A transnational crawl of the European Parliamentary Elections 2019

Technical report, Arquivo.pt
August 2020

Ivo Branco, Ricardo Basílio, Daniel Gomes
contacto@arquivo.pt

1 Summary

The European Parliamentary Elections are an event of international relevance. The strategy adopted to preserve the World Wide Web has been delegating to national institutions the responsibility of selecting and preserving information relevant to their hosting countries. However, the preservation of web pages that document transnational events is not officially assigned.

Arquivo.pt – the Portuguese web-archive is a research infrastructure that preserves historical web content. Arquivo.pt permanently selects, archives and provides public access to its web collections.

This document describes an experiment conducted by the Arquivo.pt team aimed at preserving web content that documents the European Parliamentary Elections of 2019 by applying a combination of human and automatic selection processes.

The web is a rich and enormous source of varied information of research. Thus, selecting samples of relevant web-data to be studied is mandatory. The presented selection methodology combines human expertise with automation to maximize coverage about cross-lingual events but requires a very limited amount of resources. Thus, we believe that it can be applied to easily select and generate highly relevant samples of web-data about any kind of transnational events.

2 Semi-automatic selection of relevant online content

We identified relevant terms in Portuguese about the 2019 European Parliamentary Elections, automatically translated them to 24 official languages of the European Union, reviewed the translation in collaboration with the Publications Office of the European Union and then automatically queried a web search engine to get a total of 12147 URLs to seed the crawls.

The automation of the selection process enabled the expansion of the information coverage about the event to multiple countries and languages without significantly increasing the amount of resources required.

In parallel, we launched a collaborative list to gather contributions of relevant seeds from the international community. This collaborative initiative was disseminated through Portuguese and international contacts, like Arquivo.pt social media or IIPC mailing lists. We received 608

contributions from 16 countries. Slovakia and Portugal were the countries that suggested the highest number of seeds (114).

3 Crawling web content

We iteratively ran 6 crawls using different configurations and crawling software (Heritrix 3.3.0, Brozzler and Browsertrix) to maximize the quality of the collected content. One crawl was executed before the elections and 5 afterwards. These crawls were performed between May and July of 2019 and resulted in the collection of 99 million URLs (4.8 TB).

This web-data was aggregated into one special collection and became searchable and accessible through Arquivo.pt in July 2020 on https://arquivo.pt/ee2019. Notice that this collection is also available for automatic processing through the Arquivo.pt API. Collaborations with researchers interested in studying this web collection are welcome.

4 Pre-election crawls

Before the elections we ran a crawling on Heritrix using seeds from a collaborative list and seeds from an automatic extraction of a search engine, more details below.

4.1 Collaborative list before elections

We created a collaborative list of web addresses before elections that was sent to our Portuguese and international contacts, like Arquivo.pt social media pages, Arquivo.pt email groups or IIPC mailing lists.

We closed a first list before the elections and created a new version of the list after the elections.

We have received multiple suggestions. The following table groups those contributions in categories.

Table 1: Contributions per categories before the European Elections 2019.

Type	**Rest of europe**	**Portuguese**
Generic links	-	19
Official sites	25	23
Blogs and opinion articles	17	2
Independent and/or commercial channels	7	2
Social media news	101	11
Satire	2	2
Social networks links	13	8

Political Parties and Associations	82	28
Candidates	95	-
EU institutions, agencies and bodies	30	-
European Elections 2019 News	17	-
Total	389	95

Some suggestions have been categorized by country, it was as optional information. On the next table we show the number of contributions per country.

Table 2: Contributions per country before the European Elections 2019.

Country	**Number of contributions**
Slovakia	114
Portugal	95
Hungary	67
Croatia	18
Greece	17
Norway	9
Ireland	2
Romania	2
Spain	2
Ukraine	1
Not mentioned	157
Total	484

4.2 Automatic search-engine retrieval of seeds

To increase the number of seeds we decided to run a script that uses the Microsoft Azure Cognitive Services Bing Web Search API (more documentation about this service on Microsoft documentation site).

