A first attempt to archive the .EU domain
Technical report

Daniel Bicho
daniel.bicho@fccn.pt

João Miranda
joao.miranda@fccn.pt

Daniel Gomes
daniel.gomes@fccn.pt

9 March 2015

1 Summary

The .EU domain is commonly used to reference sites related to Europe. The strategy adopted to archive the World Wide Web has been delegating the responsibility of each domain to the respective national archiving institutions. However, the .EU domain fails to fit in this model because it covers multiple nations. Thus, the preservation of .EU sites has not been yet assigned and undertaken by any institution.

The Portuguese Web Archive mainly preserves online content relevant to the Portuguese community. It periodically crawls information from the web and provides a public search service over the preserved information available at archive.pt. RESAW is an European network that aims to create a Research Infrastructure for the Study of Archived Web Materials (resaw.eu).

This report describes a first attempt to crawl and preserve web sites hosted under the .EU domain performed by the Portuguese Web Archive within the scope of RESAW activities. The objective was to gain insight about how to archive the .EU domain. We describe the applied methodology, crawling process and obtained results. We detail the problems found and adaptations required during the crawl. Finally, we estimated resources and adjustments required for the following crawl of the .EU domain. Technical knowledge about the Heritrix crawler is required to fully interpret the presented results.

This report is complemented with the following files:

Heritrix crawl log original crawl log generated by Heritrix available at `http://arquivo.pt/crawlreport/heritrix.crawlog.tar.gz`;

Heritrix reports Heritrix generated reports available at `http://arquivo.pt/crawlreport/heritrix.reports.tar.gz`;

NotebookHMTL: analysis sheet generated using the Notebook Python library available at `http://arquivo.pt/crawlreport/crawleu.html` to process the Hertrix report files.

This first crawl began on the 21 November 2014 and finished on the 16 December 2014. As future work we intend to perform 2 more crawls of the .EU domain. Each one of performed .EU crawls shall be indexed and become searchable through archive.pt one year after its finish date.

2 Methodology

2.1 Initial Seeds

A pool of seeds to the home pages of sites hosted under the .EU domain was gathered from the following sources:

- Google search query site:.EU = 555 URLS;
- Google search querys site:.EU sites list = 122 URLS;
- Domain Typer - `https://domaintyper.com` = 4 551 URLS;
- Google search querys site:.EU blogs = 119 URLS;
- Google search querys site:.EU news = 178 URLS;
- Google search querys site:.EU "european union" = 226 URLS;
- dmoz - `http://www.dmoz.org/docs/en/rdf.html` = 8 292 URLS;
- Crawl Logs, seeds .EU catched by other crawls = 2 264 URLS;
- Alexa Top Sites `http://www.alexa.com/topsites` = 4 582 URLS;
- HttpArchive `http://www.httparchive.org/downloads.php` = 20 244 URLS;
- WebSiteisworth `http://www.websiteisworth.com/domain-by-extension/eu&page=1` = 14 249 URLS.

The seeds from these sources were merged and cleaned, removing duplicates and malformed seeds. The crawl was launched using a total of 34 138 unique seeds.

2.2 Crawl Configuration

The Crawler used was *Heritrix* version 1.14.3 respecting the robots exclusion protocol rules imposed by the visited sites.

Configurations used at start of the crawl are below, for any values omitted, consider the *Heritrix* default configurations[1]:

- Toe threads: 200
- Scope: DecidingScope
  - RejectByDefault
  - SurfPrefixedDecideRule: true
  - TooManyHopsDecideRule: 5 hops
  - TransclusionDecideRule: 2 max-trans-hops
  - PathologicalPathDecideRule
  - TooManyPathSegmentsDecideRule: 10 max-path-depth
  - PrerequesiteAcceptDecideRule
- OnDomainsDecideRule: ACCEPT
- Number of max hops: 5
- BdbFrontier
  - min-delay-ms: 10 000
  - queue-total-budget: 10 000
  - cost-policy: UnitCostAssignmentPolicy
- Fetch-Processors midfetch-decide-rules
  - timeout-seconds: 300
  - max-length-bytes: 10 000 000
- Write-processor Archiver
  - pool-max-active: 400

[1] Default configuration - `http://sourceforge.net/p/archive-crawler/code/HEAD/tree/release-branches/Heritrix-1.14.3/Heritrix/src/conf/profiles/default/order.xml`

3 Crawling Process

The resources available to store this crawl were 9 TB of disk space. The crawl started on the 21 November 2014. It ran for 20 days without any noticeable problem. On the 11 December 2014, the crawl was paused because we identified the existence of spam sites that were overloading the Frontier with abnormally large amounts of URLs queued to be visited. These spam sites were poorly designed online shops (e.g. `autobazar.eu`), link farm sites (e.g. `in-links.eu`) or large number of sub-domains that referenced multilingual versions of the same site (e.g. `en.myface4u.eu`, `pt.myface4u.eu`, `dk.myface4u.eu`). Most of the identified spam sites were related to pornography. The main spam domains identified were:

- dbquanti.eu
- autobazar.eu
- in-links.eu
- myface4u.eu
- share-with.eu
- prace-jobs.eu
- cutegirls.eu
- bongacams.eu

We also detected an overloaded queue in the Frontier corresponding to a site that provided a web analytics service hosted under the .COM domain (fr.sitestat.com). The crawl was restricted to .EU domains. However, sites hosting content embedded on .EU pages or redirected from .EU domains were also crawled.

