

# A survey of web archive search architectures

Miguel Costa, **Daniel Gomes**  
(Portuguese Web Archive@FCCN)  
Francisco Couto, Mário J. Silva  
(University of Lisbon)

# The Internet Archive was founded in 1996

Web-archived  
page of the first  
Web Archive

## Building IA INTERNET ARCHIVE a digital library for the future



[Acknowledgements](#)  
[Board Members](#)  
[Finding Us](#)  
[In the News](#)  
[Webmasters](#)

### *Our Mission*

Internet Archive is collecting and storing public materials from the Internet such as the World Wide Web, Netnews, and downloadable software which have been donated by [Alexa Internet](#).

The Archive will provide historians, researchers, scholars, and others access to this vast collection of data (reaching ten terabytes), and ensure the longevity of this information.

For more information about our philosophy and objectives, please read [Archiving the Net](#) by the Archive's founder, [Brewster Kahle](#).

Visit: [The '96 US Presidential Election Web Archive](#). This archive was created in affiliation with the [Smithsonian Institution](#).

✉ [info@archive.org](mailto:info@archive.org)

# Web archiving has been growing

The screenshot shows the Wikipedia article "List of Web archiving initiatives". At the top right, there are links for "Create account" and "Log in". Below these are tabs for "Article" and "Talk", and buttons for "Read", "Edit", and "View history". A search box is located on the right. The article title is "List of Web archiving initiatives", with the subtitle "From Wikipedia, the free encyclopedia". A yellow warning box states: "This article **uses bare URLs for citations**. Please consider adding full citations so that the article remains verifiable. Several templates and the [Reflinks tool](#) are available to assist in formatting. ([Reflinks documentation](#)) (February 2012)". The main text reads: "This page contains a list of Web archiving initiatives worldwide. For easier reading, the information is divided in three tables: web archiving initiatives, archived data and access methods. This Wikipedia page was originally generated from the results obtained for the research paper *A survey on web archiving initiatives*<sup>[1]</sup> published by the Portuguese Web Archive <sup>[2]</sup> team." A "Contents" box lists: "1 Web archiving initiatives", "2 Archived data", "3 Access methods", and "4 References". On the right, a map titled "Map of Web archiving initiatives worldwide in March, 2012." shows green highlights on various continents, including North America, Europe, and parts of Asia and Africa. The left sidebar contains navigation links: "Main page", "Contents", "Featured content", "Current events", "Random article", "Donate to Wikipedia", "Interaction" (with sub-links: "Help", "About Wikipedia", "Community portal", "Recent changes", "Contact Wikipedia"), and "Toolbox".

- 77 web archiving initiatives
- 282 billion web-archived files

# Web archives must be searchable

- Users demand “Google-like” search
  - Searchable means at least full-text search
- **Unsearchable=Useless**

# **How to enable web archive search?**

Our pursued answer since 2001...

# Research on web archiving (2001)



## Digital Deposit

[Back to xldb Home Page](#)

### Research Team:

- [Mário Gaspar da Silva](#)  
(coordinator)
- [Ana Paula Afonso](#)
- [António Ferreira](#)
- [Daniel Gomes](#)
- [João Campos](#)
- [Norman Noronha](#)

Publications are changing from the traditional formats, like paper magazines, to digital media, such as online news feeds. In addition everyone with a connected computer is now a potential publisher. This is increasing the number of new publications, making the management of deposits more complex. The archival of publications has a significant role in preserving the historical past. Publications in traditional media have been archived since ancient times. However, archiving publications on the Internet with techniques designed to allow their long term preservation is a non trivial task. The tools available for building digital publications weren't designed with preservation in mind and so don't meet most of the requirements involved.

- A web archive of online publications
- Project between the University of Lisbon and the National Library of Portugal

