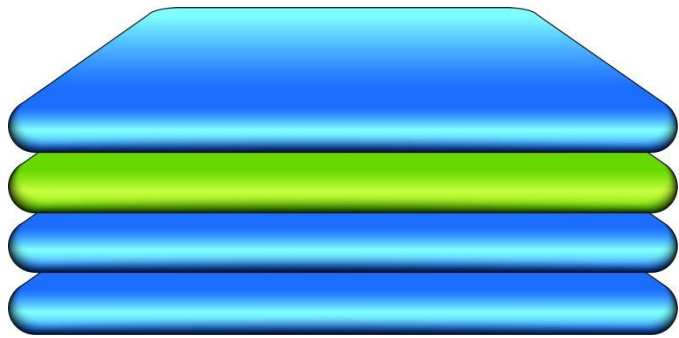


Panel:

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)



PORTUGUESE
WEB ARCHIVE

Lessons learned

The Portuguese Web Archive project started in 2008

[Site Map](#) [Accessibility](#) [Contact](#)

☐ only in current section

[Home](#) [Crawler](#) [Team](#)

You are here: [Home](#) [English](#) [Português](#)

Portuguese Web Archive

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).

Contents

1. [Welcome to the Tomba project: the Portuguese web archive](#)

It provides **version history** like the Internet Archive Wayback Machine



Search the Archive

between: and:

[Advanced search](#)


Did you want to see webpages with the text: <http://www.ul.pt?>

Versions of the archived the Web pages

We archived 347 versions of the Web page <http://www.ul.pt> from 1 January, 1996 and 3 May, 2013.

1996 1	1997 1	1998 4	1999 3	2000 21	2001 12	2002 9	2003 15	2004 76	2005 107	2006 45	2007 39	2008 6	2009 6	2010 2
13 Oct	15 Jul	25 Jan	25 Jan	9 Apr	18 Jan	27 May	2 Feb	12 Feb	2 Jan	1 Jan	3 Jan	15 Feb	25 Jun	9 Jun
		11 Nov	8 Feb	9 Apr	2 Feb	2 Jun	12 Feb	19 May	4 Jan	4 Jan	4 Jan	15 Feb	25 Jun	9 Jun
		6 Dec	30 Apr	10 May	2 Feb	3 Jun	2 Apr	7 Jun	5 Jan	6 Jan	8 Jan	14 Mar	26 Sep	
		12 Dec		10 May	2 Mar	28 Sep	11 Apr	8 Jun	6 Jan	14 Jan	12 Jan	14 Mar	26 Sep	
				11 May	8 Mar	30 Sep	28 May	10 Jun	8 Jan	15 Jan	17 Jan	22 Oct	18 Dec	
				11 May	31 Mar	13 Oct	2 Jun	11 Jun	9 Jan	6 Feb	22 Jan	22 Oct	18 Dec	
				11 May	4 Apr	24 Nov	21 Jun	12 Jun	17 Jan	9 Feb	24 Jan			
				20 May	5 Apr	29 Nov	18 Jul	14 Jun	21 Jan	25 Apr	2 Feb			
				15 Jun	5 Apr	7 Dec	4 Aug	15 Jun	23 Jan	13 Jun	2 Feb			
				15 Jun	5 May		10 Aug	16 Jun	29 Jan	29 Jun	7 Feb			
				19 Jun	16 May		20 Sep	18 Jun	4 Feb	3 Jul	9 Feb			
				20 Jun	6 Oct		1 Oct	19 Jun	5 Feb	5 Jul	16 Feb			
				20 Jun			2 Dec	22 Jun	9 Feb	6 Jul	22 Feb			
				21 Jun			22 Dec	23 Jun	10 Feb	7 Jul	1 Mar			

But also **full-text search** over 1.2 billion web files archived since 1996



Português English Help

US elections

Search the Archive

Advanced search

between: 01/01/1996 and: 31/12/2007

Decemb 2007

Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

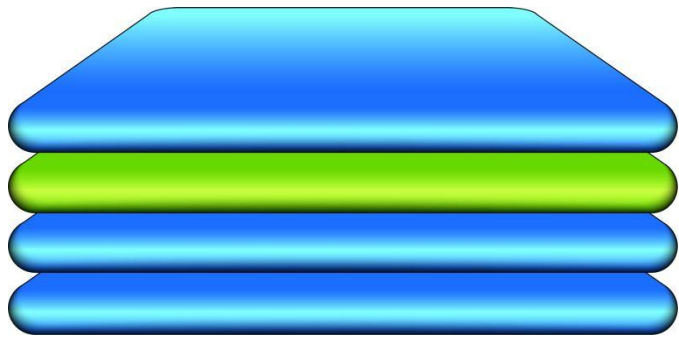
OK Cancel

Results 1 to 10 from 16,627

US Elections
6 January, 2007 - other dates
US Elections The Embassy Consular Service in Lisbon is What direction for America after the 2006 elections become clear, many people are speculating about the future of the U.S. Foreign Policy U.S. Government
http://www.american-embassy.pt/Operations/11/2007/US_mid-term_elections_have_a_clear_message_for_America.html

U.S. Embassy Lisbon, Portugal - Government & Other Links
18 April, 2004 - other dates
Primary and Caucus System in **U.S. Elections** , Foreign Press Center Briefing with Curtis Gans, Director of ...
ASKED QUESTIONS FAQs about **U.S. elections** FAQs About Voting System Standards ELECTION PROCESS
... techniques. March 3, 2004 - African Americans and the 2004 **U.S. Elections** . An Interview with Dr ...
<http://www.american-embassy.pt/elections'04.html>

Die Zeit - Politik : The US Elections: What Europeans expect...
23 January, 2006 - other dates
Die Zeit - Politik : The **US Elections**: What Europeans expect ZEIT.DE » POLITIK » **us-wahl us-wahl** The **US Elections**:
What Europeans expect Von Constanze Stelzenmüller Timothy Garton Ash has called it „the most important American
election in living memory“. Even to those of **us** who don't think that the ...
http://www.zeit.de/2004/45/european_us_wahl



PORTUGUESE
WEB ARCHIVE

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



The screenshot shows the homepage of the Portuguese Web Archive (PWA) website. The header features the PWA logo and the text 'pwa-technologies' and 'PWA preserves today's knowledge for future generations.' Below the header is a navigation bar with links: 'Project Home', 'Downloads', 'Wiki', 'Issues', 'Source', and 'Administer'. A search bar is located on the right side of the header. Below the navigation bar is a section for 'New page' and 'Search' with a dropdown menu for 'Current pages' and a search button. The main content area is titled 'Related Projects' and includes a list of projects and software related to the development of the Portuguese Web Archive search engine. The list includes 'HTTrack2Arc' and 'Roteiro2Arc'.

