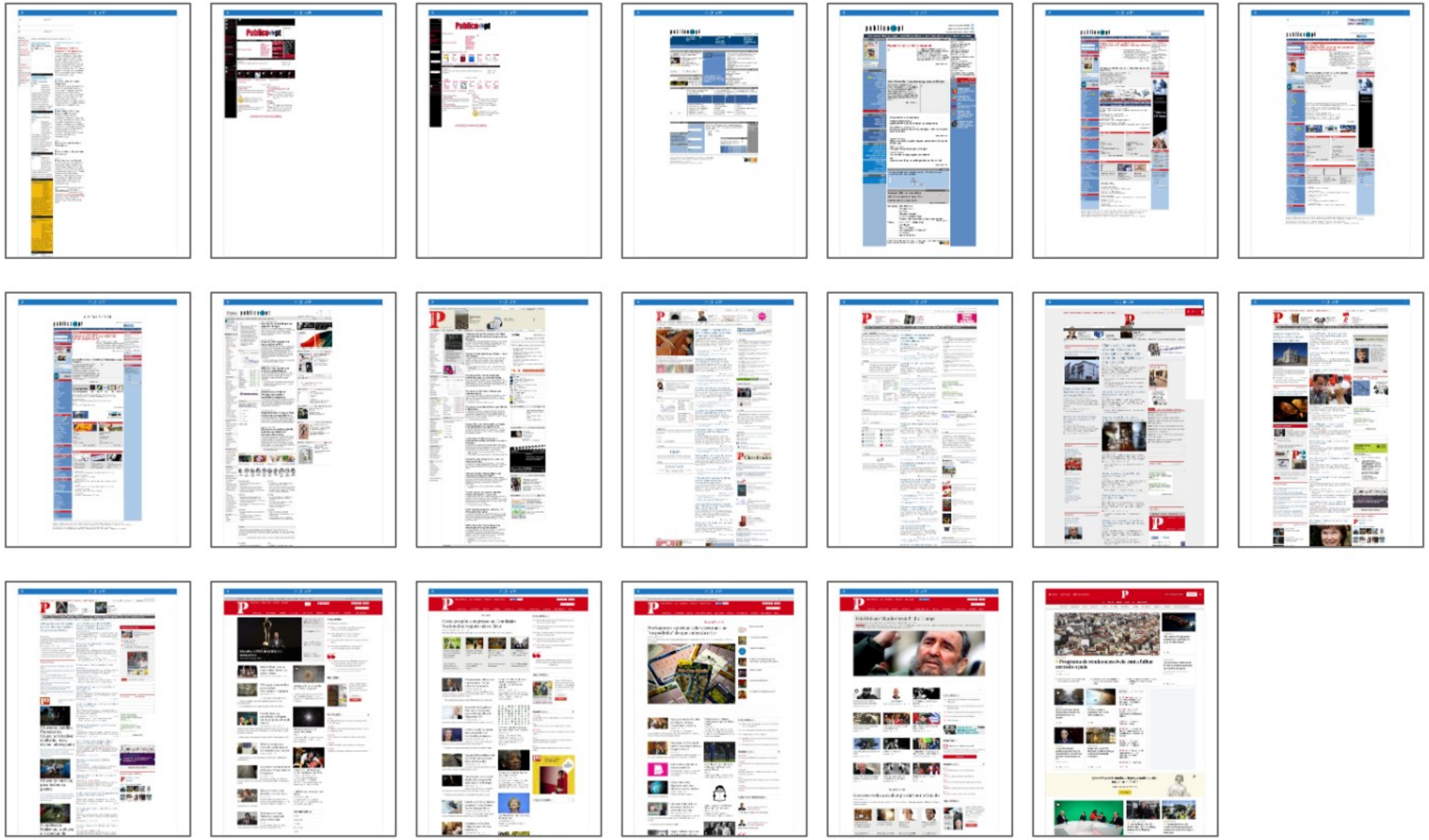

YouTube PT
Search
Sign in
FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR
ROSSIO
ARQUIVO.PT
Europa em 1914
Daniela Major
2:05 / 26:35

The commemoration of the Centennial of World War I, by Daniel Major

Arquivo.pt training initiative - Café with Arquivo.pt

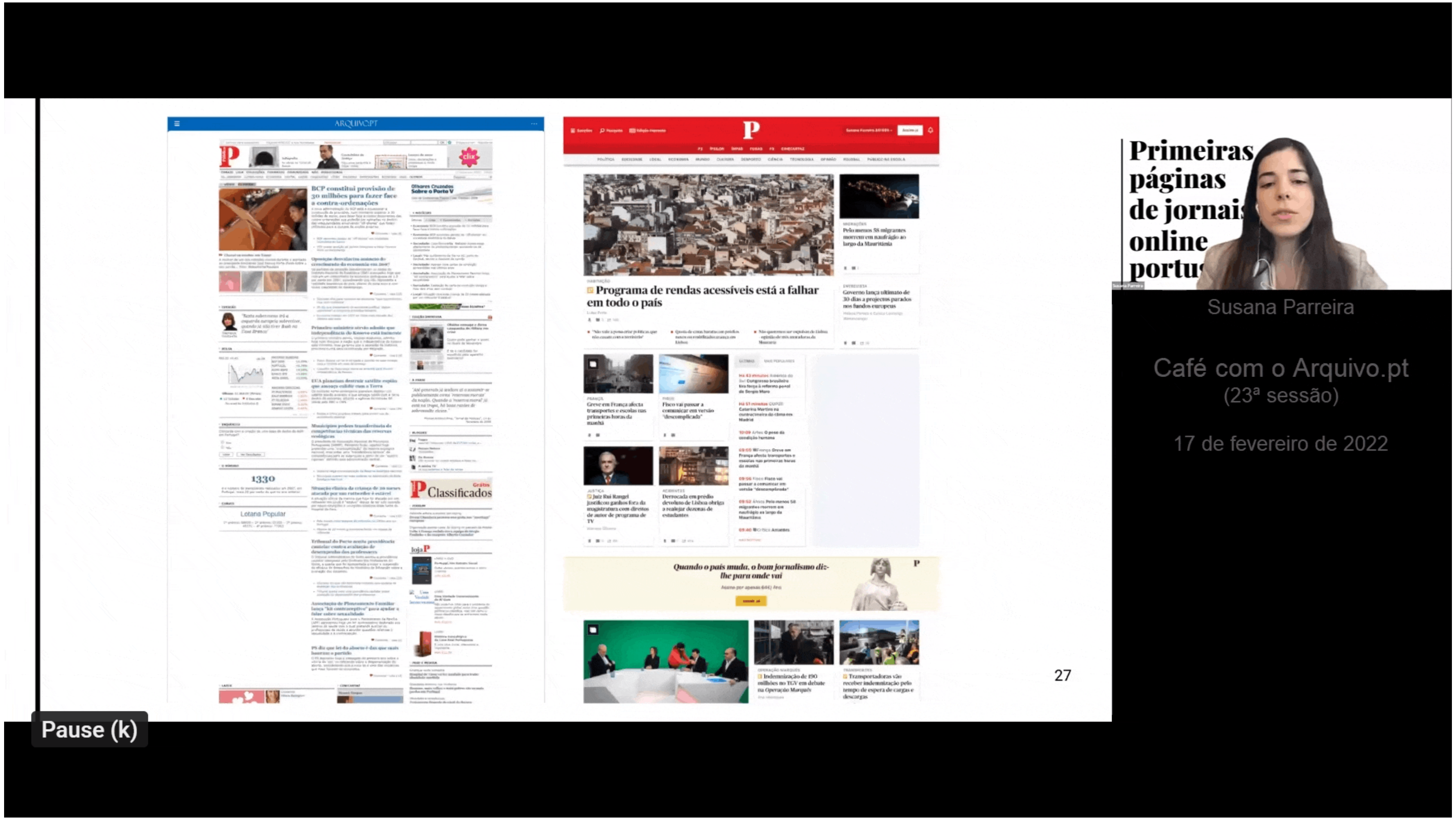
Arquivo.pt Award
2023
TRAVEL IN
TIME
and win 10.000€
arquivo.pt/award2023

