Search the Past with the

PORTUGUESE WEB ARCHIVE

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
80% Disappear or change within 1 year
77 web archiving initiatives across the world work to preserve Humanity’s cultural heritage available online.
PORTUGUESE WEB ARCHIVE
The Portuguese Web Archive project started in 2008

Welcome to the Tomba project: the Portuguese web archive

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

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

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

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

Portugal is now beginning its national web archiving initiative with the Tomba project at FCCN (National Foundation for Scientific Computing).
It was announced last year (2012)

• Public and free at archive.pt
Provides **URL search** like the Internet Archive Wayback Machine

![Screenshot of the Portuguese Web Archive](image)

### Versions of the archived the Web pages


<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Month</td>
<td>13 Oct</td>
<td>15 Jul</td>
<td>25 Jan</td>
<td>11 Nov</td>
<td>6 Dec</td>
<td>12 Dec</td>
<td>25 Jan</td>
<td>18 Jan</td>
<td>27 May</td>
<td>2 Feb</td>
<td>12 Feb</td>
<td>2 Jan</td>
<td>1 Jan</td>
<td>3 Jan</td>
<td>9 Jun</td>
</tr>
<tr>
<td></td>
<td>2 Apr</td>
<td>2 Feb</td>
<td>2 Feb</td>
<td>8 Feb</td>
<td>30 Apr</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>11 May</td>
<td>4 Jul</td>
<td>28 Sep</td>
<td>11 Apr</td>
<td>26 May</td>
<td>10 Jun</td>
<td>8 Jan</td>
<td>14 Jan</td>
<td>17 Jan</td>
<td>5 Jan</td>
<td>15 Jan</td>
<td>22 Oct</td>
<td>10 Dec</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 May</td>
<td>5 Apr</td>
<td>24 Nov</td>
<td>21 Jun</td>
<td>12 Jun</td>
<td>17 Jan</td>
<td>9 Feb</td>
<td>24 Jan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>15 Jun</td>
<td>5 May</td>
<td>7 Dec</td>
<td>4 Aug</td>
<td>10 Aug</td>
<td>16 Jun</td>
<td>29 Jan</td>
<td>28 Jun</td>
<td>8 Jul</td>
<td>22 Jun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>19 Jun</td>
<td>16 May</td>
<td>20 Sep</td>
<td>19 Jun</td>
<td>5 Feb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 Jun</td>
<td>6 Oct</td>
<td>1 Oct</td>
<td>20 Jun</td>
<td>9 Feb</td>
<td>5 Jul</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>20 Jun</td>
<td>22 Dec</td>
<td>23 Jun</td>
<td>10 Feb</td>
<td>7 Jul</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21 Jun</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problem with URL search

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

- 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

• Publications, test collections and collaboration proposals
The source code is free and open.
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
The Library of Congress also had one WWW site in 1996

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

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.
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.

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!

Thanks.

www.archive.pt
daniel.gomes@fccn.pt