



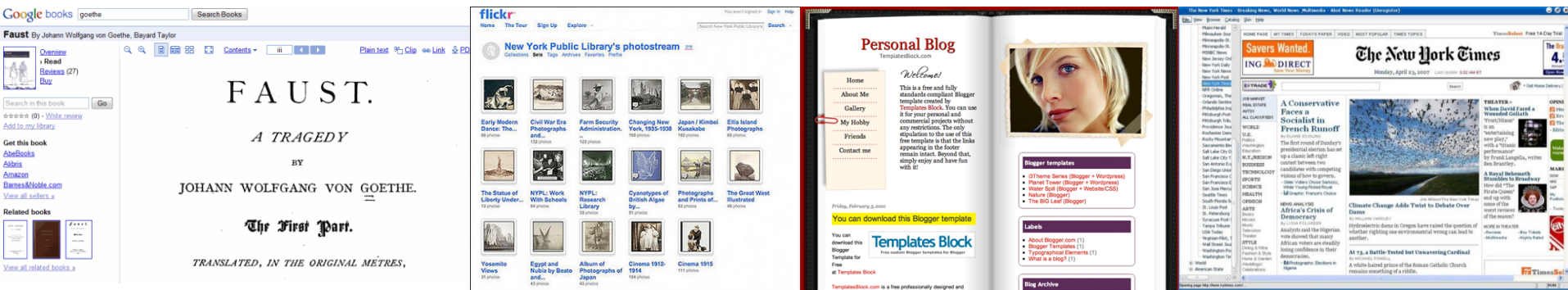
ARCHIVE.PT

# A Memory Infrastructure

By Daniel Gomes

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

# Fact: The Web has been replacing printed media



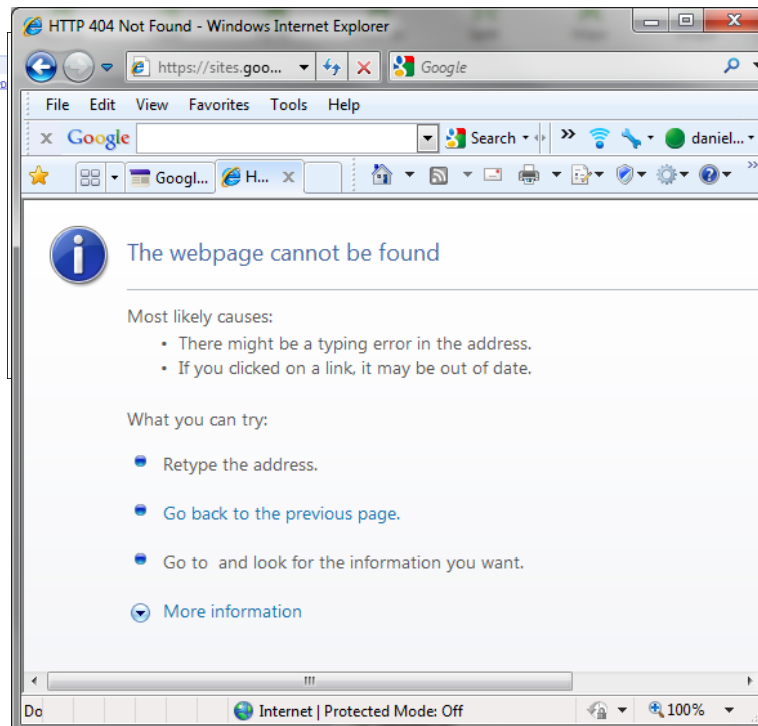
eBooks

Photo  
galleries

Blogs

News

# However, all these valuable information quickly disappears



80%

Disappear or change within  
1 year



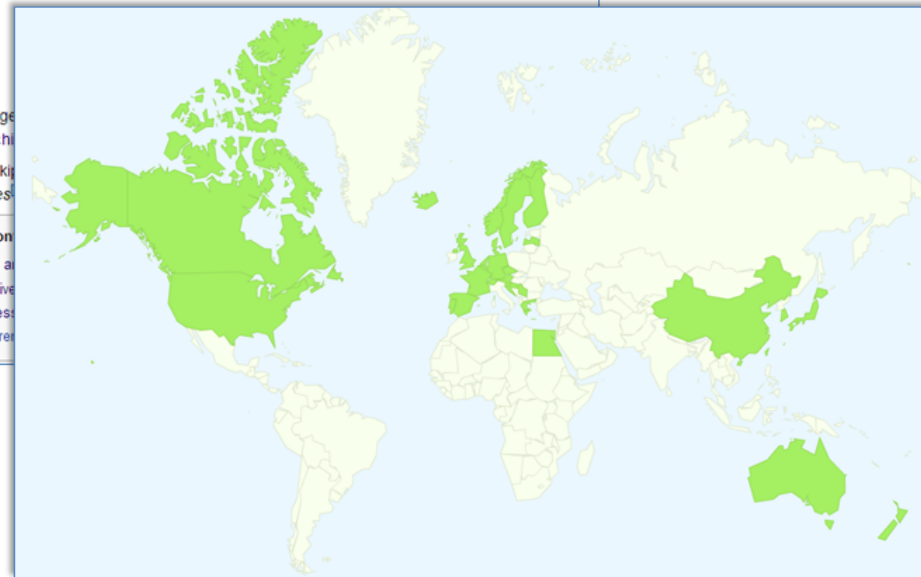
The right to be  
**forgotten** is  
naturally granted  
on the Web