Related Projects
List of projects and software related to the development of the Portuguese Web Archive search engine

Updated Sep 21, 2011 by [whispsil](#)

Related projects and software

This page list several projects and software developed by or relevant to the Portuguese Web Archive.

Software developed by the Portuguese Web Archive

- [HTTrack2Arc](#) — is a tool that converts HTTrack crawls(<http://www.httrack.com/>) to ARC files, the file format used by the Internet Archive.
- [Roteiro2Arc](#) — is a tool that converts to Internet Archive ARC files the local archive of the Portuguese Web present in the CD-ROM featured with the book "Novo Roteiro Prático da Internet" by José Magalhães.

- “Dead” archived collections became searchable and accessible
- **Specific conversion tools per collection were required but baseline software could be reused**

Oldest Library of Congress site (October 1996)



*The Library
of Congress*
Founded in 1800

Choose a topic below, see [what's new](#), or [search](#) our Web pages and Gopher menus.

[General Information and Publications](#)

Find out about the Library and its mission, special programs and services, information for visitors, publications (including Library Associates and *Civilization Magazine*), employment opportunities, and other general information.

[Government, Congress, and Law](#)

Search THOMAS (legislative information), access services of the Law Library of Congress (including the Global Legal Information Network), or locate government information.

[Research and Collections Services](#)

Browse historical collections for the National Digital Library (American Memory), visit Library Reading Rooms, access special services for persons with disabilities, and read about Library of Congress cataloging, acquisitions, and preservation operations, policies, and related standards.

- **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

The Joomla SEO Book © Alledia Inc. 2007

The default Joomla robots.txt looks like this:

```
User-agent: *  
Disallow: /administrator/  
Disallow: /cache/  
Disallow: /components/  
Disallow: /editor/  
Disallow: /help/  
Disallow: /images/  
Disallow: /includes/  
Disallow: /language/  
Disallow: /mambots/  
Disallow: /media/  
Disallow: /modules/  
Disallow: /templates/  
Disallow: /installation/
```

- **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**

DEDUPPLICATOR



LANDSBÓKASAFN ÍSLANDS
HÁSKÓLABÓKASAFN

Last Published: May 29, 2012

[Heritrix](#) | [Lucene](#) | [SourceForge](#)

DeDuplicator

Welcome

[FAQ](#)
[Releases](#)
[License](#)
[Getting started](#)
[Javadoc](#)

Project Documentation

► [Project Information](#)
► [Project Reports](#)

SOURCEFORGE.NET



The DeDuplicator (Heritrix add-on module)

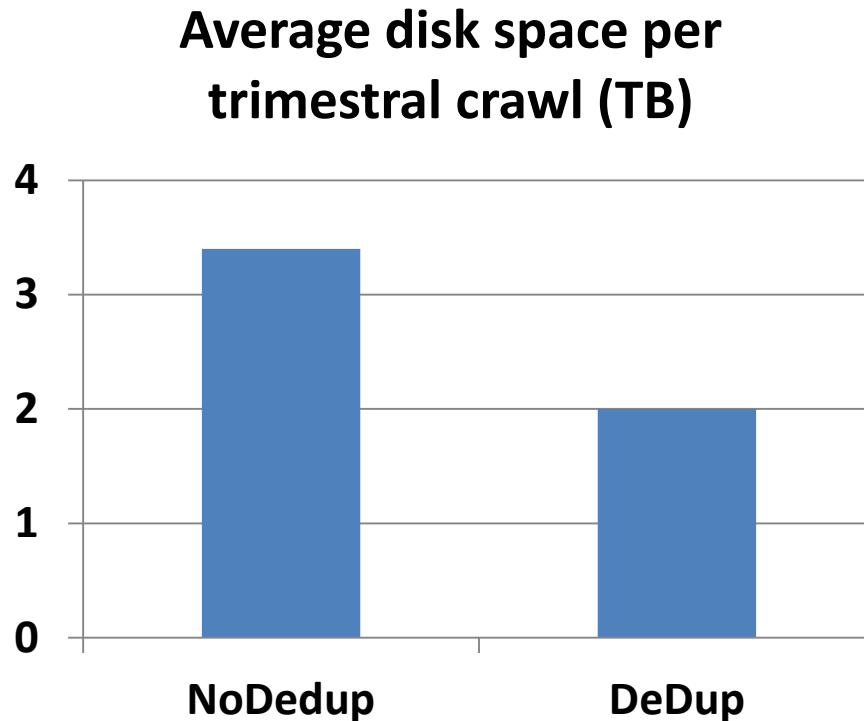
Release information

Current stable release is [0.4.0](#).

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

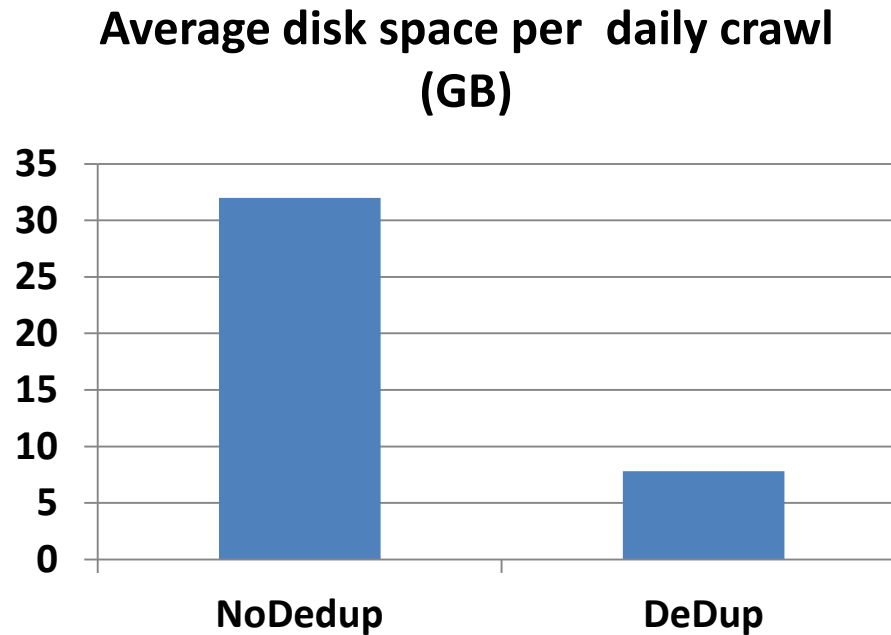
- Downloads content, computes checksum and compares it with version from the previous crawl
 - Unchanged->Discarded
 - Changed->Stored
- **No impact on download rate**