Arquivo.pt training initiative - Café with Arquivo.pt



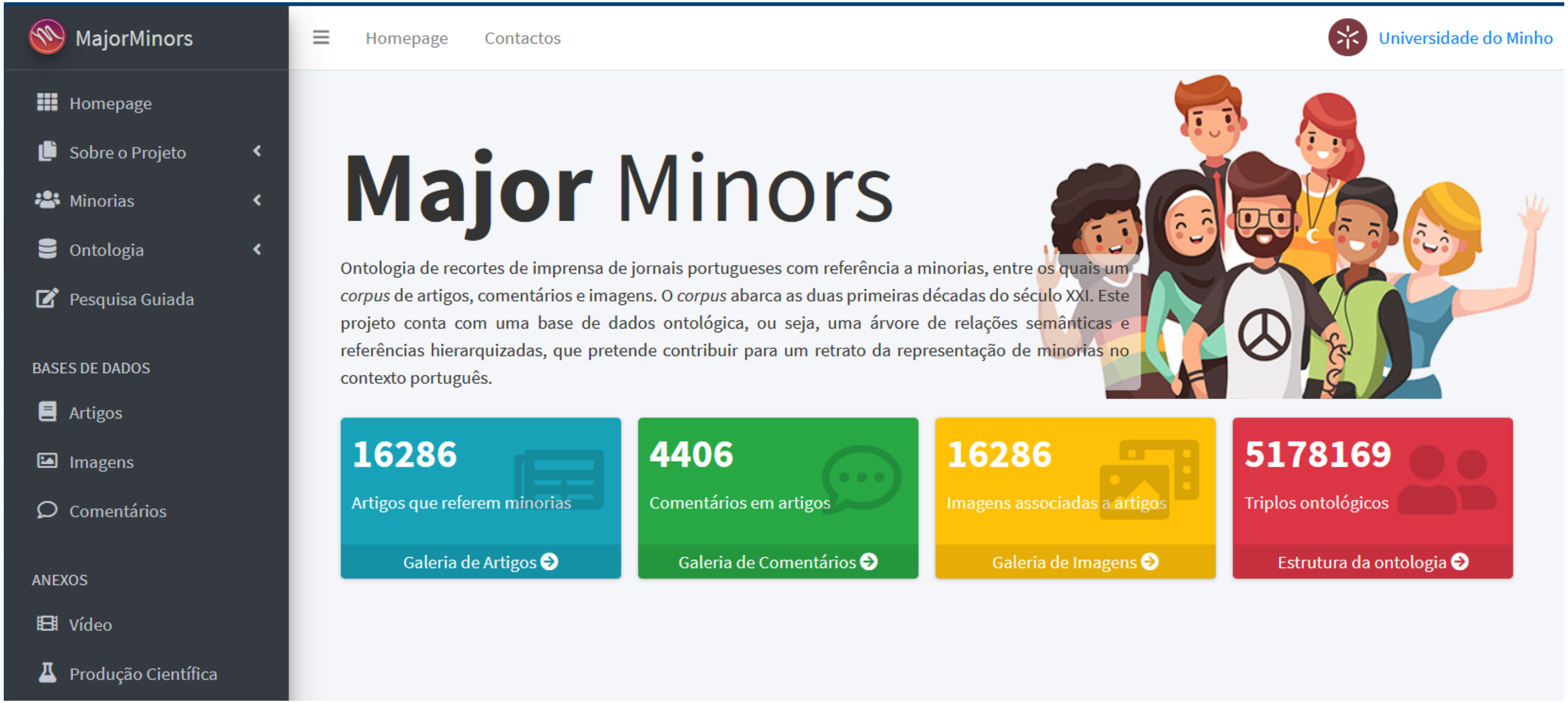
Front pages of Portuguese online newspapers, Arquivo.pt Award 2021 by Susana Parreira

Arquivo.pt training initiative – Café with Arquivo.pt


Primeiras páginas de jornais online portu
Susana Parreira
Café com o Arquivo.pt
(23ª sessão)
17 de fevereiro de 2022
27
Pause (k)
Programa de rendas acessíveis está a falhar em todo o país
Classificados

Front pages of Portuguese online newspapers, Arquivo.pt Award 2021 by Susana Parreira

Arquivo.pt training initiative - Café with Arquivo.pt



Major Minors

Ontologia de recortes de imprensa de jornais portugueses com referência a minorias, entre os quais um *corpus* de artigos, comentários e imagens. O *corpus* abarca as duas primeiras décadas do século XXI. Este projeto conta com uma base de dados ontológica, ou seja, uma árvore de relações semânticas e referências hierarquizadas, que pretende contribuir para um retrato da representação de minorias no contexto português.

