Secondments@Arquivo.pt and ***new*** research tools available

daniel.gomes@fccn.pt

Who are we?

ARQUIVO.PT

Free online service to research the Past Web

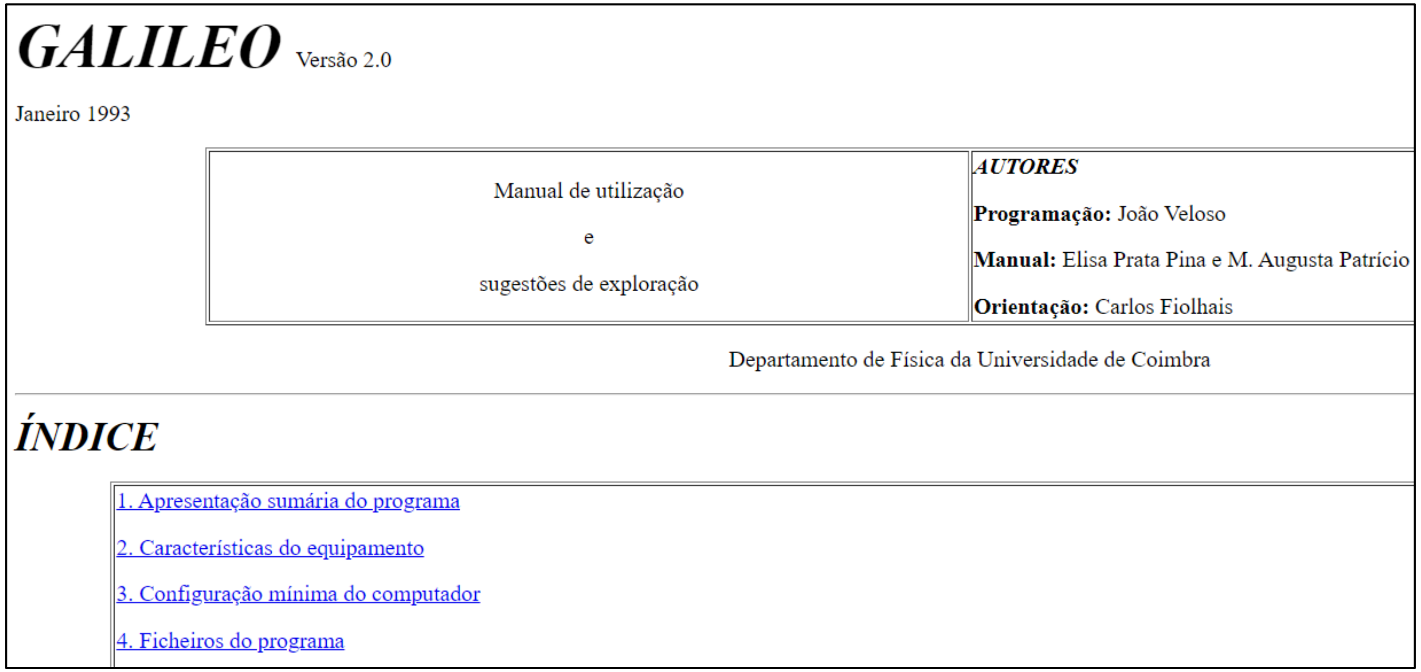
Preserves **publicly accessible** information related with:

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

Governmental service provided by
the Ministry for Science and Technology (Portugal)

A digital research infrastructure

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 Patrício **Orientação:** Carlos Fiolhais

Departamento de Física da Universidade de Coimbra

ÍNDICE

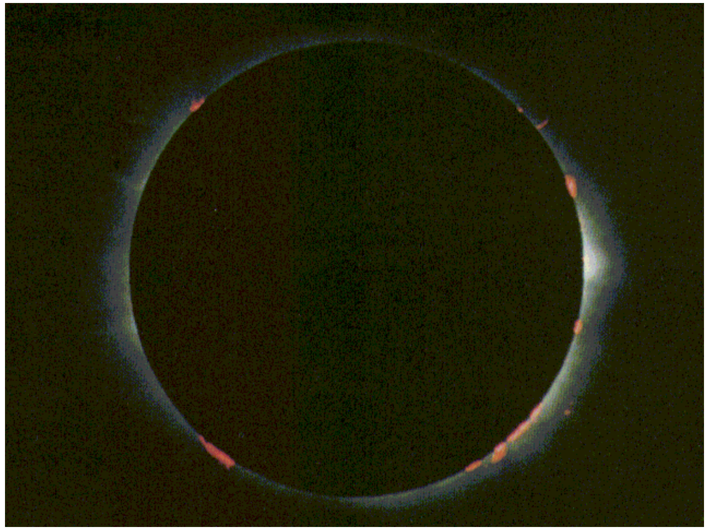
1. Apresentação sumária do programa

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)

CLEOPATRA Secondments

- Diego Alves (ESR14), “Robustness of Corpus-based Typological Strategies for Dependency Parsing”, August 2022
- Swati (ESR12), “Cross-lingual news reporting bias”, December 2022

➢ Get feedback from researchers with real questions about our research tools to improve them

Robustness of Corpus-based Typological Strategies for Dependency Parsing

ESR14 Secondment

From 06/2022 to 08/2022

Diego Alves
Daniel Gomes

Funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 812997

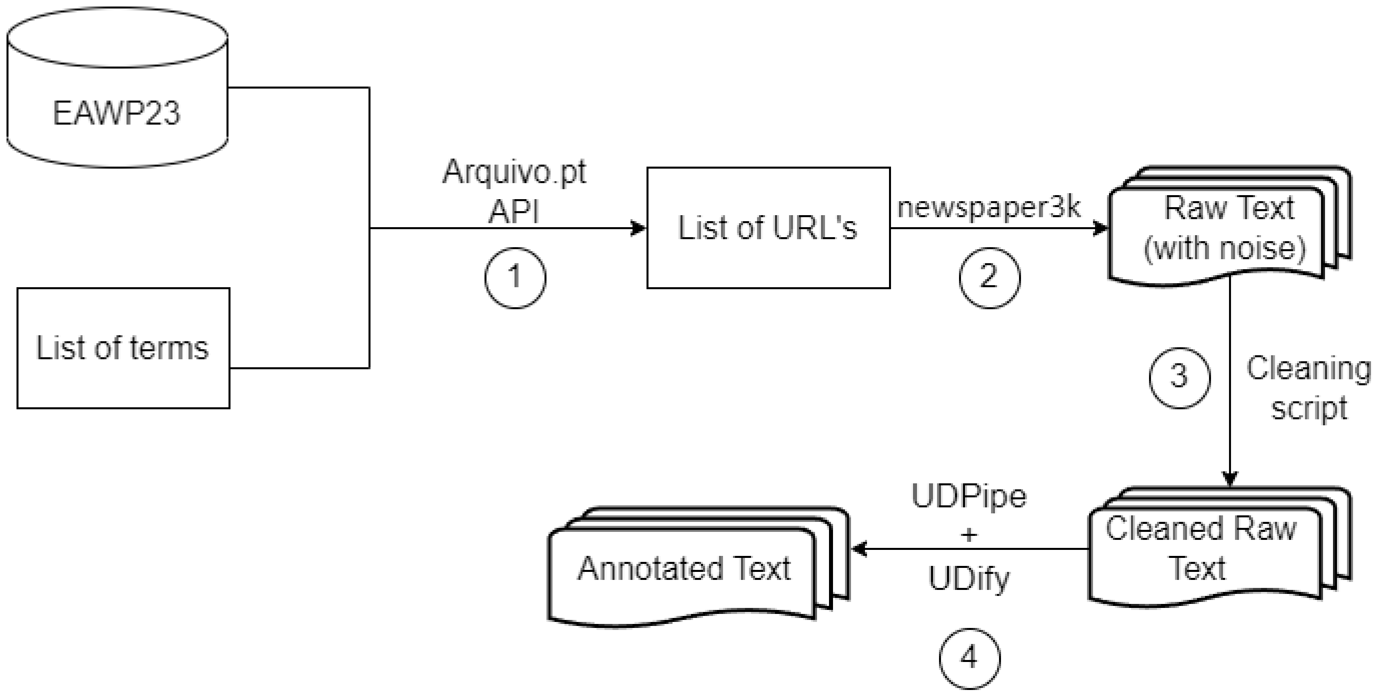
Objectives

- Creation of datasets from texts extracted from the EAWP23 collection of the Arquivo.pt:
  - 2019 European Parliamentary Elections collection
  - 24 European Union official languages
- Conduct corpus-based typological analysis with syntactic information extracted from this dataset and compare to the state-of-the-art (i.e.: parallel corpora - PUD)