# Research on search engines (2001)



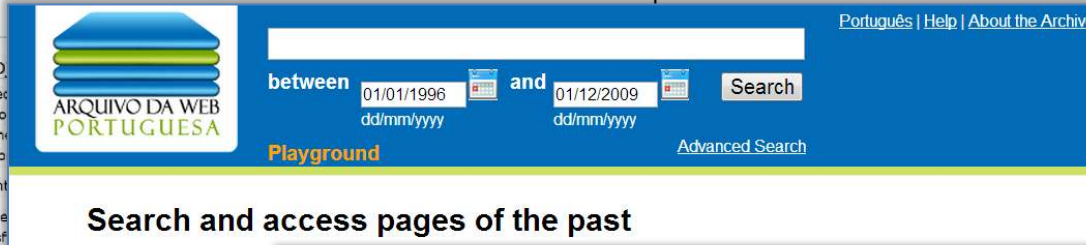
- Portuguese-web search engine

# Portuguese Web Archive = Web search + Web archiving

2008

2010

2013



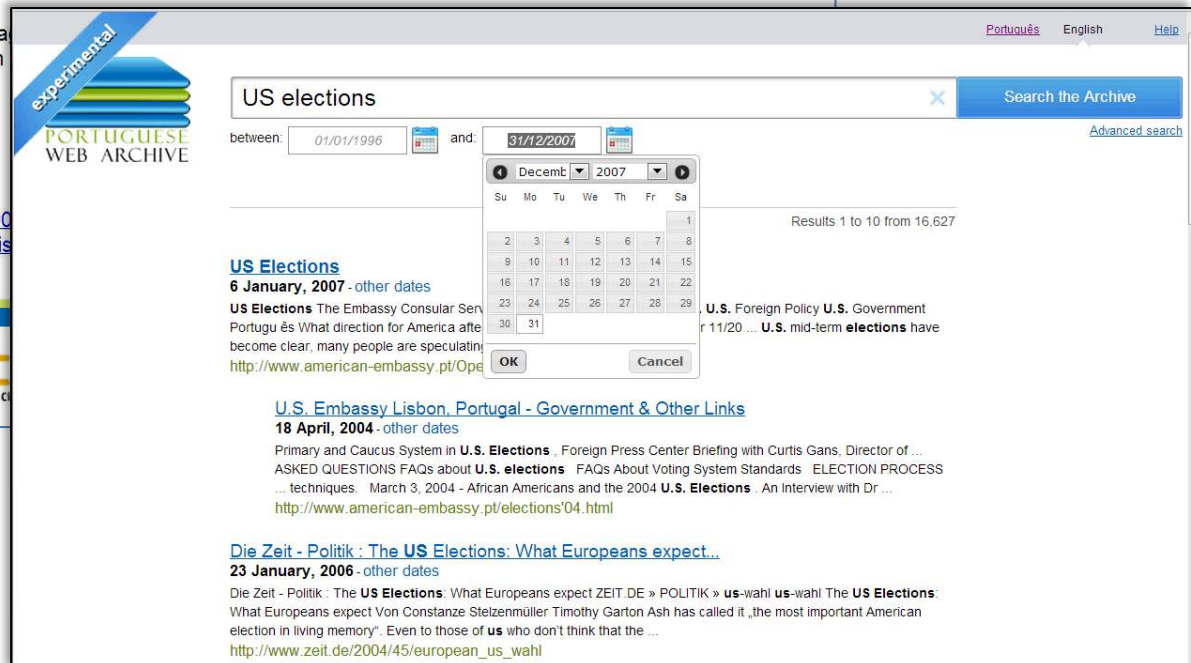
## Search and access pages of the past

See or rediscover pages of the past  
There are more than 100 million pages archived

Search examples:

- [Expo 98](#)
- [Euro 2004](#)
- [Elections in 2004](#)
- [University of Lisbon](#)

**FCC**  
Fundação para a Computação Científica





# Survey about web archiving initiatives (2011)

## A survey on web archiving initiatives

Daniel Gomes, João Miranda, and Miguel Costa

Foundation for National Scientific Computing (FCCN)\*\*

Av. do Brasil, 101

1700-066 Lisboa, Portugal

(daniel.gomes, joao.miranda, miguel.costa)@fccn.pt

**Abstract.** Web archiving has been gaining interest and recognized importance for modern societies around the world. However, for web archivists it is frequently difficult to demonstrate this fact, for instance, to funders. This study provides an updated and global overview of web archiving. The obtained re-

- URL search: 89%
- Meta-data search: 79%
- **Full-text search: 67% (28 initiatives)**
  - The knowledge is out there

