

# Search the Past with the



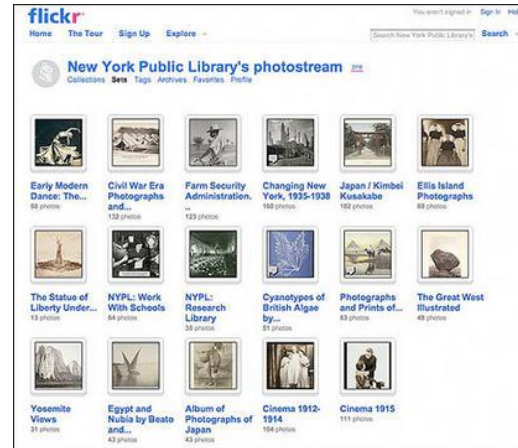
**Daniel Gomes**, Miguel Costa,  
David Cruz, João Miranda and Simão Fontes  
Foundation for National  
Scientific Computing

# The Web has been replacing printed media

## eBooks



## Photo galleries



## Blogs



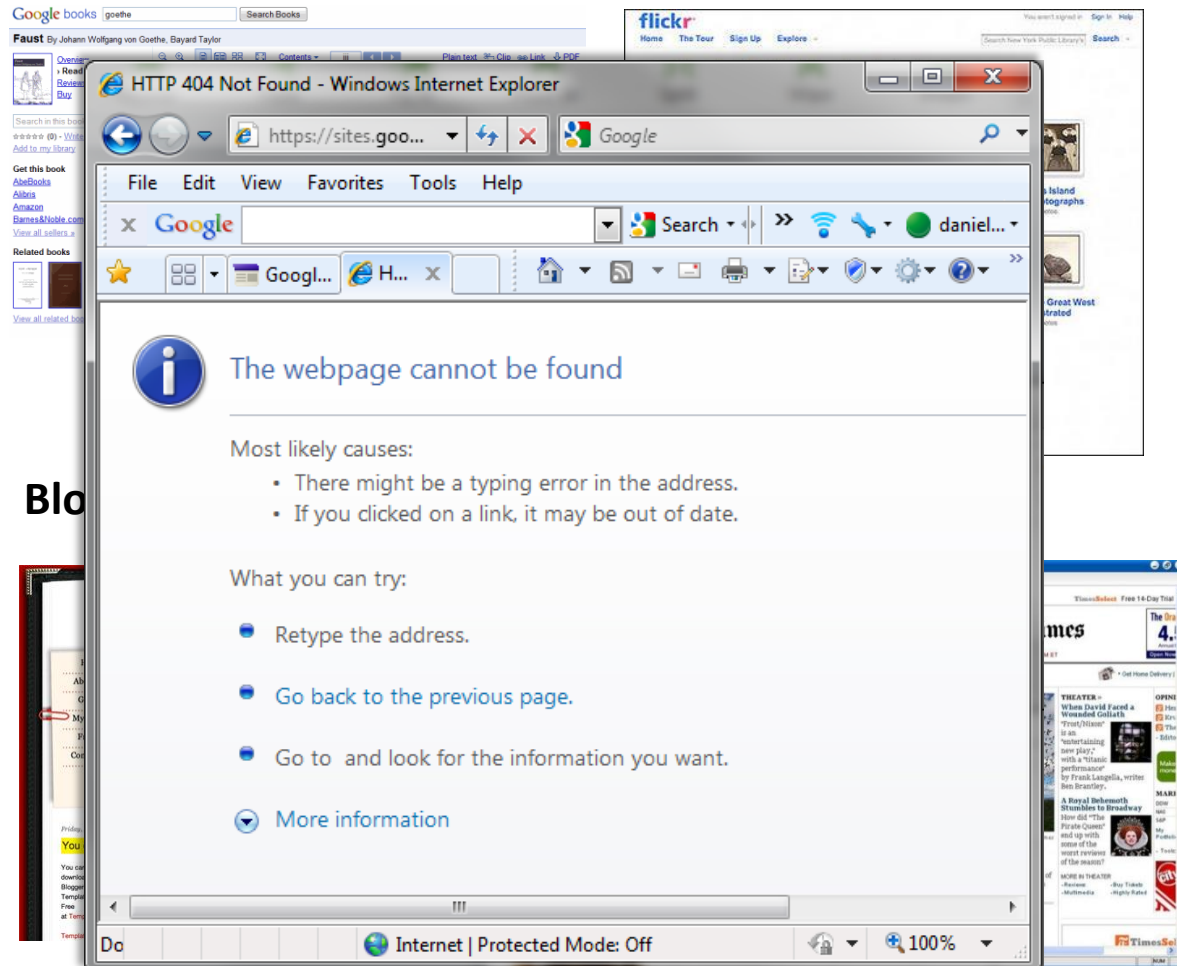
## News



# However, all these valuable information quickly disappears

eBooks


Photo galleries



80%

Disappear or change  
within 1 year

# 77 web archiving initiatives across the world work to preserve Humanity's cultural heritage available online



**WIKIPEDIA**  
The Free Encyclopedia

- Main page
- Contents
- Featured content
- Current events
- Random article
- Donate to Wikipedia

▼ Interaction

- Help
- About Wikipedia
- Community portal
- Recent changes
- Contact Wikipedia

► Toolbox

Article [Talk](#)