Dataset creation

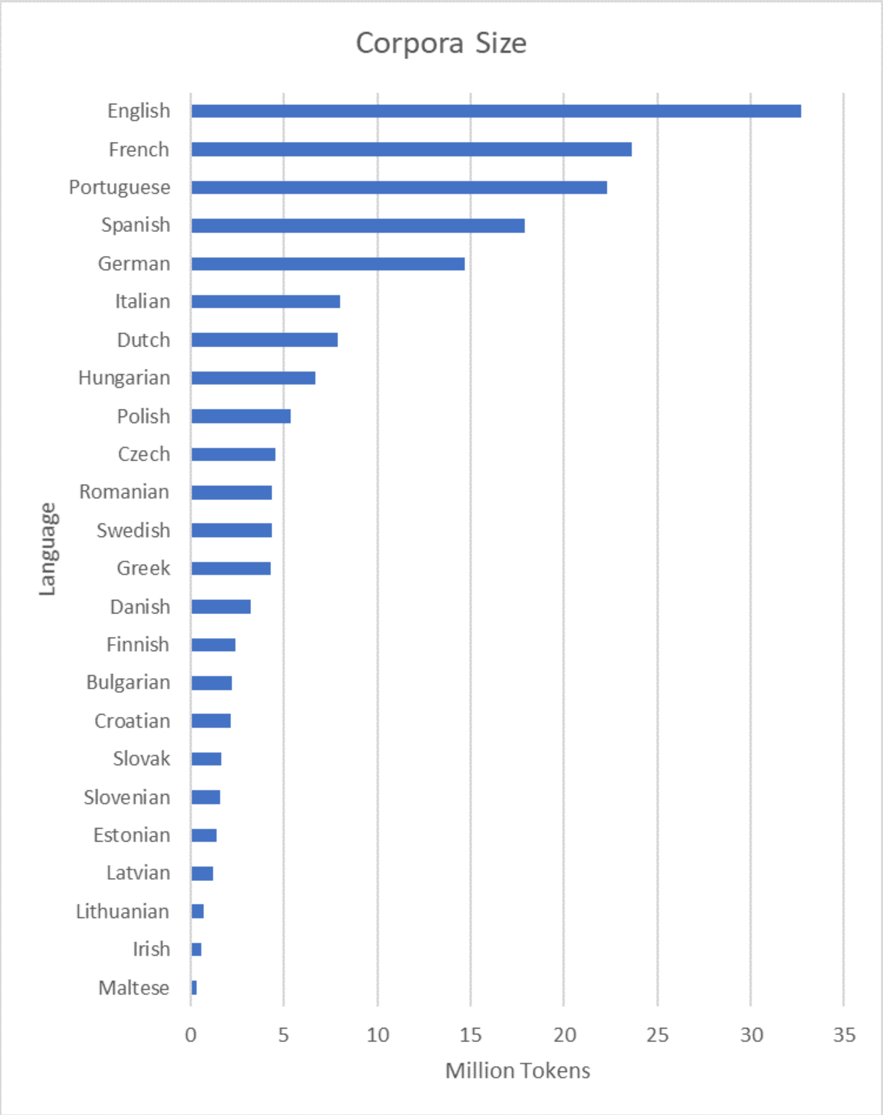
- EAWP23
  - Semi-automatic selection of relevant online content (40 relevant terms + collaborative list)
  - Crawling web content (6 crawls using different configurations)
  - 99 million URLs (4.8 TB)

Dataset creation



Raw text corpus:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/WCGNHU
Annotated corpus:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ULWS1K

Dataset Creation

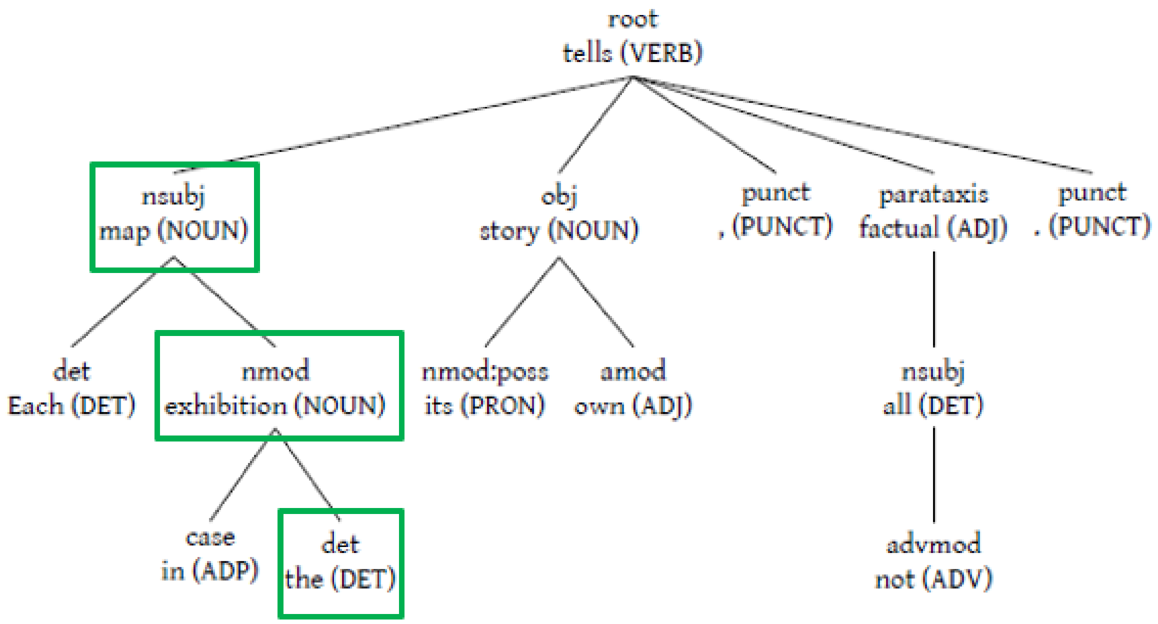


Typological Analysis

- Best identified quantitative approach for dependency parsing improvement:
  - MarsaGram linear patterns
- MarsaGram is a tool for exploring treebanks, it extracts context-free grammars (CFG) from annotated data-sets
- Linear property:
  - C_A_precedes_B
  - element A (POS - deprel) precedes element B (POS - deprel) in a subtree ruled by element C (POS)
- Language vectors:
  - Features: MarsaGram patterns
  - Values: Frequency in the corpus

Typological Analysis

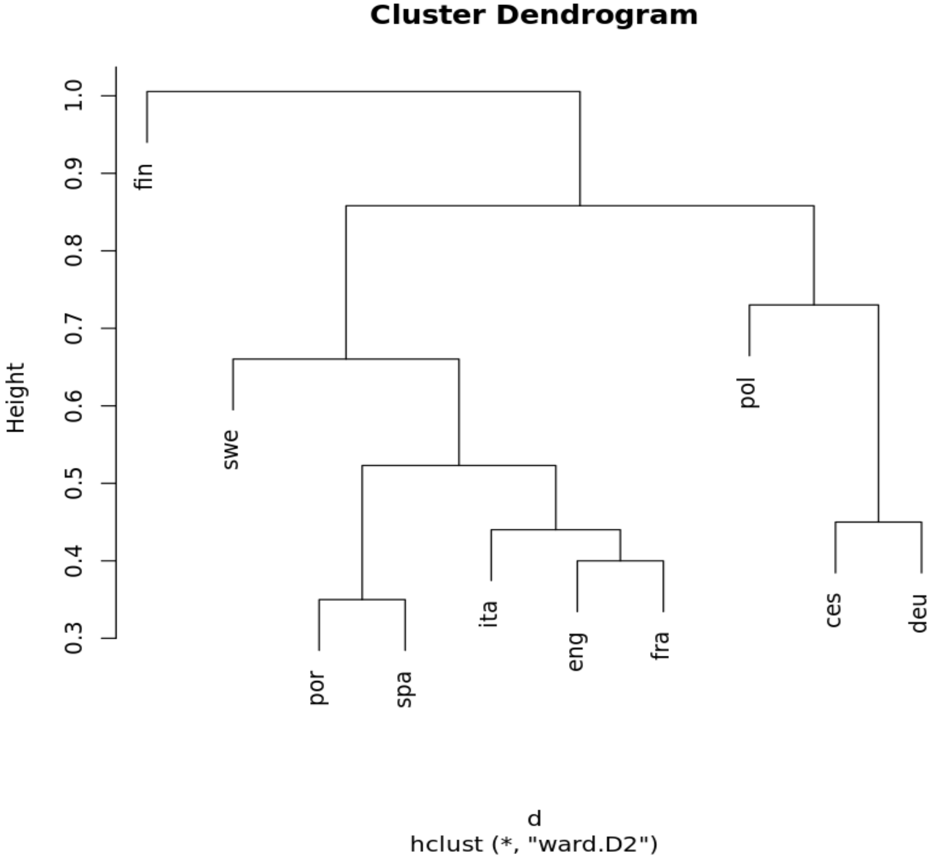
- NOUN_precede_DET-det_NOUN-nmod



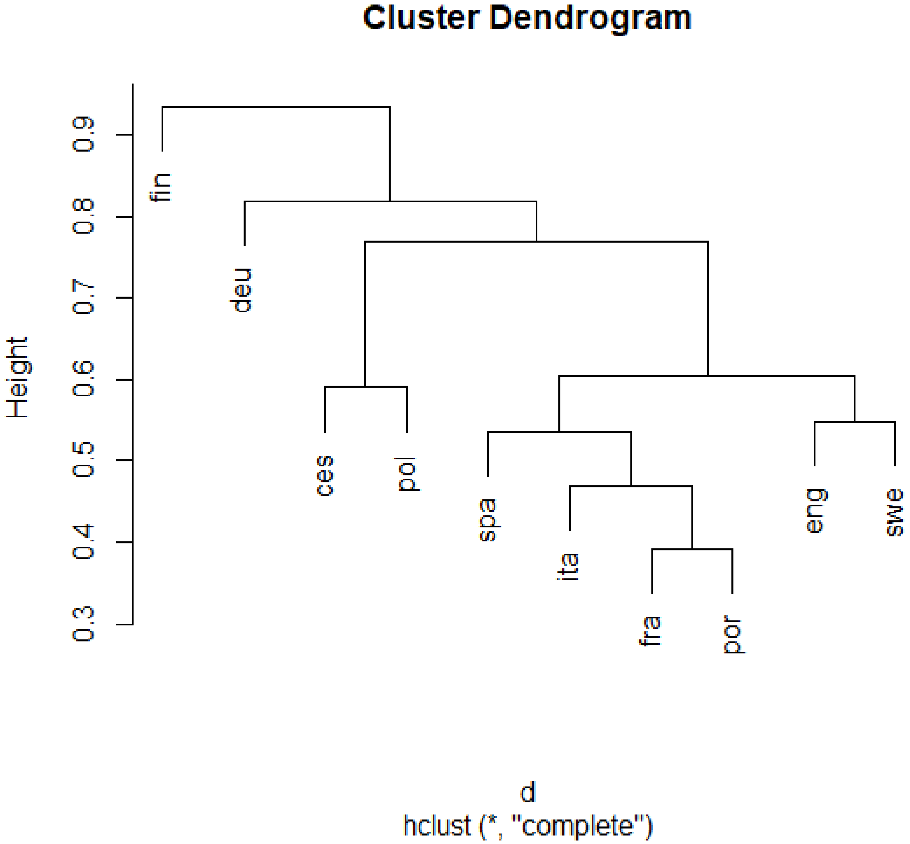
"Each map in the exhibition tells its own story, not all factual."