The right to be  
**remembered** on  
the Web is a  
challenge



# 77 web archiving initiatives across the world to preserve Humanity's online memory





Most web archives  
are part  
of the IIPC



INTERNATIONAL  
INTERNET  
PRESERVATION  
CONSORTIUM



ARCHIVE.PT



PORTUGUESE  
WEB ARCHIVE

# 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.

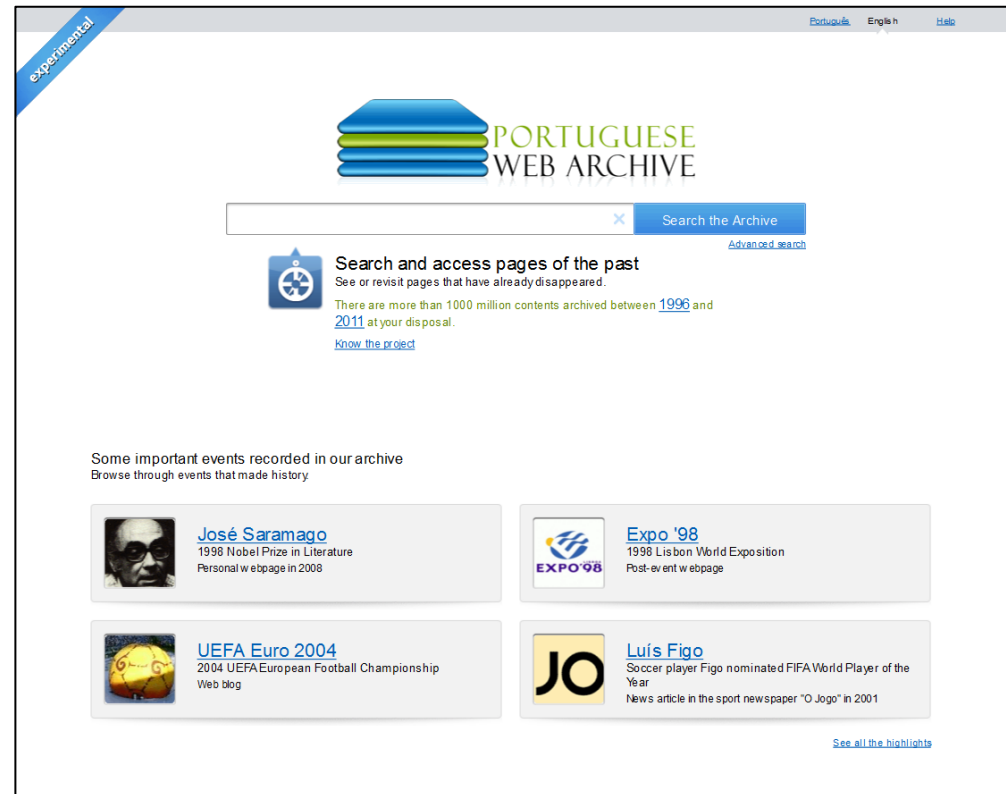
Portugal is now beginning its national web archiving initiative with the Tomba project at [FCCN](#) (National Foundation for Scientific Computing).

#### Contents

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



It was announced  
in 2012

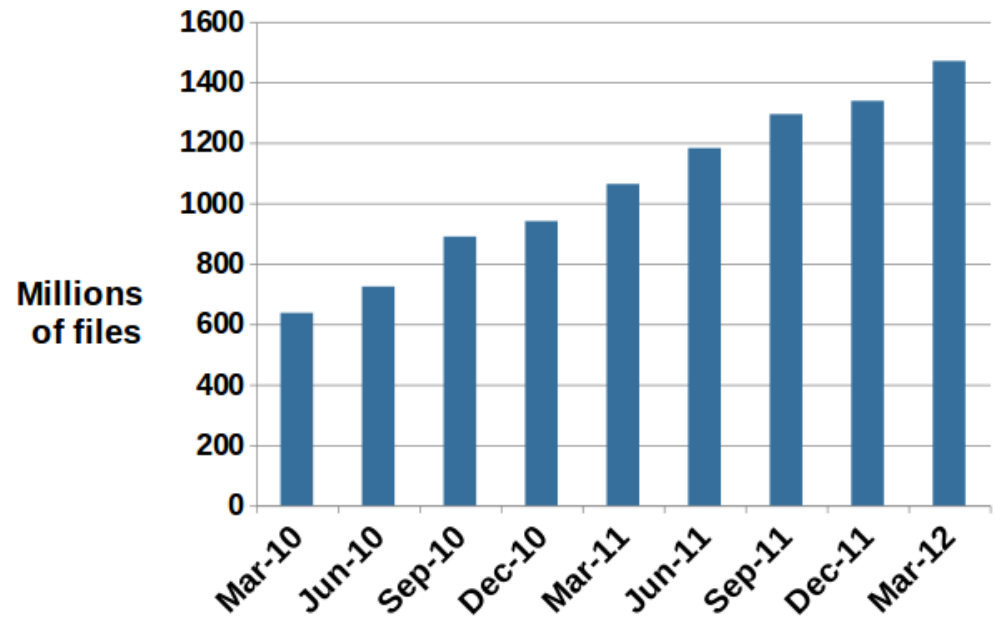


# Archive.pt in a nutshell

- A free preservation service provided to web authors
  - ◆ Only preserves publicly available information
  - ◆ Mainly related to the Portuguese community
  - ◆ Respects imposed exclusions ([robotstxt.org](http://robotstxt.org))
  - ◆ Opt-out option for authors
- A past-web search engine available to web users
  - ◆ A complement for live-web search engines
  - ◆ A “Google” for the past!