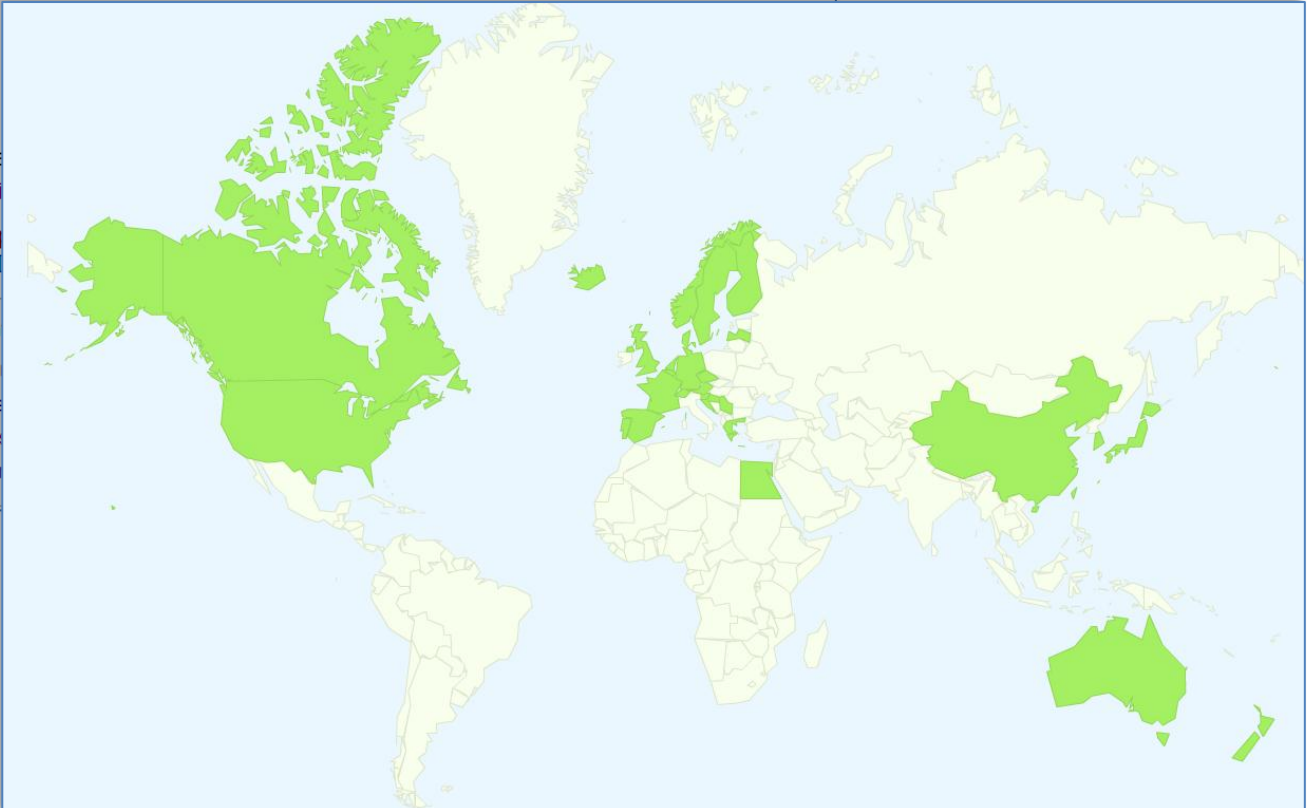
## List of Web archiving initiatives

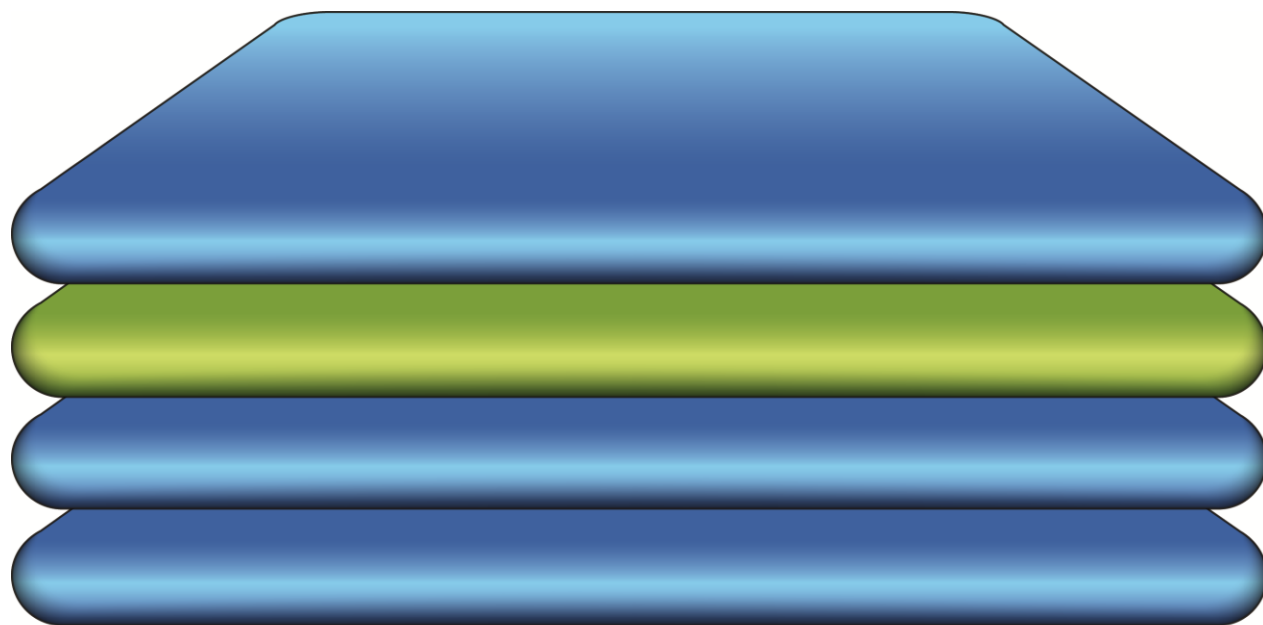
From Wikipedia, the free encyclopedia

This page is about web archiving initiatives.

This Wikipedia article lists the following initiatives:

Country
1 Web archive
2 Archive
3 Access
4 Reference





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.