Typological Analysis

PUD



Arquivo.pt



Conclusions and Perspectives

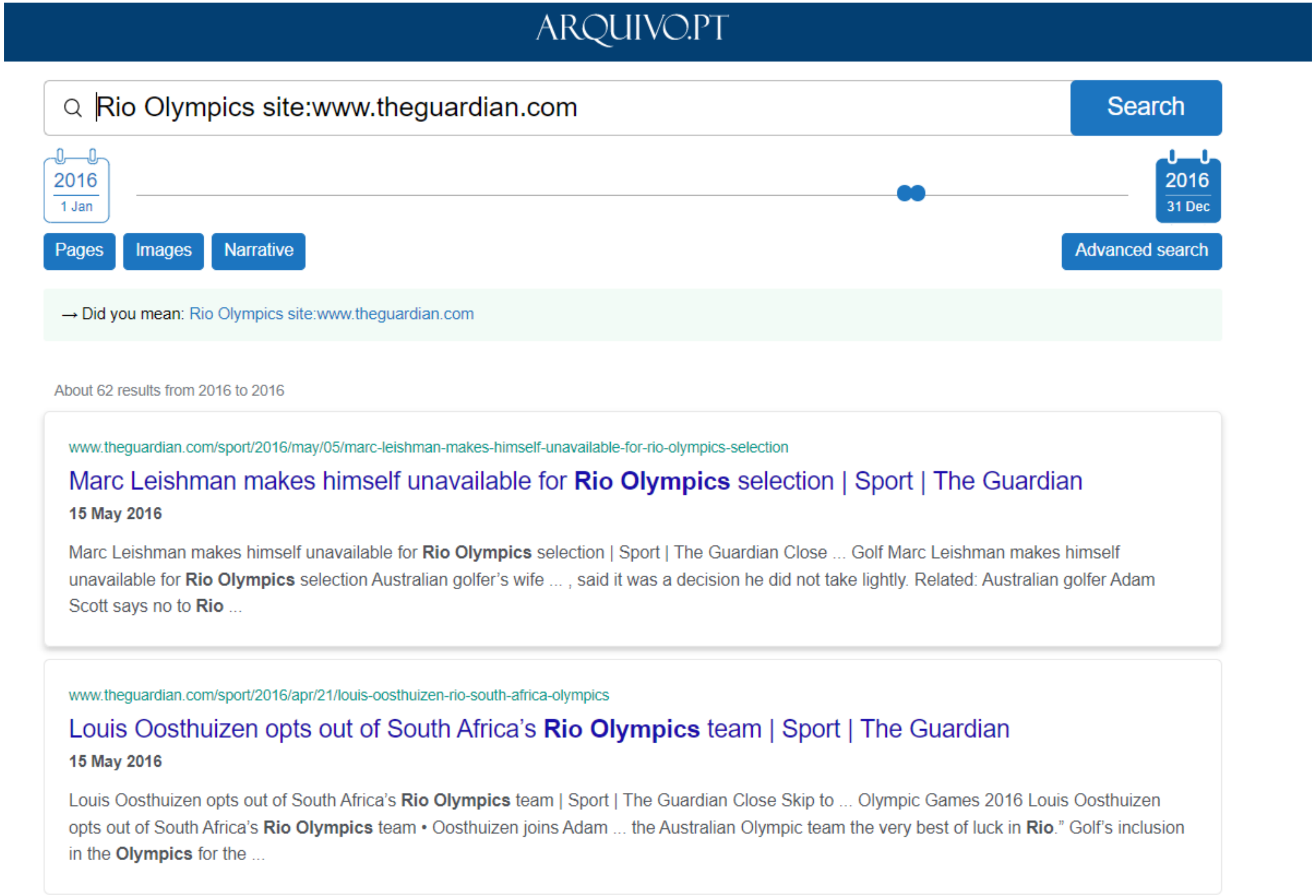
- Differences in the typological classification specially concerning Germanic languages

- To be evaluated:
  - Deeper analysis of the bias introduced by the automatic annotation
  - Impact on the choice of best language pairs for dependency parsing improvement

New research tools available!

daniel.gomes@fccn.pt

Search **texts** from the past in any language or about any event



Rio Olympics site:www.theguardian.com

→ Did you mean: Rio Olympics site:www.theguardian.com

About 62 results from 2016 to 2016

www.theguardian.com/sport/2016/may/05/marc-leishman-makes-himself-unavailable-for-rio-olympics-selection

Marc Leishman makes himself unavailable for **Rio Olympics** selection | Sport | The Guardian

15 May 2016

Marc Leishman makes himself unavailable for **Rio Olympics** selection | Sport | The Guardian Close ... Golf Marc Leishman makes himself unavailable for **Rio Olympics** selection Australian golfer's wife ... , said it was a decision he did not take lightly. Related: Australian golfer Adam Scott says no to **Rio** ...

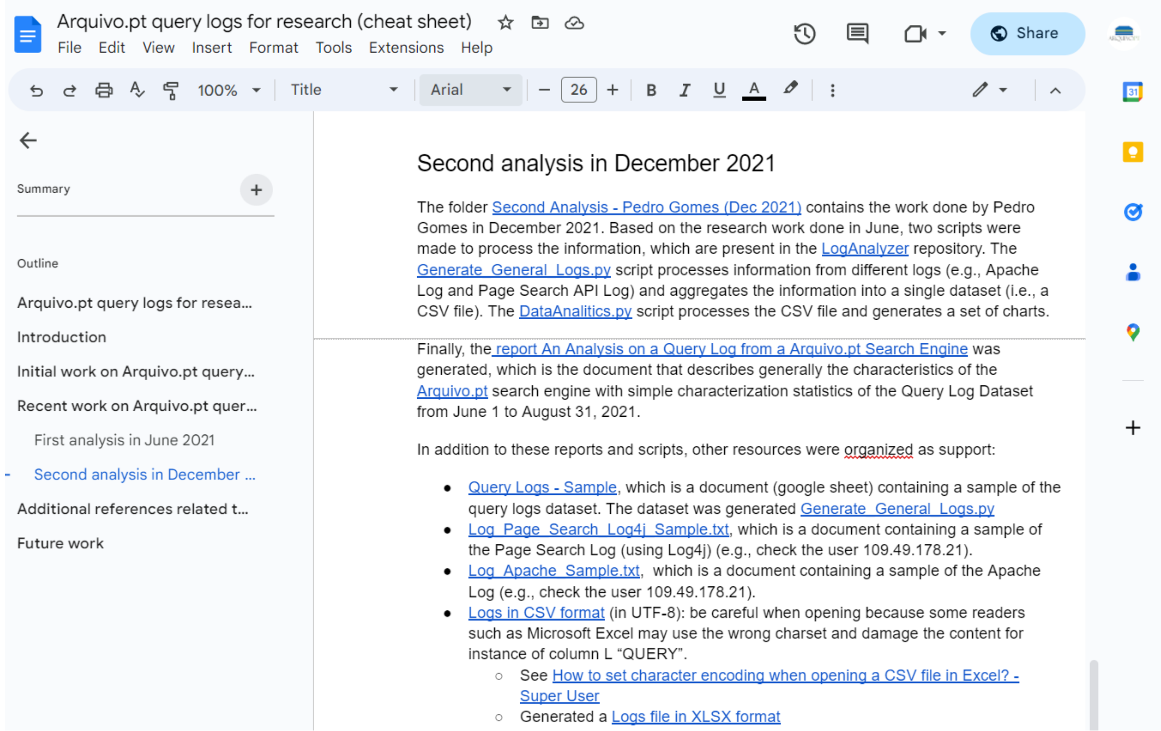
www.theguardian.com/sport/2016/apr/21/louis-oosthuizen-rio-south-africa-olympics

Louis Oosthuizen opts out of South Africa's **Rio Olympics** team | Sport | The Guardian

15 May 2016

Louis Oosthuizen opts out of South Africa's **Rio Olympics** team | Sport | The Guardian Close Skip to ... Olympic Games 2016 Louis Oosthuizen opts out of South Africa's **Rio Olympics** team • Oosthuizen joins Adam ... the Australian Olympic team the very best of luck in **Rio**." Golf's inclusion in the **Olympics** for the ...

Arquivo.pt search query logs available for research ***new!***


Arquivo.pt query logs for research (cheat sheet)
File Edit View Insert Format Tools Extensions Help
Share
100%
Title
Arial
26
Summary
Outline

Arquivo.pt query logs for resea...
Introduction
Initial work on Arquivo.pt query...
Recent work on Arquivo.pt quer...
First analysis in June 2021
Second analysis in December ...
Additional references related t...
Future work

Second analysis in December 2021

The folder Second Analysis - Pedro Gomes (Dec 2021) contains the work done by Pedro Gomes in December 2021. Based on the research work done in June, two scripts were made to process the information, which are present in the LogAnalyzer repository. The Generate_General_Logs.py script processes information from different logs (e.g., Apache Log and Page Search API Log) and aggregates the information into a single dataset (i.e., a CSV file). The DataAnalitics.py script processes the CSV file and generates a set of charts.

