Research the Past Web using Web archives

Daniel.Gomes@fccn.pt

Daniel.Bicho@fccn.pt

Fernando.Melo@fccn.pt

Tutorial outline

Research the Past Web using Web archives

1. **Search and access**
   - The Past Web: examples and use cases
   - Public online services
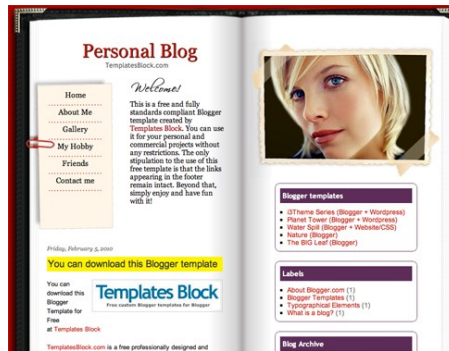
2. **Publish and preserve**
   - Recommendations to publish preservable information
   - How to preserve information collected from the Web

3. **Automatic processing**
   - Interoperability protocols
   - Application Programming Interfaces (API)

4. Your precious feedback: please take your notes

arquivo.pt/training

Fact:
Web replaced Press

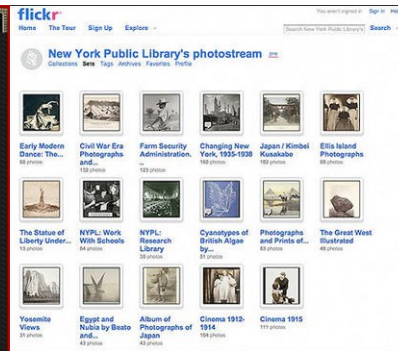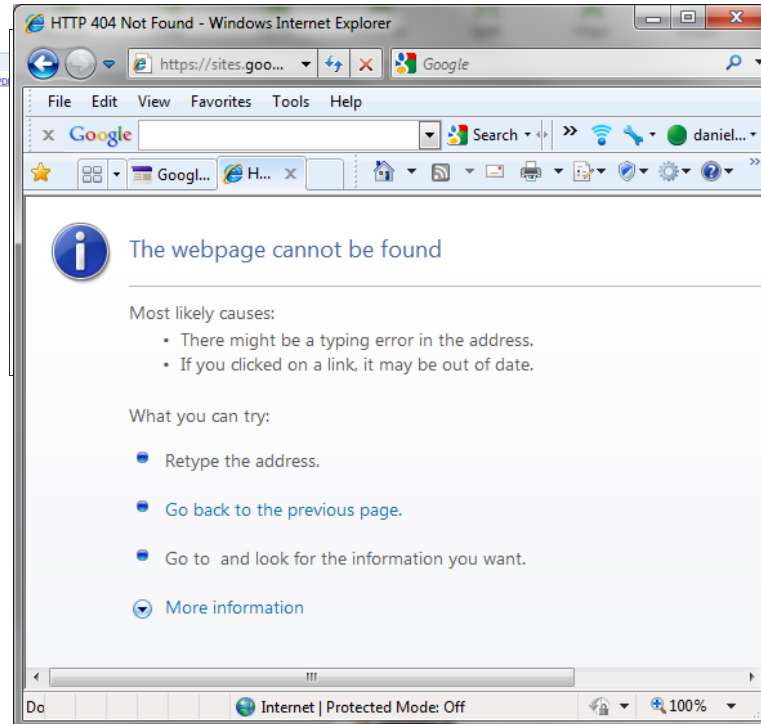Personal blogs   Photo galleries   E-books   News

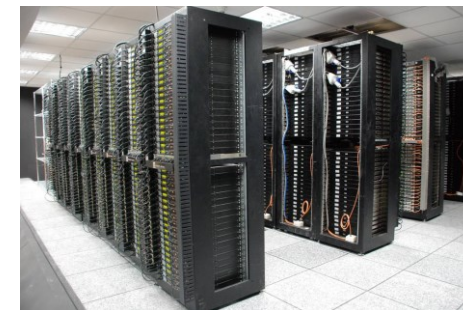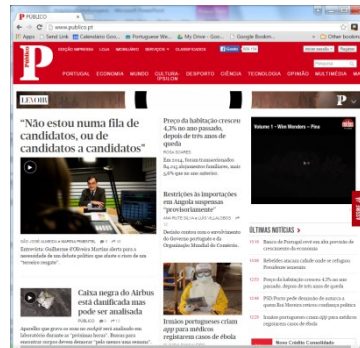# However, this valuable information quickly vanishes