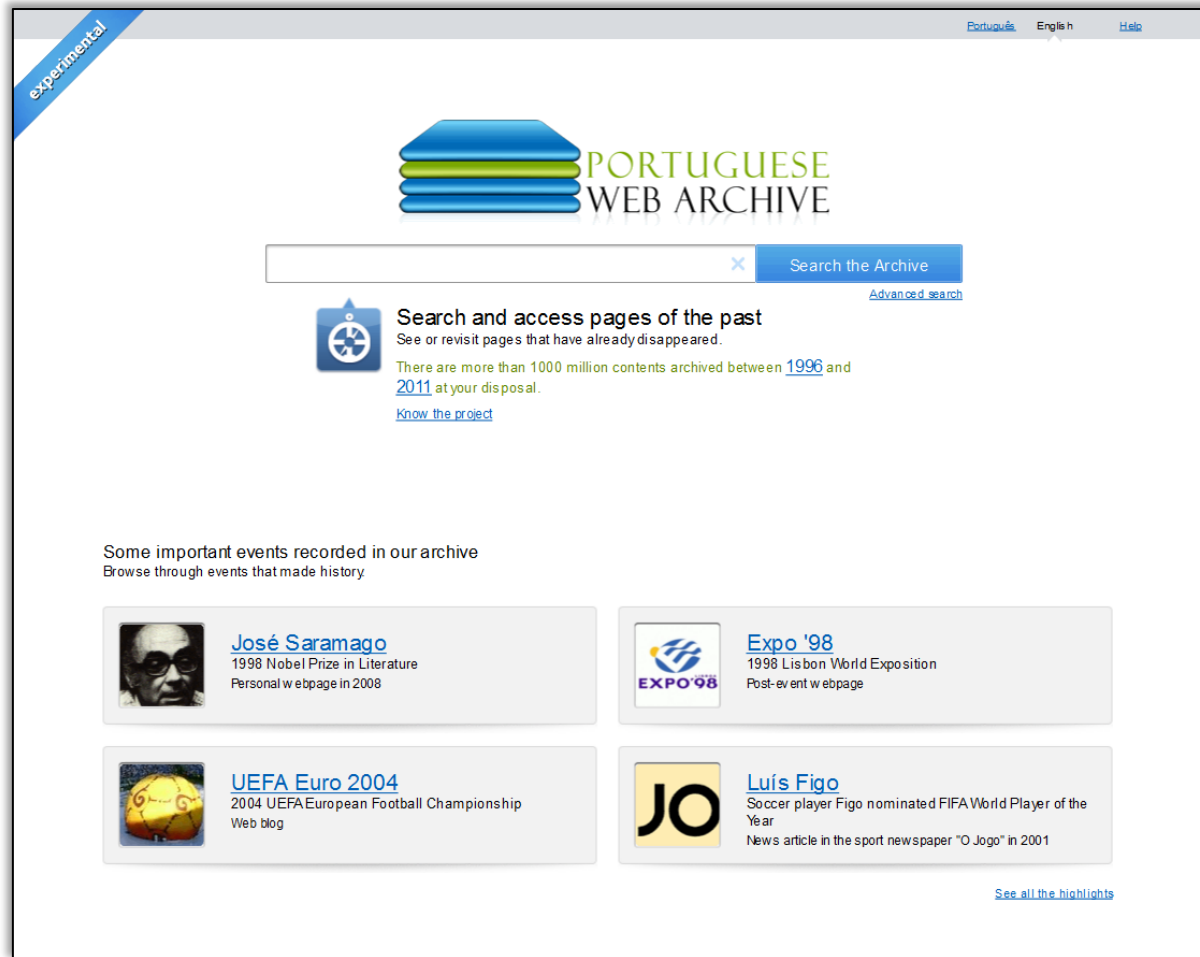
As a result, several countries are creating their own national archives to ensure the preservation of contents of historical relevance to their cultures.

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 last year (2012)



- Public and free at [archive.pt](http://archive.pt)



# Provides URL search like the Internet Archive Wayback Machine



[Search the Archive](#)

between:  and: [Advanced search](#)

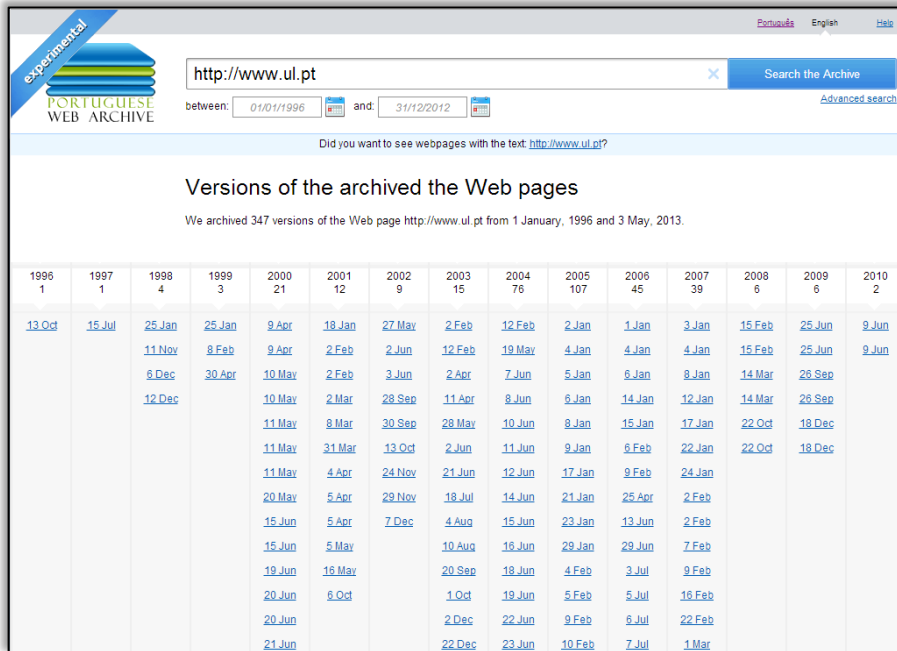
Did you want to see webpages with the text: [http://www.ul.pt?](#)