Finally, the report An Analysis on a Query Log from a Arquivo.pt Search Engine was generated, which is the document that describes generally the characteristics of the Arquivo.pt search engine with simple characterization statistics of the Query Log Dataset from June 1 to August 31, 2021.

In addition to these reports and scripts, other resources were organized as support:

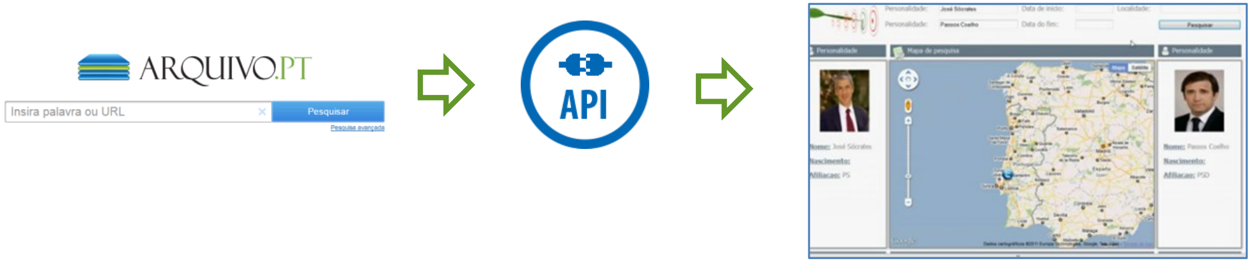
- Query Logs - Sample, which is a document (google sheet) containing a sample of the query logs dataset. The dataset was generated Generate_General_Logs.py
- Log_Page_Search_Log4j_Sample.txt, which is a document containing a sample of the Page Search Log (using Log4j) (e.g., check the user 109.49.178.21).
- Log_Apache_Sample.txt, which is a document containing a sample of the Apache Log (e.g., check the user 109.49.178.21).
- Logs in CSV format (in UTF-8): be careful when opening because some readers such as Microsoft Excel may use the wrong charset and damage the content for instance of column L "QUERY".
  - See How to set character encoding when opening a CSV file in Excel? - Super User
  - Generated a Logs file in XLSX format

- Unique resource for innovative research
- Information retrieval over historical web data is unexplored ground
- Only web archive that supports full-text search
- Only web archive that supports image search

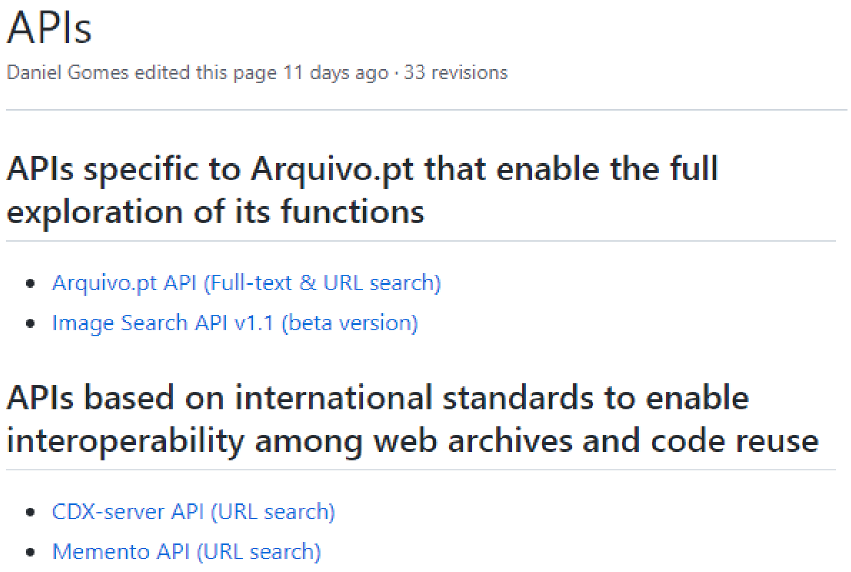
APIs and Bulk download of web-archived content ***new!***

Four **A**pplication **P**rogramming **I**nterfaces 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

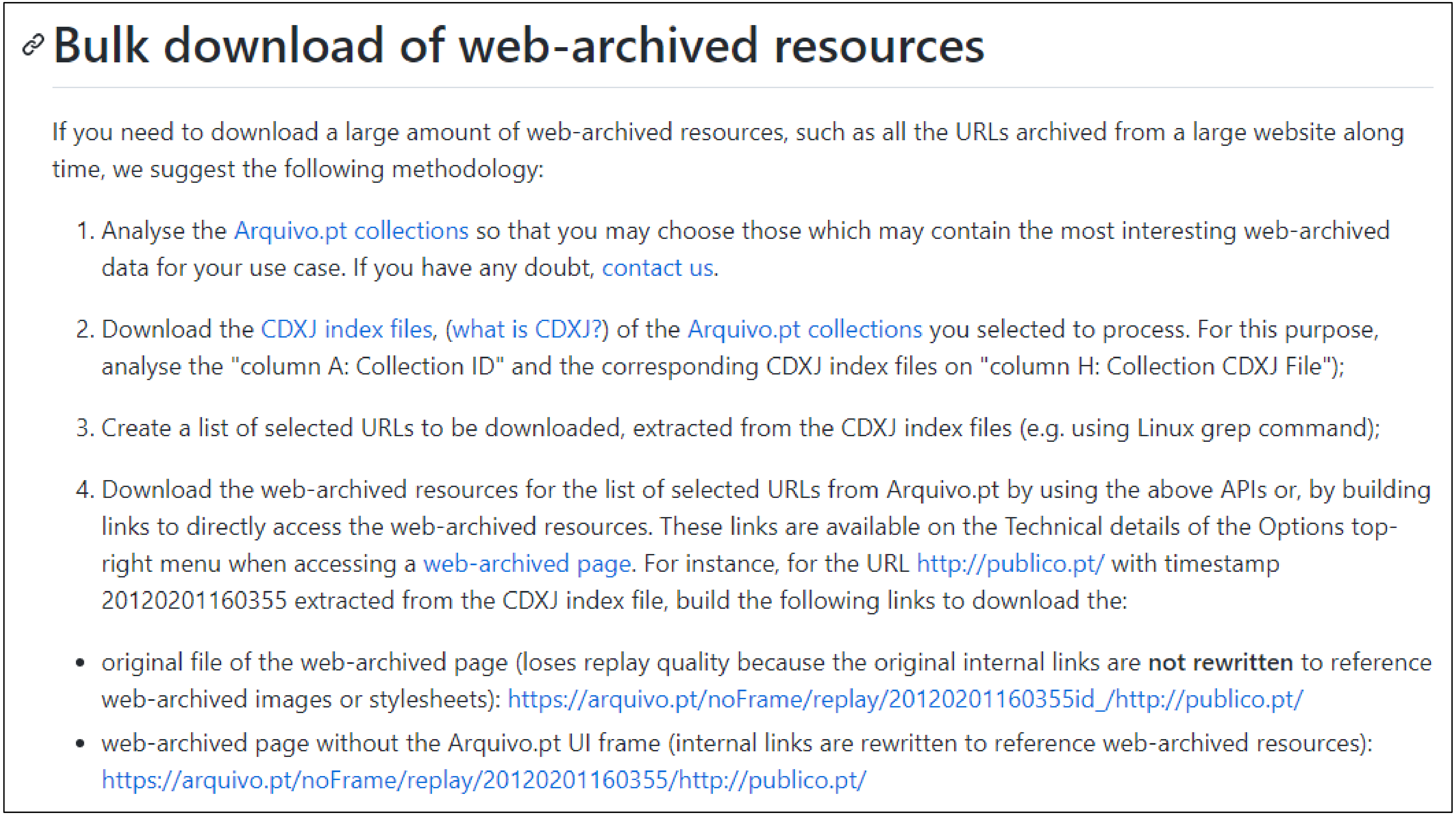
APIs specific to Arquivo.pt that enable the full exploration of its functions

- 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

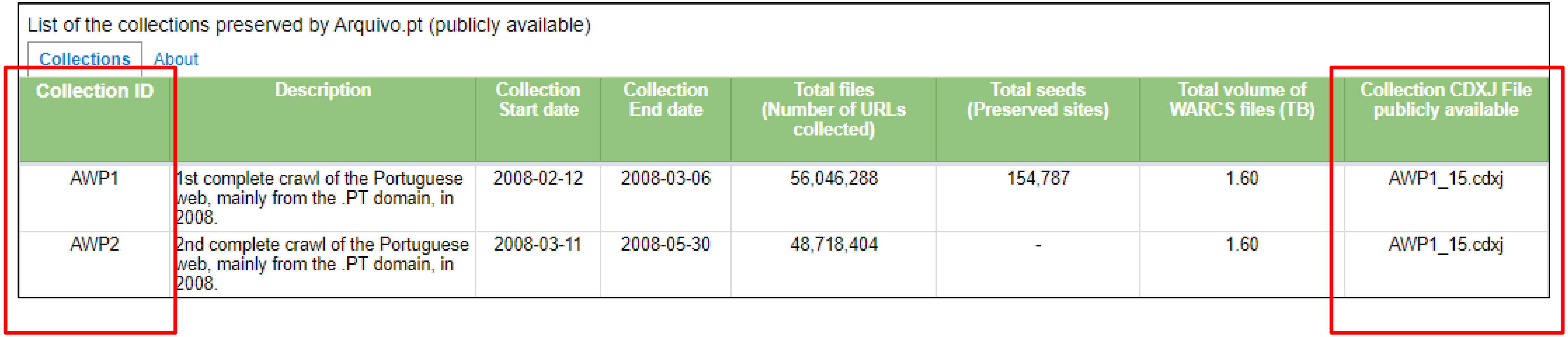
If you need to download a large amount of web-archived resources, such as all the URLs archived from a large website along time, we suggest the following methodology:

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

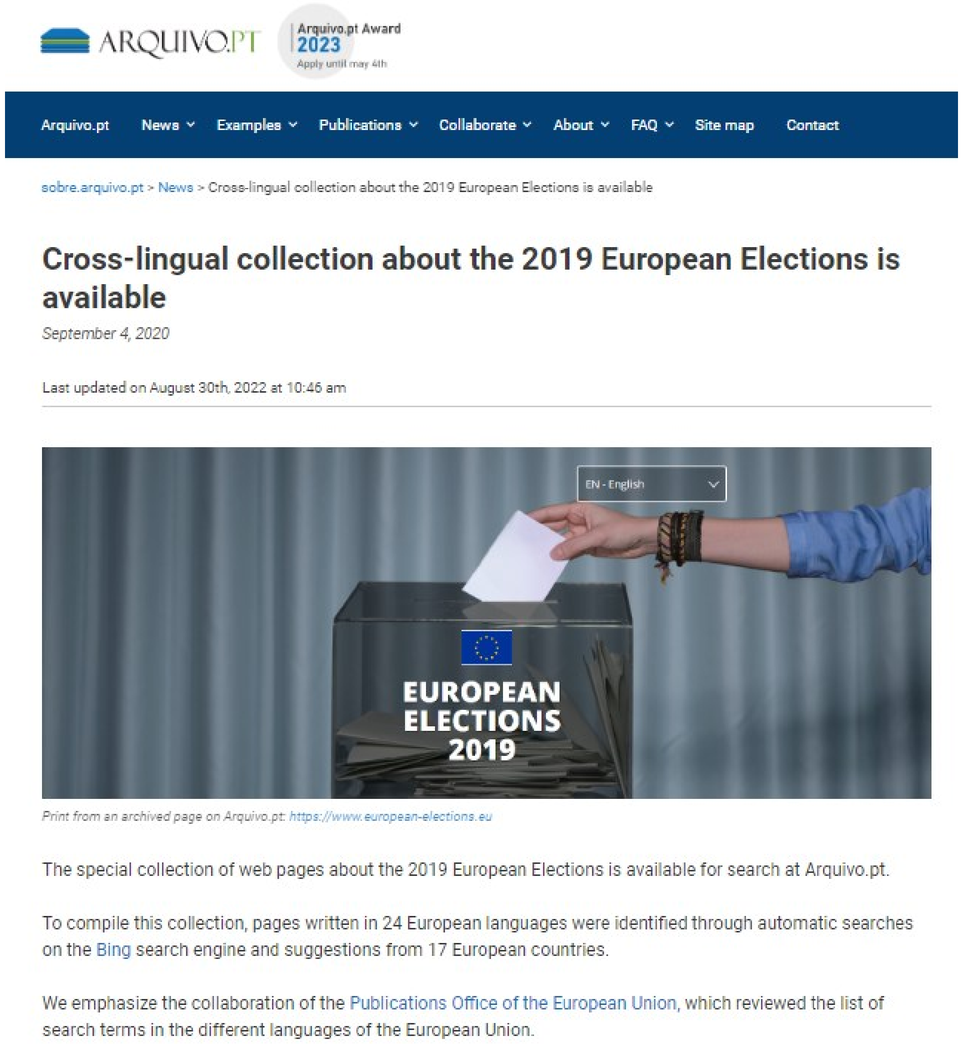
1. Analyze the Arquivo.pt collections to choose the most relevant



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

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 CDXJ File publicly available
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	AWP1_15.cdxj
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	AWP1_15.cdxj

Cross-lingual collection on 2019 EU celections (EAWP23)



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

Last updated on August 30th, 2022 at 10:46 am

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.

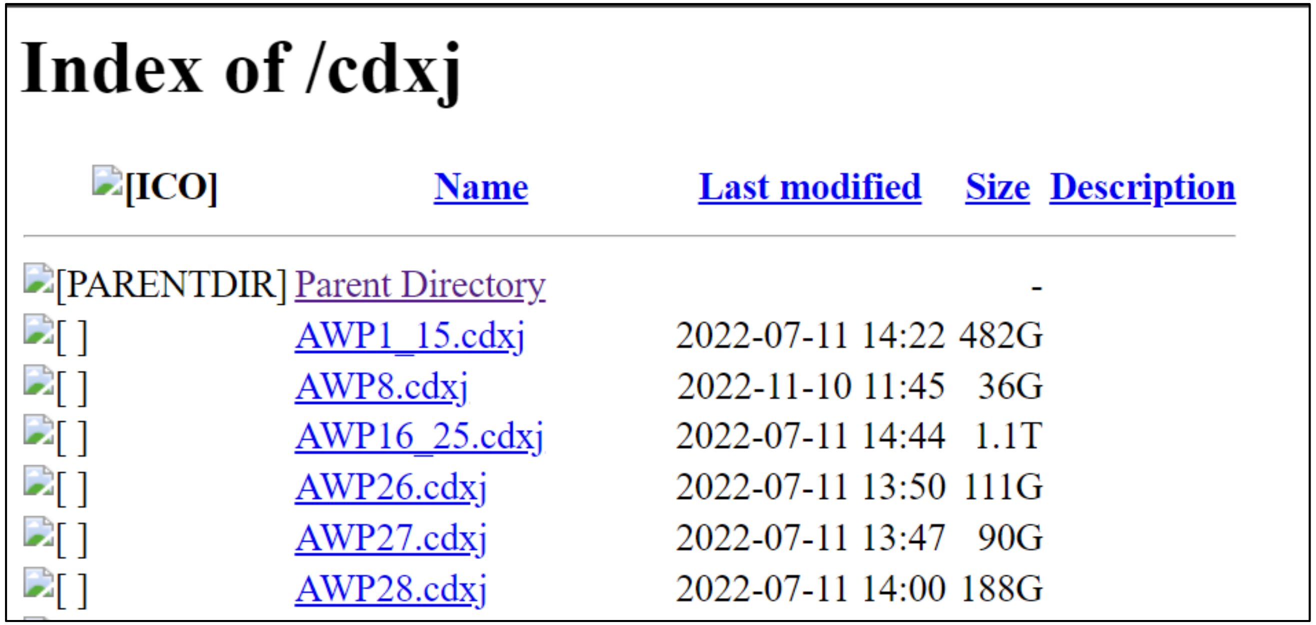
To compile this collection, pages written in 24 European languages were identified through automatic searches on the Bing search engine and suggestions from 17 European countries.

We emphasize the collaboration of the Publications Office of the European Union, which reviewed the list of search terms in the different languages of the European Union.

Created as a use case for cross lingual research

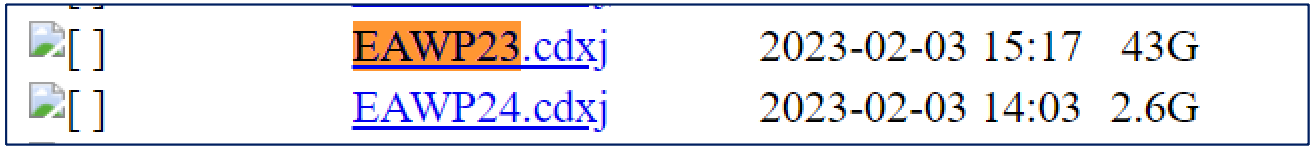
- Useful for under-resourced languages
- Parallel corpora because EU official websites are carefully translated

2. Download the correspondent CDXJ index files



Index of /cdxj

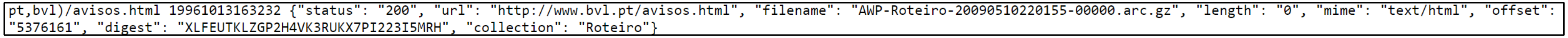
[ICO]	Name	Last modified	Size	Description
[PARENTDIR]	Parent Directory		-	
[ ]	AWP1_15.cdxj	2022-07-11 14:22	482G	
[ ]	AWP8.cdxj	2022-11-10 11:45	36G	
[ ]	AWP16_25.cdxj	2022-07-11 14:44	1.1T	
[ ]	AWP26.cdxj	2022-07-11 13:50	111G	
[ ]	AWP27.cdxj	2022-07-11 13:47	90G	
[ ]	AWP28.cdxj	2022-07-11 14:00	188G	



