A transnational and cross-lingual crawl of the European Parliamentary Elections 2019

Ivo.Branco@fccn.pt

RicardoBasilio@fcsh.unl.pt

Daniel.Gomes@fccn.pt

Present Arquivo.pt

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.

Motivation

European Parliamentary Elections

- International relevance
- preserve the World Wide Web has been delegated to national institutions the responsibility of selecting and preserving information relevant to their hosting countries.
- preservation of web pages that document a transnational event is not officially assigned.

Collaboration

A collaborative initiative was disseminated through Portuguese and international contacts to request sites to be preserved.

- Arquivo.pt
  - mailing lists
  - social networks
- IIPC mailing lists

Result:

- 680 contributions
- From 16 countries

Country	# Contributions
Slovakia	114
Portugal	114
France	78
Other countries	226
Not mentioned	91
Total	**680**

Automatic translation

- Started with terms in Portuguese
- Automatically translated them to 24 official languages
  - Using a google sheet with *GOOGLETRANSLATE* formula like:
    - `=GOOGLETRANSLATE(A3, "pt", "en")`
  - Words in Portuguese on 1st column
  - Add more 23 columns each one per other language
- Publications Office of the European Union reviewed the translations

Automatic query the search engine

From each word (expression) ⇒ automatically queried a web search engine and use first 10 results for each search

- Python script ≅ 50 lines of code
- use Bing Web Search API

This process enable the expansion of the coverage of the information about the event to multiple countries and languages without significantly increasing the amount of resources.

Crawls

1. Pre-elections crawl using Heritrix
   - collaborative list
   - couple of sentences on a search engine
2. Post-elections automatic translation + query search engine
   - 12 terms per language on 24 languages
3. Manual revision of the translations + query search engine

Review post-elections crawl:

- increase the number from 12 to 40 search terms per language
- run a high quality crawl for each seed

Crawls

4. Brozzler ⇒ to increase quality
   - run a single page crawl on each seed
   - uses a real browser to better preserve CSS+JS
5. Browsertrix
   - for youtube videos
   - because lack compatibility on Brozzler crawler and pywb replay
6. Heritrix
   - final crawl

EE2019

arquivo.pt/ee2019

- full text page search
- image search
- page replay

Numbers:

- Total seeds/preserved sites: 12147
- Disk space: 4.8 TB
- From: 2019-05-22
- Until: 2019-07-17
- 99 millions collected files

arquivo.pt/api

- available for automatic processing

Collaborations with researchers interested in studying this web collection are welcome.

Conclusion

Using this method we have combined:

Human expertise + Automation

⇒ Maximize coverage of a cross-lingual event

⇒ using very limited amount of resources

This process can be applied easily to any kind of transnational event.

Coronavirus COVID-19 outbreak

- IIPC Collaborative Collection
- Didn’t use any automatic translation
- Automatically queried a web search engine
  - For each portuguese news site search “coronavirus site:news-site.pt“
  - use 10 first search results
  - Result on a spreadsheet with
    - query input
    - position 1-10
    - page URL
    - page title
  - Manual revision by web curator team member
    - remove some sites
    - by default single page crawl
    - other increase the depth to + 1 hop

Thank you. Questions?

More information

arquivo.pt/ee2019 > About this project