Panel:
Web Archiving –
Lessons and Potential

Abbie Grotke (Library of Congress)
Barbara Signori (Swiss National Library)
Clément Oury (Bibliothèque nationale de France)
Daniel Gomes (Portuguese Web Archive)
Mário J. Silva (INESC-ID)
Nuno Freire (European Library)
Lessons learned
The Portuguese Web Archive project started in 2008.

Welcome to the Tomba project: the Portuguese web archive

Publishing tools, such as Blogger, enabled people with limited technical skills to become web publishers. Never before in the history of mankind so much information was published. However, it was never so ephemeral. Web documents such as news, blogs or discussion forums are valuable descriptions of our times, but most of them will not last longer than one year.

If we do not archive the current web contents, the future generations could witness an information gap in our days.

The Internet Archive collects and stores contents from the world-wide web. However, it is difficult for a single organization to archive the web exhaustively while satisfying all needs, because the web is permanently changing and many contents disappear before they can be archived.

As a result, several countries are creating their own national archives to ensure the preservation of contents of historical relevance to their cultures.

Portugal is now beginning its national web archiving initiative with the Tomba project at FCCN (National Foundation for Scientific Computing).
It provides **version history** like the Internet Archive Wayback Machine.
But also **full-text search** over 1.2 billion web files archived since 1996.
Acquiring web data
We needed to integrate third-party collections archived before 2007

- An archive must have “old stuff”
- Integration of historical collections
  - 1.9 TB from the Internet Archive between 1996 and 2007
  - 600 MB CD ROM with sites published in 1996
Tools to convert saved web files to ARC format

“Dead” archived collections became searchable and accessible

Specific conversion tools per collection were required but baseline software could be reused
The integration effort was worth to save few but valuable information.
Crawling the live-web since 2007

- Trimestral broad crawls: 78 million files per crawl
- Daily selective crawls: 764 000 files per day
- Heritrix 1.14.3 initially configured based on previous experience crawling the Portuguese Web
  - Trial-error process until final configuration
- Must recheck configurations periodically
The URLs of the publications crawled daily change frequently

- Expresso newspaper had 5 different domains since 2008
- Seed list of daily crawls must be periodically validated by humans
Default Robots.txt of Content Management Systems forbid crawling images

• Developers of popular Content Management Systems are not aware of web archiving
  – Joomla forbids images since 2007
Attempt to raise awareness

• Contacted webmasters of the selected publications by email
  – Only 10% returned feedback

• None, raised any objection, just questions.
• Some, did not know they had robots exclusion rules on their sites.
• Most, did not know what was a “web archive”.
• All, were satisfied from being selected as representatives of our cultural heritage
The DeDupicator (Heritrix add-on module)

Release information

Current stable release is 0.4.0.

All releases, including interim (potentially unstable) releases can be found here:
Release History of DeDupicator for Heritrix 1 and here: Release History of DeDupicator for Heritrix 3

- Downloads content, computes checksum and compares it with version from the previous crawl
  - Unchanged->Discarded
  - Changed->Stored
- No impact on download rate
Savings on Trimestral crawls

Average disk space per trimestral crawl (TB)

41% less disk space to store content
Savings on Daily crawls

76% less disk space to store content
Total savings from using DeDuplicator

26.5 TB/year

• Using DeDuplicator saved space without performance degradation.
PORTUGUESE WEB ARCHIVE

Ranking the past Web
NutchWAX as baseline for full-text search

Introduction

NutchWAX ("Nutch + Web Archive eXtensions") searches web archive collections. The Web Archive eXtensions (WAX) include adaptation of the Nutch fetcher step to go against web archives rather than crawl the open net -- adaptation currently does Internet Archive & ARC files only -- and plugins to add extra fields to the index that return an Archive Records' location in the repository, its collection name, etc.

Project Sponsors

IIPC

The International Internet Preservation Consortium (IIPC) is a consortium of twelve National Libraries and the Internet Archive. The mission of the IIPC is to acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations.

NWA

The Nordic Web Archive (NWA) is the Nordic National Libraries' forum for co-ordination and exchange of experience in the fields of harvesting and archiving web documents.

