A Comparison Between the Performance of Wayback Machines

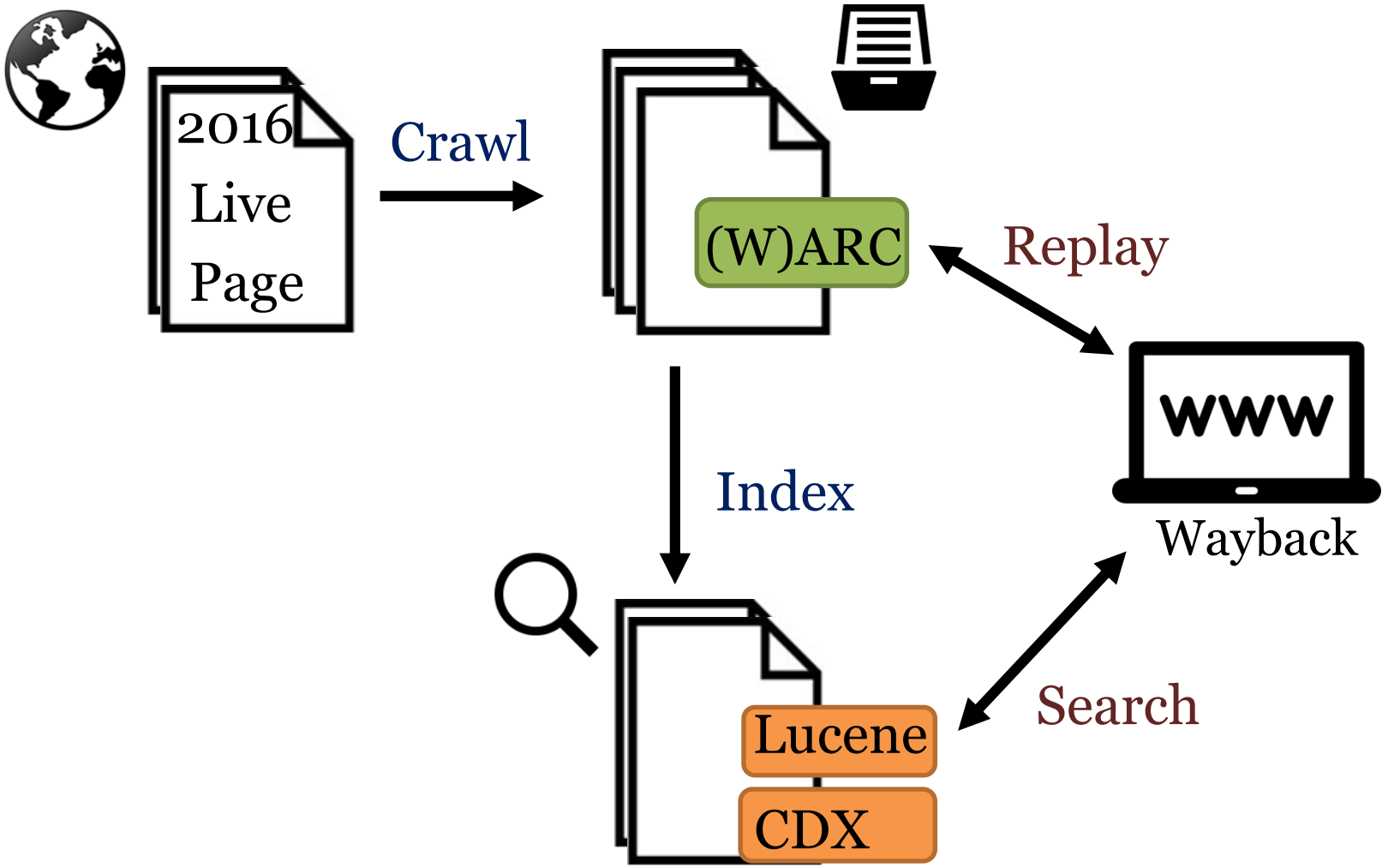
Fernando Melo fernando.melo@fccn.pt

Main reasons for this study

Outdated Wayback

Evaluate possible alternatives

How does a Web archive work?


2016
Live
Page
Crawl
(W)ARC
Replay
WWW
Wayback
Index
Lucene
CDX
Search

What is a Wayback Machine?

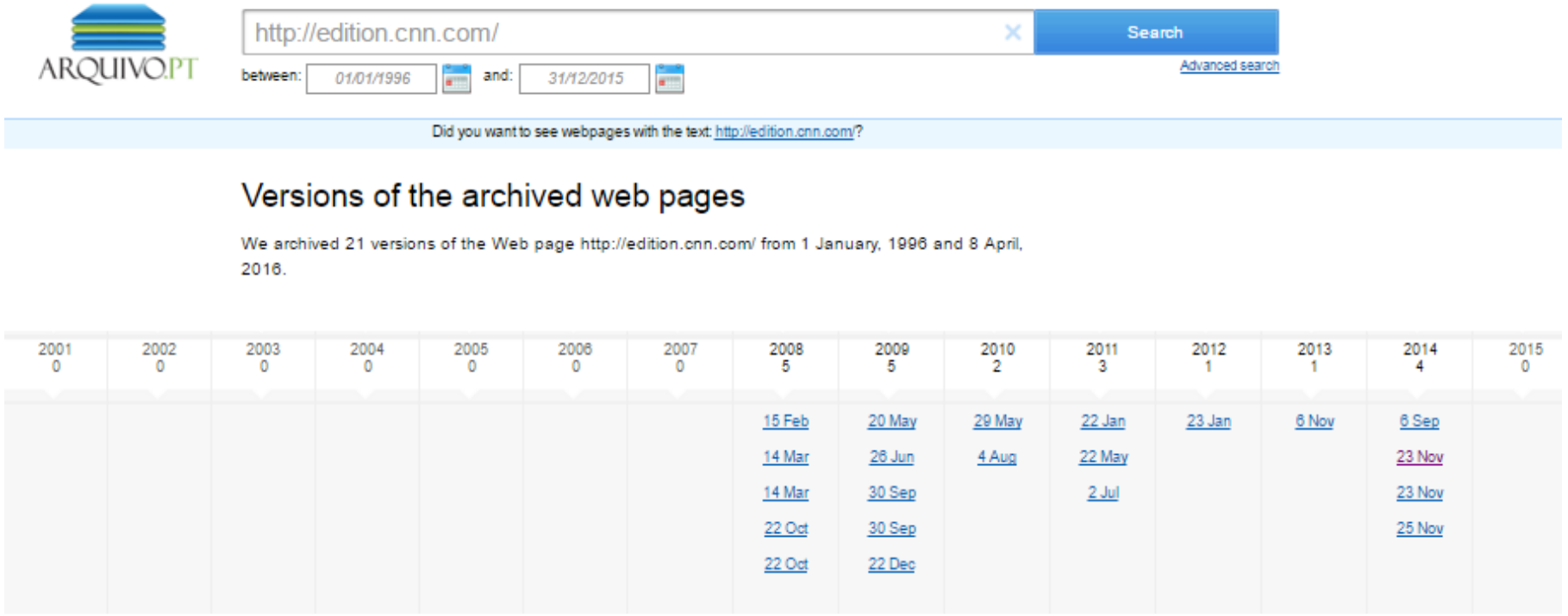
What is a Wayback Machine?