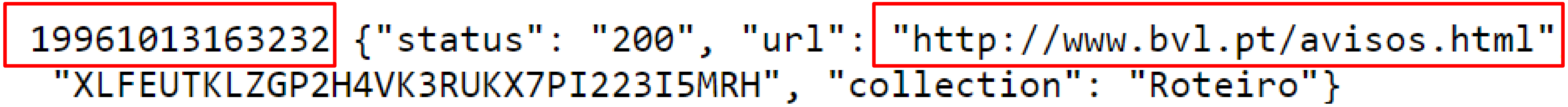
[ ]	EAWP23.cdxj	2023-02-03 15:17	43G
[ ]	EAWP24.cdxj	2023-02-03 14:03	2.6G

https://arquivo.pt/cdxj

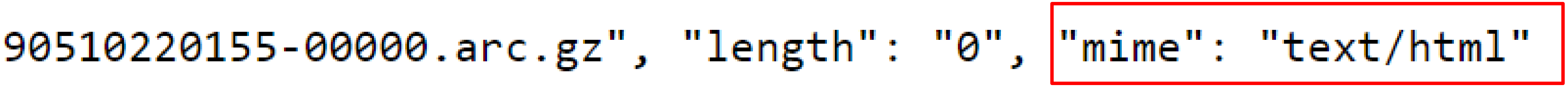
3. Select URLs to be downloaded from the CDXJ index files



```
pt,bvl)/avisos.html 19961013163232 {"status": "200", "url": "http://www.bvl.pt/avisos.html", "filename": "AWP-Roteiro-20090510220155-00000.arc.gz", "length": "0", "mime": "text/html", "offset":
"5376161", "digest": "XLFEUTKLZGP2H4VK3RUKX7PI223I5MRH", "collection": "Roteiro"}
```

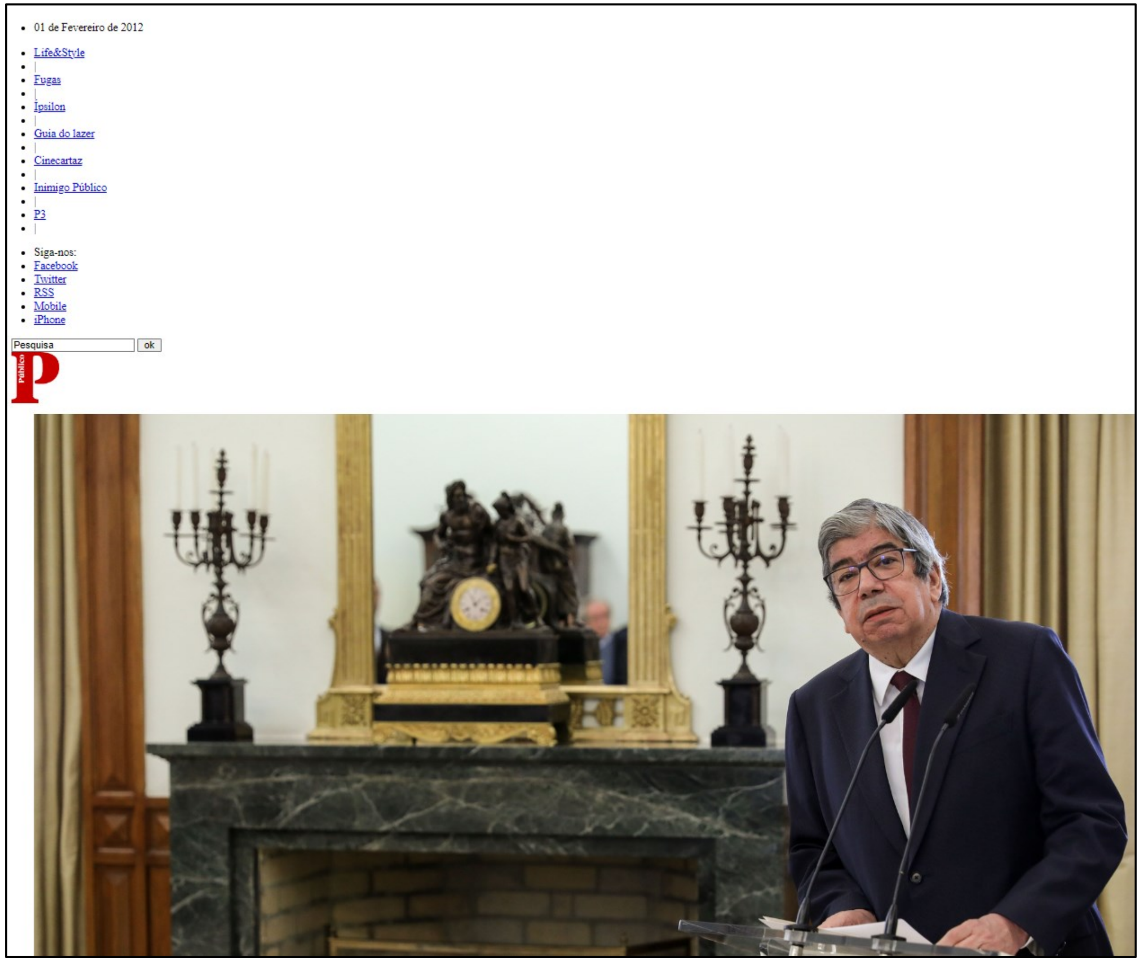


```
 19961013163232 {"status": "200", "url": "http://www.bvl.pt/avisos.html"
  "XLFEUTKLZGP2H4VK3RUKX7PI223I5MRH", "collection": "Roteiro"}
```



```
90510220155-00000.arc.gz", "length": "0", "mime": "text/html"
```

4. Download original file of the web-archived page



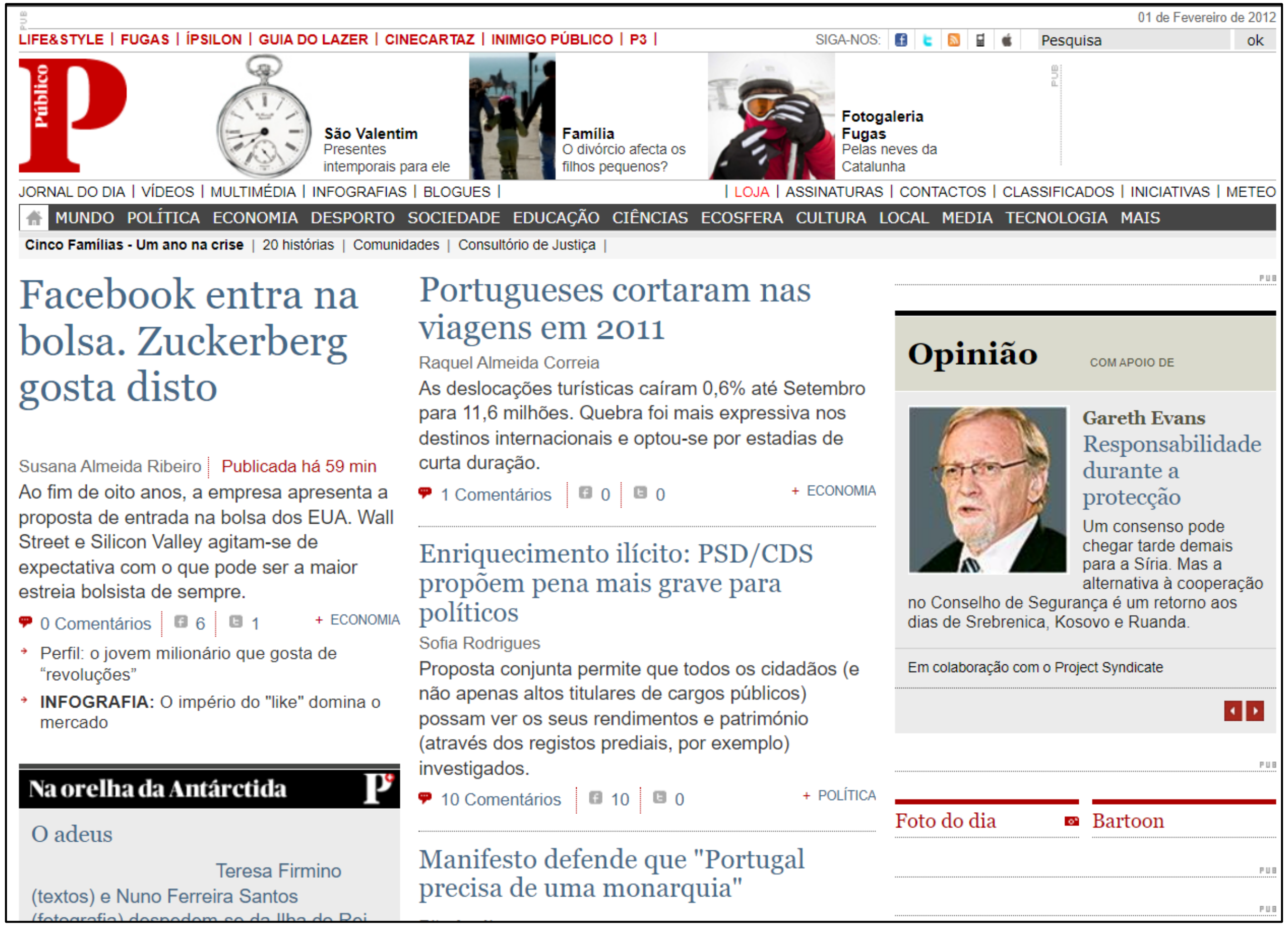
- 01 de Fevereiro de 2012

- Life&Style
- |
- Fugas
- |
- Ípsilon
- |
- Guia do lazer
- |
- Cinecartaz
- |
- Inimigo Público
- |
- P3
- |

- Siga-nos:
- Facebook
- Twitter
- RSS
- Mobile
- iPhone

https://arquivo.pt/noFrame/replay/**20120201160355**id_/**http://publico.pt/**

4. Download web-archived page without the Arquivo.pt UI frame



01 de Fevereiro de 2012

LIFE&STYLE | FUGAS | ÍPSILON | GUIA DO LAZER | CINECARTAZ | INIMIGO PÚBLICO | P3 |

SIGA-NOS: Pesquisa ok

São Valentim
Presentes intemporais para ele

Família
O divórcio afecta os filhos pequenos?

