A Comparison Between The Performance of Wayback Machines

Fernando Melo, Daniel Bicho and Daniel Gomes
Arquivo.pt - The Portuguese Web Archive
{fernando.melo, daniel.bicho, daniel.gomes}@fccn.pt

April 13, 2016

Abstract

It is crucial to archive Web pages, since only 20% of today's Web pages will remain unchanged for more than one year [28]. One component of Web archiving software is the Wayback Machine, responsible for displaying Web pages as similar as they were when collected. The Arquivo.pt team decided to study the performance of Wayback Machine software since theirs Wayback software component was outdated, providing poor reproduction quality for files that were correctly stored and indexed, and with *leaks* to live Web resources, instead of replaying the proper archived content. This technical report presents preliminary experimental results about the replay quality and response speed of OpenWayback, PyWb, and Arquivo.pt Wayback machine software.

1 Introduction

The Web is a key means of communication in today's developed societies [18]. Since its creation, the Web has grown exponentially not only in the number of users, but also in content [29]. However the Web information is ephemeral, only 20% of Web pages remain unchanged after one year [28]. It is thus fundamental to preserve the Web information, as it is a part of human history.

Multiple archiving initiatives have been created in the past two decades [15]. Most of these initiatives attempt to preserve national or regional Web pages. Such is the case of the Portuguese Web Archive, which is responsible for the Arquivo.pt research infrastructure [1], that preserves not only the Web sites in the Portuguese domain, but also other culturally relevant sites. We have found out that some of our archived Web pages weren't being displayed correctly, although the contents of those pages were properly stored and indexed in the servers. This discovery led us to revise our Wayback software (Arquivo.pt Wayback), which is based on an old open-source version (1.2.1) of Internet Archive's Wayback [22].

The Wayback component of a Web archiving collection is responsible for allowing users to view past versions of Web sites, and one of its functions is to replay the Web sites as similar as they were in the date they were collected.

On the 21st of November 2014 the Arquivo.pt team made a first attempt to archive the .EU domain, storing a total of 250 163 776 documents from 34 138 initial seeds [14]. We used this collection to compare the Wayback replay mechanisms of PyWb [23], OpenWayback [20] and Arquivo.pt [17] open-source Waybacks.

2 Wayback Machines

A Wayback Machine (or a Web archival replay tool) is a software component that allows users to search by URL and date, and to play back an archived version of a Website in a Web browser. It enables the replay of historic collections, that are invaluable to everyone that needs to retrieve a past version of a Web site. Some possible use cases include a journalist revisiting past cases, an historian analysing digital historic documents, or even a user that discovers a broken link of a live Web page and tries to recover the missing content from a Web archive.

2.1 Arquivo.pt Wayback

Wayback is an open source Java application released in September 2005 by the Internet Archive, that allows searching and replaying archived Web pages. This open source Wayback machine (OSWM) [13] became widely used by entities members of the International Internet Preservation Consortium (IIPC) and became the most used rendering software for Web archives [19].

The Arquivo.pt Wayback derives from version 1.2.1 of OSWM from 2008 and is used as a replay mechanism for the Website `http://arquivo.pt`, a Web archiving initiative that is concerned not only with the preservation of the .pt domain, but also with other culturally relevant Websites. We have made several adaptations to the version 1.2.1 of OSWM, including the integration with NutchWax that allows full-text search.

Arquivo.pt Wayback has the following features:

- URL-search, supported by Lucene indexes;
- Capability to replay video and audio content (if embedded on the Web page);
- Indexing and Replaying GET requests;
- HTTP proxy mode support;
- Built in indexer tools.

2.2 PyWb

PyWb is a full rewrite (in python as the name suggests) of Wayback machine functionality. It is currently used by Webrecorder [26], a service that allows people to record live Web pages, and to store and replay an archived version of those Web pages. PyWb is also used by Perma [6] which creates permanent links to online sources cited, and by Rhizome [10] to provide replay access to archived captures from live Web.