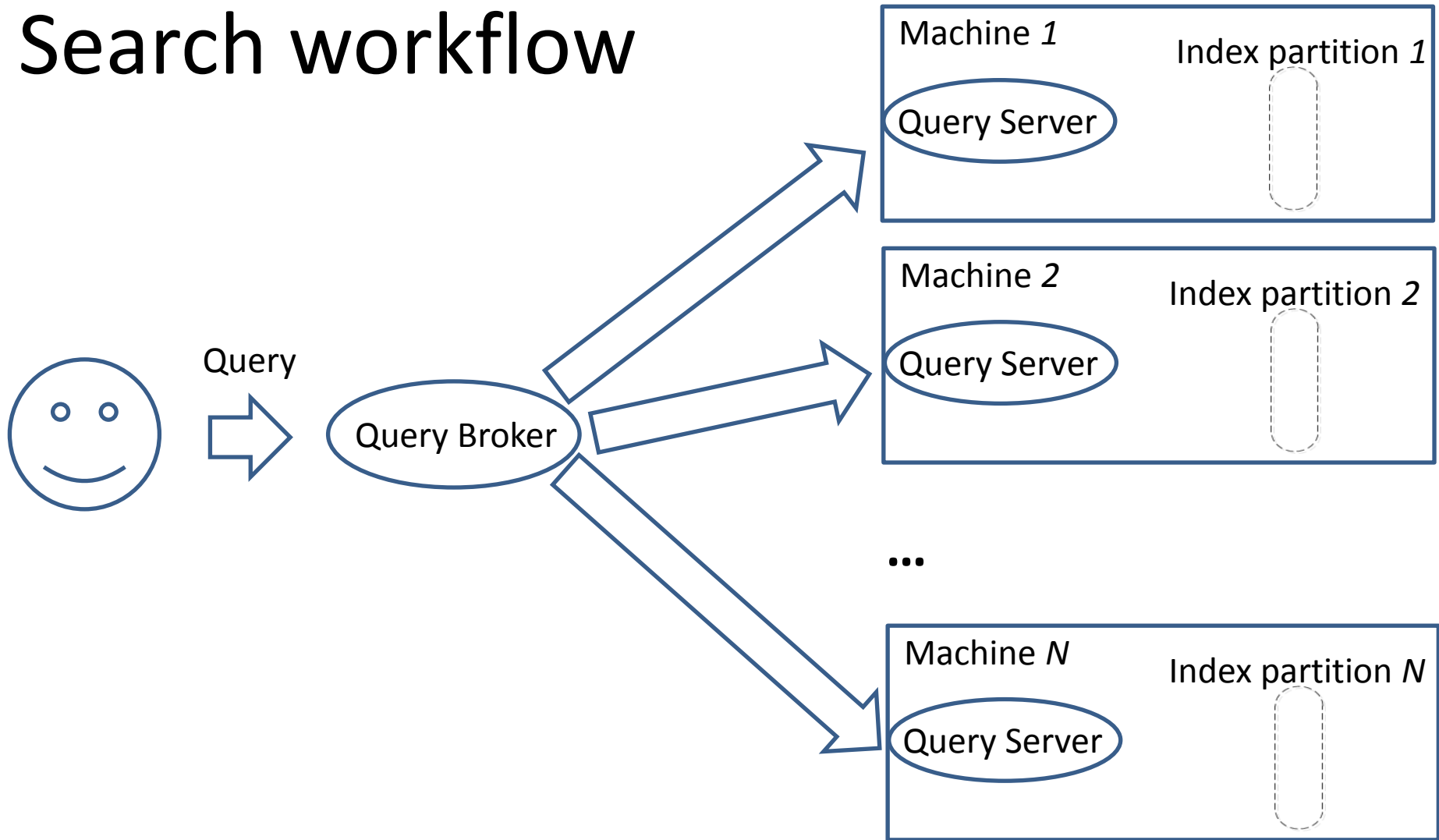
# For this survey of web archive search architectures

- Identified prevalent search architectures
- Compared main features based on:
  - Available publications (still few)
  - Our experience

# Portuguese Web Archive

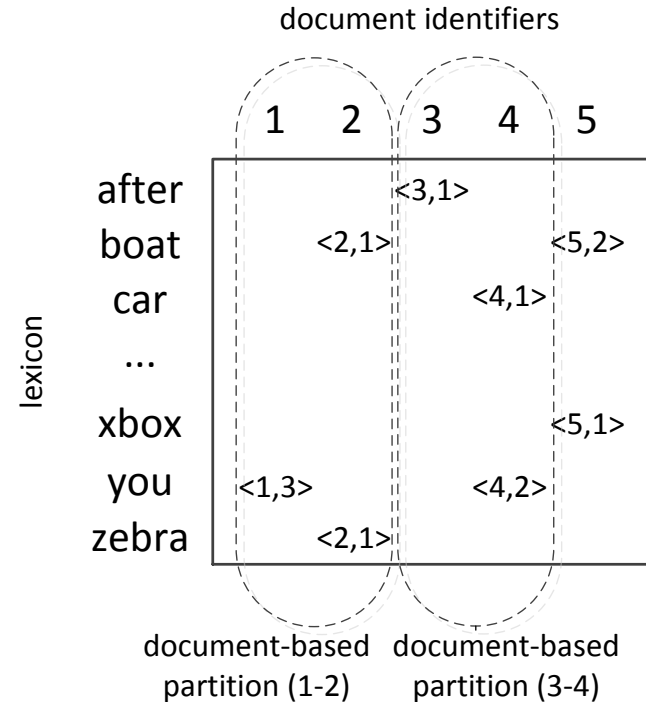
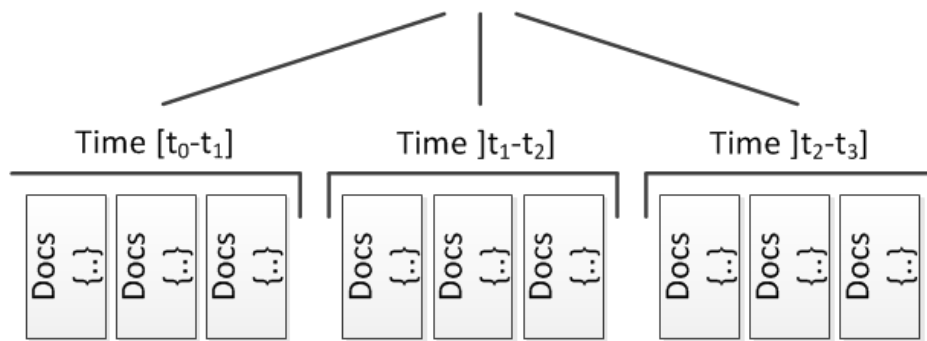
- Based on NutchWAX
  - Archive-access tools are widely used to support search
- Full-text search over 1.2B docs at [archive.pt](http://archive.pt)