Fotogaleria Fugas
Pelas neves da Catalunha

JORNAL DO DIA | VÍDEOS | MULTIMÉDIA | INFOGRAFIAS | BLOGUES | | LOJA | ASSINATURAS | CONTACTOS | CLASSIFICADOS | INICIATIVAS | METEO

MUNDO POLÍTICA ECONOMIA DESPORTO SOCIEDADE EDUCAÇÃO CIÊNCIAS ECOSFERA CULTURA LOCAL MEDIA TECNOLOGIA MAIS

Cinco Famílias - Um ano na crise | 20 histórias | Comunidades | Consultório de Justiça |

Facebook entra na bolsa. Zuckerberg gosta disto

Susana Almeida Ribeiro | Publicada há 59 min

Ao fim de oito anos, a empresa apresenta a proposta de entrada na bolsa dos EUA. Wall Street e Silicon Valley agitam-se de expectativa com o que pode ser a maior estreia bolsista de sempre.

0 Comentários | 6 | 1 + ECONOMIA

- Perfil: o jovem milionário que gosta de "revoluções"
- **INFOGRAFIA:** O império do "like" domina o mercado

Na orelha da Antárctida

O adeus

Teresa Firmino (textos) e Nuno Ferreira Santos (fotografia) despedem-se da Ilha do Rei

Portugueses cortaram nas viagens em 2011

Raquel Almeida Correia

As deslocações turísticas caíram 0,6% até Setembro para 11,6 milhões. Quebra foi mais expressiva nos destinos internacionais e optou-se por estadias de curta duração.

1 Comentários | 0 | 0 + ECONOMIA

Enriquecimento ilícito: PSD/CDS propõem pena mais grave para políticos

Sofia Rodrigues

Proposta conjunta permite que todos os cidadãos (e não apenas altos titulares de cargos públicos) possam ver os seus rendimentos e património (através dos registos prediais, por exemplo) investigados.

10 Comentários | 10 | 0 + POLÍTICA

Manifesto defende que "Portugal precisa de uma monarquia"

Opinião COM APOIO DE

Gareth Evans
Responsabilidade durante a protecção

Um consenso pode chegar tarde demais para a Síria. Mas a alternativa à cooperação no Conselho de Segurança é um retorno aos dias de Srebrenica, Kosovo e Ruanda.

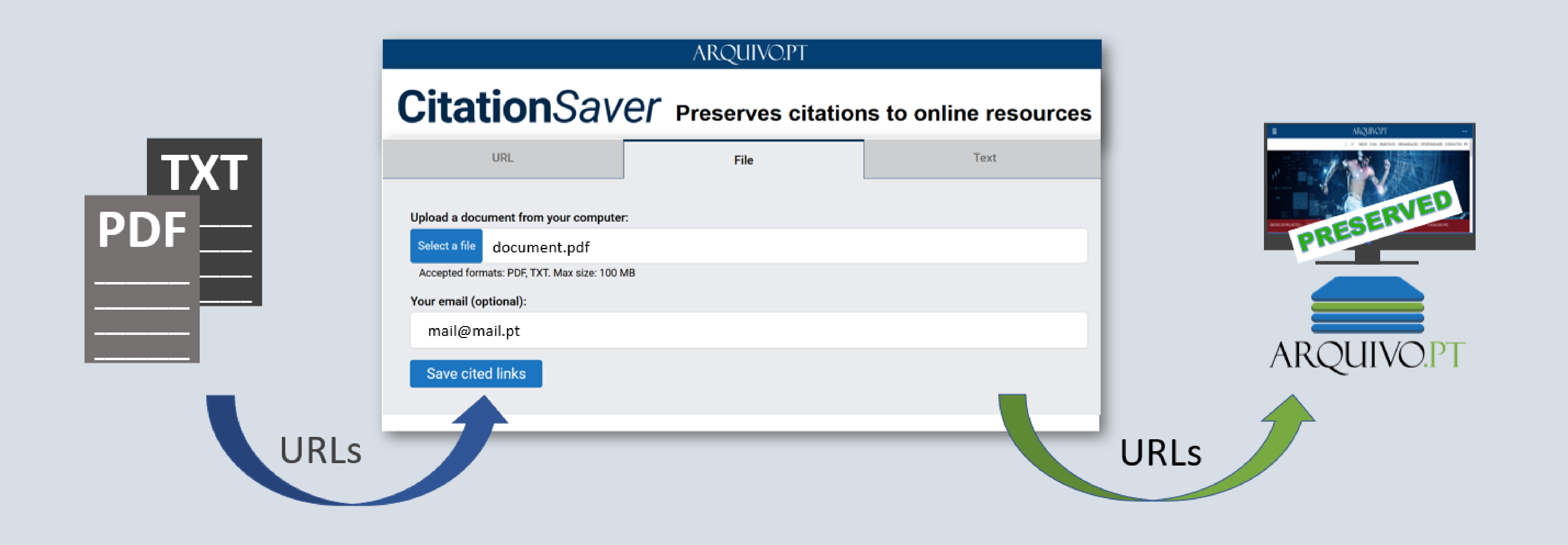
Em colaboração com o Project Syndicate

Foto do dia | Bartoon

https://arquivo.pt/noFrame/replay/**20120201160355/http://publico.pt/**

Tools to support the preservation of research data available online

CitationSaver preserves citations to web resources ***new!***



ARQUIVO.PT

CitationSaver Preserves citations to online resources

URL | File | Text

Upload a document from your computer:

Select a file document.pdf

Accepted formats: PDF, TXT. Max size: 100 MB

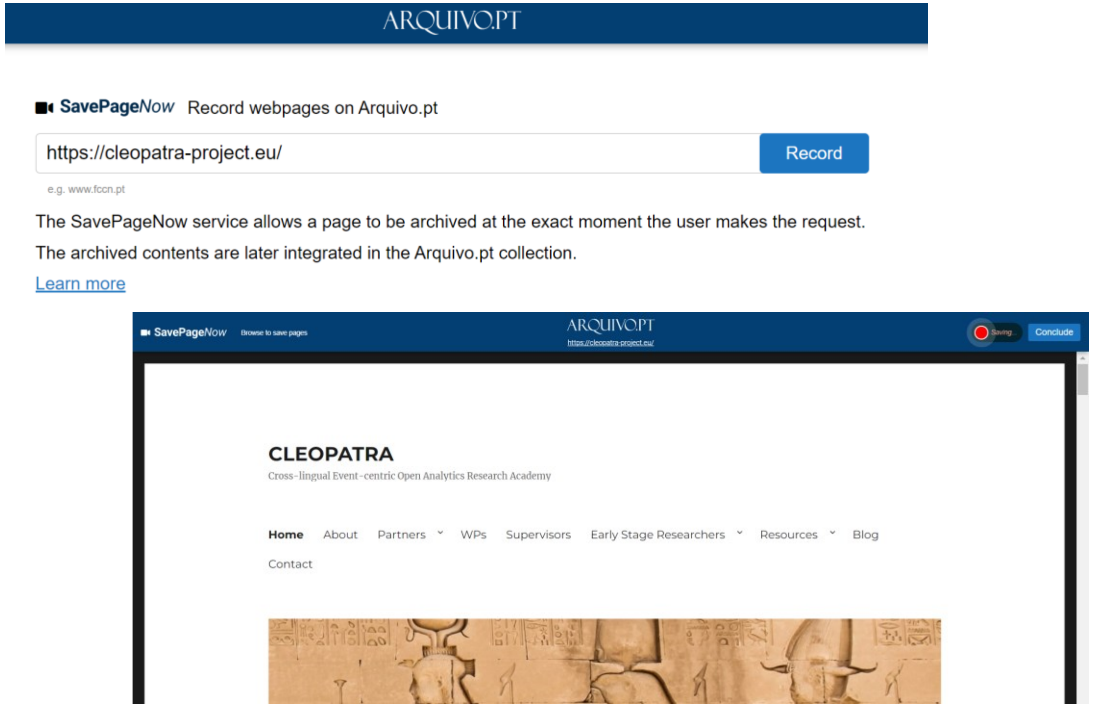
Your email (optional):