80%

of web pages
change their content or disappear
within 1 year

Archives are also require to preserve information published online

Preserve the Press

Preserve the Web

Archive.org: the first web archive

Brewster Kahle founded the **Internet Archive** in 1996, a non-profit organization with the mission of providing "Universal access to all knowledge".

Web archiving initiatives around the World (August 2018):
86

Who are we?

Free online service to research the Past Web

Preserves **publicly accessible** information related with:

- Portugal
- **Research** and **Education** (international)

Governmental service provided by Foundation for Science and Technology (Portugal)

Arquivo.pt is used word-wide

	Country	Users	% Users
1.	Portugal	46,891	46.56%
2.	United States	26,373	26.19%
3.	Brazil	2,266	2.25%
4.	Russia	2,234	2.22%
5.	United Kingdom	2,231	2.22%
6.	Japan	2,172	2.16%
7.	Canada	1,237	1.23%
8.	Mozambique	1,213	1.20%
9.	India	902	0.90%
10.	Germany	894	0.89%

53% of our users are international

User Interfaces and documentation also in English

Combined with Google Translate enables **cross-lingual access** to preserved content

The Past Web

Examples of web-archived pages

Historical web pages: Tim Berners Lee personal page and W3C in the 90s

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 compa[...] mission to realize the full potential of the Web. He is a Principal Research[...] Computer Science (LCS)at the Massachusetts Institute of Technology (M[...]

Before going to CERN, Tim was a founding director of Image Computer S[...] consultant in hardware and software system design, real-time communica[...] processing; and a principal engineer with Plessey Telecommunications, in[...] of Oxford University.