Savings on Trimestral crawls



41% less disk space to store content

Savings on Daily crawls

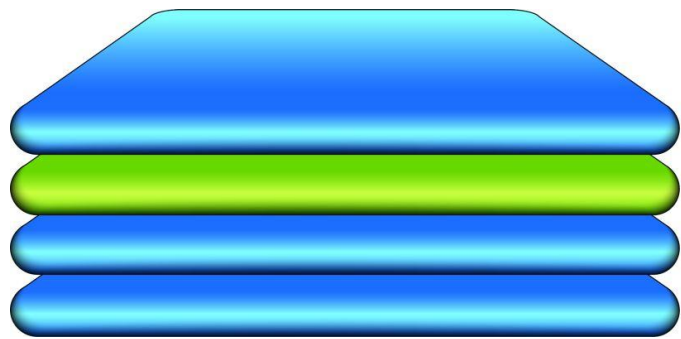


76% less disk space to store content

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



Last Published: 08 Mar 2009

Sourceforge | Heritrix | Archive Access | Internet Archive | Home

NutchWAX
[Home](#)
[Downloads](#)
[Getting Started](#)
[Building from Source](#)
[User Query-time Help](#)
[Regression Test Suite](#)
[Wayback-NutchWAX](#)
[Praxis](#)
[FAQ](#)
Project Documentation
► [Project Information](#)
► [Project Reports](#)
built by:  **maven**

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

 international internet preservation consortium	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.
	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.
	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

Recolha AWP02



ARQUIVO DA WEB
PORTUGUESA

Pesquisa

eleições date:20041204000000-20091204000000 Localizar Help

Search took 10.533 seconds. Resultados 1-10 (de um total de 359.901 documentos):

[Água Lisa \(1\): IRAQUE](#)
» fevereiro 01, 2005 IRAQUE As **eleições** do Iraque terão espalhado desilusões a esmo. Paciência. O anti ... fevereiro 1, 2005 09:05 PM É suposto que as **eleições** sirvam para alguma coisa. É pensável que aqui na Europa se realizassem **eleições** com o quadro existente no Iraque? Penso que estará de acordo comigo ...
<http://agualisa.blogs.sapo.pt/arquivo/471037.html> [html] (10523 bytes) - 2008-04-11 18:01:26 - [other versions](#) - [explain](#)

[EXPRESSO — Notícias, opinião, blogues, fóruns, podcasts. O semanário de referência português.](#)
<http://aeiou.expresso.pt/gen.pl?sid=ex.sections/24895> [html] (108632 bytes) - 2008-03-11 16:33:20 - [other versions](#) - [explain](#)

[Eleições - AlãoQUER](#)
Eleições - AlãoQUER AlãoQUER Aquele que procura a verdade corre o risco de a encontrar « post anterior | home | post seguinte » Terça-feira, 1 de Fevereiro de 2005 **Eleições** O Governo da maioria PSD ... **eleições** com maioria absoluta, o que é realmente importante é saber se o número de deputados que ...
<http://alaoquer.blogs.sapo.pt/8081.html> [html] (16547 bytes) - 2008-04-11 22:12:04 - [other versions](#) - [explain](#)

- 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

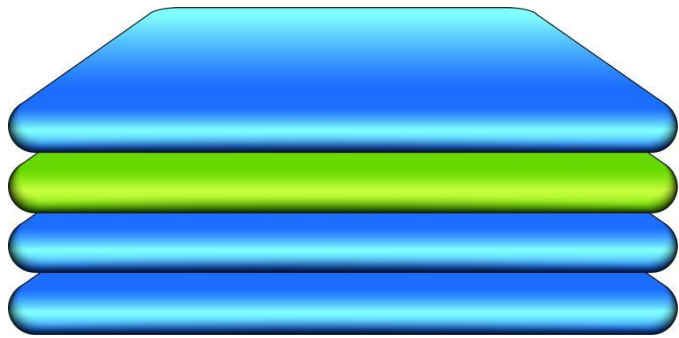
Had to build a Web Archive Information Retrieval Test Collection: PWA9609

- 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

Metric	Time-unaware ranking models	Time-aware ranking models (our proposals)		
	NutchWAX	TVersions	TSpan	MdRankBoost (L2R)
nDCG@1	0.250	0.430	0.450	0.550
nDCG@10	0.174	0.202	0.193	0.555
Precision@1	0.320	0.500	0.520	0.600
Precision@10	0.168	0.172	0.158	0.194

More details: Miguel Costa, Mário J. Silva, [Evaluating Web Archive Search Systems](#), WISE'2012




PORTUGUESE
WEB ARCHIVE

Designing user interface

NutchWAX (2007) vs. PWA (2012)