16286
Artigos que referem minorias
Galeria de Artigos

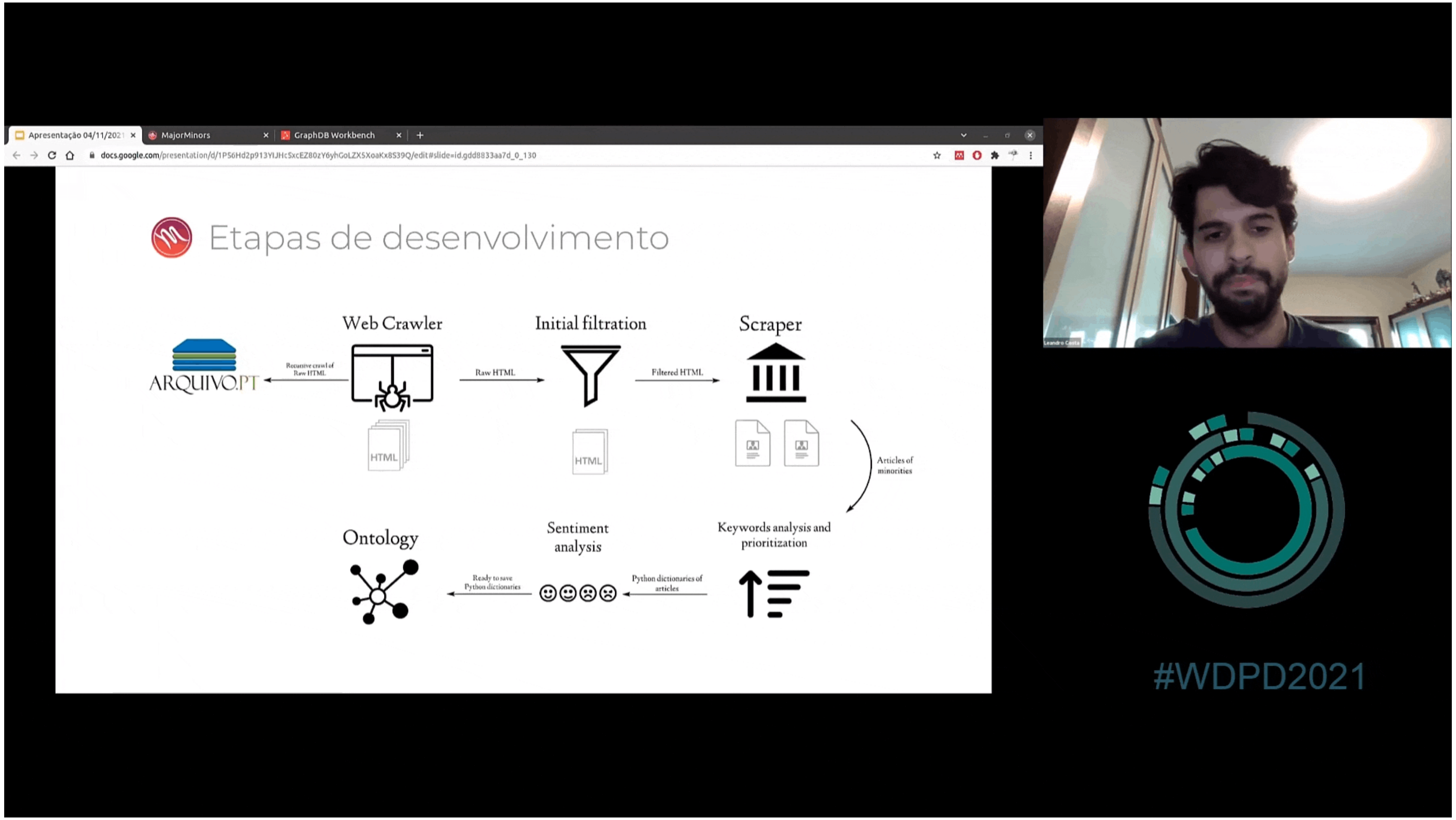
4406
Comentários em artigos
Galeria de Comentários

16286
Imagens associadas a artigos
Galeria de Imagens

5178169
Triplos ontológicos
Estrutura da ontologia

Major Minors, Arquivo.pt Award 2021 by Paulo Martins e Leandro Costa

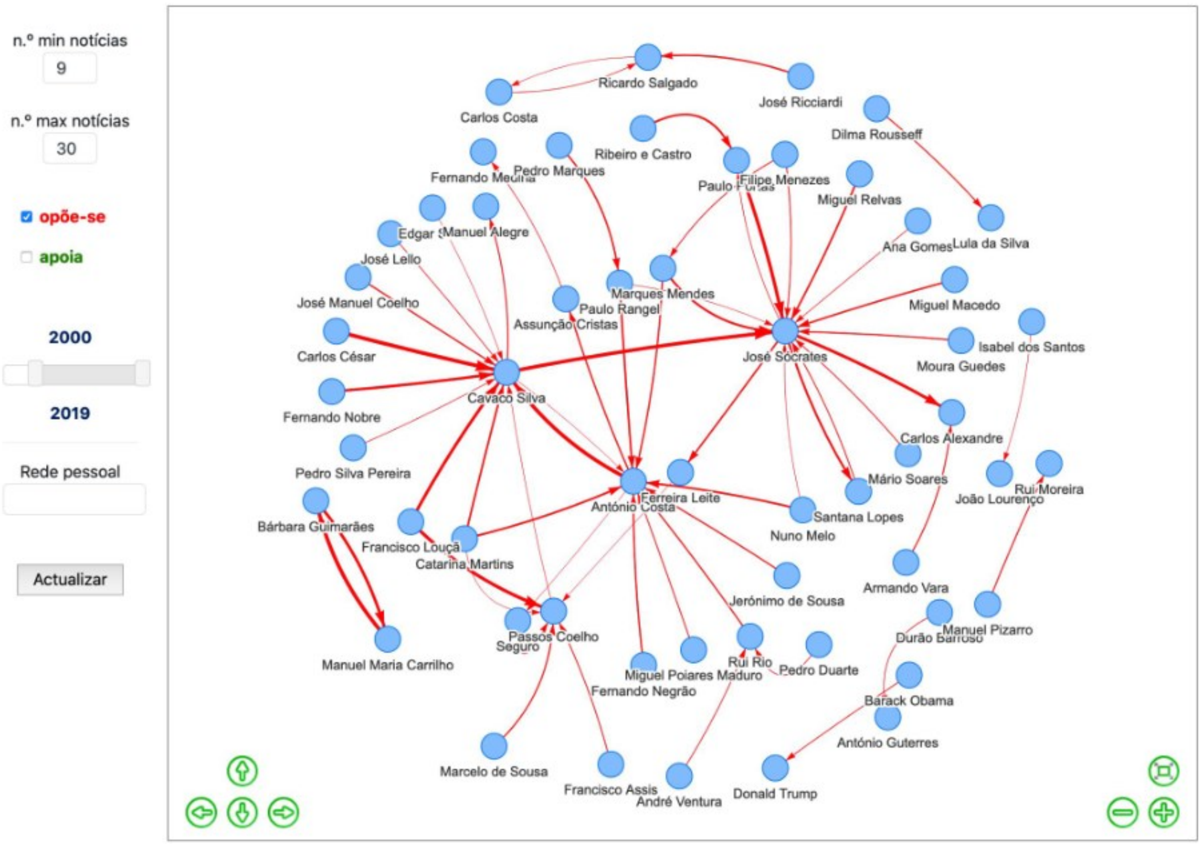
Arquivo.pt training initiative – Café with Arquivo.pt