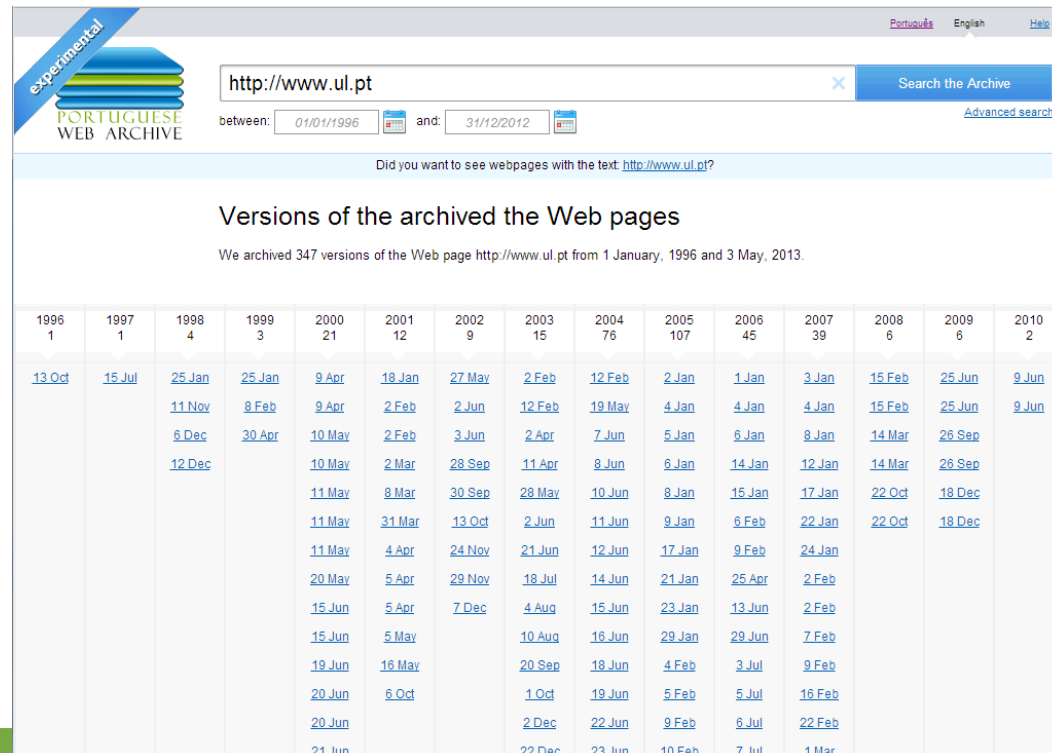
1 700 million archived files  
since 1996 (57 TB)  
...and growing



How can **anyone**  
find web archived  
information?



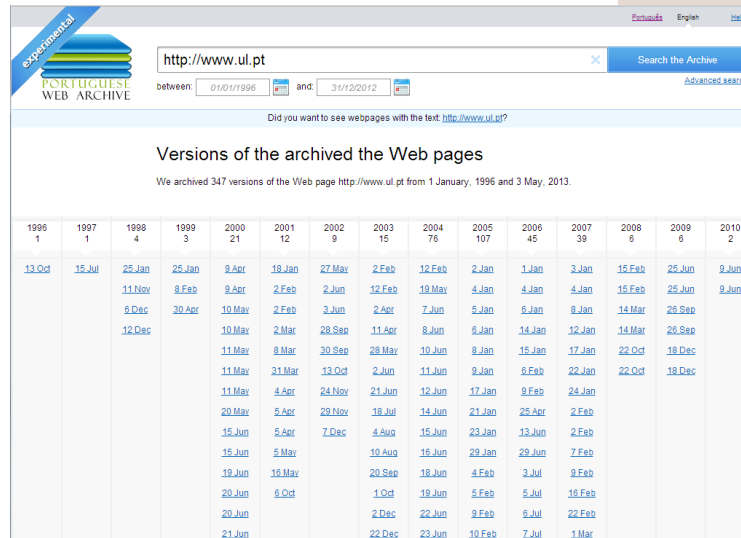
# Using URL search like the Internet Archive Wayback Machine



The screenshot shows the Portuguese Web Archive search interface. At the top, there is a search bar with the URL <http://www.ul.pt> and a "Search the Archive" button. Below the search bar, there are filters for "between:" and "and:" with date ranges. The main content area displays "Versions of the archived the Web pages" for the URL <http://www.ul.pt>. It states that 347 versions of the web page were archived from 1 January, 1996 to 3 May, 2013. A table below shows the dates of the archived versions, organized by year and month.