Recolha AWP02



ARQUIVO DA WEB
PORTUGUESA

Pesquisa

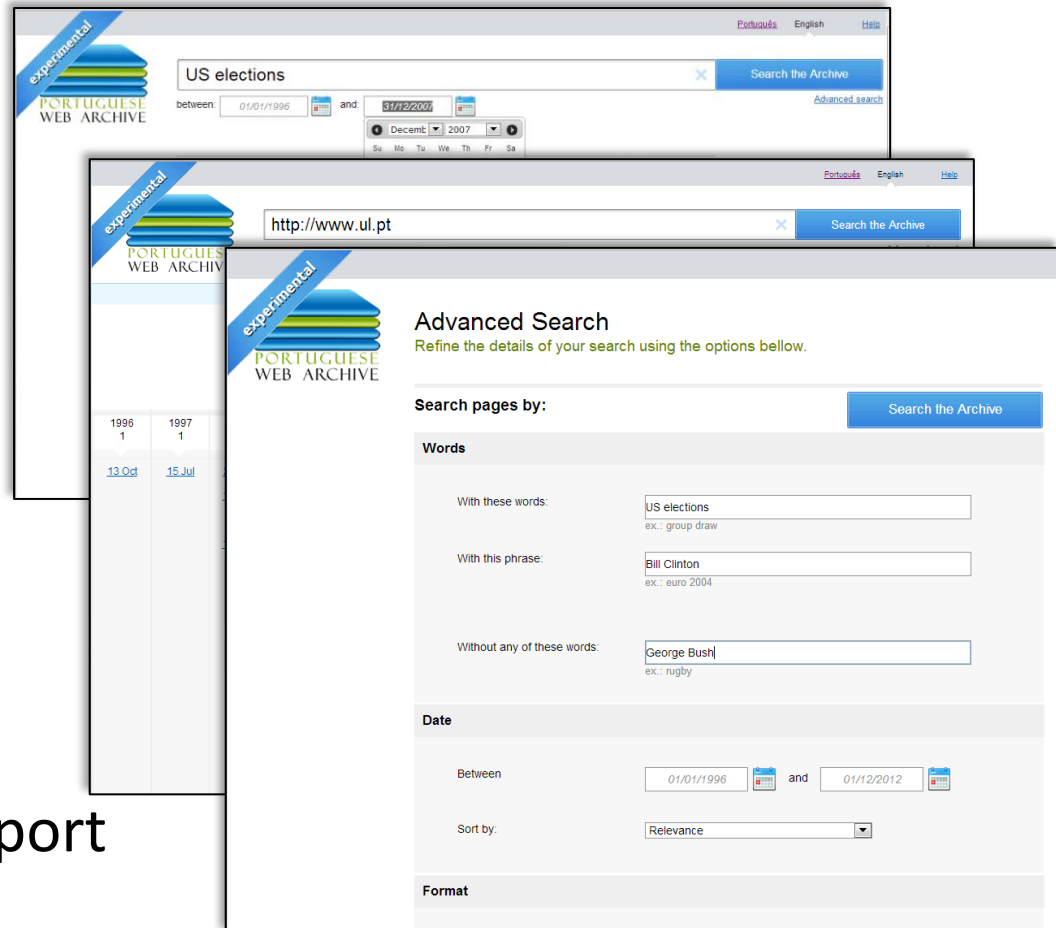
eleições date:20041204000000-20091204000000 Localizer Help

Search took 10.533 seconds. Resultados 1-10 (de um total de 359.901 documentos):

[Água Lisa \(1\): IRAQUE](#)
» fevereiro 01, 2005 IRAQUE As **eleições** do Iraque terão espalhado desilusões a esmo. Paciência. O anti ... fevereiro 1, 2005 09:05 PM É suposto que as **eleições** sirvam para alguma coisa. É pensável que aqui na Europa se realizassem **eleições** com o quadro existente no Iraque? Penso que estará de acordo comigo ...
<http://agualisa.blogs.sapo.pt/arquivo/471037.html> [html] (10523 bytes) - 2008-04-11 18:01:26 - [other versions](#) - [explain](#)

[EXPRESSO — Notícias, opinião, blogues, fóruns, podcasts. O semanário de referência português.](#)
<http://setou.expresso.pt/gen.pl?sid=ex.sections/24895> [html] (108632 bytes) - 2008-03-11 16:33:20 - [other versions](#) - [explain](#)

[Eleições - AlãoQUER](#)
Eleições - AlãoQUER AlãoQUER Aquele que procura a verdade corre o risco de a encontrar « post anterior | home | post seguinte » Terça-feira, 1 de Fevereiro de 2005 **Eleições** O Governo da maioria PSD ... **eleições** com maioria absoluta, o que é realmente importante é saber se o número de deputados que ...
<http://alaoquer.blogs.sapo.pt/8081.html> [html] (16547 bytes) - 2008-04-11 22:12:04 - [other versions](#) - [explain](#)



The image shows two overlapping screenshots of the Portuguese Web Archive (PWA) interface. The top screenshot shows a search for "US elections" with filters for date (between 01/01/1996 and 01/12/2007) and a calendar view. The bottom screenshot shows the "Advanced Search" interface with fields for "Words", "Date", and "Format".

US elections

between 01/01/1996 and 01/12/2007

Search the Archive

Advanced search

Advanced Search
Refine the details of your search using the options below.