Apresentação 04/11/2021
MajorMinors
GraphDB Workbench
docs.google.com/presentation/d/1PS6Hd2p913YIJHcSxcEZ80zY6yhGoLZX5XoaKx8S39Q/edit#slide=id.gdd8833aa7d_0_130
ARQUIVO.PT
Web Crawler
Recursive crawl of Raw HTML
Raw HTML
Initial filtration
Filtered HTML
Scraper
HTML
HTML
Articles of minorities
Keywords analysis and prioritization
Python dictionaries of articles
Sentiment analysis
Ready to save Python dictionaries
Ontology
#WDPD2021

Etapas de desenvolvimento

[Major Minors, Arquivo.pt Award 2021 by Paulo Martins e Leandro Costa]

Arquivo.pt training initiative – Café with Arquivo.pt



n.º min notícias: 9

n.º max notícias: 30

☑ opõe-se

□ apoia

2000 – 2019

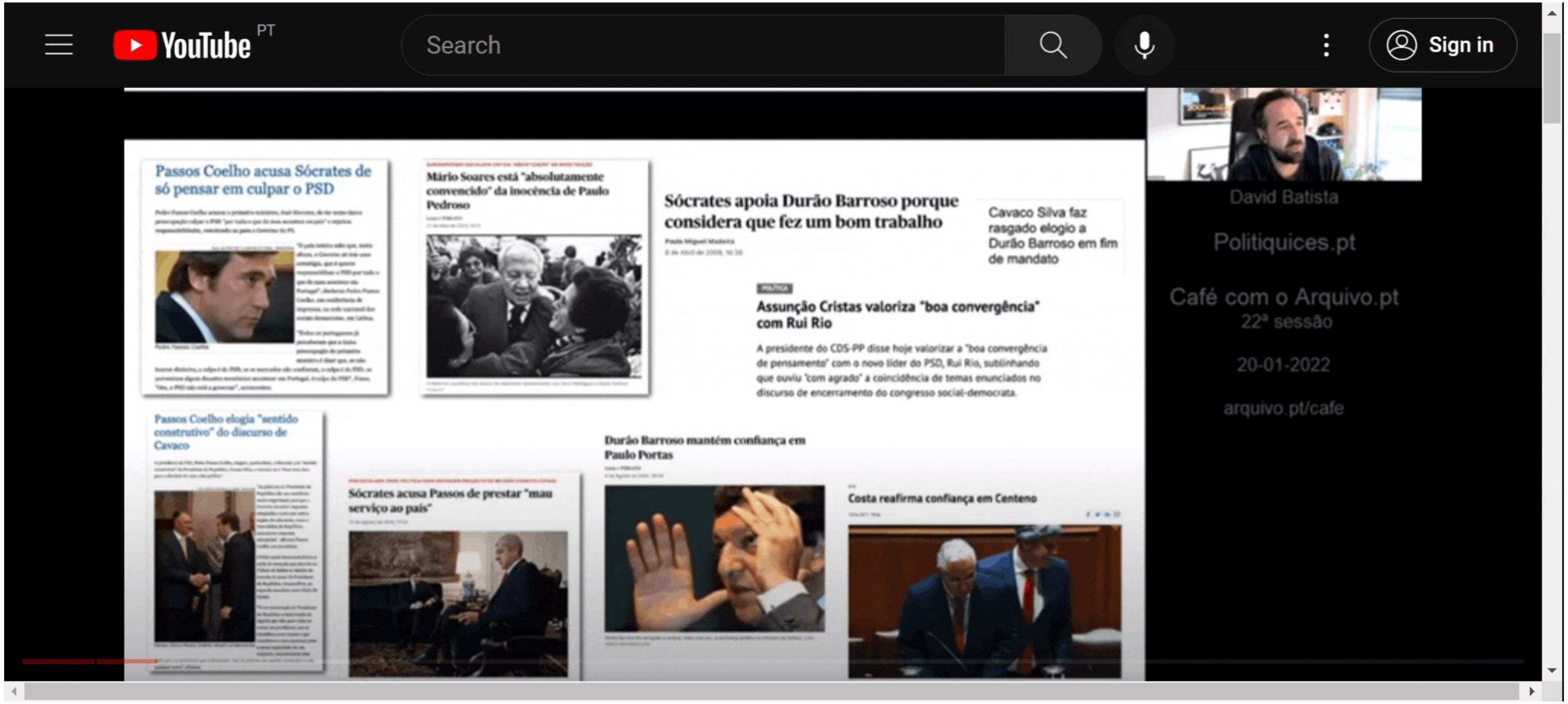
Rede pessoal

Actualizar

WIKIDATA

www.politiquices.pt, Arquivo.pt Award 2021, by David Batista

Arquivo.pt training initiative – Café with Arquivo.pt



Passos Coelho acusa Sócrates de só pensar em culpar o PSD

Mário Soares está "absolutamente convencido" da inocência de Paulo Pedroso

Sócrates apoia Durão Barroso porque considera que fez um bom trabalho

Cavaco Silva faz rasgado elogio a Durão Barroso em fim de mandato

Assunção Cristas valoriza "boa convergência" com Rui Rio

A presidente do CDS-PP disse hoje valorizar a "boa convergência de pensamento" com o novo líder do PSD, Rui Rio, sublinhando que ouviu "com agrado" a coincidência de temas enunciados no discurso de encerramento do congresso social-democrata.