## 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
<a href="#">13 Oct</a>	<a href="#">15 Jul</a>	<a href="#">25 Jan</a>	<a href="#">25 Jan</a>	<a href="#">9 Apr</a>	<a href="#">18 Jan</a>	<a href="#">27 May</a>	<a href="#">2 Feb</a>	<a href="#">12 Feb</a>	<a href="#">2 Jan</a>	<a href="#">1 Jan</a>	<a href="#">3 Jan</a>	<a href="#">15 Feb</a>	<a href="#">25 Jun</a>	<a href="#">9 Jun</a>
		<a href="#">11 Nov</a>	<a href="#">8 Feb</a>	<a href="#">9 Apr</a>	<a href="#">2 Feb</a>	<a href="#">2 Jun</a>	<a href="#">12 Feb</a>	<a href="#">19 May</a>	<a href="#">4 Jan</a>	<a href="#">4 Jan</a>	<a href="#">4 Jan</a>	<a href="#">15 Feb</a>	<a href="#">25 Jun</a>	<a href="#">9 Jun</a>
		<a href="#">6 Dec</a>	<a href="#">30 Apr</a>	<a href="#">10 May</a>	<a href="#">2 Feb</a>	<a href="#">3 Jun</a>	<a href="#">2 Apr</a>	<a href="#">7 Jun</a>	<a href="#">5 Jan</a>	<a href="#">6 Jan</a>	<a href="#">8 Jan</a>	<a href="#">14 Mar</a>	<a href="#">26 Sep</a>	
		<a href="#">12 Dec</a>		<a href="#">10 May</a>	<a href="#">2 Mar</a>	<a href="#">28 Sep</a>	<a href="#">11 Apr</a>	<a href="#">8 Jun</a>	<a href="#">6 Jan</a>	<a href="#">14 Jan</a>	<a href="#">12 Jan</a>	<a href="#">14 Mar</a>	<a href="#">26 Sep</a>	
				<a href="#">11 May</a>	<a href="#">8 Mar</a>	<a href="#">30 Sep</a>	<a href="#">28 May</a>	<a href="#">10 Jun</a>	<a href="#">8 Jan</a>	<a href="#">15 Jan</a>	<a href="#">17 Jan</a>	<a href="#">22 Oct</a>	<a href="#">18 Dec</a>	
				<a href="#">11 May</a>	<a href="#">31 Mar</a>	<a href="#">13 Oct</a>	<a href="#">2 Jun</a>	<a href="#">11 Jun</a>	<a href="#">9 Jan</a>	<a href="#">6 Feb</a>	<a href="#">22 Jan</a>	<a href="#">22 Oct</a>	<a href="#">18 Dec</a>	
				<a href="#">11 May</a>	<a href="#">4 Apr</a>	<a href="#">24 Nov</a>	<a href="#">21 Jun</a>	<a href="#">12 Jun</a>	<a href="#">17 Jan</a>	<a href="#">9 Feb</a>	<a href="#">24 Jan</a>			
				<a href="#">20 May</a>	<a href="#">5 Apr</a>	<a href="#">29 Nov</a>	<a href="#">18 Jul</a>	<a href="#">14 Jun</a>	<a href="#">21 Jan</a>	<a href="#">25 Apr</a>	<a href="#">2 Feb</a>			
				<a href="#">15 Jun</a>	<a href="#">5 Apr</a>	<a href="#">7 Dec</a>	<a href="#">4 Aug</a>	<a href="#">15 Jun</a>	<a href="#">23 Jan</a>	<a href="#">13 Jun</a>	<a href="#">2 Feb</a>			
				<a href="#">15 Jun</a>	<a href="#">5 May</a>		<a href="#">10 Aug</a>	<a href="#">16 Jun</a>	<a href="#">29 Jan</a>	<a href="#">29 Jun</a>	<a href="#">7 Feb</a>			
				<a href="#">19 Jun</a>	<a href="#">16 May</a>		<a href="#">20 Sep</a>	<a href="#">18 Jun</a>	<a href="#">4 Feb</a>	<a href="#">3 Jul</a>	<a href="#">9 Feb</a>			
				<a href="#">20 Jun</a>	<a href="#">6 Oct</a>		<a href="#">1 Oct</a>	<a href="#">19 Jun</a>	<a href="#">5 Feb</a>	<a href="#">5 Jul</a>	<a href="#">16 Feb</a>			
				<a href="#">20 Jun</a>			<a href="#">2 Dec</a>	<a href="#">22 Jun</a>	<a href="#">9 Feb</a>	<a href="#">6 Jul</a>	<a href="#">22 Feb</a>			
				<a href="#">21 Jun</a>			<a href="#">22 Dec</a>	<a href="#">23 Jun</a>	<a href="#">10 Feb</a>	<a href="#">7 Jul</a>	<a href="#">1 Mar</a>			