Software Component

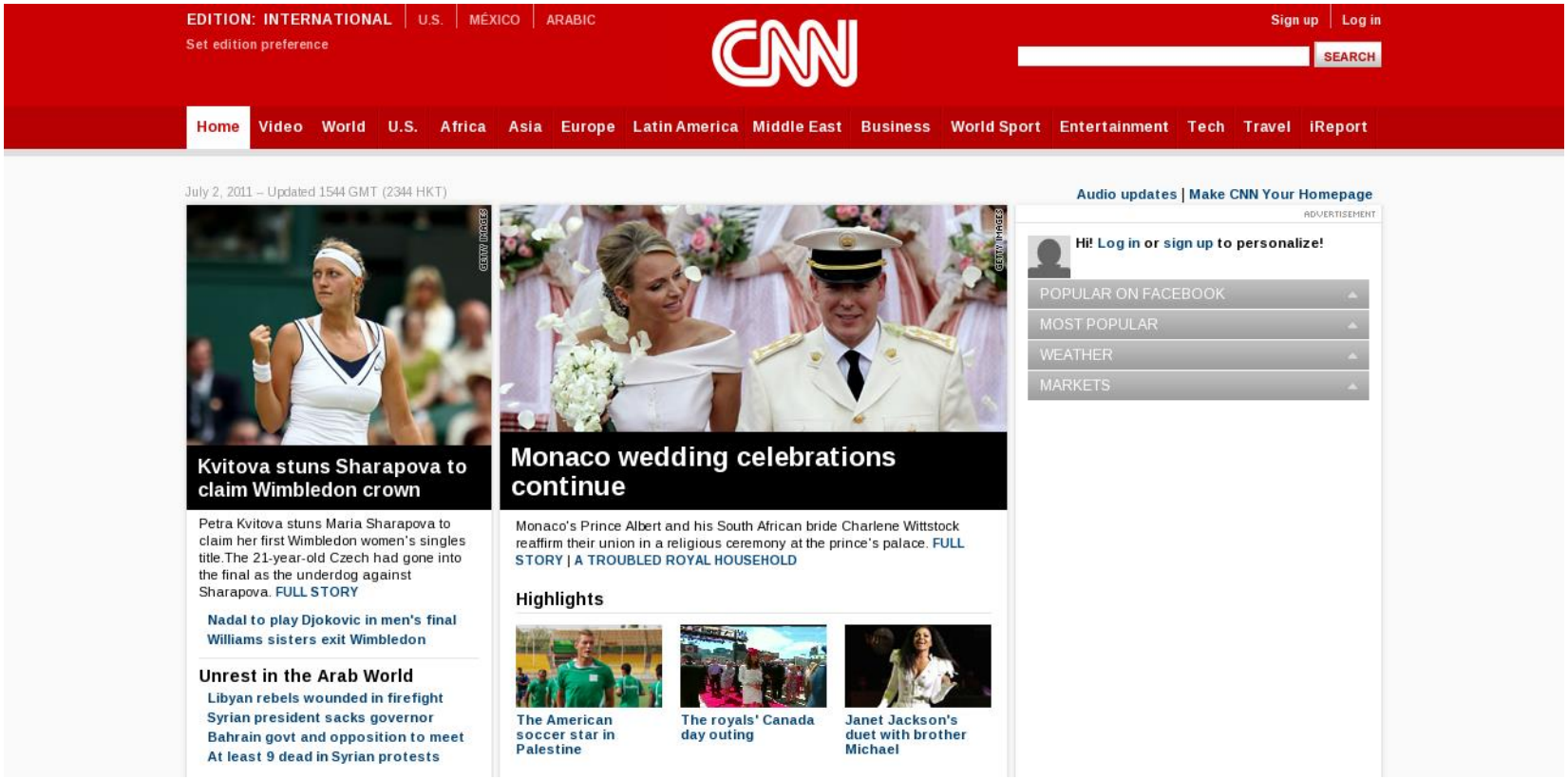
Replay Archived Web Pages

Search by URL and Date

What is a Wayback Machine?



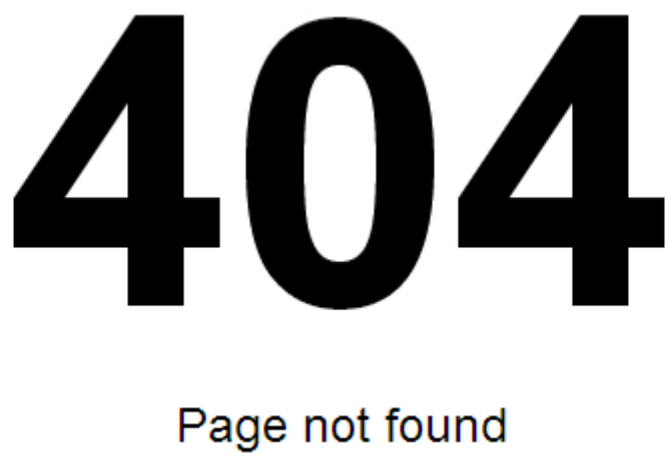
What is a Wayback Machine?



