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Outdated Wayback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?
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?


<table>
<thead>
<tr>
<th>Year</th>
<th>Month/Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>2001</td>
<td>16 Feb</td>
</tr>
<tr>
<td>2001</td>
<td>14 Mar</td>
</tr>
<tr>
<td>2001</td>
<td>14 Mar</td>
</tr>
<tr>
<td>2001</td>
<td>22 Oct</td>
</tr>
<tr>
<td>2001</td>
<td>22 Oct</td>
</tr>
<tr>
<td>2002</td>
<td>20 May</td>
</tr>
<tr>
<td>2002</td>
<td>22 May</td>
</tr>
<tr>
<td>2002</td>
<td>23 Jan</td>
</tr>
<tr>
<td>2003</td>
<td>23 Nov</td>
</tr>
<tr>
<td>2003</td>
<td>23 Nov</td>
</tr>
<tr>
<td>2003</td>
<td>25 Nov</td>
</tr>
<tr>
<td>2003</td>
<td>30 Sep</td>
</tr>
<tr>
<td>2003</td>
<td>30 Sep</td>
</tr>
<tr>
<td>2003</td>
<td>22 Dec</td>
</tr>
<tr>
<td>2004</td>
<td>4 Aug</td>
</tr>
<tr>
<td>2004</td>
<td>22 May</td>
</tr>
<tr>
<td>2004</td>
<td>2 Jul</td>
</tr>
<tr>
<td>2005</td>
<td>8 Nov</td>
</tr>
<tr>
<td>2005</td>
<td>23 Nov</td>
</tr>
<tr>
<td>2005</td>
<td>23 Nov</td>
</tr>
<tr>
<td>2005</td>
<td>25 Nov</td>
</tr>
</tbody>
</table>
What is a Wayback Machine?
Common Wayback Machine Issues
Slow Replay

LOADING
Not Found Errors

404

Page not found
Not Found Errors
Not Found Errors
Live-Web Leaks

2010 Archived Page \rightarrow \text{link} \rightarrow 2016 Live Page
Live-Web Leaks

2010 Archived Page → link → 2016 Live Page

X
Live-Web Leaks

2010 Archived Page → link → 2010 Archived Page

 correctamente
Live-Web Leaks

Original Web Page  
July 14th, 2012

Archived Web Page  
July 14th, 2012

<table>
<thead>
<tr>
<th>Lisbon</th>
</tr>
</thead>
</table>
| Today  | 33°  
|        | 17°  
| Tomorrow| 33°  
|        | 19°  |
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Outdated - Presents several replay issues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

Test a website's performance

Enter a Website URL

Test Location: Dulles, VA USA (IE 8-11, Chrome, Firefox, Android iOS 9) Select from Map

Browser: Chrome

Advanced Settings

Test Settings: Advanced, Chrome, Auth, Script, Block, SPOF, Custom

Connection: Cable (5/1 Mbps 28ms RTT)
Methodology


- Blocked
- DNS
- SSL/TLS
- Connect
- Send
- Wait
- Receive

- HTML/Text
- JavaScript
- CSS
- Image
- Flash
- Others

- Downloaded
- Partial
- From Cache

Run 1, First View for http://p27.arquivo.pt:8282/replay/20141121205453/lu.overnightprints.eu

<table>
<thead>
<tr>
<th>Request</th>
<th>Status</th>
<th>Size</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>GET lu.overnightpr</td>
<td>302</td>
<td>0</td>
<td>2.81s</td>
</tr>
<tr>
<td>GET lu.overnightpr</td>
<td>200</td>
<td>46.2 K</td>
<td>3.07s</td>
</tr>
<tr>
<td>GET wombat.js</td>
<td>200</td>
<td>68.5 K</td>
<td>3.02s</td>
</tr>
<tr>
<td>GET wb.css</td>
<td>200</td>
<td>1 K</td>
<td>3.11s</td>
</tr>
<tr>
<td>GET reset.css?stc=</td>
<td>302</td>
<td>0</td>
<td>3.23s</td>
</tr>
<tr>
<td>GET general.css?st</td>
<td>302</td>
<td>0</td>
<td>3.16s</td>
</tr>
<tr>
<td>GET formate.css?st</td>
<td>302</td>
<td>0</td>
<td>3.17s</td>
</tr>
<tr>
<td>GET css_de.css?stc</td>
<td>302</td>
<td>0</td>
<td>3.3s</td>
</tr>
<tr>
<td>GET styles.css?stc</td>
<td>302</td>
<td>0</td>
<td>3.44s</td>
</tr>
<tr>
<td>GET productpage.c</td>
<td>302</td>
<td>0</td>
<td>3.59s</td>
</tr>
<tr>
<td>GET anythingslider</td>
<td>302</td>
<td>0</td>
<td>3.6s</td>
</tr>
<tr>
<td>GET anythingslider</td>
<td>302</td>
<td>0</td>
<td>6.58s</td>
</tr>
<tr>
<td>GET anythingslider</td>
<td>302</td>
<td>0</td>
<td>3.83s</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Wayback</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arquivo Pwa Lucene</td>
<td>2008</td>
</tr>
<tr>
<td>PyWb CDX</td>
<td>2015</td>
</tr>
<tr>
<td>PyWb Pwa Lucene</td>
<td>2015</td>
</tr>
<tr>
<td>OpenWayback CDX</td>
<td>2015</td>
</tr>
</tbody>
</table>
Replay Quality – HTTP Status and Error Codes
Results – Live Web Leaks

Number of URLs

- Arquivo Pwa Lucene: 16000
- PyWb CDX: 0
- PyWb Pwa Lucene: 0
- OpenWayback CDX: 1000
Results – Timeout Error

Number of URLs

- Arquivo Pwa Lucene
- PyWb CDX
- PyWb Pwa Lucene
- OpenWayback CDX
Results – 200 OK Status Code

- Arquivo Pwa Lucene
- PyWb CDX
- PyWb Pwa Lucene
- OpenWayback CDX
Results – 404 Error HTTP Code

Number of URLs

<table>
<thead>
<tr>
<th>Service</th>
<th>Number of URLs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arquivo Pwa Lucene</td>
<td>1000</td>
</tr>
<tr>
<td>PyWb CDX</td>
<td>6500</td>
</tr>
<tr>
<td>PyWb Pwa Lucene</td>
<td>4000</td>
</tr>
<tr>
<td>OpenWayback CDX</td>
<td>2500</td>
</tr>
</tbody>
</table>
## Results – Summary Table

<table>
<thead>
<tr>
<th>Wayback</th>
<th>Success</th>
<th>Error</th>
<th>Success/Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arquivo</td>
<td>3,930</td>
<td>17,711</td>
<td>0.22</td>
</tr>
<tr>
<td>PyWb CDX</td>
<td>19,415</td>
<td>7,082</td>
<td>2.74</td>
</tr>
<tr>
<td>PyWb Pwa Lucene</td>
<td>11,087</td>
<td>4,652</td>
<td>2.38</td>
</tr>
<tr>
<td>OpenWayback</td>
<td>13,068</td>
<td>4,668</td>
<td>2.80</td>
</tr>
</tbody>
</table>
Response Speed
Results – Average Load Time

<table>
<thead>
<tr>
<th>Service</th>
<th>Average Load Time (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arquivo Pwa</td>
<td>15</td>
</tr>
<tr>
<td>PyWb CDX</td>
<td>20</td>
</tr>
<tr>
<td>PyWb Pwa Lucene</td>
<td>35</td>
</tr>
<tr>
<td>OpenWayback CDX</td>
<td>40</td>
</tr>
</tbody>
</table>
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