Passos Coelho elogia "sentido construtivo" do discurso de Cavaco

Sócrates acusa Passos de prestar "mau serviço ao país"

Durão Barroso mantém confiança em Paulo Portas

Costa reafirma confiança em Centeno

David Batista

Politiquices.pt

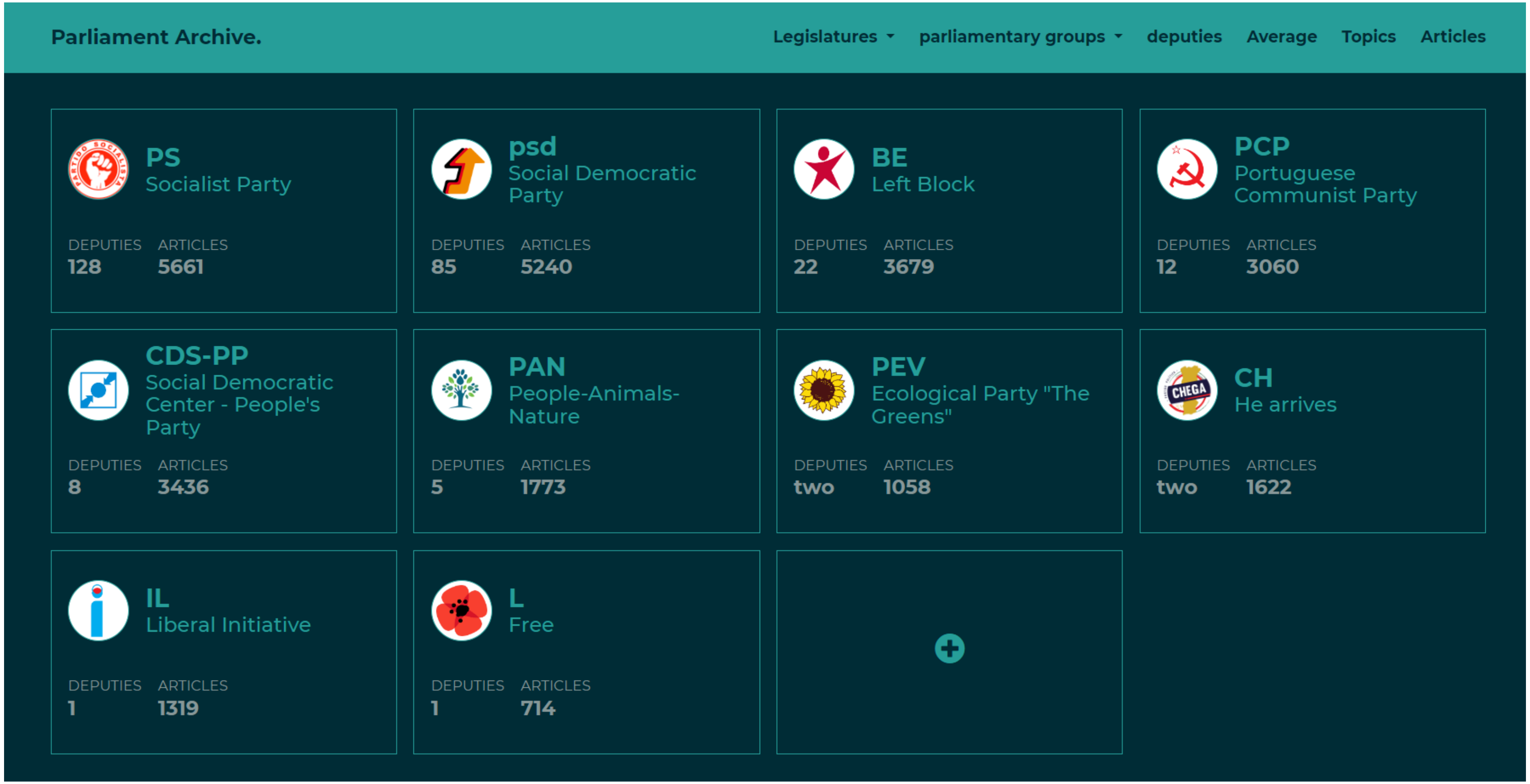
Café com o Arquivo.pt
22ª sessão

20-01-2022

arquivo.pt/cafe

www.politiquices.pt, Arquivo.pt Award 2021, by David Batista

Arquivo.pt training initiative – Café with Arquivo.pt



PS
Socialist Party
DEPUTIES 128
ARTICLES 5661

psd
Social Democratic Party
DEPUTIES 85
ARTICLES 5240

BE
Left Block
DEPUTIES 22
ARTICLES 3679

PCP
Portuguese Communist Party
DEPUTIES 12
ARTICLES 3060

CDS-PP
Social Democratic Center - People's Party
DEPUTIES 8
ARTICLES 3436

PAN
People-Animals-Nature
DEPUTIES 5
ARTICLES 1773

PEV
Ecological Party "The Greens"
DEPUTIES two
ARTICLES 1058

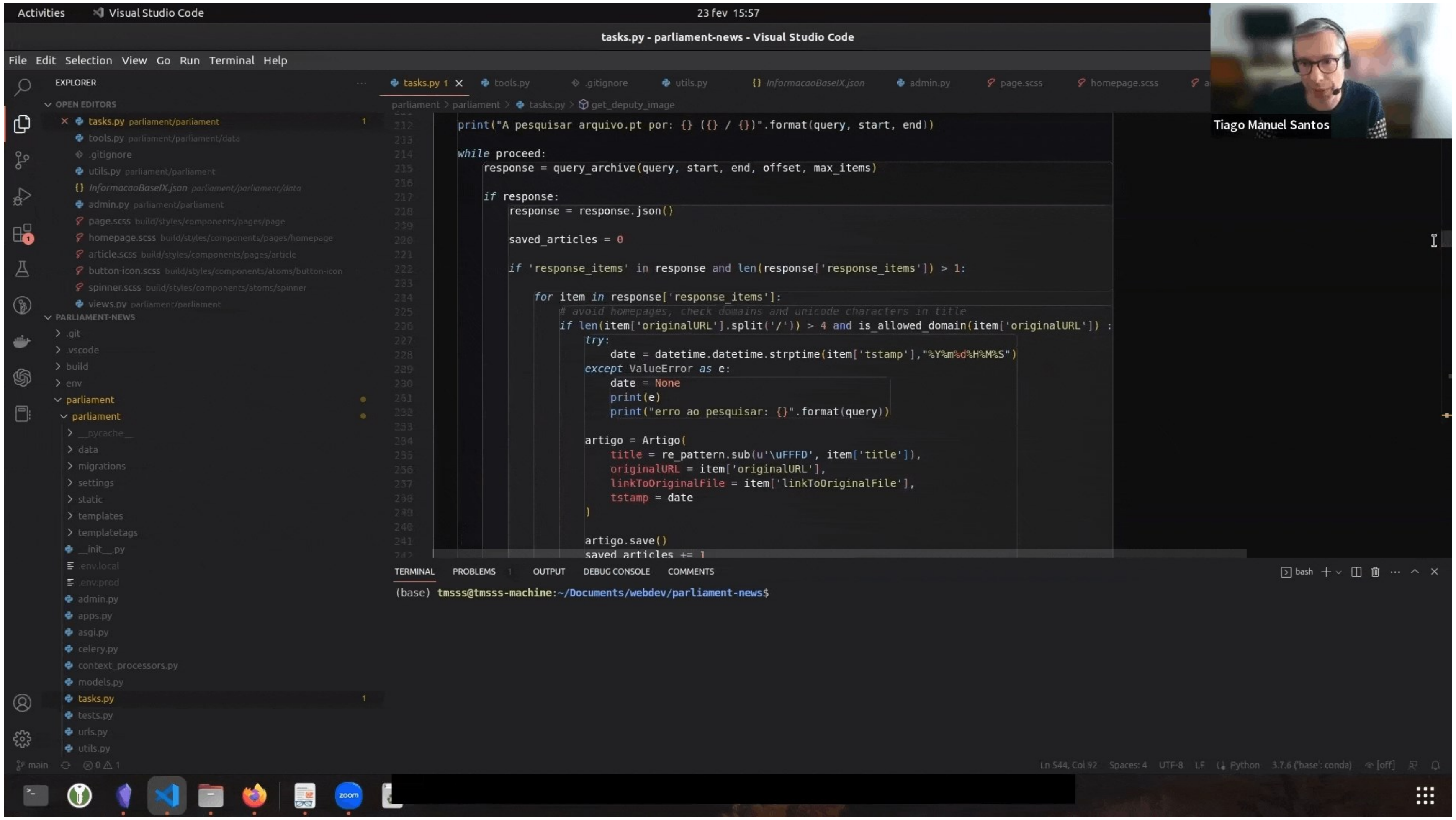
CH
He arrives
DEPUTIES two
ARTICLES 1622