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

# Problem with URL search



experimental

PORTUGUESE WEB ARCHIVE

http://www.ul.pt

between 01/01/1996 and 31/12/2012

Search the Archive

Advanced search

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

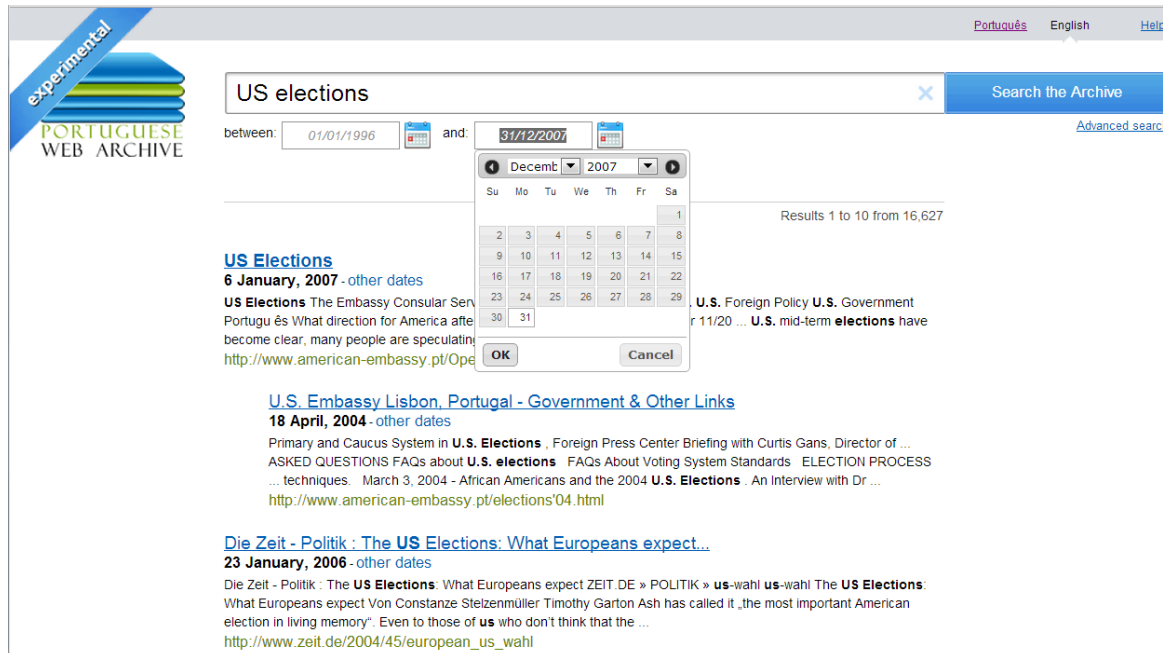
Versions of the archived the Web pages

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 Mar	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 Mar	4 Jan	4 Jan	4 Jan	15 Feb	25 Jun	9 Jun
		6 Dec	20 Apr	10 May	2 Feb	2 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 Mar	8 Mar	30 Sep	28 Mar	19 Jun	8 Jan	15 Jan	17 Jan	22 Oct	18 Dec	
				11 Mar	31 Mar	13 Oct	2 Jun	11 Jun	9 Jan	6 Feb	22 Jan	22 Oct	18 Dec	
				11 Mar	4 Apr	24 Nov	21 Jun	12 Jun	17 Jan	9 Feb	24 Jan			
				20 Mar	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	15 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			

→ Users **do not know** the URL from the past that **contained** the information that they need.

Archive.pt also provides **full-text search**  
over 1.2 billion web files  
archived  
since 1996



The screenshot shows the Archive.pt search interface. At the top, there's a search bar with the text "US elections" and a "Search the Archive" button. Below the search bar, there are date filters: "between: 01/01/1996" and "and: 31/12/2007". A calendar widget is open, showing the month of December 2007. The search results are displayed below the date filters, showing "Results 1 to 10 from 16,627". The first result is titled "US Elections" and dated "6 January, 2007 - other dates". The snippet for this result mentions "The Embassy Consular Service" and "What direction for America after 11/20...". The second result is titled "U.S. Embassy Lisbon, Portugal - Government & Other Links" and dated "18 April, 2004 - other dates". The snippet for this result mentions "Primary and Caucus System in U.S. Elections" and "Foreign Press Center Briefing with Curtis Gans, Director of ...". The third result is titled "Die Zeit - Politik : The US Elections: What Europeans expect..." and dated "23 January, 2006 - other dates". The snippet for this result mentions "Die Zeit - Politik : The US Elections: What Europeans expect ZEIT.DE » POLITIK » us-wahl us-wahl The US Elections: What Europeans expect Von Constanze Steizenmüller Timothy Garton Ash has called it 'the most important American election in living memory'".

experimental  
PORTUGUESE  
WEB ARCHIVE

Português English Help

US elections

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

Search the Archive

Advanced search

Results 1 to 10 from 16,627

**US Elections**  
6 January, 2007 - other dates  
US Elections The Embassy Consular Service  
Portugal és What direction for America after  
become clear, many people are speculating  
<http://www.american-embassy.pt/Open>

**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 Steizenmüller 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](http://www.zeit.de/2004/45/european_us_wahl)



Archive.pt  
is used by  
international  
users

Country / Territory	Sessions	% Sessions
1.  Portugal	88,298	77.57%
2.  Brazil	6,868	6.03%
3.  United States	4,522	3.97%
4.  Japan	3,615	3.18%
5. (not set)	1,609	1.41%
6.  United Kingdom	1,316	1.16%
7.  Germany	1,154	1.01%
8.  France	809	0.71%
9.  Spain	614	0.54%
10.  Angola	488	0.43%

- Archived web pages written in several languages
- Includes all Portuguese speaking domains (.AO, .MZ, .CV) except Brazil
- 22% of the Sessions were not from Portugal

# Web archives.

## For Who? For What?

- Journalist revisits past news
- Historian analyses digital documents
- Professor recovers presentation
- Manager documents project
- Web user recovers page from broken link
- ...
- All of us.


Access  
mechanisms for  
automatically  
processing web-  
archived data





# OpenSearch programming interface access to extend functionality

My favorites ▼ | Sign in

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

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

Search  Current pages ▼ for

**OpenSearch**  
*OpenSearch API.*

Updated Mar 14, 2012 by [migoo...@](#)

## 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.

## Query

The OpenSearch URL must contain the query and search parameters.