Common Wayback Machine Issues

Slow Replay

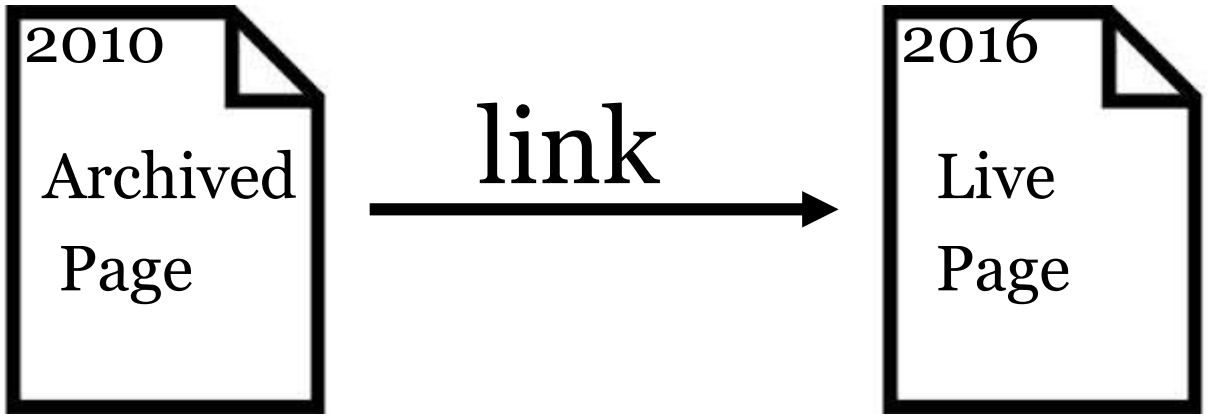
Not Found Errors



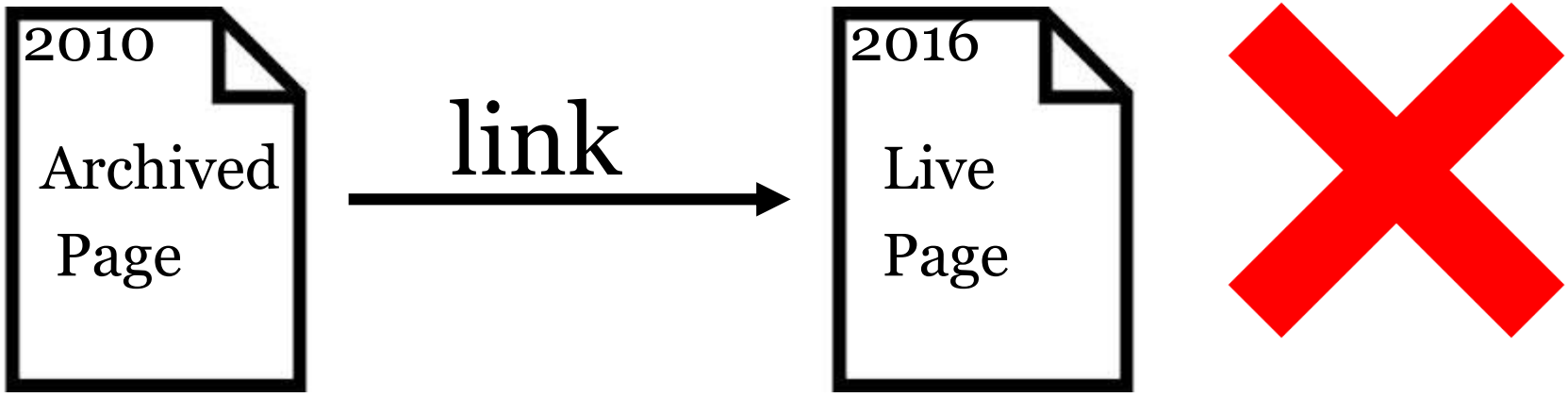
Not Found Errors

Not Found Errors

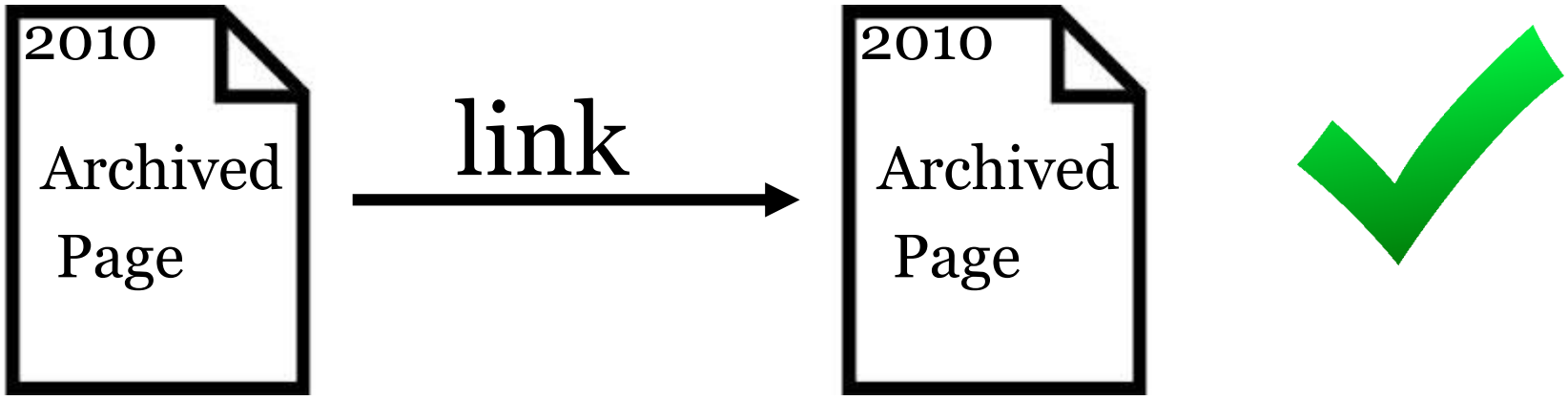
Live-Web Leaks



Live-Web Leaks

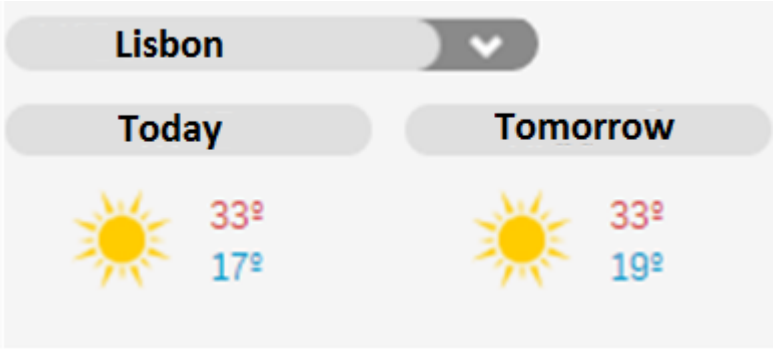


Live-Web Leaks

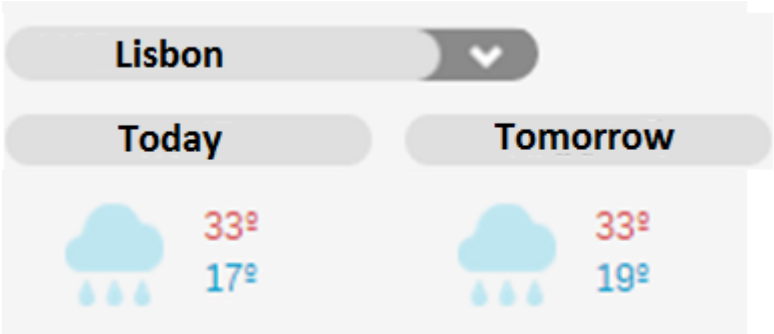


Live-Web Leaks

Original Web Page
July 14th, 2012

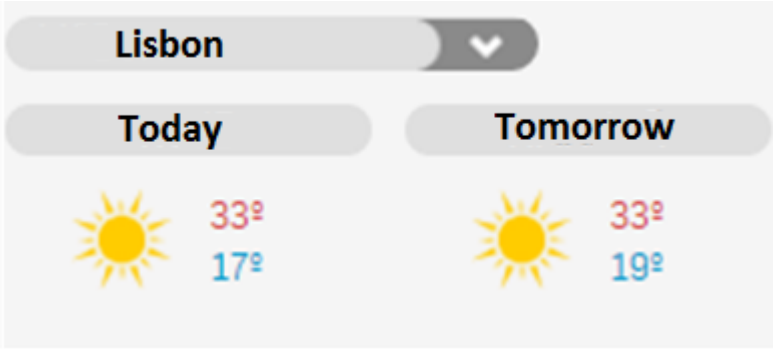


Archived Web Page
July 14th, 2012

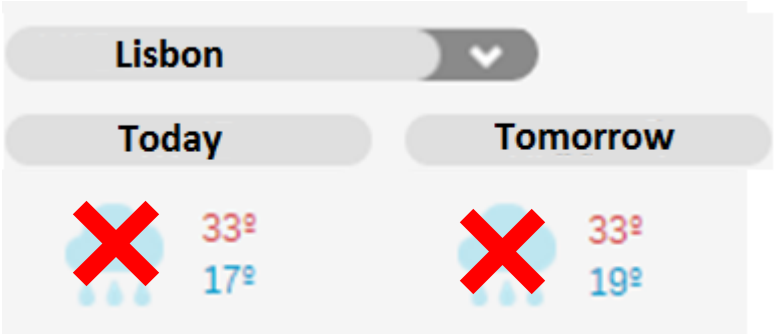


Live-Web Leaks

Original Web Page
July 14th, 2012

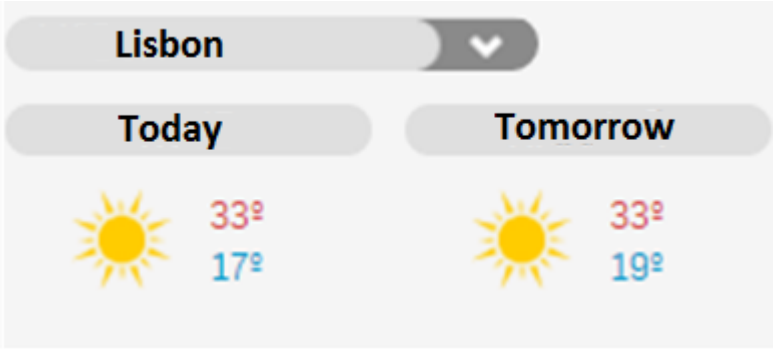


Archived Web Page
July 14th, 2012

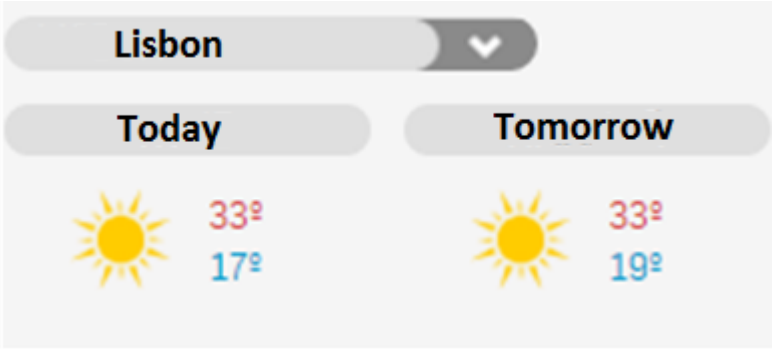


Live-Web Leaks

Original Web Page
July 14th, 2012



Archived Web Page
July 14th, 2012



Let's evaluate the **performance** of Wayback Machine Software!

Wayback Machines

Arquivo.pt Wayback

OpenWayback

PyWb

Wayback Machines

Arquivo.pt Wayback

Derives from version 1.2.1 of Open Source Wayback Machine (2008)

Java

Used by Arquivo.pt

Outdated - Presents several replay issues

PyWb Wayback

Developed by Ilya Kreymer

Python

Used by

- http://rhizome.org
- http://webrecorder.io
- http://perma.cc

OpenWayback

Released by the Internet Archive

Maintained by the IIPC

Java

OpenWayback - Users

National and University Library of Iceland
The British Library
Archive-It Mirror @ ODU
Stanford Web Archive Portal
The Library of Congress
Bibliotheca Alexandrina
York University Digital Library
Bibliothèque nationale de France
University of North Texas Libraries