IL
Liberal Initiative
DEPUTIES 1
ARTICLES 1319

L
Free
DEPUTIES 1
ARTICLES 714

Parlamientary archive (Portugal), Arquivo.pt Award 2022, by Tiago Santos - arquivo-parlamento.pt

Arquivo.pt training initiative – Café with Arquivo.pt

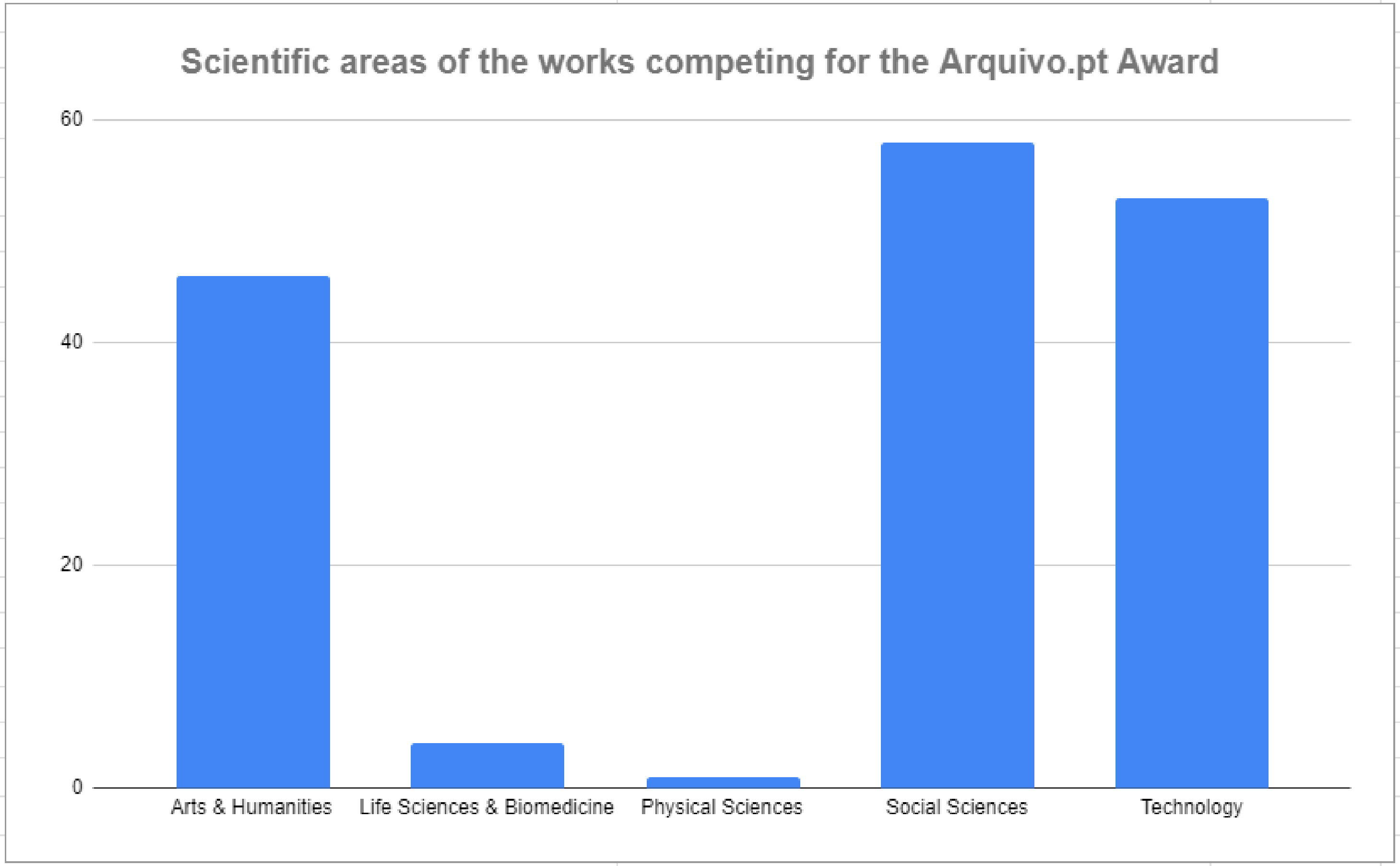


```
print("A pesquisar arquivo.pt por: {} ({} / {})".format(query, start, end))

while proceed:
    response = query_archive(query, start, end, offset, max_items)

    if response:
        response = response.json()

        saved_articles = 0

        if 'response_items' in response and len(response['response_items']) > 1:

            for item in response['response_items']:
                # avoid homepages, check domains and unicode characters in title
                if len(item['originalURL'].split('/')) > 4 and is_allowed_domain(item['originalURL']) :
                    try:
                        date = datetime.datetime.strptime(item['tstamp'],"%Y%m%d%H%M%S")
                    except ValueError as e:
                        date = None
                        print(e)
                        print("erro ao pesquisar: {}".format(query))

                    artigo = Artigo(
                        title = re_pattern.sub(u'\uFFFD', item['title']),
                        originalURL = item['originalURL'],
                        linkToOriginalFile = item['linkToOriginalFile'],
                        tstamp = date
                    )

                    artigo.save()
                    saved_articles += 1
```

```
(base) tmsss@tmsss-machine:~/Documents/webdev/parliament-news$
```