The following regular expression filters to reject the identified spam sites were applied to the *Heritrix* configuration.

Reject Match Pattern:

```
(^http://fr.sitestat.com/.*)|
(^http://[^\/]+\.dbquanti.EU/.*)|(^http://dbquanti.EU/.*)|
(^http://[^\/]+\.autobazar.EU/.*)|(^http://autobazar.EU/.*)|
(^http://[^\/]+\.in-links.EU/.*)|(^http://in-links.EU/.*)|
(^http://[^\/]+\.bongacams.EU/.*)|(^http://bongacams.EU/.*)|
(^http://[^\/]+\.myface4u.EU/.*)|(^http://myface4u.EU/.*)|
(^http://[^\/]+\.share-with.EU/.*)|(^http://share-with.EU/.*)|
(^http://[^\/]+\.prace-jobs.EU/.*)|(^http://prace-jobs.EU/.*)|
(^http://[^\/]+\.cutegirls.EU/.*)|(^http://cutegirls.EU/.*)
```

The state of the crawling versus the available resources were analyzed. The queue budget in the frontier was reduced from 10 000 to 50 URLS per site to ensure that the crawl could reach all the seeds and the crawl was resumed on the 12 December 2014.

- BdbFrontier
  - queue-total-budget: 50

On the 15 December 2014 the crawl was stuck at 99% harvesting the site `www.partymenu.eu`. The following changes at *Heritrix* configuration were made as an attempt to gracefully finish the crawl:

- delay-factor: 1.0 from 4.0
- max-retries: 10 from 30
- retry-delay-seconds: 20 from 900

However, these changes did not yield the desired effect. Even changing the parameter for the *delay-factor*, the time between each fetch stood the same (300 seconds). After further investigation we spotted that *Heritrix* could not successfully download some resources, reporting at the logs a timeout (timeTrunc). The *midfetch-decide-rule* timeout configuration was 300 secs. *Heritrix* was trying to download content of type text/html. Thus, 300 secs is a large timeout to download most files of this type of content. We concluded that it was malformed dynamically generated content and reduced the timeout to 10 seconds to finish the crawl. The crawl ended on the 16 December 2014.

4 Crawling Results

Heritrix generated report:

Crawl Name: EAWP6

Crawl Status: Finished

Duration Time: 23 days 16 hours and 22 minutes

Total Seeds Crawled: 51 164

Total Seeds not Crawled: 4 338

Total Hosts Crawled: 1 084 605

Total Documents Crawled: 250 163 776

Processed docs/sec: 80.27

Bandwidth in Kbytes/sec: 4 826

Total Raw Data Size in Bytes: 10 113 875 430 284 (9.2 TB)

Novel Bytes: 10 113 875 404 802 (9.2 TB)

Not-modified Bytes: 25 482 (25 KB)

At the end of the crawl, the number of seeds was 51 164, from an initial pool of 34 138 seeds. 17 026 new seed URLs were added by *Heritrix* due to redirects. Cleaning the logs file from the identified spam, 135 907 unique domain URLs were extracted and can be used as seeds on the following crawl of the .EU domain.

The following results for the crawl were obtained:

- 5.8 TB disk space was used to store the crawled content using the compressed ARC format;
- 3 255 redirects were identified to other top level domains;
- Average size per document was 230 KB;
- Average site size was 9 MB;
- Average number of URLs per site was 337 URLs.
- 173 482 212 URLs remained to be crawled;

5 Conclusions

This report documents our first attempt to crawl and archive the sites hosted under the .EU domain.

Crawling .EU we identified the major spam sites presented in the domain, their are composed mainly by sub domains spam and link farms. Building a list of filters to the next crawl, we can mitigate the impact of these spam sites on the crawling process.

It's fairly common in .EU top level domain to find redirects to other top level domains like .com. In about 17 028 redirected seeds, 3 789 where redirected outside .EU.

Based on the total documents crawled, the 173 482 212 URLs that remained to be crawled and the number of seeds extracted, we estimated that 23 TB of disk space should be required for the following crawl of the .EU domain (without performing deduplication). For a configuration with 3 days between each checkpoint, about 3 TB of disk space should be needed to checkpoints.

With this set of configurations and with the insight discovered, a completed crawl at .EU can be made in 38 days. In this estimations, are not included performance gains with the spam filters optimization in the next crawl. So the crawling can take less time to complete and the resource usage can also be lower.

For the next crawl, the starting configuration are fine enough, but we will have to add the filters built from this crawl to optimize the crawl.

According to the Heritrix generated reports, some sites exceeded the maximum budget allowed per site of 10 000 URLs, reaching almost 20 000 URLs. This situation must be further investigated.

As future work we intend to perform 2 more crawls of the .EU domain. Each one of performed .EU crawls shall be indexed and become searchable through `archive.pt` one year after its finish date. Collaborations with researchers interested on studying the collected web data sets or crawl logs are welcome.