Search pages by: Search the Archive

Words

With these words: US elections
ex.: group draw

With this phrase: Bill Clinton
ex.: euro 2004

Without any of these words: George Bush
ex.: rugby

Date

Between 01/01/1996 and 01/12/2012

Sort by: Relevance

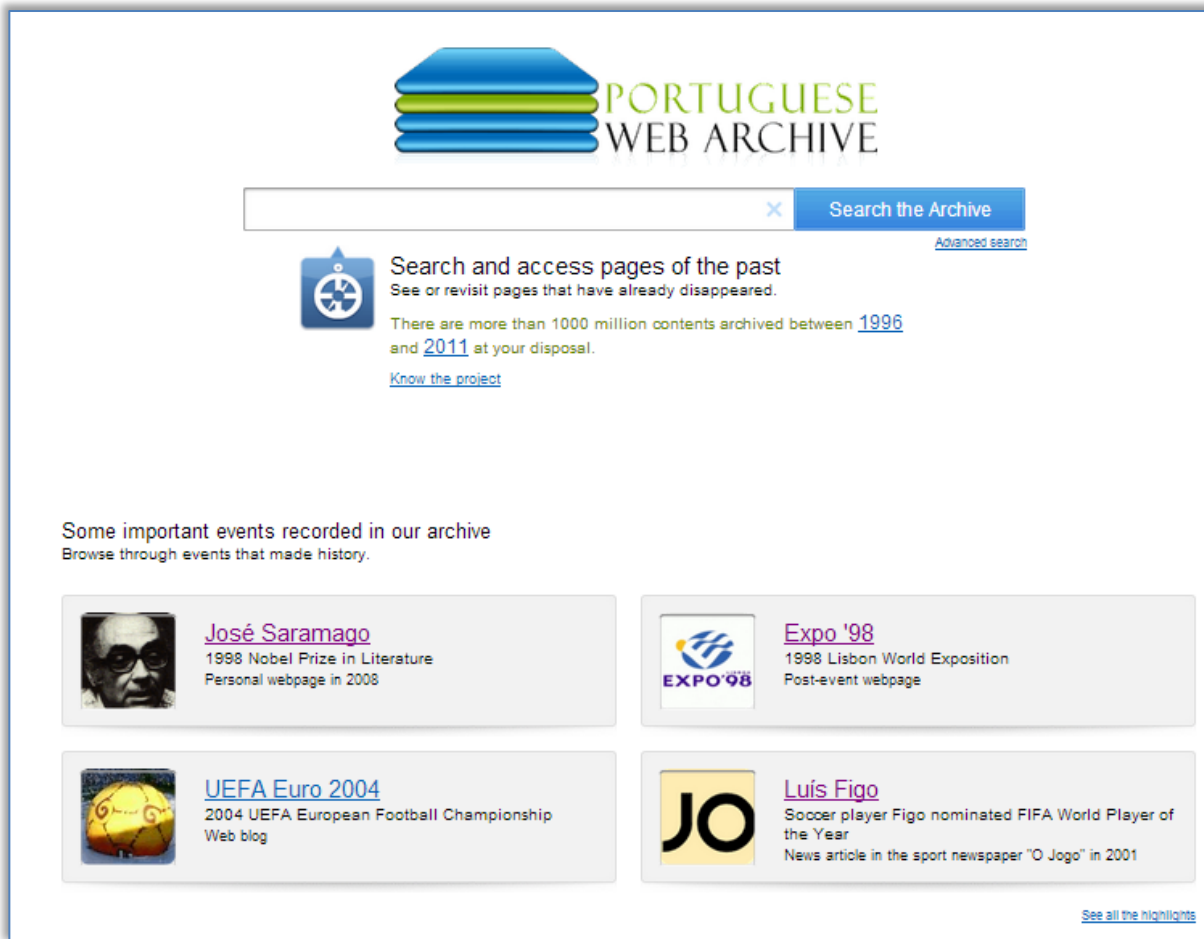
Format

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

Observations from usability testing

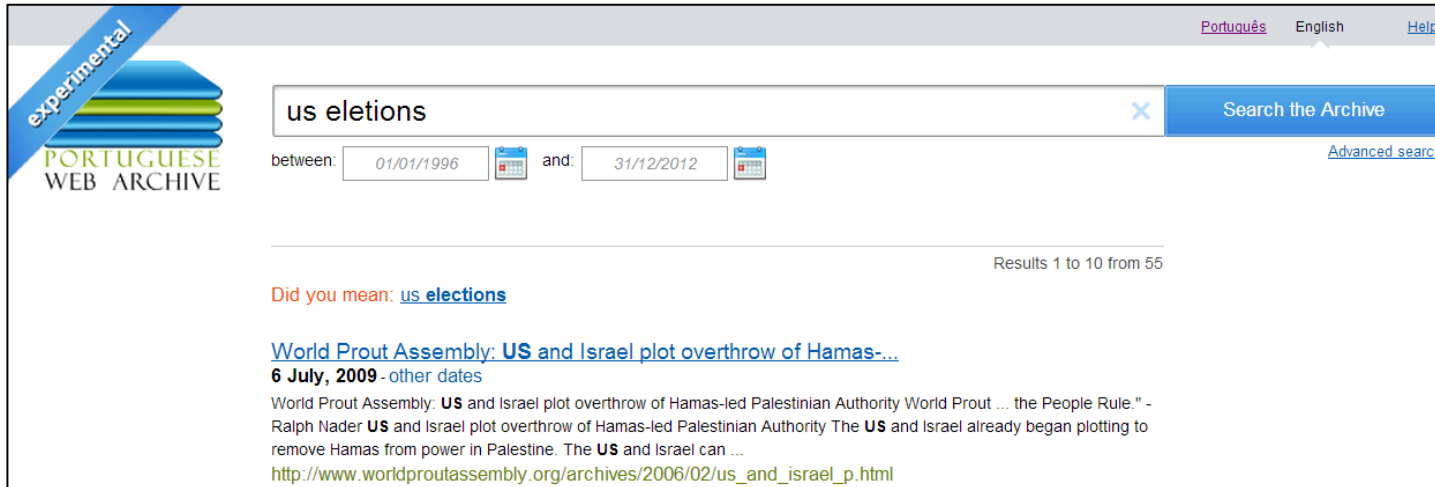


Searching the past web is a confusing concept



- Understanding web archiving requires being techie
- **Must provide examples of web-archived pages**

Users are addicted to query suggestions

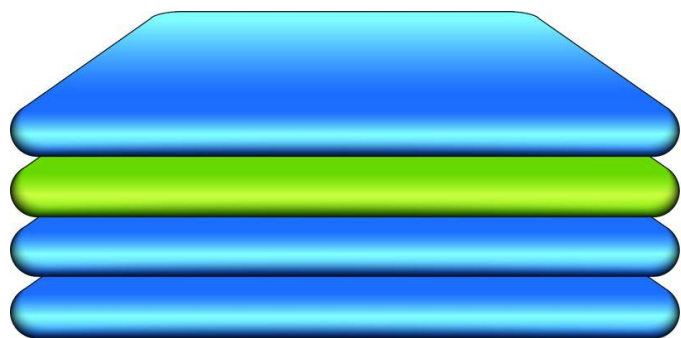


The screenshot shows the Portuguese Web Archive search interface. At the top left, there is a blue banner with the word "experimental" and the Portuguese Web Archive logo. The search bar contains the text "us eletions" (misspelled). Below the search bar, there are date filters: "between: 01/01/1996" and "and: 31/12/2012". To the right of the search bar is a blue button labeled "Search the Archive". Below the search bar, there is a link for "Advanced search". The search results show "Results 1 to 10 from 55". A suggestion is displayed: "Did you mean: [us elections](#)". Below this, a search result is shown with the title "World Prout Assembly: [US and Israel plot overthrow of Hamas-...](#)" and the date "6 July, 2009 - other dates". The snippet of the result reads: "World Prout Assembly: **US** and Israel plot overthrow of Hamas-led Palestinian Authority World Prout ... the People Rule." - Ralph Nader **US** and Israel plot overthrow of Hamas-led Palestinian Authority The **US** and Israel already began plotting to remove Hamas from power in Palestine. The **US** and Israel can ...". The URL of the result is http://www.worldproutassembly.org/archives/2006/02/us_and_israel_p.html.