We first made an experiment through a list of compound expressions in Portuguese, each one pointed to an aspect of elections (campaign, candidates, parties, political debates and trends, etc). We also added some expressions in other languages, as English, French,

Spanish, Italian and German, and we found pages from those countries closely related to European Elections.

The script has called the Azure Web Search API 84 times and retrieved the first 10 results for each search. We retrieved 1 557 search results (seeds).

Table 3: Number of keywords used on Bing.

	Rest of europe	**Portuguese**
Keywords (search criteria) used on Bing search engine	26	58

4.3 Pre-election 1st crawl

The crawl started with 2 290 unique seeds on the 22th May 2019. It ran for 5 days without any noticeable problem.

Heritrix 3 configuration:

- Max hops: 5
- Queue Budget: 5000
- alsoCheckVia: true

Heritrix generated report:

- Crawl name: Europeias2019
- Crawl status: Finished
- Duration time: 4 days 14 hours 33 minutes
- Total seeds crawled: 2 461
- Total seeds uncrawled: 12
- Total hosts visited: 156 106
- URIs processed: 10 962 805
- URI successes: 10 922 604
- URI failures: 38 544
- URI disregards: 1 657
- Novel URIs: 10 922 604
- Total crawled bytes: 833 354 621 329 (776 GiB)
- Novel crawled bytes: 833 354 621 329 (776 GiB)
- URIs/sec: 27.44
- KB/sec: 2 044

The following results of the crawl were obtained:

- Number of collected files: 8 903 551
- Number of WARC files (100 MB each file): 5 032
- Disk space: 546G

5 Post-elections crawls

We ran 5 crawls after the European elections.

5.1 Collaborative list post elections

New contributions from Europe were added to a new version of the collaborative list, opened after elections. Table 5 shows the origin country and number of contributions, e.g., we received 114 seed contributions from Slovakia. No additional contributions from Portugal were received in this 2nd round and values are the same as before the elections.

Tabel 4: Summary of contributions per categories after elections

Type	**Rest of europe**	**Portuguese**
Generic links		19
Official sites	48	23
Blogs and opinion articles	27	2
Independent and/or commercial channels	9	2
Social media news	135	11
Satire	3	2
Social networks links	15	8
Political Parties and Associations	197	28
Candidates	86	-
EU institutions, agencies and bodies	29	-
European Elections 2019 News	17	
Total	566	114

Tabel 5: Summary of contributions per country after elections

Country	**Number of contributions**
Slovakia	114
Portugal	114
France	78

Hungary	67
Czech Republic	65
Luxembourg	53
Greece	17
Norway	9
Netherlands	3
Denmark	3
Ireland	2
Romania	2
Finland	2
Spain	2
Ukraine	1
Not mentioned	91
Total	680

Merged all unique links received on the European election 2019 collaboration list about Portugal (114) and from Europe (566) sections, with other more links queried on Bing (2868) explained below.

5.2 Post elections 2nd crawl automatic translation + search engine

On the 2nd crawl we increased the approach that we have just experimented on the first crawl, by covering some european languages. Starting with 12 expressions in Portuguese related to European elections we translated them automatically to all the 24 official European languages. Using a google sheet and the GOOGLETRANSLATE formula.

Example: “=GOOGLETRANSLATE(A3, "pt", "en")”

This translates the A3 cell from portuguese to english. A similar formula to translate 12 different expressions to the 24 official European languages.

This gave us about 288 terms to search on Bing Web Search API. Using a python script we conducted the search and retrieved the first 10 results for each search. This gave us about 2869 results.

Given those, 2869 seeds were retrieved through an automatic translation and search engine retrieval, plus the seeds of the collaborative list of post elections.

Started on 6 of June of 2019 a second crawl after the European Elections with 2696 unique seeds. It ran for 5 days.