# Search workflow



- For large collections, indexes must be **partitioned** across several machines

# Time, document partitioning (PWA)



- Advantages

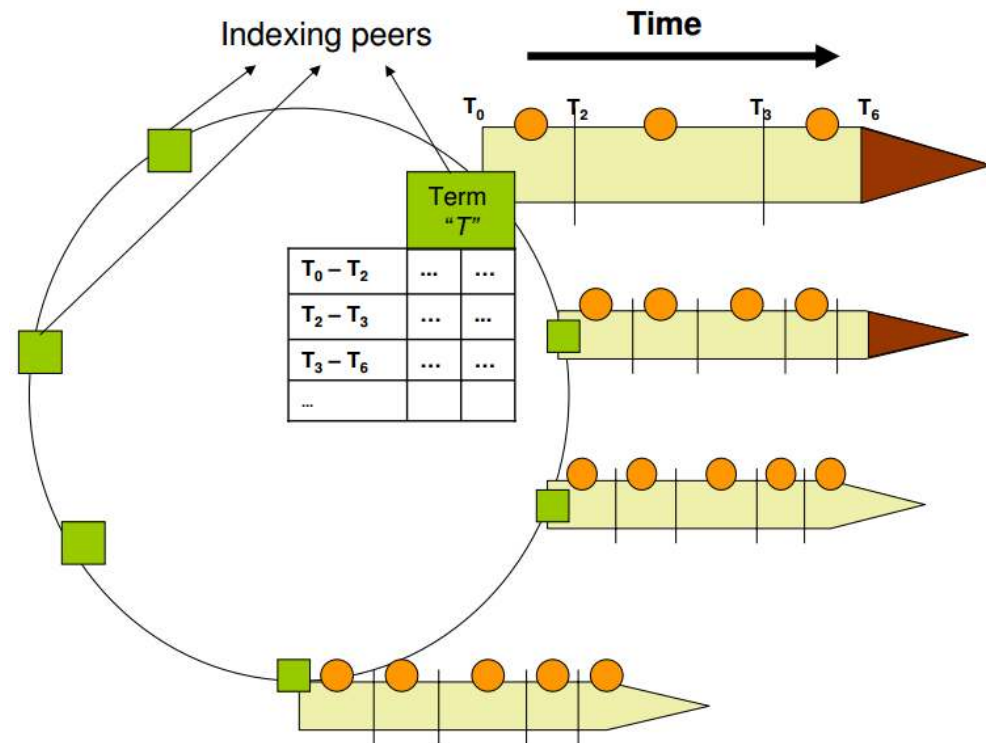
- Selects *time partitions* according to query timespan
- Progressive degradation
  - All computers have all terms
  - One partition fails, remaining respond to query term
- No index rebuilding required to add new collections