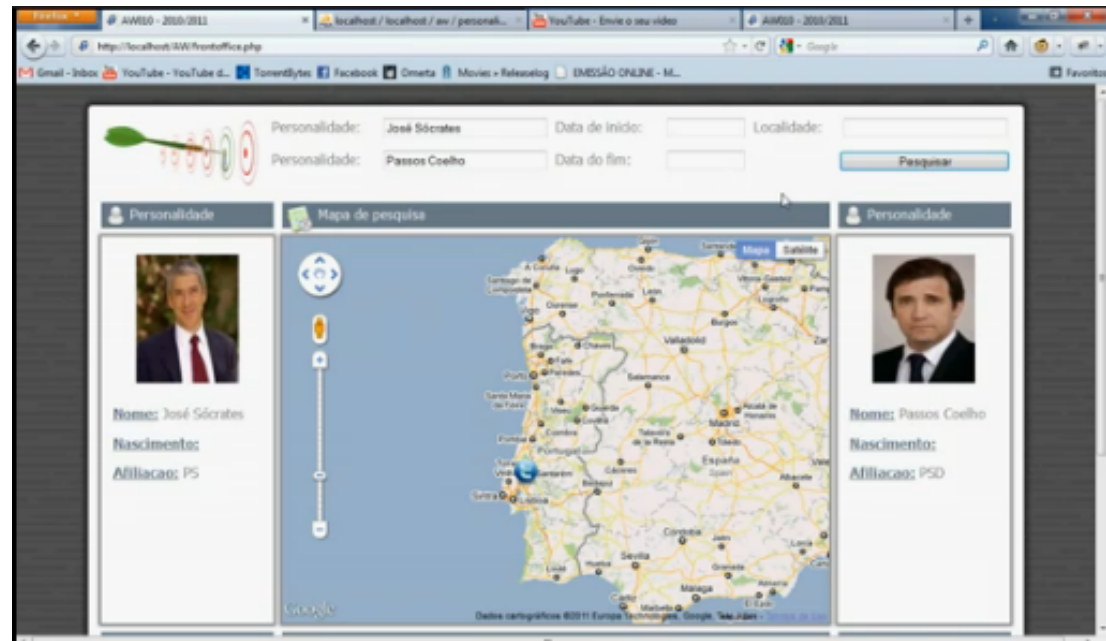
Parameters:

# OpenSearch to develop innovative services over web-archived information

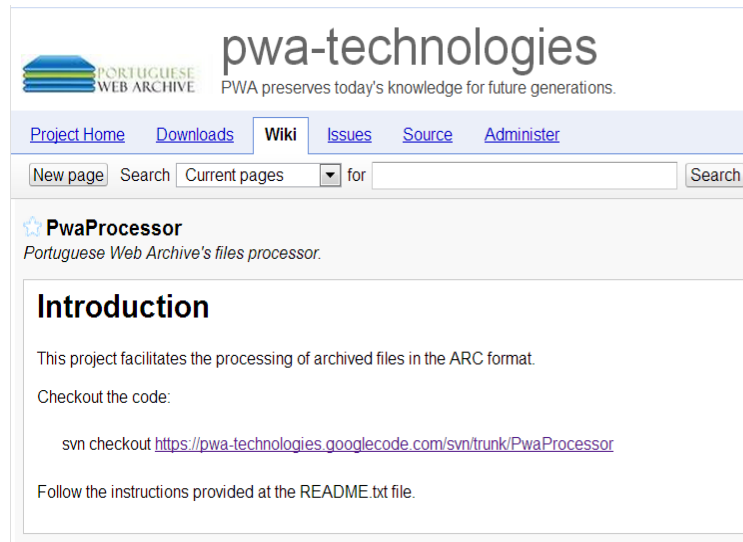
## Web Applications course to develop information

aggregators about:

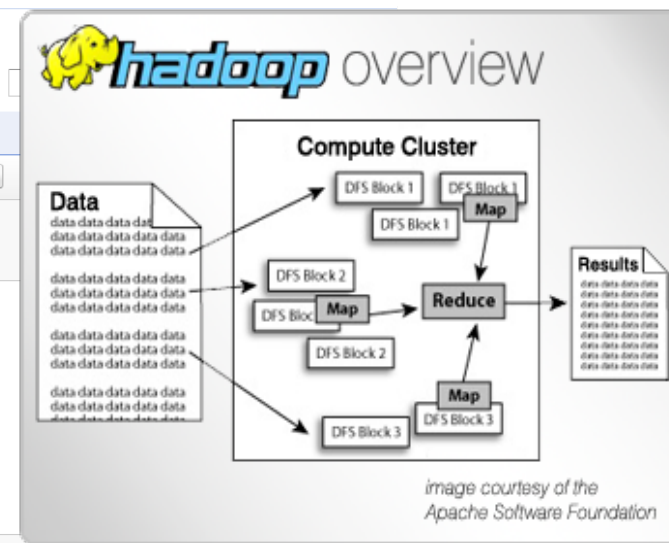
- Politicians (2011)
- Football (2012)



# Big Data analytics over web-archived data through our large-scale processing platform (Hadoop cluster)



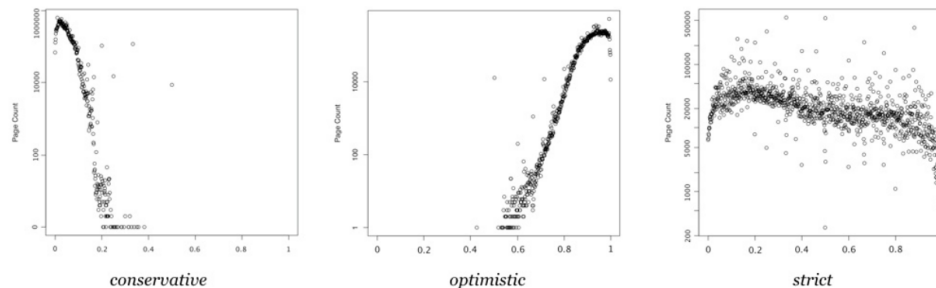
The screenshot shows the website for PwaProcessor, a project by pwa-technologies. The header includes the Portuguese Web Archive logo and the tagline "PWA preserves today's knowledge for future generations." Navigation links include Project Home, Downloads, Wiki, Issues, Source, and Administer. A search bar is present. The main content area is titled "PwaProcessor" and "Portuguese Web Archive's files processor." It features an "Introduction" section stating: "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."