- 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



PORTUGUESE
WEB ARCHIVE

Hardware

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**

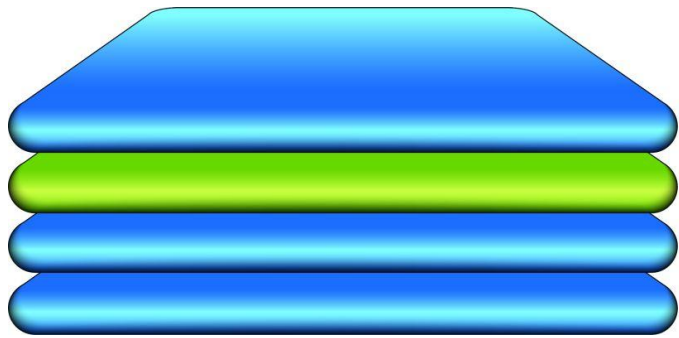


PORTUGUESE
WEB ARCHIVE

Legal issues

Just concerns


- Respect Robots Exclusion Protocol
- 1 year embargo
- Proactively remove illicit content
- Remove content on-demand by authors



PORTUGUESE
WEB ARCHIVE


*Potential as research
infrastructure*

API to process archived data using the PWA Hadoop cluster

**pwa-technologies**
PWA preserves today's knowledge for future generations.

[Project Home](#) [Downloads](#) [Wiki](#) [Issues](#) [Source](#) [Administer](#)

Search Current pages for

 **PwaProcessor**
Portuguese Web Archive's files processor. Updated Jan 3, 2012 by [migcosta](#)

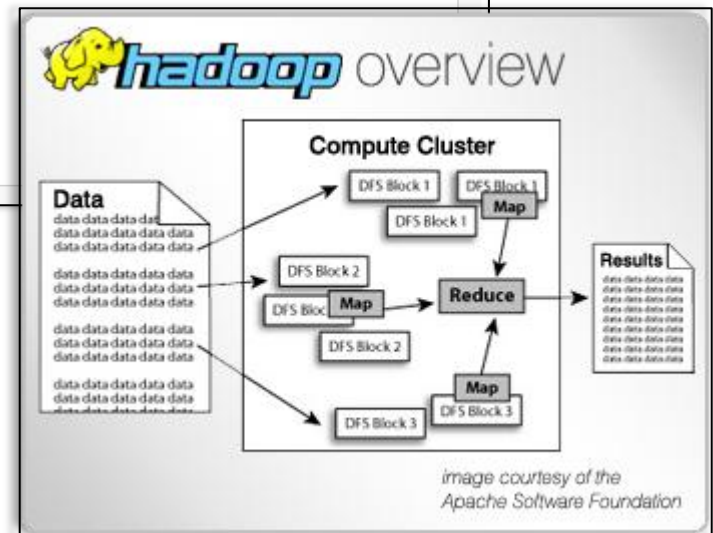
Introduction

This project facilitates the processing of archived files in the ARC format.

Checkout the code:

svn checkout <https://pwa-technologies.googlecode.com/svn/trunk/PwaProcessor>

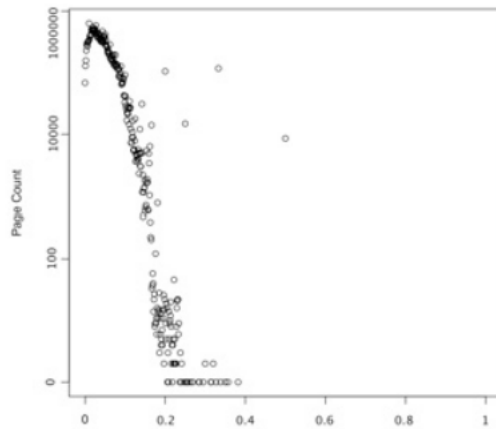
Follow the instructions provided at the README.txt file.



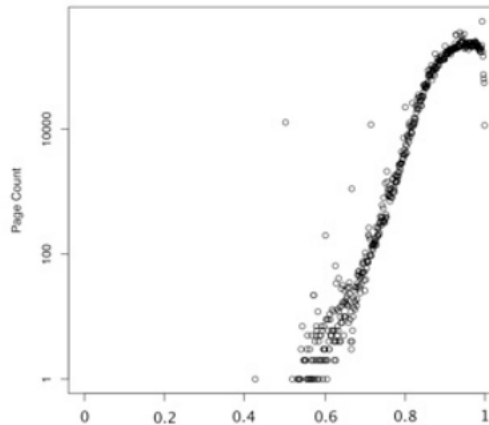
Measure web accessibility for people with disabilities