- Disadvantages

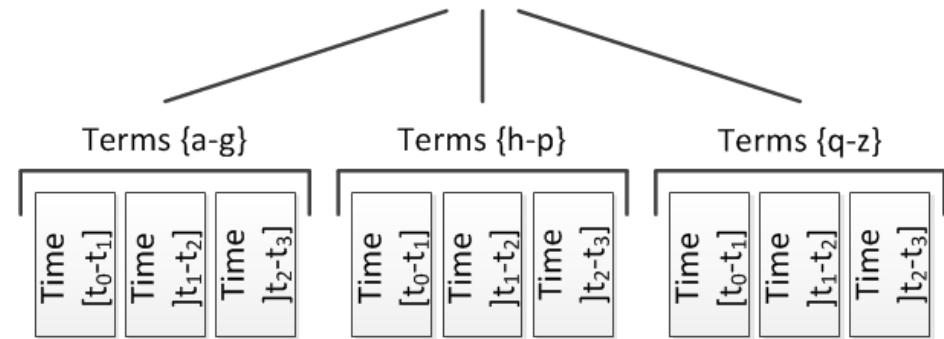
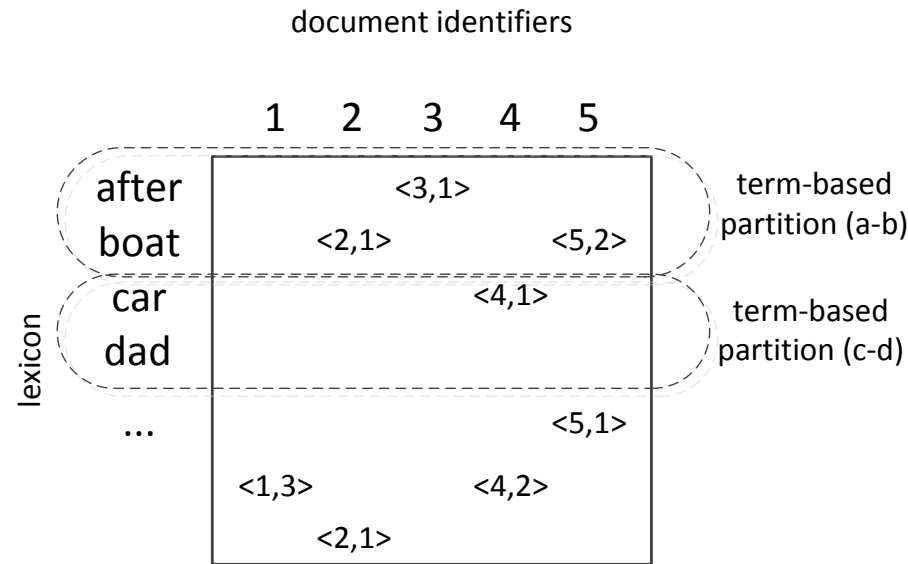
- High workload: all *document partitions* within timespan must be scanned for each query
- Centralized data center approach

# Everlast

- P2P architecture
- Different types of nodes
  - Crawlers
  - Version directories
  - Indexes
- Low cost nodes
- Full-text search
- “Unlimited” scalability
- Tested in laboratory



# Term, time partitioning (Everlast)



- Advantages

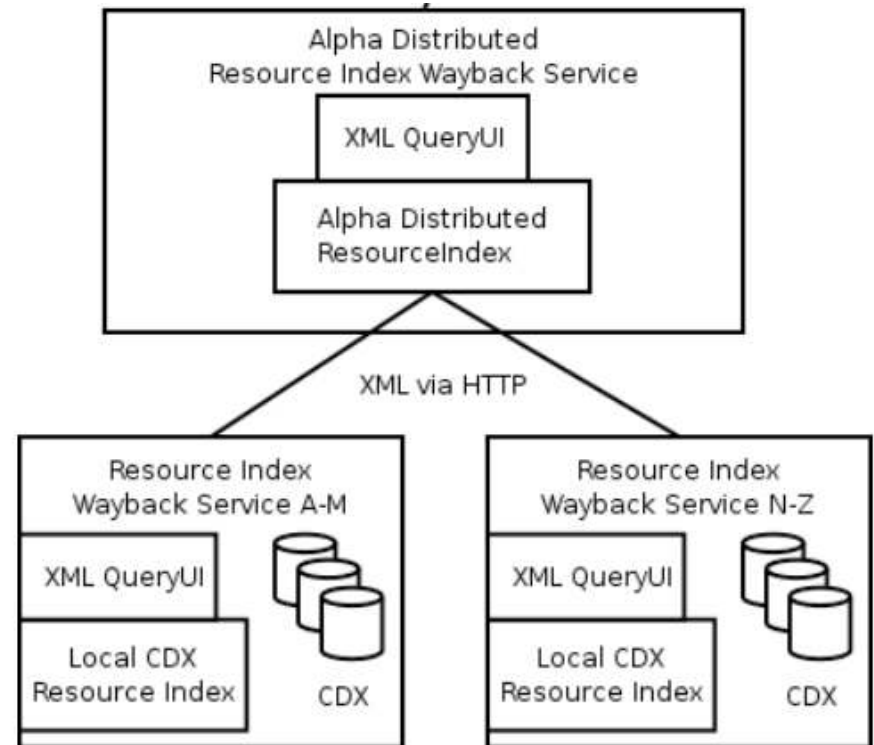
- Robustness of decentralized architecture
- Lower workload: only one *term partition* is contacted for each query

- Disadvantages

- Index updates to add new collections
- *Term partition* unreachable may prevent response to query term
  - Redundancy required
- Latency due to network

# Wayback Machine (URL search)

- Doesn't use inverted indexes
  - Flat sorted files of URLs
- URL partitioning
- Advantages
  - High throughput with millions of queries daily
  - Easy to manage: “no PhD required”
- Disadvantages
  - High communication workload because all queries are broadcasted to all index partitions
  - Limited search features