# Problem with URL search



experimental PORTUGUESE WEB ARCHIVE

Search the Archive

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

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

Versions of the archived the Web pages

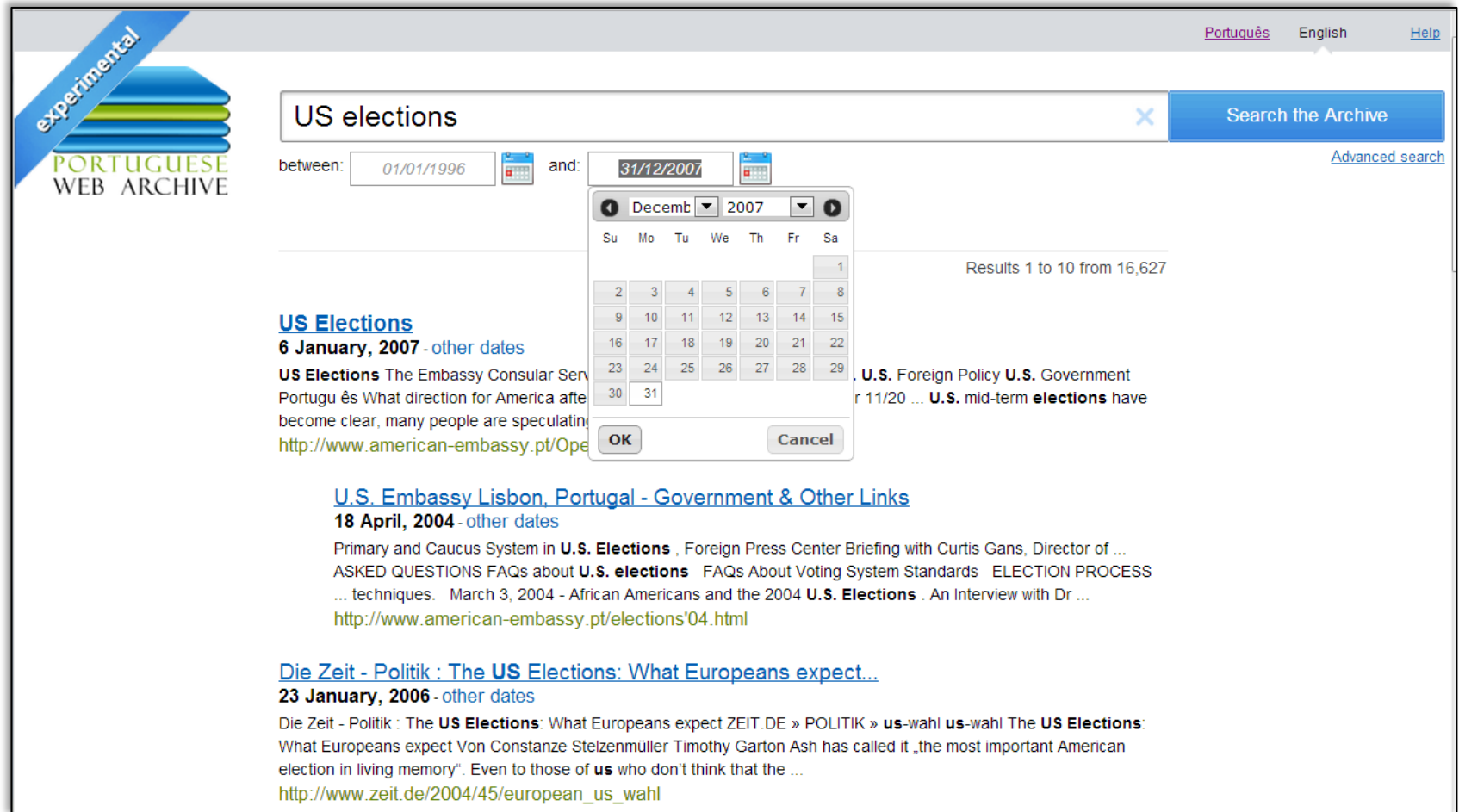
We archived 347 versions of the Web page <http://www.ul.pt> from 1 January, 1996 and 3 May, 2013.

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
<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 Mar</a> <a href="#">10 Mar</a> <a href="#">11 Mar</a> <a href="#">11 Mar</a> <a href="#">11 Mar</a> <a href="#">20 Mar</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="#">2 Mar</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 Mar</a> <a href="#">10 Jun</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="#">22 Jun</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="#">17 Jan</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="#">2 Jan</a> <a href="#">4 Jan</a> <a href="#">5 Jan</a> <a href="#">6 Jan</a> <a href="#">14 Jan</a> <a href="#">12 Jan</a> <a href="#">17 Jan</a> <a href="#">9 Feb</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="#">5 Jan</a> <a href="#">8 Jan</a> <a href="#">14 Jan</a> <a href="#">12 Jan</a> <a href="#">17 Jan</a> <a href="#">9 Feb</a> <a href="#">25 Apr</a> <a href="#">13 Jun</a> <a href="#">29 Jun</a> <a href="#">3 Jul</a> <a href="#">5 Jul</a> <a href="#">6 Jul</a> <a href="#">7 Jul</a>	<a href="#">3 Jan</a> <a href="#">4 Jan</a> <a href="#">8 Jan</a> <a href="#">14 Mar</a> <a href="#">22 Oct</a> <a href="#">22 Oct</a>	<a href="#">15 Feb</a> <a href="#">15 Feb</a> <a href="#">14 Mar</a> <a href="#">26 Sep</a> <a href="#">18 Dec</a> <a href="#">18 Dec</a> <a href="#">1 Mar</a>	<a href="#">25 Jun</a> <a href="#">25 Jun</a> <a href="#">26 Sep</a> <a href="#">18 Dec</a>	<a href="#">9 Jun</a>



- Users **do not know** the URL that contained the information that they need.

[Archive.pt](https://archive.pt) also provides **full-text search** over 1.2 billion web files archived since 1996



# New web archive search system based on open source web archiving tools (NutchWAX)



The screenshot shows the NutchWAX website interface. At the top right is the **nutchwax** logo. Below it is a navigation bar with links: [Sourceforge](#), [Heritrix](#), [Archive Access](#), [Internet Archive](#), and [Home](#). A status bar on the left indicates "Last Published: 08 Mar 2009".