Results

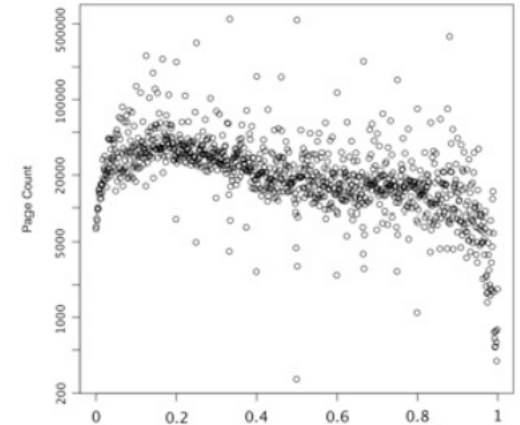
rates versus page count distribution



conservative



optimistic



strict

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

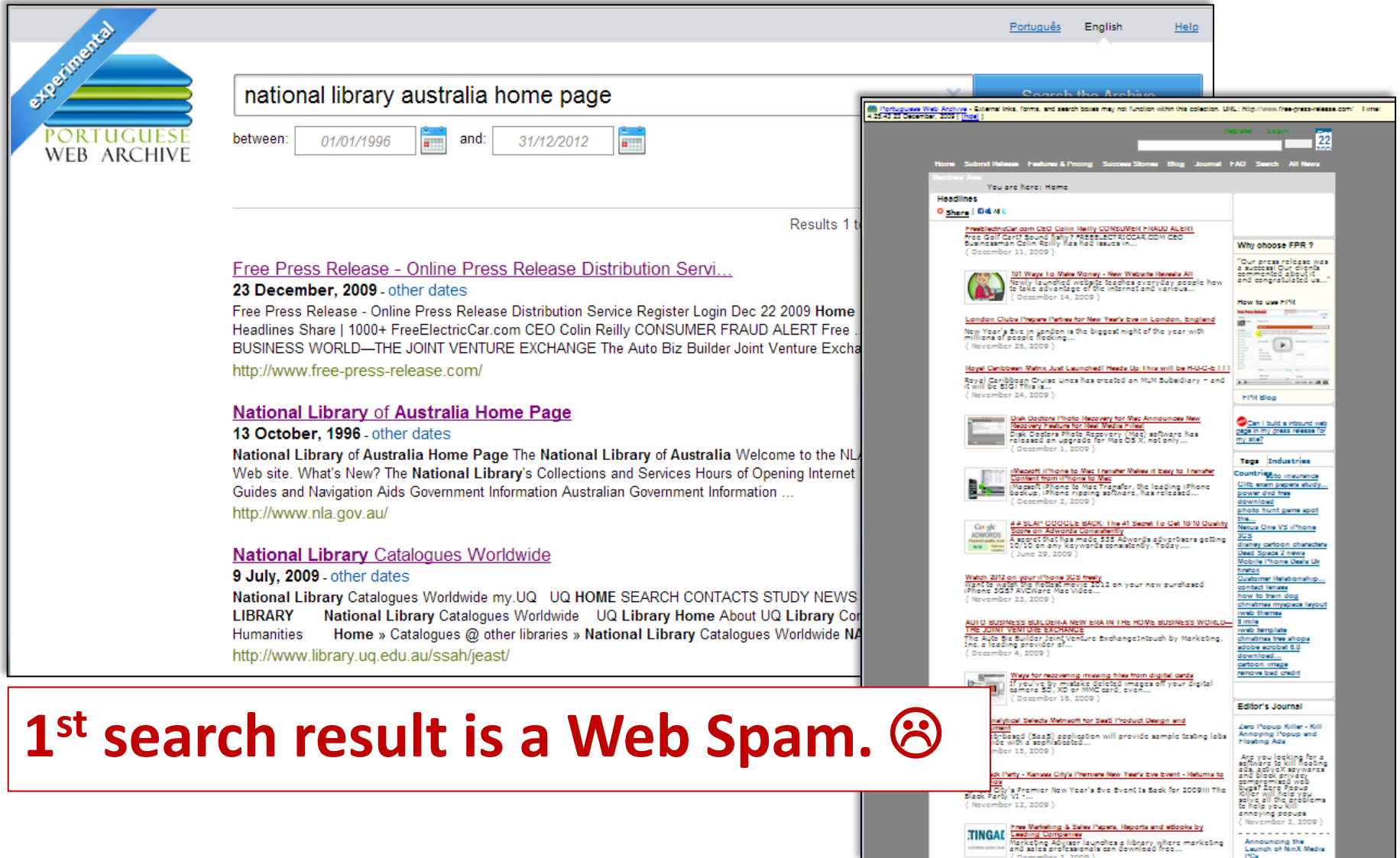
Characterizations of the Portuguese Web structure

The screenshot shows the HERIUX web crawler interface. At the top, it displays the status as of Sep. 6, 2010 07:40:23 GMT, with no alerts. Below this, it indicates 'HOLDING JOBS' with 0 jobs pending and 2 completed. The interface includes tabs for Console, Jobs, Profiles, Logs (selected), Reports, Setup, and Help. Under the Logs tab, there are options to view logs (local-errors.log, progress-statistics.log, runtime-errors.log, uri-errors.log) and filters for line number, time stamp, regular expression, and tail. A 'Refresh time' dropdown is set to 'No refresh', and 'Lines to show' is set to 50. The log content shows a crawl log for default, with entries for 2010-09-01T02:55:19.336Z, 2010-09-01T02:55:19.752Z, and 2010-09-01T02:55:21.720Z, each with a line number and a URL.

Media type	% contents 2005	% contents 2008	Trend
Text/html	61.2%	57.8%	-5.5%
Image/jpeg	22.6%	22.8%	+1.2%
Image/gif	11.4%	9.4%	-17.4%
Text/pdf	1.6%	1.9%	+18.5%
Other	3.2%	8.1%	-

In João Miranda, Daniel Gomes, Trends in Web characteristics, 2009.

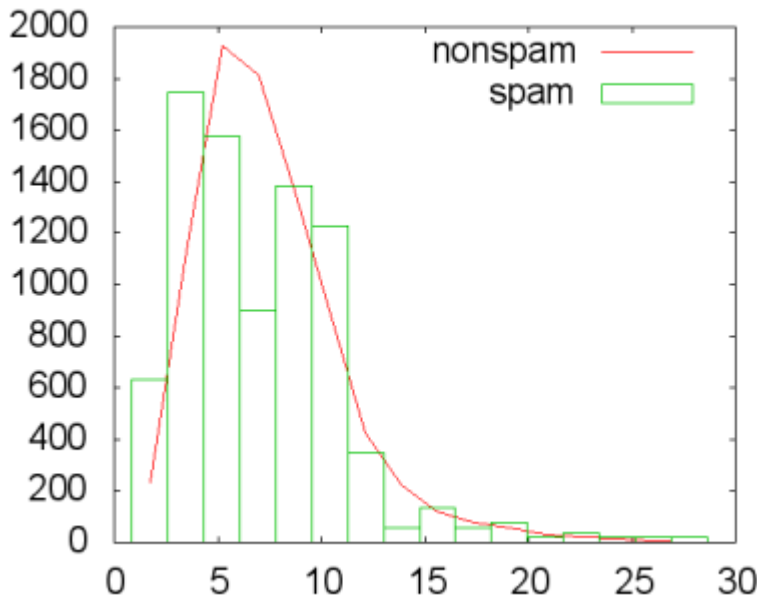
Archiving Web Spam degrades search results



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

[My favorites ▼](#) | [Sign in](#)



pwa-technologies
PWA preserves today's knowledge for future generations.