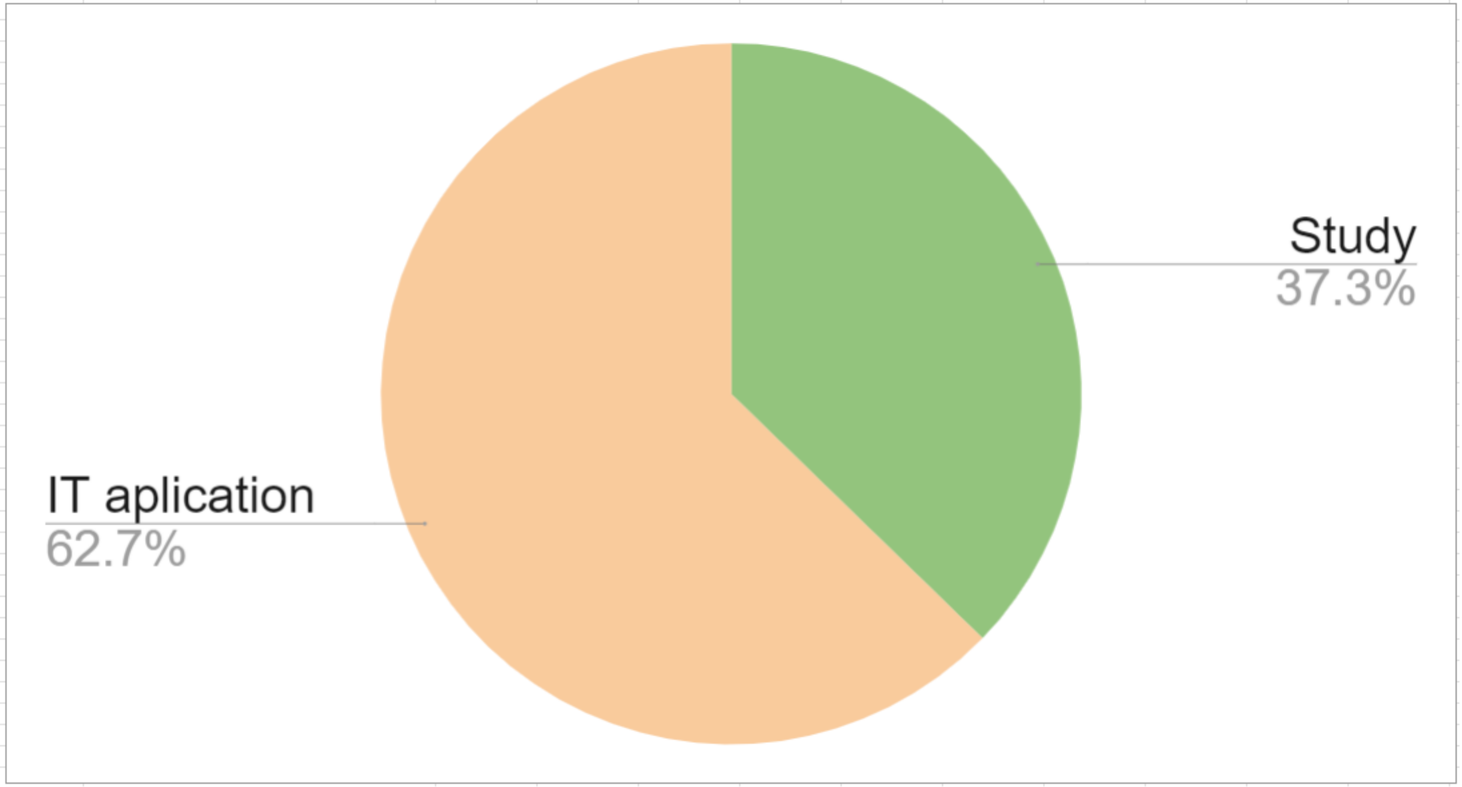
Parlamientary archive (Portugal), Arquivo.pt Award 2022, by Tiago Santos - arquivo-parlamento.pt

What a web archive is for?



Areas of study of the 160 works competing for the Arquivo.pt Award, between 2018 and 2023

What a web archive is for?



Studies versus IT applications. Of the 160 works competing for the Arquivo.pt Awards, between 2018 and 2023

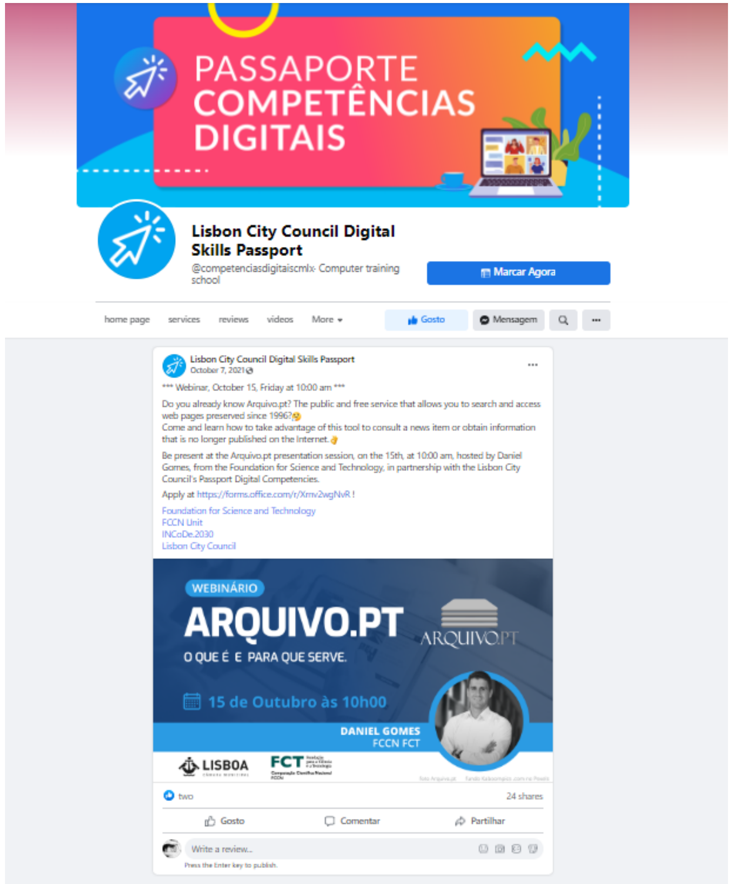
Arquivo.pt training initiative

Webinars cycle

Arquivo.pt training initiative – Webinars cycle

Webinars cycle promoted by the Lisbon City Council

Target: Common citizens



Lisbon City Council Digital Skills Passport
@competenciasdigitaiscmlx· Computer training school