# Overall comparison

<b>Search requirement</b>	<b>Wayback Machine</b>	<b>Portuguese Web Archive</b>	<b>Everlast</b>
Storage and workload scalability	High	High	Very High
Service reliability	High	High	Medium
Time-aware indexing	No	Yes	Yes
Performance of response times and throughput	High	Very High	Medium

- None is the best, just different.
- Our objective was to improve documentation about web archive search

Food-for-thought

# Problem for **existing** web archives

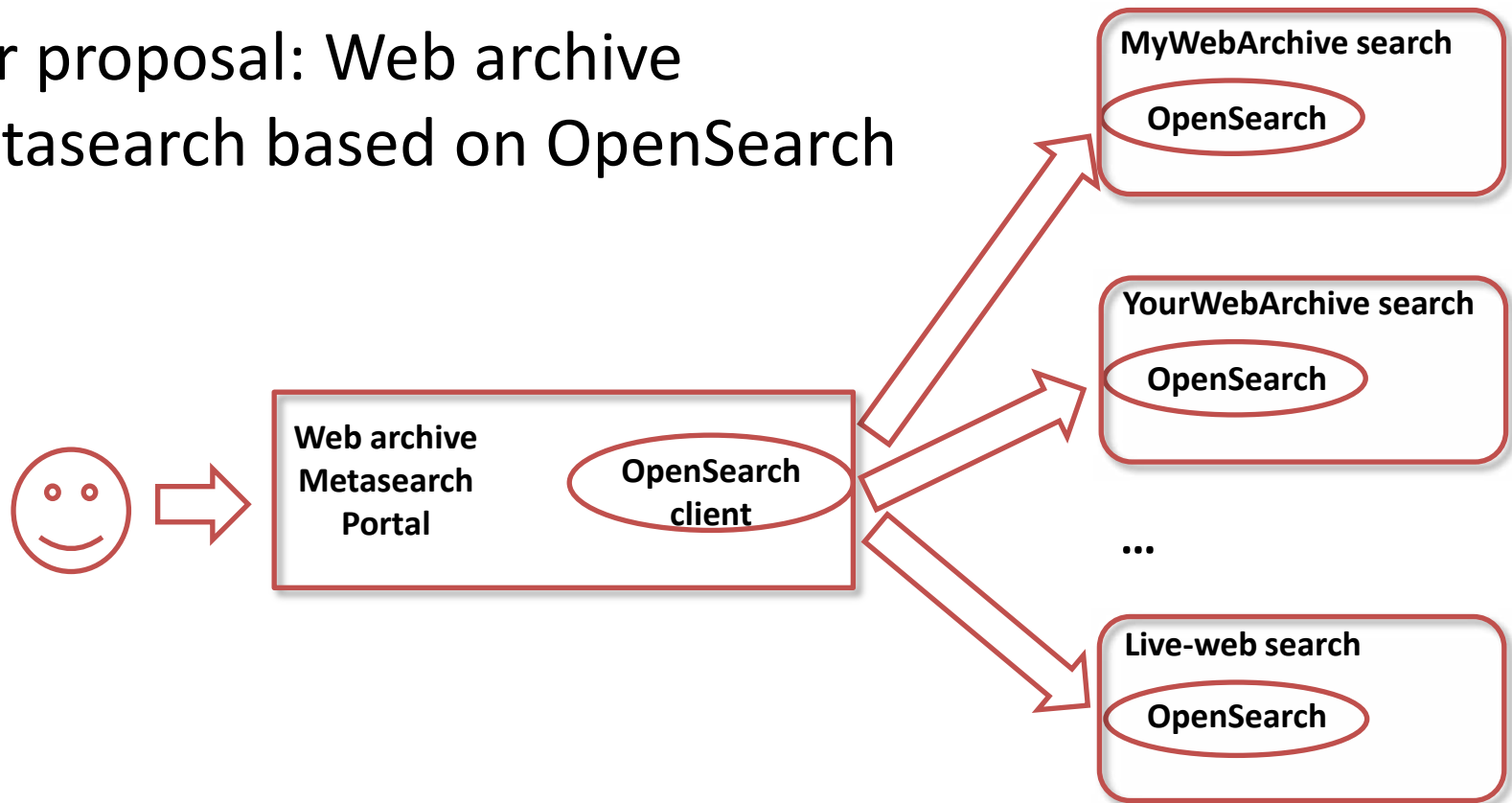
- Users don't know where to search for past web content
  - “Page unavailable” means lost forever
- Dissemination of web archive services is expensive

It would be nice to have a **single** portal for cross-web archive search but...