**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



**IIPC**  
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.

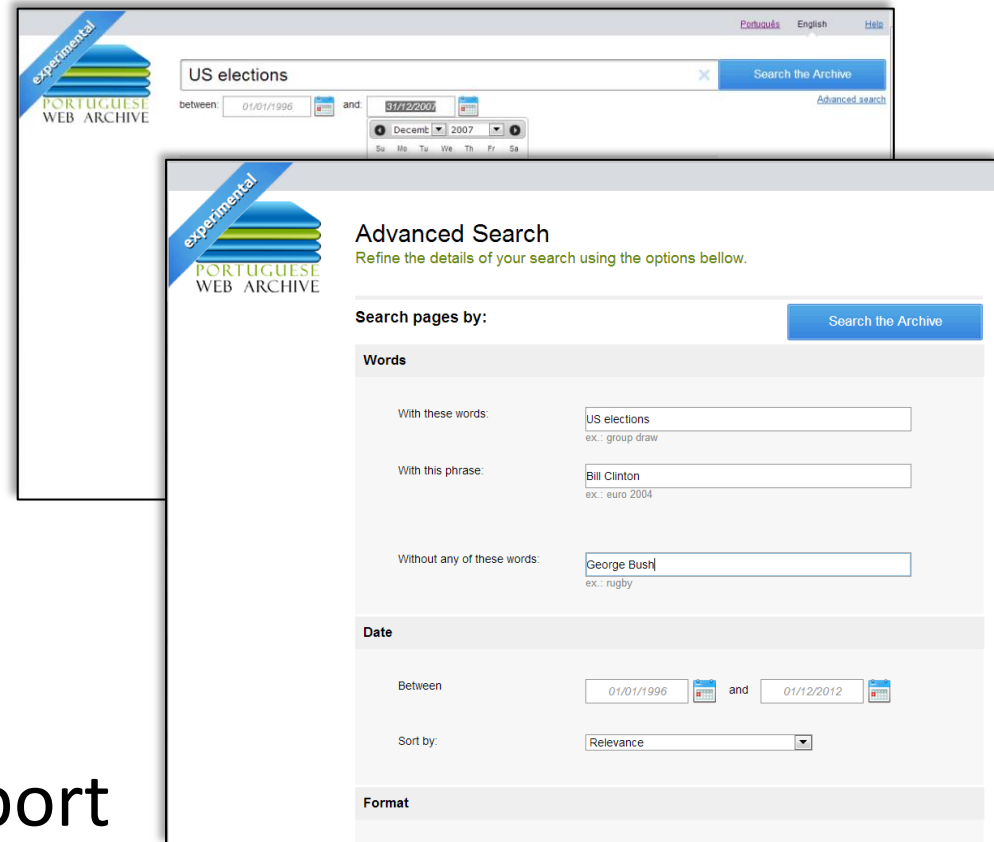
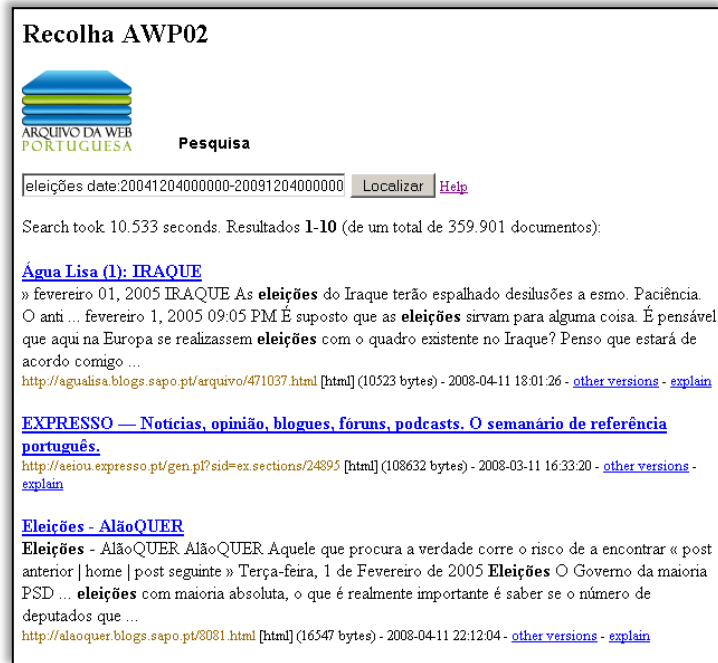


**<NWA>**

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.

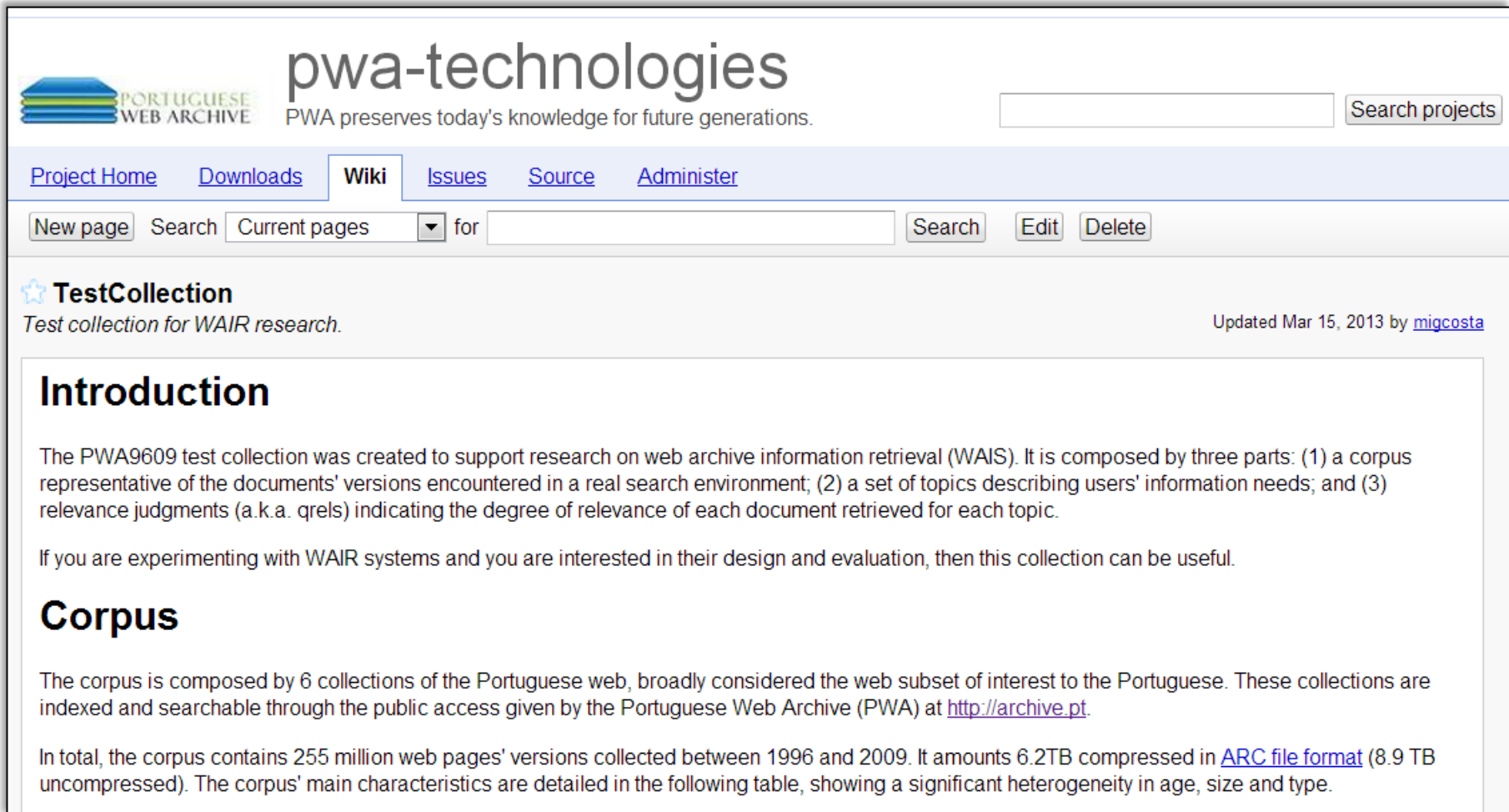
- Quicker response times
- Improved search results relevance