mail@mail.pt

Save cited links

arquivo.pt/citationsaver

SavePageNow to reference later! ***new!***



SavePageNow Record webpages on Arquivo.pt

https://cleopatra-project.eu/

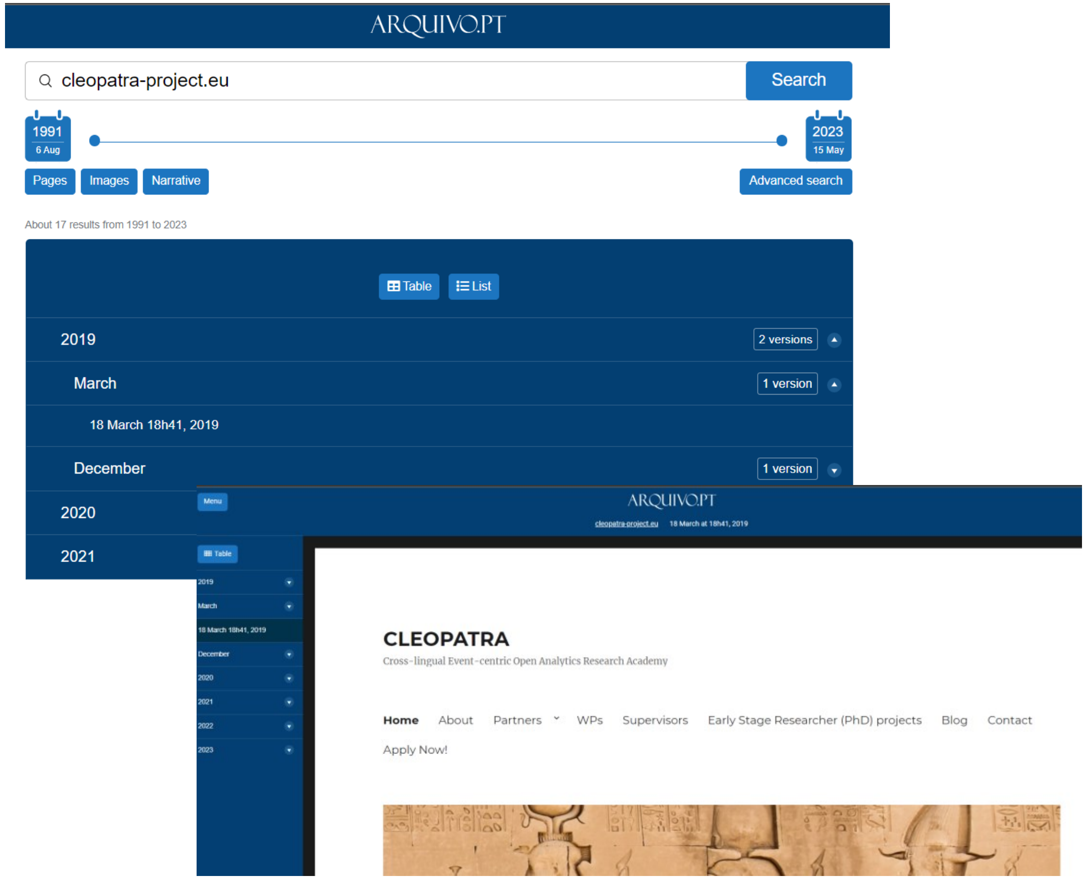
e.g. www.fccn.pt

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

https://cleopatra-project.eu/

CLEOPATRA
Cross-lingual Event-centric Open Analytics Research Academy



cleopatra-project.eu

1991 6 Aug — 2023 15 May

About 17 results from 1991 to 2023

2019 — 2 versions
March — 1 version
18 March 18h41, 2019
December — 1 version
2020
2021

cleopatra-project.eu 18 March at 18h41, 2019

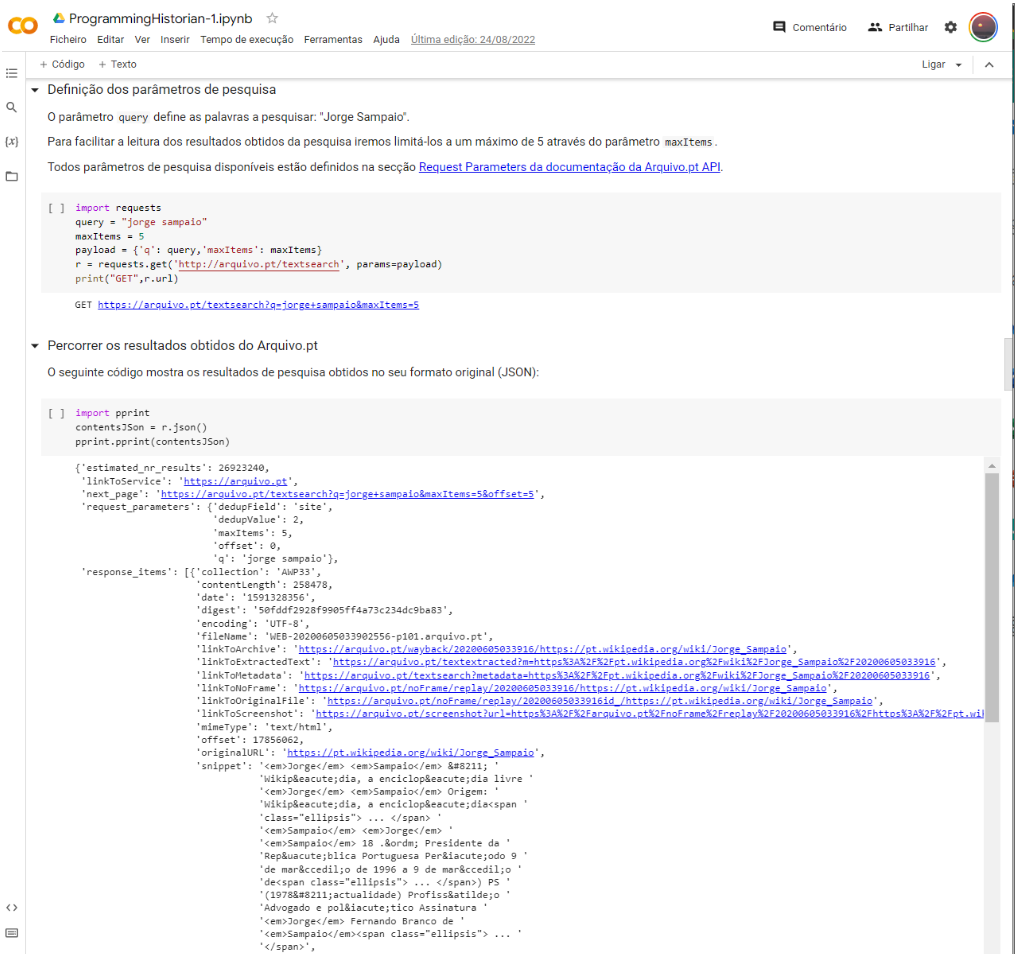
2019
March
18 March 18h41, 2019
December
2020
2021
2022
2023

CLEOPATRA
Cross-lingual Event-centric Open Analytics Research Academy

arquivo.pt/savepagenow

Learn more

Tutorial: how to explore Arquivo.pt using Python ***new!***



Definição dos parâmetros de pesquisa

O parâmetro `query` define as palavras a pesquisar: "Jorge Sampaio".

Para facilitar a leitura dos resultados obtidos da pesquisa iremos limitá-los a um máximo de 5 através do parâmetro `maxItems`.

Todos parâmetros de pesquisa disponíveis estão definidos na secção Request Parameters da documentação da Arquivo.pt API.

```
[ ] import requests
    query = "jorge sampaio"
    maxItems = 5
    payload = {'q': query,'maxItems': maxItems}
    r = requests.get('http://arquivo.pt/textsearch', params=payload)
    print("GET",r.url)
```

```
GET https://arquivo.pt/textsearch?q=jorge+sampaio&maxItems=5
```

Percorrer os resultados obtidos do Arquivo.pt

O seguinte código mostra os resultados de pesquisa obtidos no seu formato original (JSON):

```
[ ] import pprint
    contentsJSon = r.json()
    pprint.pprint(contentsJSon)
```

```
{'estimated_nr_results': 26923240,
 'linkToService': 'https://arquivo.pt',
 'next_page': 'https://arquivo.pt/textsearch?q=jorge+sampaio&maxItems=5&offset=5',
 'request_parameters': {'dedupField': 'site',
                        'dedupValue': 2,
                        'maxItems': 5,
                        'offset': 0,
                        'q': 'jorge sampaio'},
 'response_items': [{'collection': 'AWP33',
                     'contentLength': 258478,
                     'date': '1591328356',
                     'digest': '50fddf2928f9905ff4a73c234dc9ba83',
                     'encoding': 'UTF-8',
                     'fileName': 'WEB-20200605033902556-p101.arquivo.pt',
                     'linkToArchive': 'https://arquivo.pt/wayback/20200605033916/https://pt.wikipedia.org/wiki/Jorge_Sampaio',
                     'linkToExtractedText': 'https://arquivo.pt/textextracted?m=https%3A%2F%2Fpt.wikipedia.org%2Fwiki%2FJorge_Sampaio%2F20200605033916',
                     'linkToMetadata': 'https://arquivo.pt/textsearch?metadata=https%3A%2F%2Fpt.wikipedia.org%2Fwiki%2FJorge_Sampaio%2F20200605033916',
                     'linkToNoFrame': 'https://arquivo.pt/noFrame/replay/20200605033916/https://pt.wikipedia.org/wiki/Jorge_Sampaio',
                     'linkToOriginalFile': 'https://arquivo.pt/noFrame/replay/20200605033916id_/https://pt.wikipedia.org/wiki/Jorge_Sampaio',
                     'linkToScreenshot': 'https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20200605033916%2Fhttps%3A%2F%2Fpt.wi
                     'mimeType': 'text/html',
                     'offset': 17856062,
                     'originalURL': 'https://pt.wikipedia.org/wiki/Jorge_Sampaio',
                     'snippet': '<em>Jorge</em> <em>Sampaio</em> &#8211; '
                                'Wikip&eacute;dia, a enciclop&eacute;dia livre '
                                '<em>Jorge</em> <em>Sampaio</em> Origem: '
                                'Wikip&eacute;dia, a enciclop&eacute;dia<span '
                                'class="ellipsis"> ... </span> '
                                '<em>Sampaio</em> <em>Jorge</em> '
                                '<em>Sampaio</em> 18 .&ordm; Presidente da '
                                'Rep&uacute;blica Portuguesa Per&iacute;odo 9 '
                                'de mar&ccedil;o de 1996 a 9 de mar&ccedil;o '
                                'de<span class="ellipsis"> ... </span>) PS '
                                '(1978&#8211;actualidade) Profiss&atilde;o '
                                'Advogado e pol&iacute;tico Assinatura '
                                '<em>Jorge</em> Fernando Branco de '
                                '<em>Sampaio</em><span class="ellipsis"> ... '
                                '</span>',
```

arquivo.pt/tutorial

Suggested publications

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

➢ arquivo.pt/publications

Arquivo.pt **Annual** Awards

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

Support of the President of the Republic of Portugal

- **arquivo.pt/awards**

Keep in touch!

Subscribe our channels:

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