The .EU Collection - 2014

The .EU Collection - 2014

Domains can be sold to anyone with a valid address in the European Union

European Institutions, Online Shops, and Web Spam

250 million documents from 34 thousand seeds

6TB

Methodology

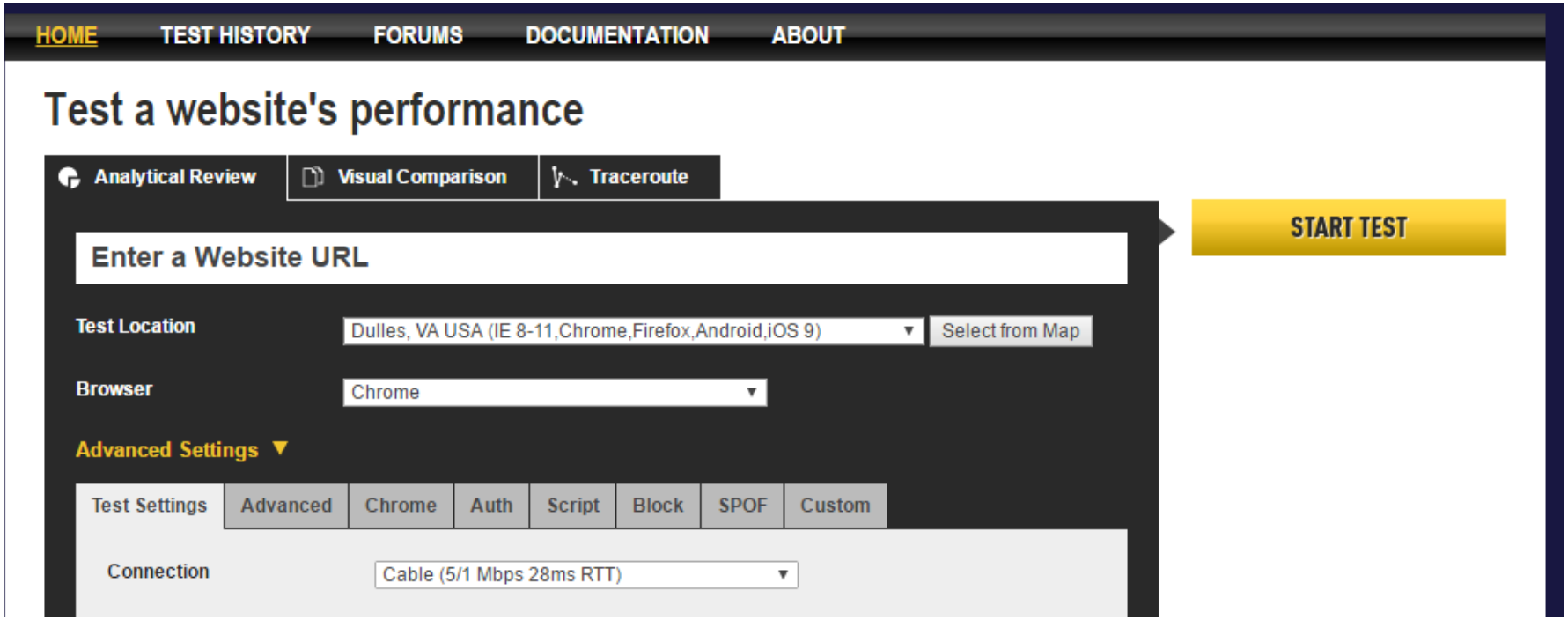
400 URLs from the .EU

WebPageTest service

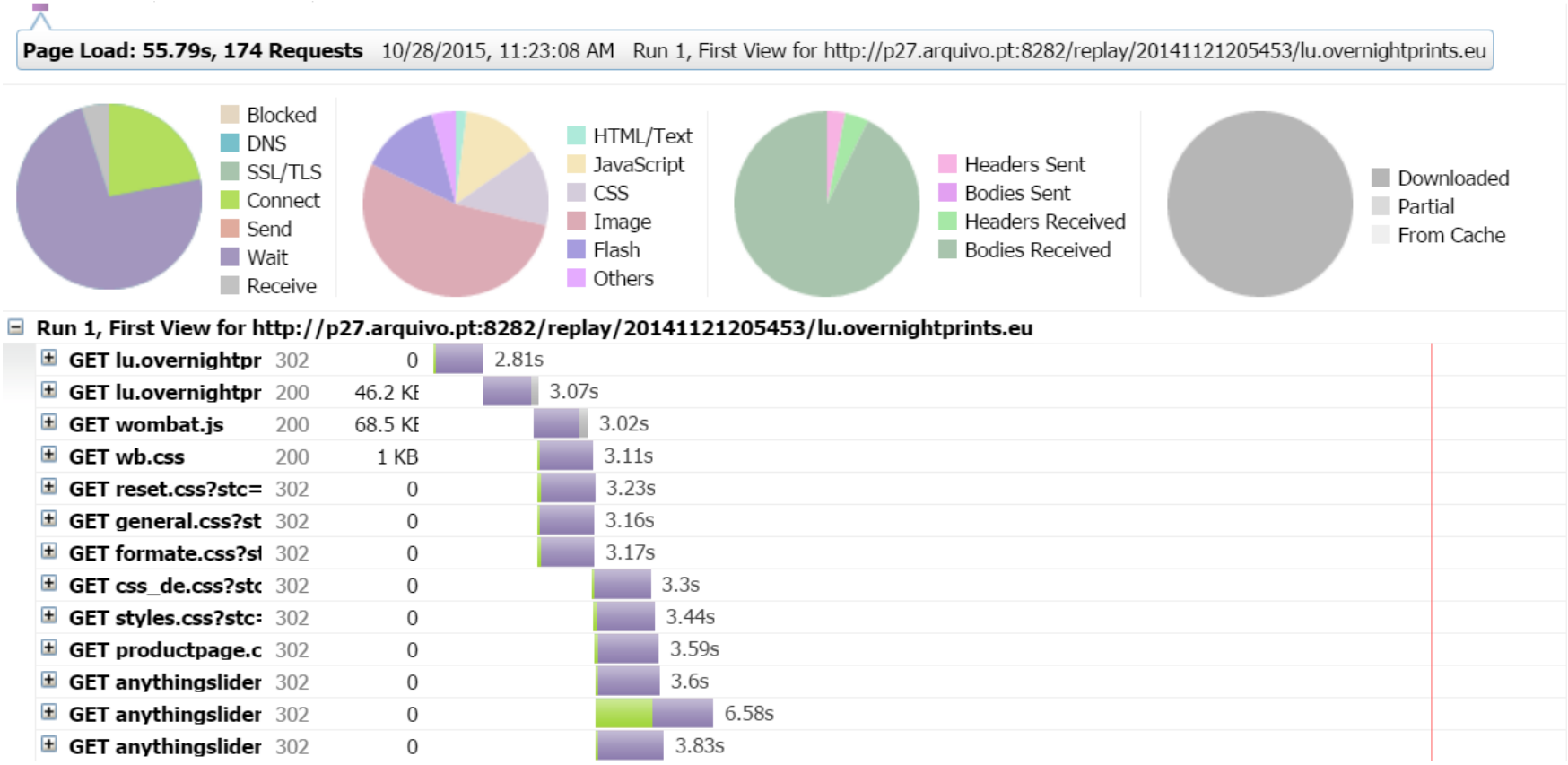
4 Wayback Configurations