IA

The Internet Archive (IA) is a 501(c)(3) non-profit organization whose mission is to build a public Internet digital library.
Users were not satisfied with NutchWAX search

- Unpolished interface
- Slow results
  - 40M URLs, >20s
- Low relevance for search results
Developed a new web archive search system

• Quicker response times
• Improve relevance for search results
Had to build a Web Archive Information Retrieval Test Collection: PWA9609

• To evaluate and improve relevance for search results

• Corpus of documents from 1996 to 2009
  – 255 million web pages (8.9 TB)
  – 6 collections: Internet Archive, PWA broad crawls, integrated collections

• Gold collection
  – Query, relevant results
Time-aware ranking models yield better search results

<table>
<thead>
<tr>
<th>Metric</th>
<th>Time-unaware ranking models</th>
<th>Time-aware ranking models (our proposals)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NutchWAX</td>
<td>TVersions</td>
</tr>
<tr>
<td>nDCG@1</td>
<td>0.250</td>
<td>0.430</td>
</tr>
<tr>
<td>nDCG@10</td>
<td>0.174</td>
<td>0.202</td>
</tr>
<tr>
<td>Precision@1</td>
<td>0.320</td>
<td>0.500</td>
</tr>
<tr>
<td>Precision@10</td>
<td>0.168</td>
<td>0.172</td>
</tr>
</tbody>
</table>


- Internationalization support
- New graphical design
- Advanced search user interface
- 71% overall user satisfaction from rounds of usability testing
Observations from usability testing

No! That's not how you're supposed to use it!
Searching the past web is a confusing concept

- Understanding web archiving requires being techie
- Must provide examples of web-archived pages
Users are addicted to query suggestions

- Developed query suggestions mechanism for web archive search
Users “google” the past and we have to comply

• Users search web archives replicating their behavior from live-web search engines
• Users input queries on the first input box that they find
  – Search system must identify query type (URL or full-text) and present corresponding results
• Must provide additional tutorials and contextual help to search the past web
Blade Systems/Storage Area Networks vs. Independent servers

- 61 computers, 1.8 TB RAM, 340 disks (370 TB)
- Blade systems and SAN are not adequate for web archiving
  - Extremely expensive
  - Single points of failure
  - Hard to manage
- **Independent servers are cheaper and more reliable**
Legal issues
Just concerns

- Respect Robots Exclusion Protocol
- 1 year embargo
- Proactively remove illicit content
- Remove content on-demand by authors
Potential as research infrastructure
API to process archived data using the PWA Hadoop cluster
Measure web accessibility for people with disabilities

Results

rates versus page count distribution

In Rui Lopes, Daniel Gomes, Luís Carriço, Web Not For All: A Large Scale Study of Web Accessibility, 2010
Characterizations of the Portuguese Web structure

Archiving Web Spam degrades search results

1st search result is a Web Spam. 😞
But archiving Web Spam is not useless for research: Improve Web Spam detectors!

In A. Garzó et al., Cross-Lingual Web Spam Classification, 2013
OpenSearch to extend functionality

Introduction

The Portuguese Web Archive provides an interface for users and tools to easily query the system. The response is a XML-based file (RSS 2.0).

Details

The PWA interface follows the OpenSearch 1.1 (Draft 5) namespace defined at http://www.opensearch.org/Specifications/OpenSearch/1.1/Draft_5.

It also follows the OpenSearch Time extension (Draft 1) namespace at http://www.opensearch.org/Specifications/OpenSearch/Extensions/Time/1.0/Draft_1 that describes how to set temporal search parameters.

The OpenSearch Description Document at http://arquivo.pt/opensearch.jsp describes the public interface and how the search client should make search requests.

Query

The OpenSearch URL must contain the query and search parameters.

Parameters:
Web archive search can be easily integrated on web browsers.
OpenSearch used by Computer Science students to create new web applications

- Web application combines information about politicians from several sources: Wikipedia, Youtube, Twitter, Portuguese Web Archive
All our source code and test collections are freely available.
Conclusions

• Web archives are crucial infrastructures for modern societies
• Must raise awareness about web archiving among users and developers
• We need to collaborate
Panel discussion

1. How is your experience related to this work?
2. How could web archives be further improved?
3. How could web archives interact with libraries/other cultural heritage organizations?
4. How to unfold the full potential of web archives as research infrastructures?
5. Which innovative collaborations could be established?
6. What is the role of web archiving in modern societies?
7. ...