PyWb features include:

- Supporting CDX and ZipNum;
- Supporting URL-search using Arquivo.pt's Lucene indexes through OpenSearchCDXServer, an experimental module which provides an example to fetch the capture index (CDX) from Arquivo.pt's OpenSearch interface[23].
- Supporting CDX-server API [24];
- Built in indexer tools;
- Command line interface for easily adding ARC and WARC files;
- Sophisticated client-side JavaScript rewriting system that allows dynamic URLs to be replayed properly;
- Capability to replay video and audio content - allows live recording of video/audio, which can then be played back from the archive [25];
- Indexing and replaying POST and GET requests (social media pages often use POST requests to scroll down the Web pages). Note that POST requests are only compatible with the WARC file format;
- Supporting HTTP and HTTPS proxy modes.

2.3 OpenWayback

In October 2013, the OpenWayback [21] was born when the Internet Archive (IA) delivered the repository of the OSWM to the IIPC, with the goals of solving shared requirements, and increasing the software stability by first testing changes across different deployment contexts across many organizations before providing an official release.

OpenWayback is used by many libraries and Web archives such as the National and University Library of Iceland [4], the British Library [9], Stanford Web Archive Portal [8], the Library of Congress [5], Bibliotheca Alexandrina [3] and York University Digital Library [12].

OpenWayback has the following features:

- CDX and ZipNum support. It supported Lucene indexes via the Nutch resource index, which is currently broken and it is going to be extinguished in the future OpenWayback versions;
- CDX-server API support;
- Built in indexer tools.
- Capability to replay video and audio content (if embedded on the Web page);
- Indexing and Replaying GET requests;
- HTTP proxy mode support.

3 The .EU Collection

European Union domains (.EU) can be sold to any person, company or organization that provides an address in the European Union, Iceland, Liechtenstein, or Norway [16]. The .EU domain includes all sort of Web sites, such as stores, spam sites, and governamental Web sites such as the European Parlament's [2].

The Arquivo.pt team begun a first attempt to archive the .EU domain on the 21st of November 2014, storing 5.8TB of unique information[14]. It collected 250 163 776 documents from 34 138 initial seeds. We estimate that the next crawl of the .EU domain according to our specifications, would take approximately 38 days to be completed and that 23 TB of disk space would be required.

4 Methodology

In this study we will evaluate the performance of Arquivo Wayback (that uses Lucene indexes); PyWb (with CDX indexes, and with the same Lucene indexes used by Arquivo Wayback); and OpenWayback (with CDX indexes).

All these indexes were generated from the same ARC files, so that we can compare the replay quality and not the quality of the crawling process. We randomly selected 400 URLs from our .EU collection and used WebPageTest service [11], an online test platform, to automatically test those URLs in each of the Waybacks. A total of 1200 (400 × 3 Waybacks) URLs were tested, and an HTTP Archive (HAR) file was automatically generated for each URL. HAR [7] is an archival format that stores performance data regarding Web pages loaded by a Web browser. We used the WebPageTest option to only test each URL once. Besides the HTTP status codes, WebPageTest also assigns the -2 error code to pages that exceeded a timeout of 2 minutes, and we assigned the -999 error to *leaks* (i.e., Web pages that made requests to the live Web). The list of the 400 tested URLs, and the output HAR files for each Wayback can be found in Github [27].

We have also manually replayed a subset of 40 of the 400 Web pages for PyWb, Arquivo and OpenWayback in order to visualize the quality of the replay of each Wayback, thus validating the obtained results from WebPageTest service.

Table 1 details the Wayback specifications, such as the version, type of indexes, and the container version. Ilya Kreymer, the developer of PyWb, created a module to enable using our Lucene indexes in PyWb [23]. The connection between our indexes and PyWb is done via Arquivo's Opensearch API.