HAR – to record performance data

Methodology



Methodology



Methodology

Only test each URL once

Tested using WebPageTest public servers

Response timeout of 2 minutes

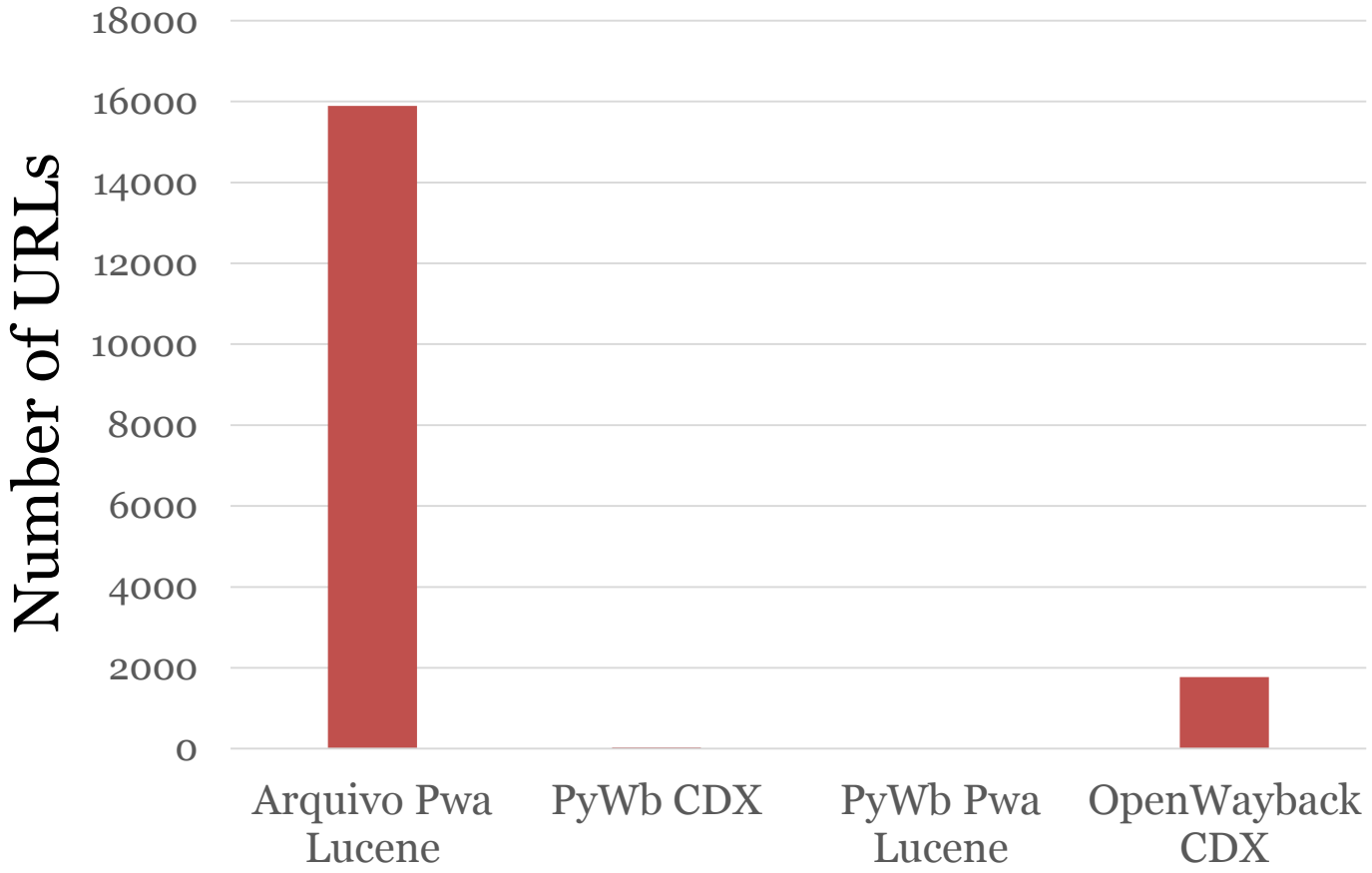
Error Code – Leak to the live Web

Wayback Specifications

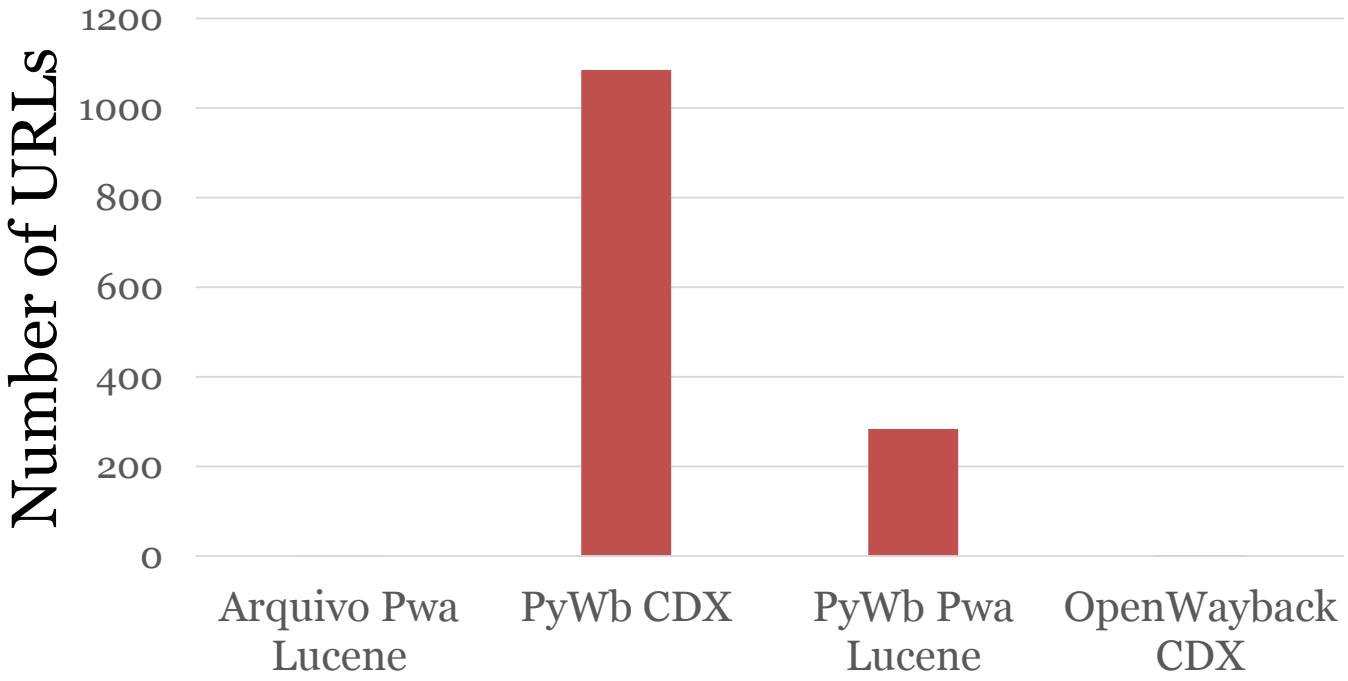
Wayback	Year
Arquivo Pwa Lucene	2008
PyWb CDX	2015
PyWb Pwa Lucene	2015
OpenWayback CDX	2015