# Designed more adequate user interfaces: NutchWAX (2007) vs. PWA (2012)



- Internationalization support
- Advanced search user interface
- Improved usability
  - 71% overall user satisfaction from usability testing

# Researched ranking algorithms for web archive information retrieval



The screenshot shows the PWA-technologies website. At the top, there is a logo for the Portuguese Web Archive (PWA) and the text "pwa-technologies" with the tagline "PWA preserves today's knowledge for future generations." Below this is a navigation bar with links: Project Home, Downloads, Wiki (selected), Issues, Source, and Administer. A search bar is also present. Below the navigation bar, there is a section for "TestCollection" with a star icon and the description "Test collection for WAIR research." The page is updated as of Mar 15, 2013 by migcosta. The main content area is titled "Introduction" and describes the PWA9609 test collection, which is composed of three parts: (1) a corpus representative of the documents' versions encountered in a real search environment; (2) a set of topics describing users' information needs; and (3) relevance judgments (a.k.a. qrels) indicating the degree of relevance of each document retrieved for each topic. It also mentions that the collection can be useful for experimenting with WAIR systems. Below the introduction is a section titled "Corpus" which states that the corpus is composed of 6 collections of the Portuguese web, indexed and searchable through the public access given by the Portuguese Web Archive (PWA) at <http://archive.pt>. It also mentions that the corpus contains 255 million web pages' versions collected between 1996 and 2009, amounting to 6.2TB compressed in ARC file format (8.9 TB uncompressed). The corpus' main characteristics are detailed in the following table, showing a significant heterogeneity in age, size and type.

**TestCollection**  
*Test collection for WAIR research.* Updated Mar 15, 2013 by [migcosta](#)

## Introduction

The PWA9609 test collection was created to support research on web archive information retrieval (WAIS). It is composed by three parts: (1) a corpus representative of the documents' versions encountered in a real search environment; (2) a set of topics describing users' information needs; and (3) relevance judgments (a.k.a. qrels) indicating the degree of relevance of each document retrieved for each topic.

If you are experimenting with WAIR systems and you are interested in their design and evaluation, then this collection can be useful.


## Corpus

The corpus is composed by 6 collections of the Portuguese web, broadly considered the web subset of interest to the Portuguese. These collections are indexed and searchable through the public access given by the Portuguese Web Archive (PWA) at <http://archive.pt>.

In total, the corpus contains 255 million web pages' versions collected between 1996 and 2009. It amounts 6.2TB compressed in [ARC file format](#) (8.9 TB uncompressed). The corpus' main characteristics are detailed in the following table, showing a significant heterogeneity in age, size and type.

- Publications, test collections and collaboration proposals

# The source code is free and open




## 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](#),  
[joacarvalhomiranda](#),  
[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.

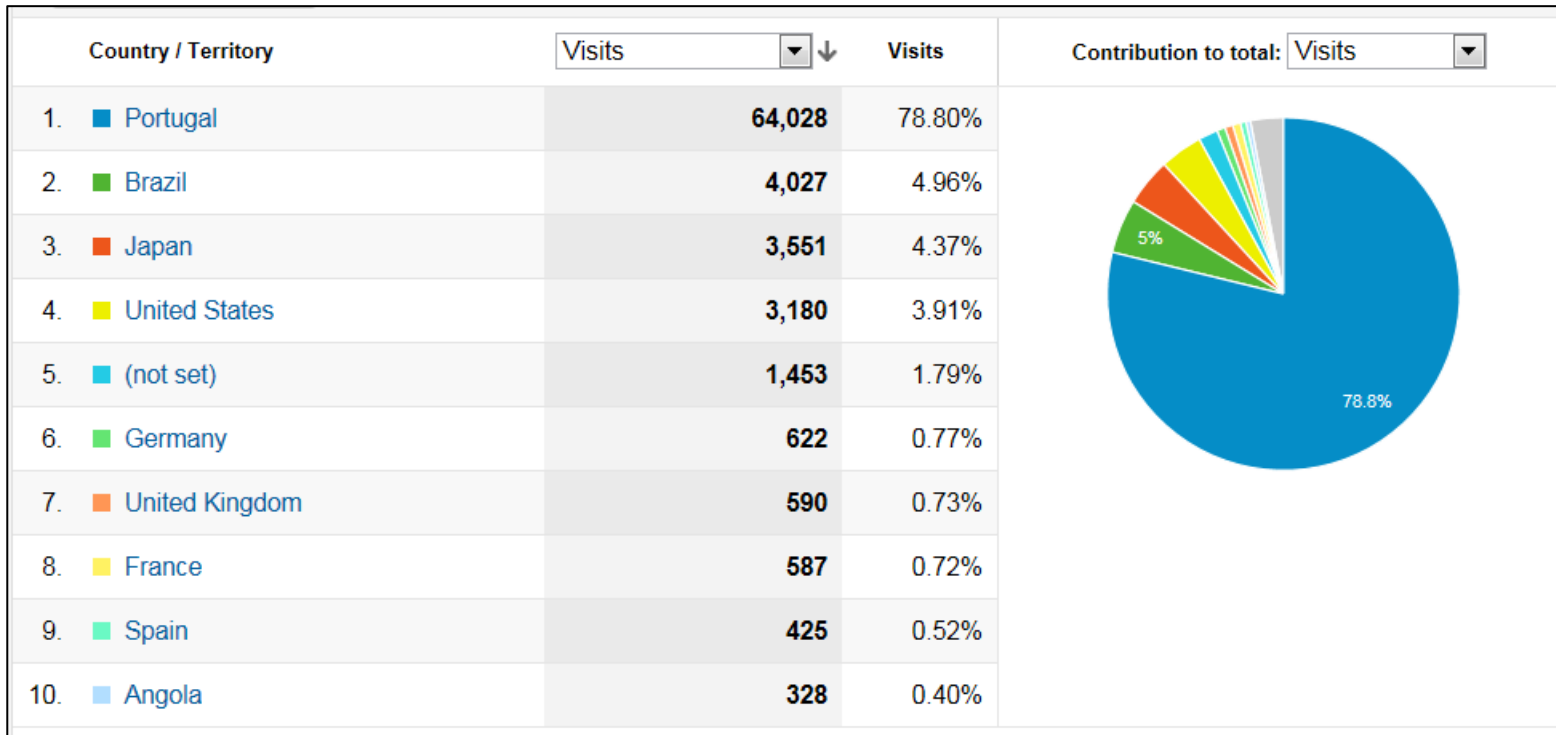
During the developing of the PWA IR (information retrieval) system we faced limitations in searching speed, quality of results, scalability and usability. To cope with this, we modified the archive-access project (<http://archive-access.sourceforge.net/>) to support our web archive IR requirements. Nutchwax, 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](#)