[Project Home](#) | [Downloads](#) | **Wiki** | [Issues](#) | [Source](#)

Search for

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

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.

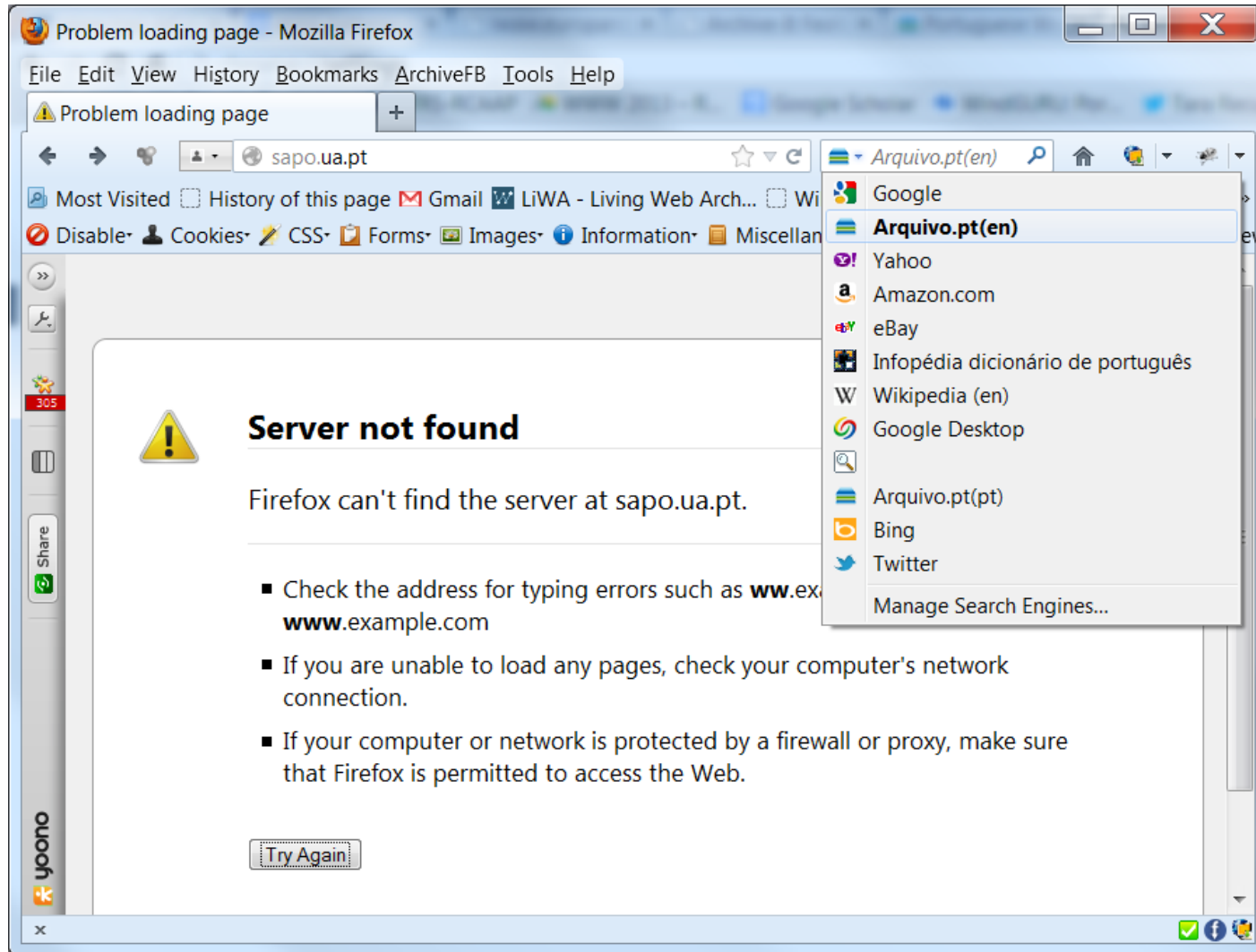
The OpenSearch Description Document at <http://arquivo.pt/opensearch.jsp> describes the public interface and how the search client should make search requests.

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





pwa-technologies
PWA preserves today's knowledge for future generations.

[Project Home](#) [Downloads](#) [Wiki](#) [Issues](#) [Source](#) [Administer](#)

Summary [People](#)


Project Information

 Recommend this on Google

 Starred by 3 users
[Project feeds](#)

Code license
[GNU Lesser GPL](#)

Labels
[Web](#), [Archive](#), [Service](#), [WebArchive](#)

 **Members**
[migcosta](#), [simaofontes](#),
[joaocarvalhomiranda](#),
[danielcoelhogomes](#), [sawfccn](#),
[devel.david@vcruz.net](#), [whispsil](#)

The Portuguese Web Archive (PWA) main goal is the preservation and access of web contents that are no longer available online.

During the developing of the PWA IR (information retrieval) system we faced limitations in searching speed, quality of results, scalability and usability. To cope with this, we modified the archive-access project (<http://archive-access.sourceforge.net/>) to support our web archive IR requirements. Nutchwax, Nutch and Wayback's code were adapted to meet the requirements. Several optimizations were added, such as simplifications in the way document versions are searched and several bottlenecks were resolved.

The PWA search engine is a public service at <http://archive.pt> and a research platform for web archiving. As it predecessor Nutch, it runs over Hadoop clusters for distributed computing following the map-reduce paradigm. Its major features include fast full-text search, URL search, phrase search, faceted search (date, format, site), and sorting by relevance and date.

The PWA search engine is highly scalable and its architecture is flexible enough to enable the deployment of different configurations to respond to the different needs. Currently, it serves an archive collection searchable by full-text with 180 million documents ranging between 1996 and 2010.

[Main features](#)

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. ...