Heritrix 3 configuration:

- Max hops: 5
- Queue Budget: 5000
- alsoCheckVia: true

Heritrix generated report:

- Crawl name: PosEuropeias2019
- Crawl status: Finished - Ended by operator
- Duration: 4 days 19 hours 49 minutes
- Seeds crawled: 2 917
- Seeds uncrawled: 37
- Hosts visited: 274 101
- URIs processed: 20 118 198
- URI successes: 20 043 907
- URI failures: 71 407
- URI disregards: 2 884
- Novel URIs: 20 043 907
- Total crawled bytes: 1 626 536 125 090 (1.5 TiB)
- Novel crawled bytes: 1 626 536 125 090 (1.5 TiB)
- URIs/sec: 48.07
- KB/sec: 3809

The following results of the crawl were obtained:

- Number of collected files: 16 538 938
- Number of WARC files (100 MB each file): 9 564
- Disk space: 1019 GB

5.3 Post elections 3rd crawl manual revision of the translations+search engine

On 14th June we received from the Publications Office of the European Union a manual revision of the automatic translations. We got a more accurate list of expressions to search since 12 expressions were corrected.

Because the 2nd crawl finished 3 days before, we decided to crawl only the 12 revisioned expressions of the 24 official European languages. We searched on Bing and extracted the results URLs. On those we removed the links that had already been crawled. This gave 1 830 unique new links.

This crawl started on June 18 and terminated on 27th June.

Heritrix version 3.3.0 configuration:

- Max hops: 5
- Queue Budget: 5000
- alsoCheckVia: true

Heritrix generated report:

- Crawl name: PosEuropeias2019V2
- Crawl status: Finished
- Duration: 8 days 9 hours 55 minutes
- Seeds crawled: 1 865
- Seeds uncrawled: 33
- Hosts visited: 1 947 825
- URIs processed: 79 473 392
- URI successes: 79 256 499
- URI failures: 206 580
- URI disregards: 10 313
- Novel URIs: 79 256 499
- Total crawled bytes: 4 254 584 635 358 (3.9 TiB)
- novel crawled bytes: 4 254 584 635 358 (3.9 TiB)
- URIs/sec: 109.03
- KB/sec: 5715

The following results of the crawl were obtained:

- Number of collected files: 68 588 799
- Number of WARC files (100 MB each file): 28 849
- Disk space: 2.8TB

After this crawl we analysed the small seeds and the dimension of the crawled information. We concluded that our Heritrix configuration was too broad. Thus, we decided to disable the *alsoCheckVia* crawl parameter to make the Heritrix crawls more compact and focused on its seeds.

5.4 Review post-elections crawl

On 15th July, after reviewing what has been done. We decided to run again the post elections crawling to increase coverage. The mains decisions were:

1. increase the number from 12 to 40 search terms per language, and
2. run a high quality crawl for each seed.

40 keywords x 24 languages x use first 10 results = 9 600 expected search results

By increasing the number of keywords to be translated, we tried to crawl more sites about the elections. We detected that the search engine results started to return more duplicated results. This was an expected result because of similarities between search terms among different languages (eg. Spanish and Portuguese, names of candidates).

Additionally some of the suggestions submitted to the collaborative list included some pages retrieved on the search engine results. We crawled around 7 171 unique seeds on the last 3 crawls.

After analysing the seeds, we detected a relevant number of Youtube links. There is a current incompatibility of captured videos on Brozzler and the pyWB replay software. This situation occurs because there isn't a standard way to capture and replay videos on this kind of platform. To fix this problem we decided to crawl Youtube seeds on Browsertrix, this crawling software is compatible with pywb replay software.

Table 6: Seeds per crawling software

Link type	**Seeds**	**Crawling software**
Youtube links	504	Browsertrix
Remaining links	6667	Brozzler and Heritrix
Total	**7171**	

5.5 Post elections 4th crawl (high quality)

To increase the quality of the replay of this Arquivo.pt collection we decided to also use Brozzler software. We run it using a single page configuration on each seed. This crawl resulted in about 69 GB of warc files.