Wayback	Version	Release Year	Indexes	Container Version
Arquivo	1.2.1	2008	Lucene	Tomcat 5.5.25
PyWb CDX	0.10.7	2015	CDX	uWSGI 2.0.11.1
PyWb Lucene	0.10.7	2015	Lucene	uWSGI 2.0.11.1
OpenWayback	2.20	2015	CDX	Tomcat 7.0.63

Table 1: Specifications for each tested Wayback

5 Experimental Results

5.1 Replay Quality

In order to evaluate the replay quality of OpenWayback, PyWb and Arquivo.pt, we measured the total number of status codes obtained for each Wayback. The results are shown in Figure 1. Besides the HTTP status codes (200's, 300's, 400's and 500's), we also considered 2 other status codes, namely (i) the **-999** code, that represents a *leak* to the live Web, which should be seen as an error, because the Wayback should present the archived contents, instead of loading resource's from the today's Web page; and (ii) the **-2** code that is given by WebPageTest when a resource is not loaded in a certain time (i.e. a timeout of 2 minutes), and thus should also be considered as an error.

As we can see in Figure 1, the Arquivo Wayback obtained more than 15 000 requests that were *leaks* to the live Web. This behaviour occurs frequently with our Wayback. One specific *leak* scenario in our Wayback is the following: the Web browser first makes a request to the live Web, and some time after that request was made, there is a rewrite to the correct resource to be loaded. Even though in a much smaller scale, OpenWayback also presented a significant number of *leaks*, i.e. more than 1700 requests. Both PyWb with CDX and Lucene indexes almost have no *leaks*. PyWb CDX and PyWb Lucene were the Waybacks that produced more **-2** timeout requests. PyWb CDX was the Wayback with the highest number of HTTP 404 (not found) error codes, followed by PyWb Lucene, OpenWayback CDX and Arquivo Lucene. Although 404 HTTP codes are negative features, it is natural that the Wayback that presents more content is the one with the highest number of not found requests. However it imposes additional workload on the servers.

Some neutral features are the 300's HTTP status codes, that are related with redirects. The most common redirect code that occurred in this experiment was the 302 found (Moved Temporarily) code, leaded by PyWb CDX, and pursued by OpenWayback CDX, PyWb Lucene and Arquivo.

Finally, the most important HTTP code to evaluate the replay quality of a Wayback is the 200 OK response, which means the request succeeded. PyWb CDX got 21 793 HTTP status codes, whereas OpenWayback CDX, PyWb Lucene and Arquivo got 16 569, 12 332, and 4514 OK responses, respectively.

The remaining status codes were obtained in such a small number for each Wayback that dispense analysis.

Table 2 summarizes the results from Figure 1. We considered success requests all the 200's HTTP status codes, whereas all the 400's, 500's, -2, and -999 codes were grouped as error codes. We have also added a basic performance indicator, that is the division of the success requests by the error ones. The higher this indicator is, the less prone to errors a Wayback is. OpenWayback CDX has achieved the highest success/error relation followed by PyWb CDX, PyWb Lucene, and Arquivo Lucene. PyWb CDX is the Wayback with more success requests, while PyWb Lucene has the smallest number of error requests.

5.2 Response Speed

In order to evaluate the speed of the Waybacks, we measured the average time to fully load a Web page (in seconds) for each Wayback. The results are shown in table 3, where we can observe that OpenWayback achieved the best results with an average of 8 seconds to fully load a Web page, followed by Arquivo with 17 seconds, PyWb CDX with 19 seconds, and PyWb Lucene with 35 seconds. We have optimised the number of threads for the uWSGI container in the experiment with PyWb CDX, whereas we have left the default number of threads in the tests with PyWb Lucene, and that is the main reason why PyWb Lucene has almost 35 seconds of average time to fully load a Web page.

Note that the default number of threads was used for PyWb Lucene, whereas that number was manually adjusted