Lisbon City Council Digital Skills Passport
October 7, 2021

*** Webinar, October 15, Friday at 10:00 am ***

Do you already know Arquivo.pt? The public and free service that allows you to search and access web pages preserved since 1996?🤔
Come and learn how to take advantage of this tool to consult a news item or obtain information that is no longer published on the Internet.👌

Be present at the Arquivo.pt presentation session, on the 15th, at 10:00 am, hosted by Daniel Gomes, from the Foundation for Science and Technology, in partnership with the Lisbon City Council's Passport Digital Competencies.

Apply at https://forms.office.com/r/Xmv2wgNvR !

Foundation for Science and Technology
FCCN Unit
INCoDe.2030
Lisbon City Council

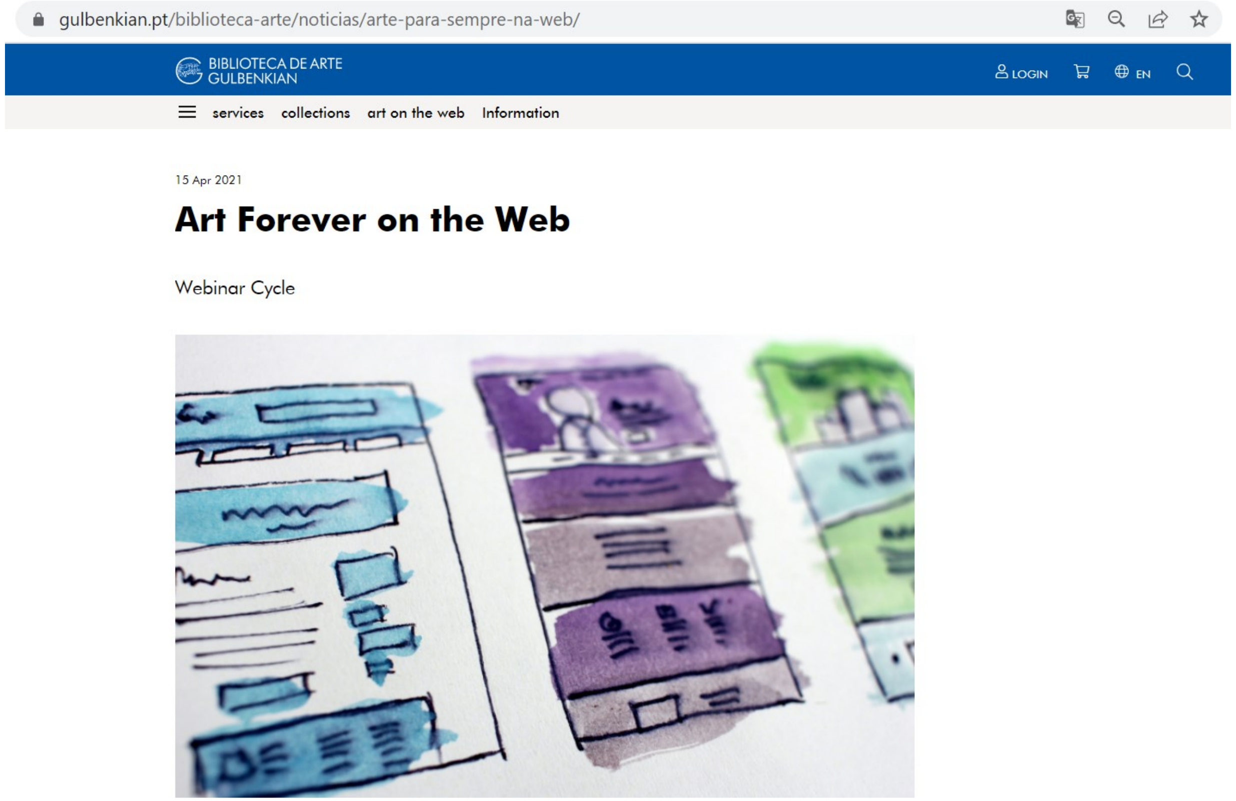
Page view through GoogleTranslate

facebook.com/competenciasdigitaiscmlx

Arquivo.pt training initiative – Webinars cycle

Webinars cycle promoted by the Fundação Calouste Gulbenkian (Art Library and Archive)

Target: artists, gallery owners and reseachers


gulbenkian.pt/biblioteca-arte/noticias/arte-para-sempre-na-web/
BIBLIOTECA DE ARTE GULBENKIAN
LOGIN
EN
services
collections
art on the web
Information

15 Apr 2021

Art Forever on the Web

Webinar Cycle

Page view through GoogleTranslate

gulbenkian.pt/biblioteca-arte/noticias/arte-para-sempre-na-web

Arquivo.pt training initiative – return to face-to-face presentations

Digital Humanities, a priority

Target: Professors, students and researchers

Workshop promoted by the Centre for Comparative Studies (CEComp) at the School of Arts and Humanities, University of Lisbon.

Arquivo.pt training initiative - Methodological challenges

«Dealing with a greater volume of data will be a common occurrence that historians will have to face from now on. This opens up **a world of possibility for historians**, but it also forces us to ask difficult questions».

Daniela MAJOR, Early Stage Researcher in Digital Humanities at the School of Advanced Study in London. It Takes a Village to Raise an Archive. How the Use of Web Sources Fosters Collaboration, Published at Web Archiving Section of the Society of American Archivists on January 18, 2023.

Conclusions

Conclusions

- 2018 - Four modules to offer to the community
- 2020 – Café and Webinars to receive contributes from the community
- 2022 – Two-way knowledge sharing (Users and Arquivo.pt team)

There are methodological challenges for Digital Humanities using

Web archives and these can be overcome with close collaboration.

Thank you!

contacto@arquivo.pt

arquivo.pt/onlinecafe

arquivo.pt/training

arquivo.pt/awards