5.6 Post elections 5th crawl videos

The 504 Youtube links were crawled on Browsertrix resulting in 30GB of WARCs.

5.7 Post elections 6th crawl

Heritrix configuration:

- Max hops: 5
- Queue Budget: 5000
- alsoCheckVia: false

Heritrix generated report:

- Crawl name: PosEuropeias2019V2
- crawl status: Finished
- duration: 6 days 22 hours 28 minutes
- seeds crawled: 6 946
- seeds uncrawled: 122
- hosts visited: 19 893
- URIs processed: 6 245 644
- URI successes: 6 241 275
- URI failures: 4 253

- URI disregards: 116
- novel URIs: 6 241 275
- total crawled bytes: 660 010 718 270 (615 GiB)
- novel crawled bytes: 660 010 718 270 (615 GiB)
- URIs/sec: 10.41
- KB/sec: 1075

The following results of the crawl were obtained:

- Number of collected files: 4 743 811
- Number of WARC files (100 MB each file): 3 428
- Disk space: 383 GB

6 Conclusions

This report documents the Arquivo.pt crawl of the European parliamentary elections of 2019.

Collaborative lists, before and after European Elections were important because they involved the community to be part of the Web preservation process. Contributions by adding seeds and reviewing search terms from our European colleagues engaged on Web archives and the European Union Publications Office were most valuable.

The success of the automation process of translation and retrieval was a nice surprise. Starting with 40 terms in Portuguese we found similar links on all the 24 official European languages. Each term was translated, queried to a web search engine and the first 10 results for each query were retrieved.

This automation process increased significantly the number of preserved pages and the number of links and to ensure that some relevant pages are preserved from each country / language.

Multiple crawling technologies have been used to improve the quality of the captured pages and to fix the replay of videos, due to the lack of a standardized way to capture and replay videos.

As future work we intend to continue to apply the described semi-automatic process of seed selection to document transnational events relevant to the Portuguese society, by combining collaborative lists to engage the community, automatic translation and search-engine link retrieval.

This collection was identified as EAWP23 and became searchable through Arquivo.pt in July 2020, one year after its finish date. Collaborations with researchers interested in studying the collected web data sets or crawl logs are most welcome.

7 To know more

- Research European Elections (shortcut): https://arquivo.pt/ee2019
- Arquivo.pt API: https://arquivo.pt/api
- European Elections 2019: We Need Your Help! https://sobre.arquivo.pt/en/european-elections-2019-we-need-your-help/
- Collaborative list of Web addresses before the European Elections 2019: https://docs.google.com/spreadsheets/d/1GhU8VHLaSVSv5Vm7E-og3a3KYv2jEO22tMEicahSoNU/
- Collaborative list of Web addresses after the European Elections 2019: https://docs.google.com/spreadsheets/d/1vwp59-8E7wZ4xPbkh_95kprz2KzhYoPSFOv9dXvRi0Q/
- Bing Web Search API: https://azure.microsoft.com/en-us/services/cognitive-services/bing-web-search-api/
- Bing search script: https://github.com/arquivo/scripts/blob/master/bing_search.py
- Keywords and seeds before the European Elections 2019 - 1st experiment, April 29, 2019: https://docs.google.com/spreadsheets/d/1ddSJFen7W0_5qEeX70Vw2w3YJUdbistXYHUaL0kBwVA/
- Keywords and seeds after the European Elections 2019 - 2nd experiment in May 29: https://docs.google.com/spreadsheets/d/1WzAQ7gxGAdy-WZiDvUMxx6XLgrBomsQOA9y_c-WnaPA/
- Brozzler: https://github.com/internetarchive/brozzler
- Heritrix: https://github.com/internetarchive/heritrix3
- Crawl Logs of European Elections 2019 collection (22 GB): https://arquivo.pt/crawlreport/ee2019/