Replay Quality – HTTP Status and Error Codes

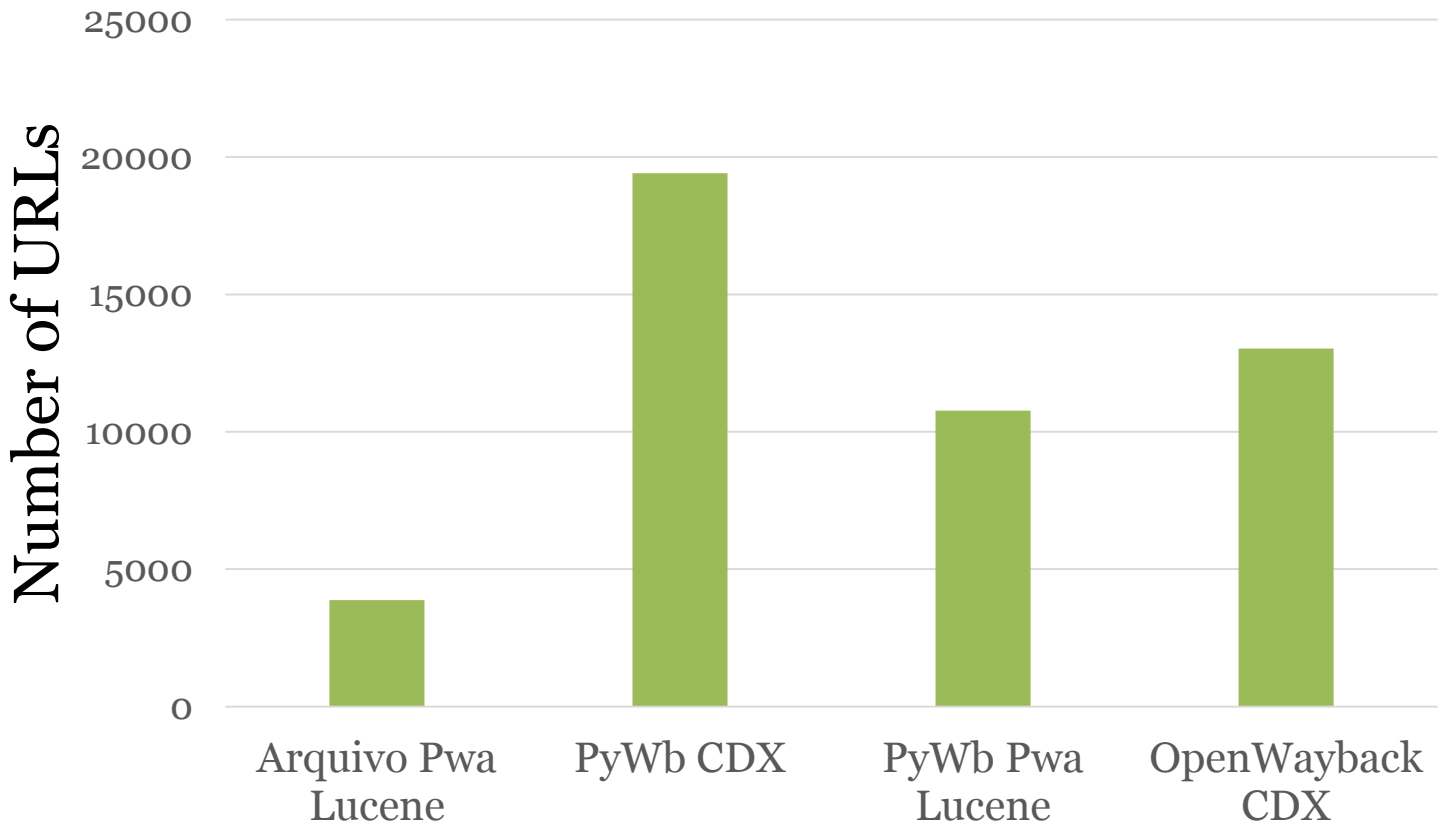
Results – Live Web Leaks



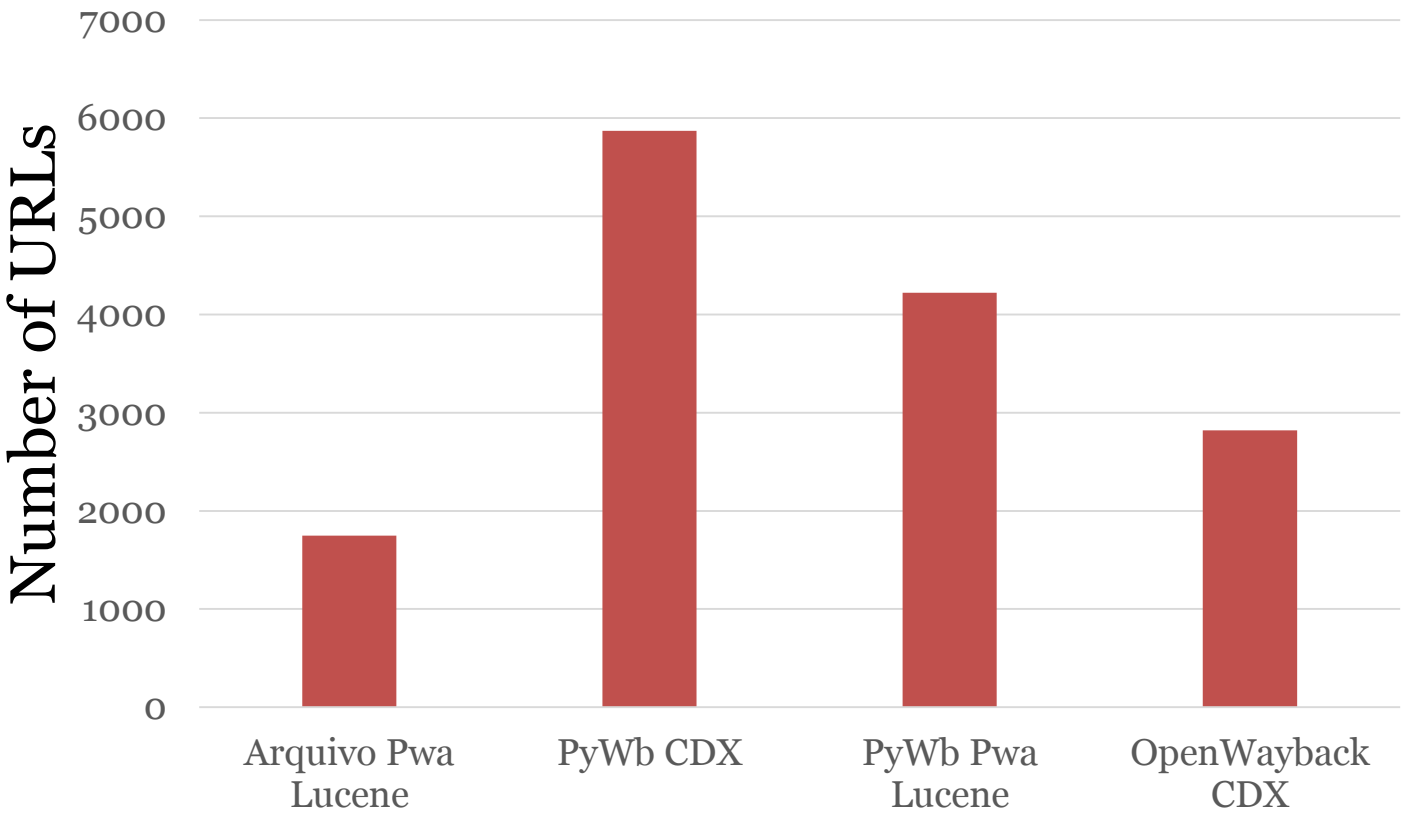
Results – Timeout Error



Results – 200 OK Status Code



Results – 404 Error HTTP Code

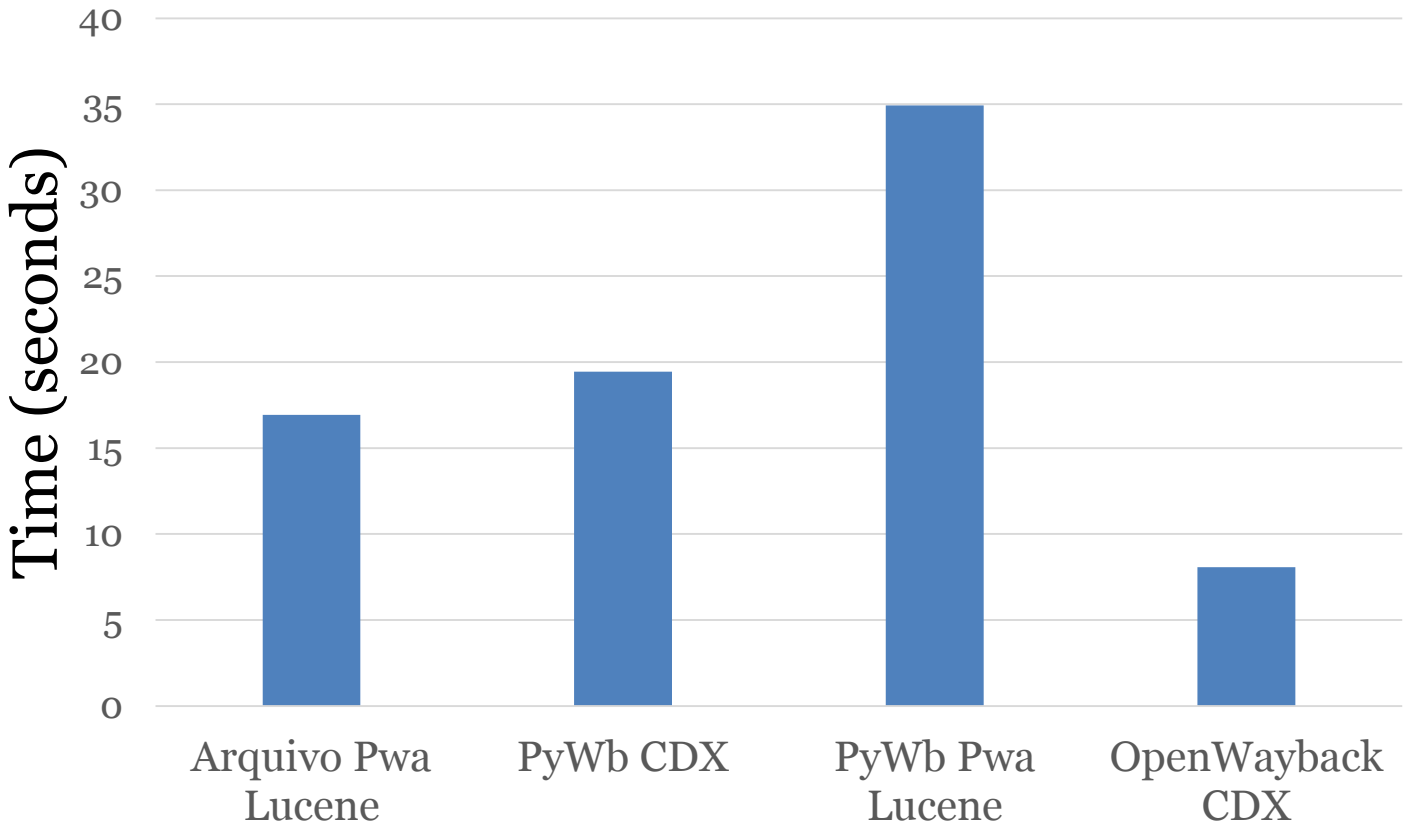


Results – Summary Table

Wayback	Success	Error	Success/ Error
Arquivo	3 930	17 711	0.22
PyWb CDX	**19 415**	7 082	2.74
PyWb Pwa Lucene	11 087	**4 652**	2.38
OpenWayback	13 068	4 668	**2.80**

Response Speed

Results – Average Load Time



Conclusions

PyWb presented the biggest number of 200 OK HTTP status codes

OpenWayback was the fastest Wayback

Replace or Update Arquivo.pt's Wayback!

Future Work

Test with older collections to evaluate the performance of Wayback Machine software

Test with private instance of WebpageTest server to be able to execute more tests and to control the server workload

References

https://github.com/Fernando-Melo/WaybackComparison