Hadoop cluster used to automatically measure the accessibility of the Web 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.



ARCHIVE.PT

# Which kind of information can we find on Archive.pt?

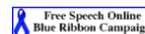
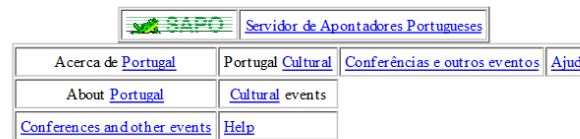
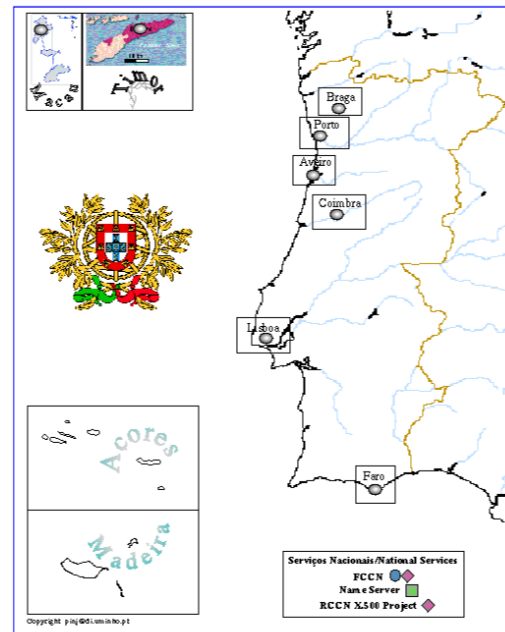
# Everything.

## Examples of web-archived pages

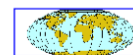


# The first Portuguese web page (90's)

«Home Page» de Portugal / Portugal Home Page



Se não possuir suporte gráfico, prim a [aqui](#) / If you don't have graphics support, click [here](#)



Europe Home Page EC Home Page World Home Page

# International events

2011 Egyptian  
revolution: a web-  
based revolution, a  
web-archived  
revolution



# Regional events

## Results for the 2001 Portuguese elections





# Open access Science: publications, test collections, data sets, tools. (1998-2010)

Linguateca project:  
resources for the  
computational  
processing of the  
Portuguese language

## Linguateca

[Structure](#)  
[Team](#)  
[Overview](#)  
[Access to resources](#)  
[Evaluation contests](#)  
**Catalogue of resources**

- [Corpora](#)
- [Lexica](#)
- [Encyclopedias](#)
- [Ontologies](#)
- [Services](#)
- [IR collections](#)
- [Course materials](#)
- [Texts in Portuguese](#)
- [Media](#)
- [Other](#)

[Catalogue of tools](#)  
[Catalogue of actors](#)  
[Catalogue of publications](#)  
[Interesting information](#)  
[FAQ](#)  
[Gramateca](#)

[Português](#)  
[Norsk](#)



Comments and  
suggestions

Please use the navigation options: on the left  
handside frame (clicking on the links: on the right  
handside does not update the left handside).

## Portuguese language resources

[Linguateca](#)

Last updated 24 July 2003

This page gives you an overview of Portuguese language resources mentioned, or directly accessible, on the Internet.








The aim is to make the resources available to everyone involved in the computational processing of Portuguese, and also to facilitate communication between the various developers and users of the resources.

We take the opportunity to thank everyone who takes the effort to contribute to this site, by providing links to other sites and/or offering resources that we in turn may make accessible.

### Conventions

All the references are presented in alphabetic order, using the names and language chosen by the respective authors.

#### Symbols used in these pages

	Resources available in CD format
	Resources available in floppy disk
	Searchable <i>online</i> resources
	Downloadable <i>online</i> resources
	Resources available <i>online</i> , fee required for the service.
	Text integrally available
	Audio resources

Go to the [Portuguese page](#) to see all the links.

[Send questions, comments and suggestions](#)

# Previous governmental portals (2011)

New  
government  
erased previous  
content



The official gazette of the Republic of Portugal is published exclusively online since 2006 (e.g. DRE.pt contains all the Laws)

- 4 copies **exclusively** prepared and **printed for preservation**
- Printed documents are usually digitized and put online for preservation?!

Define o Novo Regime de Publicação Exclusivamente Electrónica do *Diário da Assembleia da República* e Novas Regras para o Uso de Novas Tecnologias de Informação e Comunicação no Trabalho Parlamentar

Resolução da Assembleia da República n.º 68/2003, de 7 de Agosto

A Assembleia da República resolve, nos termos do n.º 5 do artigo 166.º da Constituição, o seguinte:

**Artigo 1.º**

***Diário da Assembleia da República***

1 - A partir de 15 de Setembro de 2003, a 1.ª série do *Diário da Assembleia da República* passa a ser exclusivamente disseminada em formato electrónico através do portal da Assembleia da República na Internet.

2 - A 2.ª série do *Diário da Assembleia da República* deixará igualmente de ser publicada em suporte tradicional, devendo ser adoptadas todas as medidas necessárias para que a respectiva publicação electrónica integral ocorra no mais curto prazo.

3 - A edição electrónica do *Diário da Assembleia da República* faz fé plena e a publicação dos actos através dela realizada vale para todos os efeitos legais e regimentais, devendo ser utilizado mecanismo que assinala, quando apropriado, a respectiva data e hora de colocação em leitura pública.

4 - Os serviços preparam, editam e depositam na Biblioteca da Assembleia da República e na Biblioteca Nacional quatro exemplares de uma versão impressa das duas séries do *Diário*, preparada unicamente para tal efeito.