# Archive.pt is useful to international users



- Archived web pages written in several languages
- Includes all Portuguese speaking domains (.AO, .MZ, .CV) except Brazil
- 21.2% of the Visits were not from Portugal



The Portuguese Web Archive can be  
used to document interesting stories

Let's hear one

Once upon a time in **1996...**

# There was a “mad” scientist

## Tim Berners-Lee



### Bio

Tim Berners-Lee is the inventor of the World Wide Web, an internet-based hypermedia initiative for global information sharing. He wrote the first Web clients and server and defined the URL, HTTP and HTML specifications on which the web depends while working at [CERN](#), the European Particle Physics Laboratory, in late 1990.

Tim is now the Director of the [W3 Consortium](#), an open forum of companies and organizations with the mission to realize the full potential of the Web. He is a Principal Research Scientist at the Laboratory for Computer Science ( [LCS](#)) at the Massachusetts Institute of Technology ( [MIT](#)).

Before going to CERN, Tim was a founding director of Image Computer Systems, and before that and a consultant in hardware and software system design, real-time communications graphics and text processing; and a principal engineer with Plessey Telecommunications, in Poole, England. He is a graduate of Oxford University.

Tim is married to Nancy Carlson. They have two children, born 1991 and 1994.

Who invented “the World Wide Web, an internet-based hypermedia initiative for global information sharing.”

# He founded an organization to support his invention



## The World Wide Web Consortium

The World Wide Web is the universe of network-accessible information. The [World Wide Web Consortium](#) exists to realize the full potential of the Web.

W3C works with the global community to produce [specifications](#) and [reference software](#). W3C is funded by industrial [members](#) but its products are freely available to all. The Consortium is run by [MIT LCS](#) and by [INRIA](#), in collaboration with [CERN](#) where the web originated. Please see the [list of members](#) to learn about individual members and visit their World Wide Web sites.

- [W3C Activity areas and directions](#)
- [How to contact W3C](#)
- [Frequently Asked Questions about W3C](#)
- [W3C meetings, newsletter, mailing lists \[W3C Members only\]](#)
- [Help](#)

In this document:

- [News and Updates](#)
- [Web Specifications and Development Areas](#)
- [W3C Software](#)
- [The World Wide Web and the Web Community](#)
- [Getting involved with the W3C](#)

# There were lists of WWW sites

## European Home Page

This is a very simple European Map of WWW/NIR sites. Please click on any flag to go to the (sensitive map) Home Page of that country.  
If you don't have graphics support, here is [CERN's list](#) of sites organized by geography. If you want to see another european Home Page, try this [one](#).

Try the new [Beautiful Cultural European](#) Home Pages (with travel tips).



# The Library of Congress also had one WWW site in 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.

# Since then, WWW sites became used even to publish News

INTERNACIONAL  
Expresso  
1/5/98



GUIA  
DO ESTUDANTE

 Índice

 Pesquisa

 Comentário

 Forum







## Lula vai desistir?

O PRINCIPAL líder da esquerda brasileira, Luís Ignácio da Silva («Lula»), ameaçou esta semana desistir de se candidatar à Presidência nas eleições de Outubro, deixando o campo livre para a reeleição de [Fernando Henrique Cardoso](#).

O desânimo de Lula tem origem no seu próprio partido - o Partido dos Trabalhadores (PT) -, que decidiu não se aliar ao Partido Democrático Trabalhista (PDT) no Rio de Janeiro, inviabilizando a candidatura da dupla Lula-Leonel Brizola (líder do PDT), o primeiro à chefia do Estado e o segundo à vice-presidência.

O fracasso deve-se à vitória dos radicais do PT, que conseguiram impor a candidatura do ex-líder estudantil Vladimir Palmeira ao Governo do Rio, derrotando os moderados, que se tinham comprometido a apoiar o candidato de Brizola, para garantirem que este último concorreria como «vice» de Lula.

# That now can be quickly translated to several languages

 Índice Pesquisa Comentário

## Lula will give up?

HOME The leader of the Brazilian left, Luis Ignacio da Silva ("Lula"), this week threatened to quit to run for presidency in October elections, leaving the field open for the re-election of [Fernando Henrique Cardoso](#) .

The dismay of Lula originates in his own party - the Workers Party (PT) - who decided not to ally the Democratic Labor Party (PDT) in Rio de Janeiro, preventing the application of the double-Lula Brizola (leader PDT ), the first head of state and the second vice-presidency.

The failure is due to the victory of the radicals of the PT, which succeeded in imposing the candidacy of former student leader Vladimir Palmeira Government of Rio, defeating the moderates, who had committed to support the candidate of Brizola, to ensure that the latter would compete as a 'vice' of Lula.

[Back](#)  
[Forward](#)  
[Reload](#)  
[Save as...](#)  
[Print...](#)  
[Translate to English](#)  
[View page source](#)  
[View page info](#)  
[Show Anchors](#)  
[Inspect element](#)

1998 archived news article from Portuguese publication translated with Google Translate in 2013



Would you like to search for  
more interesting pages?

Visit me at the Demo lobby  
or try it at [archive.pt](http://archive.pt)!

Thanks.



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

**daniel.gomes@fccn.pt**