- Web-archived data is spread
- Search architectures are different
- Search technologies are different
- **Interoperability is required**

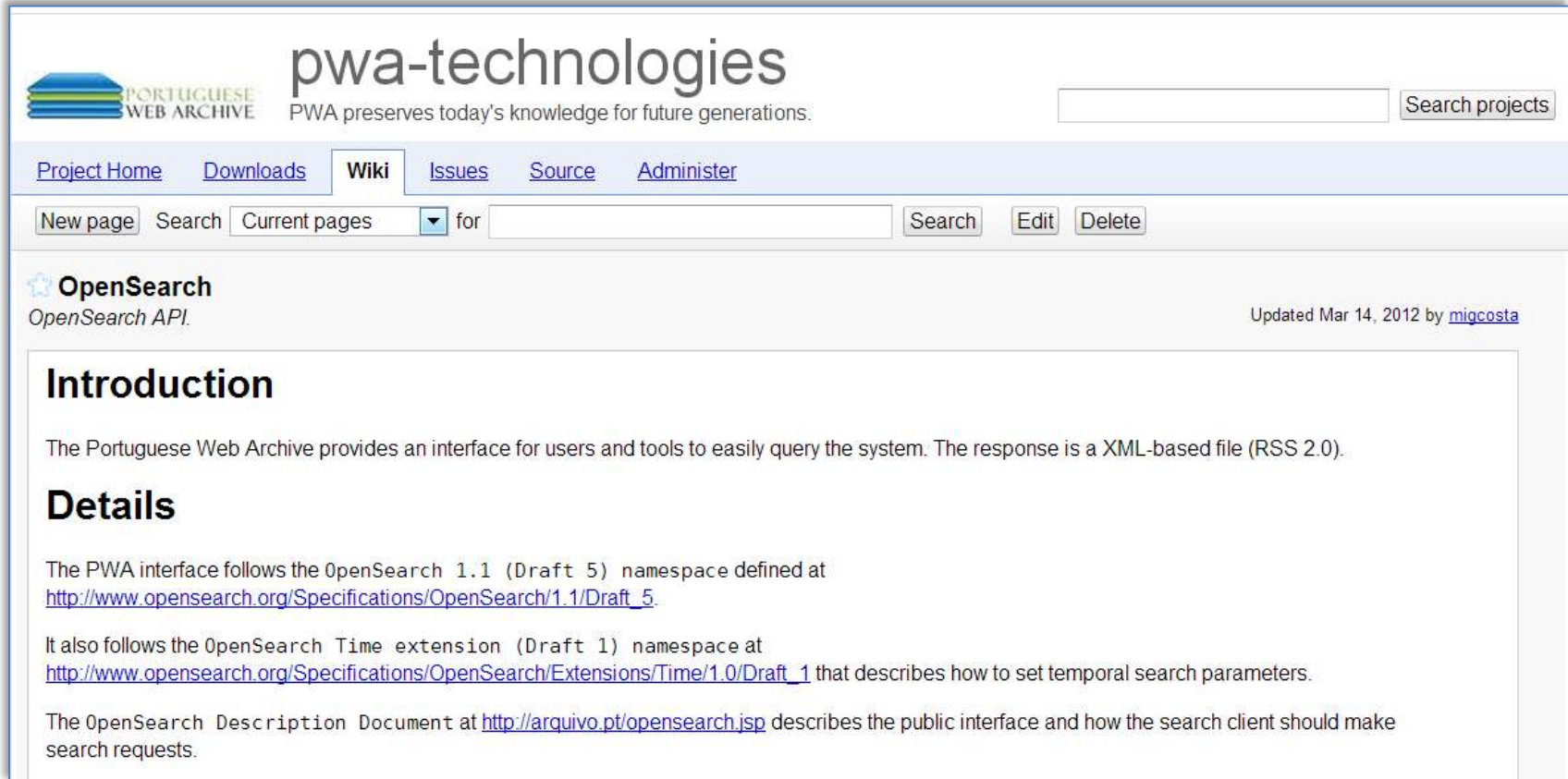
How to design a cross-web  
archive search architecture?

# Our proposal: Web archive metasearch based on OpenSearch



- OpenSearch is a widely supported and simple technology
  - Most web archives use NutchWAX and it supports OpenSearch
- Portal would be simple and cheap to implement
  - Extremely useful to web users
  - Increase visibility of web archiving initiatives
- Easily combines live-web with past-web search results

# Successfully tested by Computer Science students



The screenshot shows the website for pwa-technologies, which is part of the Portuguese Web Archive. The page features a navigation menu with links for Project Home, Downloads, Wiki, Issues, Source, and Administer. Below the navigation, there are search and editing tools, including a search box, a dropdown for current pages, and buttons for Search, Edit, and Delete. The main content area displays the OpenSearch API interface, including an introduction and details sections.