Web archived  
DRE.pt  
since 2004



The screenshot shows the homepage of the Diário da República Eletrónico (DRE.pt). At the top is a green and red banner with the Portuguese coat of arms and the text "DIÁRIO DA REPÚBLICA ELECTRÓNICO". Below this is a large central area with a faint watermark of the coat of arms. To the left, under "Acesso Gratuito", are links for "Imagens", "Sumários", "Datas de Distribuição", "Sumários por e-mail", and "Diplomas Numerados". To the right, under "Acesso por Assinatura", are links for "Legislação", "Actos da Administração Pública", "Concursos Públicos", and "Actos Societários". At the bottom, there are sections for "Portal Jurídico", "União Europeia", "Jornais Oficiais", "Novidades e Informações", "Buscas no D.R.E.", and "Língua Oficial Portuguesa".



INCM

A consulta das Bases de Dados do DRE não dispensa a consulta do D.R. original

Linha Azul: 808 200 110 | e-mail: [dre@incm.pt](mailto:dre@incm.pt) | [Precos para 2004](#)

Copyright © 1997-2004 - DRE

# Printed publications are also web archived

→ Newspaper front pages on 24 May 2009

## Em Destaque

### Capas dos Jornais de Hoje



[Diário de Notícias](#)



[Global](#)



[24 Horas](#)



[Metro](#)



[Destak](#)



[O Primeiro de Janeiro](#)



[Diário Económico](#)



[Jornal de Negócios](#)



[Jornal de notícias](#)



[Público](#)



[Correio da manhã](#)



[Jornal A Bola](#)



[Record](#)



[O Jogo](#)

Old photos from the  
pre-Digital Era that  
were published  
online (1966)



Praça do Comércio  
10 de Junho de 1966



# Historical documents with added value (meta-data)

Blog post with digitization of the republican newspaper (1915)



SEGUNDA-FEIRA, JULHO 26, 2010

A DEFESA DA REPÚBLICA - in A CORJA

ALMANAQUE REPUBLICANO

Itinerário, tábuas ou fragmentos da Alma Republicana

Artur B. Mendonça José M. Martins

Vale do Mondego, Coimbra (Portugal)

A Inaque Republicano

facebook

Cria o teu cartão de visita

ETIQUETAS

- 18 de Janeiro de 1934

ARQUIVO

ARQUIVO

E-Mail

at-m

CENTENÁRIO DA REPÚBLICA 1910-2010

BLOGS

- 100 Anos República (Alcácer)
- 100 Anos República F. Foz
- 31 de Janeiro
- Abril de Novo

**User generated meta-data**

A CORJA. Semanário republicano anti-clerical independente (Liberdade, Justiça, Verdade, Progresso) - [Ano I, nº1 (6 de Fevereiro 1915) ao nº 25 (16 de Agosto 1915)], Coimbra; Administrador, Anibal Reis [nº22, M. Simões]; Secretário, Mário de Brito [nº10, J. L. Frazão]; Director: José Peixoto de Alarcão [nº21, Fernandes Martins]; Colaboradores: A. Batista Rama, Afonso Duarte (poema), Alfredo Pimenta (poema), António Correia de Oliveira (poema), António Nobre (poema), Baldaque da Silva, Coelho Neto, Ernesto Almeida, Fernandes Martins, Fernando de Araújo, Guerra Junqueiro (poema), J. Peixoto de Alarcão, J. Pestana Júnior, João de Deus (poema), José Figueiredo Júnior [idem in, "A Revolta"], Ribeiro de Carvalho (poema); Administração, Rua Dr. João Jacinto, 38, Coimbra; Redacção, Couraça de Lisboa, 10, Coimbra;

# Even insignificant individual information. Teenager blogged: “math test went bad”

---

## Blog de uma adolescente

« [março 2006](#) | [Main](#) | [maio 2006](#) »

---

**abril 29, 2006**

### Fim de semana

Isto está uma seca, tive que estudar ciências! Pois é, a minha única nega. Tou tão deprimida... o teste de matemática correu-me tão mal!! Sabia aquilo tudo e na hora de-me uma branca. Agora é estudar, estudar, estudar, estudar..... Espero que o vosso fim de semana seja bem melhor que o meu inclusive o feriado, que também ficarei em casa a estudar!! \*\*\*\*\*Beijocas\*\*\*\*\*

Posted by IFilipinha2000 at 06:57 PM | [Comentários: \(1\)](#)

---

**abril 27, 2006**

### Testes e amigas

Estou furiosa!! Estudei bué para matemática e correu-me super mal o teste. Deu-me uma branca. Agora não posso tirar nenhuma nega às outras disciplinas. A minha colega quer esquecer o rapaz de que gosta, mas ela gosta tanto dele.... Sabem, por um lado acho que ela devia esquece-lo, mas ela já o tentou esquecer tanta vez que mais vale lutar por ele. Quem sabe se com ela ele não muda a sua atitude? Com os amiguinhos gozões dele é que é um bocadinho difícil, mas nada é impossível. \*\*\*\*\*Beijocas\*\*\*\*\*

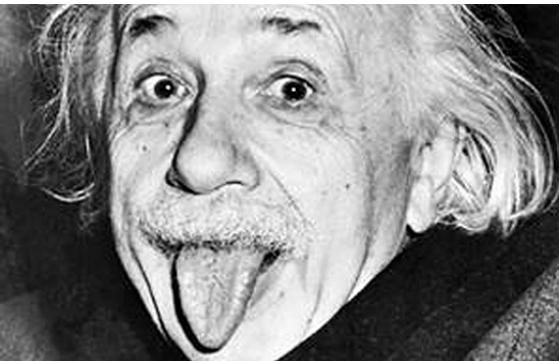


# What if this teenager will become a...

→ Famous singer?  
(who used to be a  
"nerd")