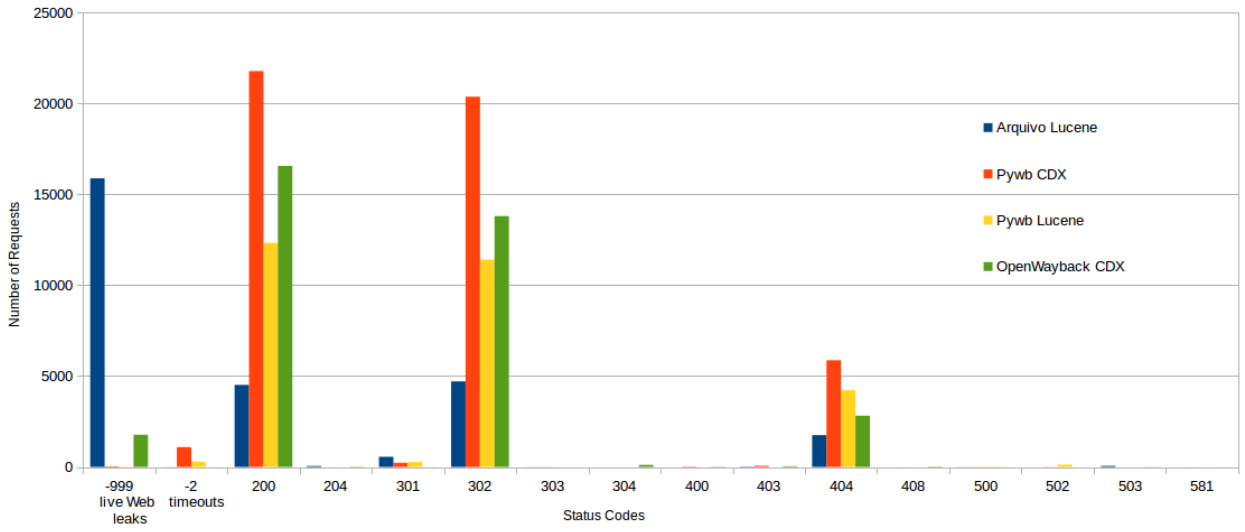

Figure 1: Number of status and error codes for each Wayback

Wayback	Success	Error	Success/Error
Arquivo Lucene	4564	17711	0.26
PyWb CDX	**21794**	7082	3.0773792714
PyWb Lucene	12333	**4652**	2.6511177988
OpenWayback CDX	16585	4668	**3.5529134533**

Table 2: Summary table for the performance of each Wayback

for PyWb CDX. It is also important to regard that these speed results depend on the WebpageTest servers and their workload at the time the experiments were run.

6 Conclusions

This technical report measured the replay quality of PyWb, OpenWayback, and Arquivo.pt Waybacks.

PyWb CDX was the Wayback that presented the highest number of 200 OK HTTP status codes, which indicates that it presented most comprehensive replay of archived pages.

Both PyWb CDX and Lucene presented an insignificant number of *leaks* to the live Web comparing with Arquivo and OpenWayback.

OpenWayback was the Wayback with the best relation between success and error codes, followed by PyWb CDX.

We confirmed our suspicions that Arquivo.pt Wayback software component is outdated, and needs to be replaced or updated. Some of the problems detected in our Wayback include redirect loops, incomplete pages, CSS style sheets not being loaded, and a high percentage of 404 (not found) status code errors.

Wayback	Avg. Load Time (seconds)
OpenWayback CDX	8.07
PyWb CDX	19.45
PyWb Lucene	34.93
Arquivo Lucene	16.93

Table 3: Average time to fully load a Web page on each Wayback

Overall the Waybacks with CDX indexes performed better in replay quality and response speed than the ones with Lucene indexes. Our manual validation allowed us to verify that there are still some problems with our NutchWax Web application, which retrieves search results from the Lucene indexes. More specifically, there are some resources that exist but aren't being returned, and some redirect loops, also caused by the results being returned by the Web app.

7 Future Work

There are many possibilities of further study on the performance of Wayback software. In the future we would like to test optimizing the number of threads in the uWSGI container for PyWb, and in Apache Tomcat for OpenWayback and Arquivo Wayback machine software. It would also be interesting to evaluate the response speeds of the Waybacks software with private instances of WebpageTestService, in our servers instead of using the WebpageTest cloud service, and to increase the number of URLs to test.

Another possibility is to expand test data set across time. The .EU crawl contains recently published Web pages and it could be biasing results against older resources of Wayback software.

8 Acknowledgments

We would like to thank Ilya Kreymer for developing PyWb-opensearch-cdx module, which enabled to test our indexes in PyWb, and for all his support. We would also like to thank the OpenWayback development group for the help given answering our questions.

References

[1] Arquivo.pt - pesquise páginas do passado. **http://arquivo.pt/** [Online; accessed 14-December-2015].