**OpenSearch**  
OpenSearch API. Updated Mar 14, 2012 by [migcosta](#)

## 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

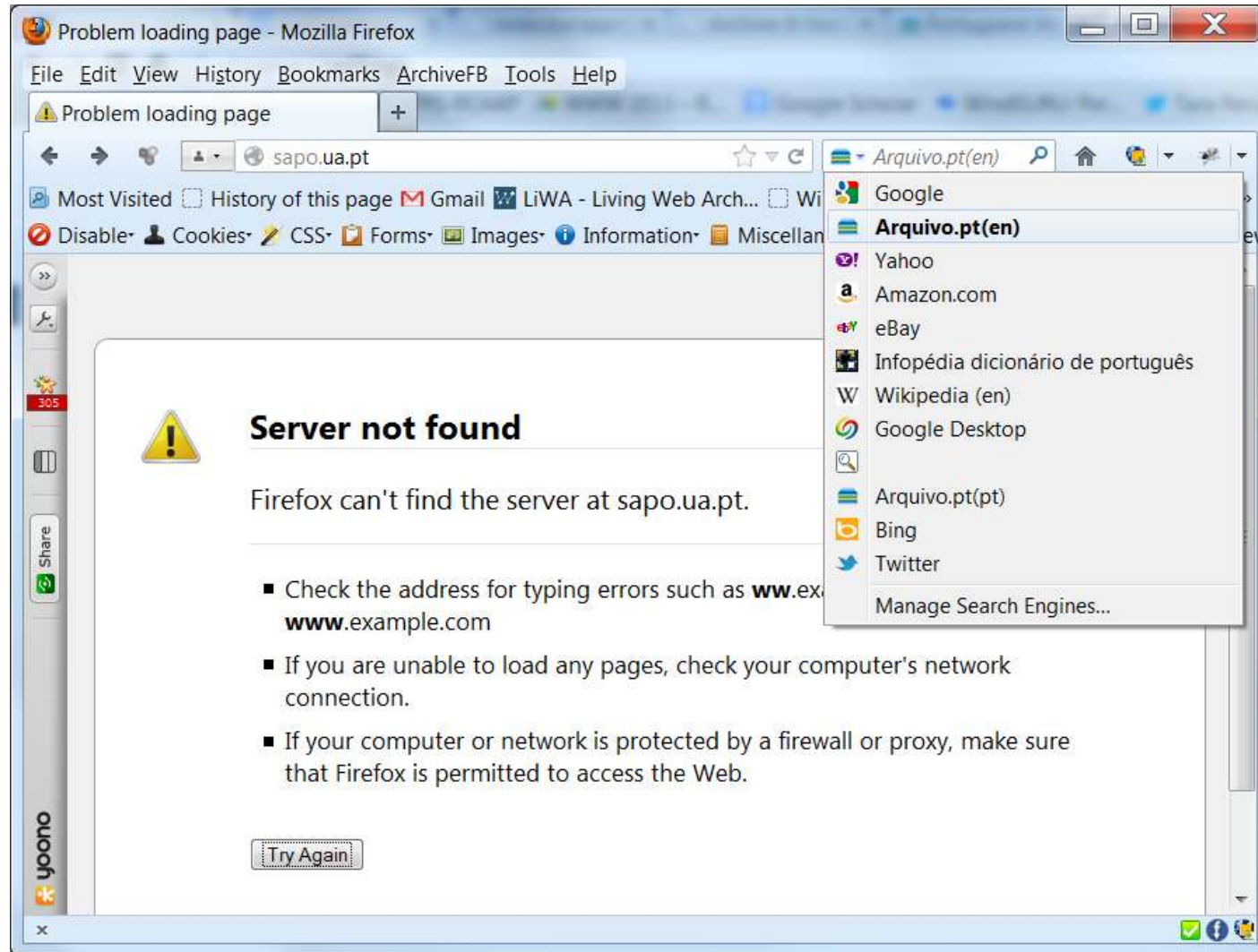
The PWA interface follows the OpenSearch 1.1 (Draft 5) namespace defined at [http://www.opensearch.org/Specifications/OpenSearch/1.1/Draft\\_5](http://www.opensearch.org/Specifications/OpenSearch/1.1/Draft_5).

It also follows the OpenSearch Time extension (Draft 1) namespace at [http://www.opensearch.org/Specifications/OpenSearch/Extensions/Time/1.0/Draft\\_1](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.

- Web applications that gather information about politicians from several sources: Wikipedia, Youtube, Twitter, **Portuguese Web Archive**

# Web archive search easily integrated on web browsers





# Required research to cross-web archive search

- Cross-web archive ranking algorithms
  - How to rank search results?
- User interface design
  - How to adequately present results from different sources?

# Conclusions

- Web archives must support full-text search
- Web archive search architectures are different but search interoperability should be a requirement
- OpenSearch has potential to quickly enable cross-web archive search
  - What do you think?

Contact us whenever  
you like  
Thanks.



[www.archive.pt](http://www.archive.pt)

[daniel.gomes@fccn.pt](mailto:daniel.gomes@fccn.pt)