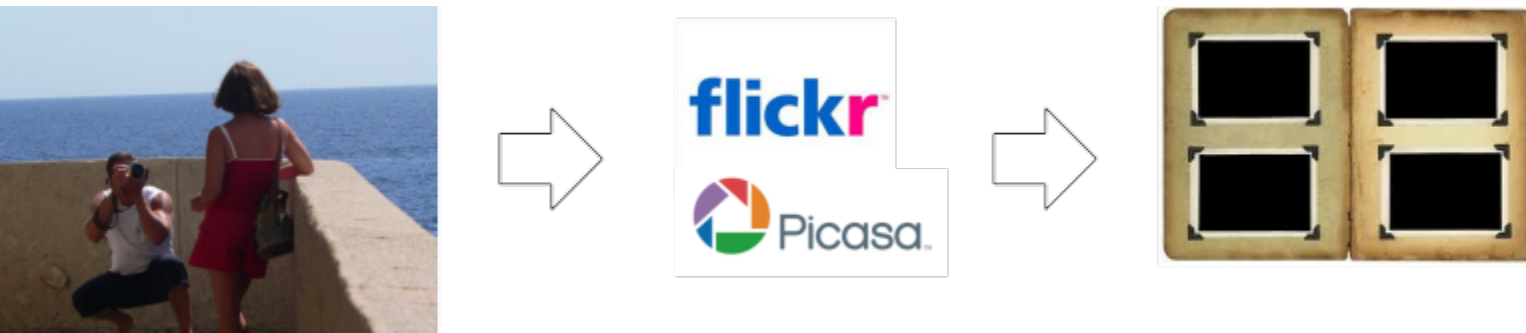


→ Physics Nobel?  
(that was not that  
good at math)



# Alert!

People are not preserving their digital data



Web archives may be the **only** source of memories to many people

# Collaborate with Archive.pt



We need **your** collaboration!



For non-techies:  
suggest the archive  
of interesting sites

[arquivo.pt/sugerir](http://arquivo.pt/sugerir)

A screenshot of the "Suggest a Portuguese site to be archived" page on the Portuguese Web Archive website. The page has a light blue header with the site's logo and navigation links. The main content area is white and contains a form for submitting suggestions. The form includes a text input for the site address, a text area for a description, and a "Like" button. A sidebar on the right contains a "News" section with recent updates and a "More news..." link.

PORTUGUESE  
WEB ARCHIVE

About the project  Search

only in current section ☐

Home Collaboration FAQ About Tools Dissemination Search the past

You are here: [Home](#) → [Collaboration](#) → Suggest a Portuguese site to be archived

### Suggest a Portuguese site to be archived

Please suggest a website containing information relevant to the Portuguese community that should be archived.

We particularly need your help, to identify relevant Portuguese sites hosted outside the country domain .PT (e.g. .COM, .NET, .EU).

If you wish to suggest a list of several sites, for instance from your Bookmarks, [contact us](#).

**Address of the site •**  
Please submit only the address of the home page of the site. The Portuguese Web Archive crawler will automatically find and archive the remaining contents of the site.

**Description •**  
Short text about the main subjects of the site.

**News**

Nov 10, 2014  
[Portuguese Web Archive: a Memory Infrastructure @DLM2014](#)

Jun 27, 2014  
[Our new article on Web Archiving](#)

Jun 27, 2014  
[System administrator for the Portuguese Web Archive](#)

[More news...](#)

Portuguese Web Archive  
 You like this.



For techies: All our software is  
freely available and you can  
improve it

pwa-technologies at  
[code.google.com](mailto:pwa-technologies@code.google.com)

The screenshot shows the Google Code project page for "pwa-technologies". At the top, the "PORTUGUESE WEB ARCHIVE" logo is on the left, and the project name "pwa-technologies" is in the center, with the tagline "PWA preserves today's knowledge for future generations." to its right. A search bar is on the far right. Below the header is a navigation menu with links: "Project Home", "Downloads", "Wiki", "Issues", "Source", and "Administer". The "Summary" tab is selected, showing a "Project Information" section on the left and a main content area on the right. The "Project Information" section includes a "Recommend this on Google" button, a "Starred by 3 users" indicator with a link to "Project feeds", a "Code license" section pointing to "GNU Lesser GPL", and a "Labels" section with "Web, Archive, Service, WebArchive". The "Members" section lists several names and email addresses. The main content area contains three paragraphs of text describing the project's goals, development challenges, and search engine capabilities, along with a link to "Main features".

**PORTUGUESE WEB ARCHIVE** **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](#),  
[joaocarvalhoiranda](#),  
[danielcoelhoogomes](#), [sawfcon](#),  
[devel.david@ivcruz.net](#), [whispall](#)

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](#)

# For web authors: follow recommendations to publish preservable sites

## Recommendations for authors to enable web archiving

This page presents recommendations to create web pages that can be efficiently archived and accessed across time. It is impossible to preserve the web without the authors collaboration.

*Last update on September 16, 2010.*

A web archive works similarly to a web search engine like Google. Therefore, the presented recommendations contribute to create contents that can be correctly processed of both these types of systems.

The following recommendations were classified in *fundamental* and *advisable*. They were written with the purpose of helping any author. [Please contact us if you have any doubt.](#)

### Web site organization

- [One link for each content](#) (fundamental)
- [Crawler-friendly homepage](#) (fundamental)
- [Maintain address to the same content across time](#) (advisable)
- [Robots Exclusion Protocol to indicate access restrictions](#) (advisable)

[arquivo.pt/recommendations](http://arquivo.pt/recommendations)



For everybody:  
disseminate  
Archive.pt

[arquivo.pt/news](http://arquivo.pt/news)



# Final message

Web archives are crucial  
infrastructures for  
Humanity's Memory







ARCHIVE.PT



Thank you!

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