[2] European parliament - official website. **http://www.europarl.europa.eu/** [Online; accessed 15-December-2015].

[3] International school of information science - web archive. **http://www.bibalex.org/isis/frontend/archive/archive_web.aspx** [Online; accessed 15-December-2015].

[4] Landsbókasafns Íslands - háskólabókasafn. **http://vefsafn.is/** [Online; accessed 15-December-2015].

[5] Library of congress - archived web sites. **http://www.loc.gov/websites/collections/** [Online; accessed 15-December-2015].

[6] Perma - websites change perma links don't. **https://perma.cc/** [Online; accessed 15-December-2015].

[7] Software is hard - har specifications. **http://www.softwareishard.com/blog/har-12-spec/** [Online; accessed 15-December-2015].

[8] Stanford web archive portal - a searchable collection of websites archived by stanford university. **https://swap.stanford.edu/** [Online; accessed 15-December-2015].

[9] Uk web archive - preserving uk websites. **http://www.webarchive.org.uk/ukwa/** [Online; accessed 15-December-2015].

[10] webenact - rhizome's server for re-enacting captures from the live web. **http://webenact.rhizome.org/** [Online; accessed 15-December-2015].

[11] Webpagetest - test a website's performance. **http://www.webpagetest.org/** [Online; accessed 15-December-2015].

[12] York - web archive. **http://digital.library.yorku.ca/wayback/** [Online; accessed 15-December-2015].

[13] I. Archive. Sourceforge - wayback repository. **http://archive-access.sourceforge.net/projects/wayback/** [Online; accessed 15-December-2015].

[14] D. Bicho, J. Miranda, and D. Gomes. A first attempt to archive the .EU domain. Technical report, Portuguese Web Archive (FCT-FCCN), March 2015.

[15] J. M. Daniel Gomes and M. Costa. A survey on web archiving initiatives. In *International Conference on Theory and Practice of Digital Libraries 2011*, Berlin, Germany, September 2011.

[16] Eurid. Get a .eu. **https://www.eurid.eu/en/get-eu** [Online; accessed 18-January-2016].

[17] FCCN. Github - arquivo.pt repository. **https://github.com/arquivo/pwa-technologies** [Online; accessed 15-December-2015].

[18] D. Gomes and M. Costa. The importance of web archives for humanities. *International Journal of Humanities and Arts Computing*, 8(1):106–123, 2014.

[19] IIPC. Github - openwayback general overview. **https://github.com/iipc/openWayback/wiki/General-overview** [Online; accessed 15-December-2015].

[20] IIPC. Github - openwayback repository. **https://github.com/iipc/openWayback** [Online; accessed 15-December-2015].

[21] IIPC. Github - openwayback wiki. **https://github.com/iipc/openWayback/wiki** [Online; accessed 15-December-2015].

[22] B. Kahle. Internet archive - search the history of over 452 billion pages on the internet. **https://archive.org/** [Online; accessed 15-December-2015].

[23] I. Kreymer. Github - pywb repository. **https://github.com/ikreymer/pywb** [Online; accessed 15-December-2015].

[24] I. Kreymer. Github - pywb repository, cdx-server api. **https://github.com/ikreymer/pywb/wiki/CDX-Server-API** [Online; accessed 15-December-2015].

[25] I. Kreymer. Github - pywb repository, video replay and recording. **https://github.com/ikreymer/pywb/wiki/Video-Replay-and-Recording** [Online; accessed 15-December-2015].

[26] I. Kreymer. Webrecorder - a web archiving platform and service for all! **https://webrecorder.io/** [Online; accessed 15-December-2015].

[27] F. Melo. Github - repository with the outputs from this report. **https://github.com/Fernando-Melo/WaybackComparison** [Online; accessed 15-December-2015].

[28] A. Ntoulas, J. Cho, and C. Olston. What's new on the web? the evolution of the web from a search engine perspective. pages 1–12. ACM Press, 2004.

[29] A. Trotman and J. Zhang. Future web growth and its consequences for web search architectures. *CoRR*, abs/1307.1179, 2013.