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

www.w3.org/pub/WWW/People/Berners-Lee/, 1996

@Arquivo.pt

www.w3.org, 1996
@Arquivo.pt

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

No Google, no Facebook, you had to navigate through the links
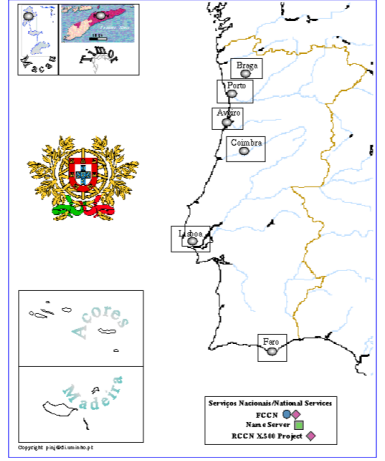
s700.uminho.pt/europa.html, 1996
@Arquivo.pt
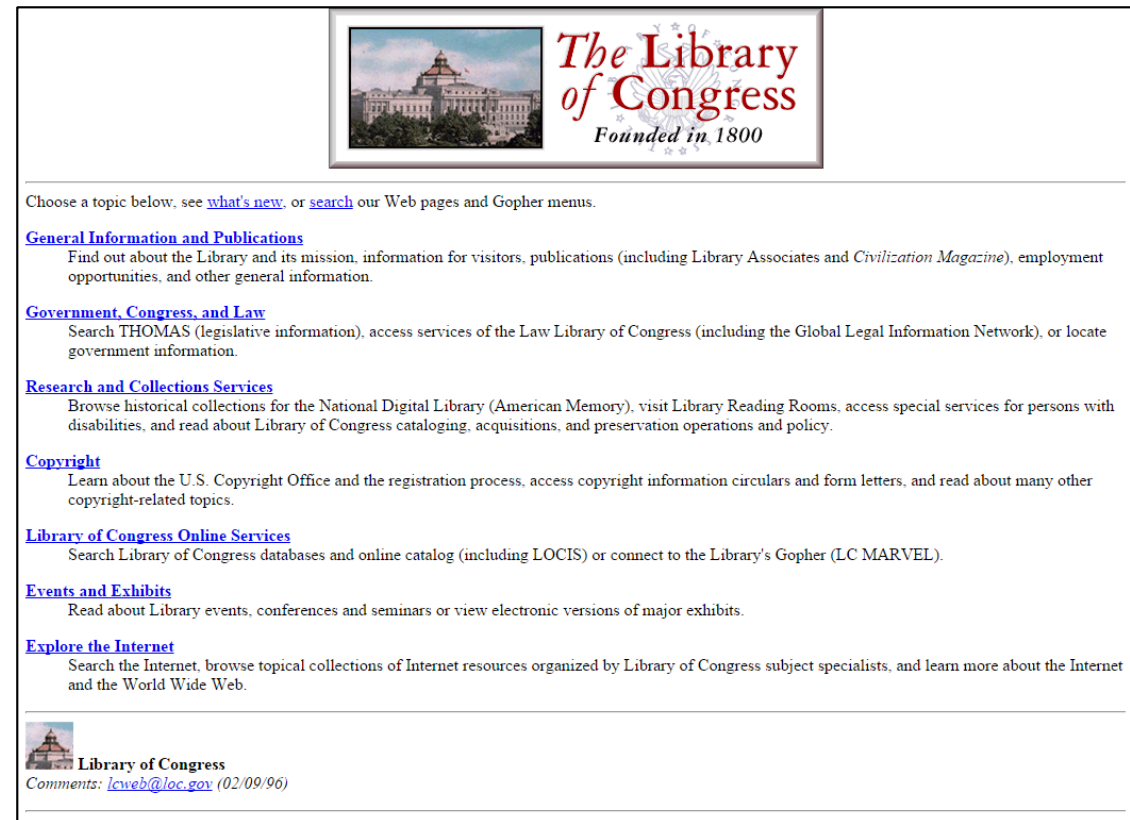
s700.uminho.pt/homepage-pt.html, 1996
@Arquivo.pt

Web-born information **NOT** collected online: web pages from CD-ROMs (digital archeology)

First web page of the Library of Congress (1996).

Obtained from a CD-ROM attached to a printed book.

Printed information, re-published online

Old printed information, published online, enriched with user-generated meta-data

Published
on print in 1915

Enriched
online in 2010

Open-access
research data and publications

arcomem.eu, 2014
@Internet Archive

The Past Web

Use cases

Fix citations to broken links
that jeopardize Science in all areas

Data sets and tools applied in experiments are online.

When cited resources disappear, experiments cannot be reproduced

- Not Reproducible,
  Not Science.

Web archives preserve online scientific information.

Malaysian Journal of Library & Information Science, Vol. 16, no. 3, December 2011: 17-29

Death of web citations: a serious alarm for authors

The FASEB Journal

The Journal of the Federation of American Societies for Experimental Biology

HOME | CURRENT ISSUE | NEW ARTICLES | ARCHIVE | ALERTS | RSS | SUBMIT | HELP

Unavailability of online supplementary scientific information from articles published in major journals

« Previous | Next Article »
Table of Contents

P.A. Ioannidis[*],[†‡]

This Article

doi:
10.1096/fj.05-4784lsf
December 2005
The FASEB Journal
vol. 19 no. 14 1943-1944

» Abstract
Full Text
Full Text (PDF)

[+] Classifications
[+] Services

and Epidemiology,
reece. E-mail:

JOURNAL OF THE ASSOCIATION FOR INFORMATION SCIENCE AND TECHNOLOGY

Research Article
Link decay in leading information science journals

Dion Hoe-Lian Goh and Peng Kin Ng

Article first published online: 3 NOV 2006

DOI: 10.1002/asi.20513

Copyright © 2007 Wiley Periodicals, Inc., A Wiley Company

Issue

Persistence and Decay of Web Citations Used in Theses and Dissertations Available at the Sokoine National Agricultural Library, Tanzania

ACADEMIC JOURNAL ARTICLE

By Sife, Alfred S.; Bernard, Ronald

International Journal of Education and Development using Information and Communication Technology , Vol. 9, No. 2 , July 1, 2013

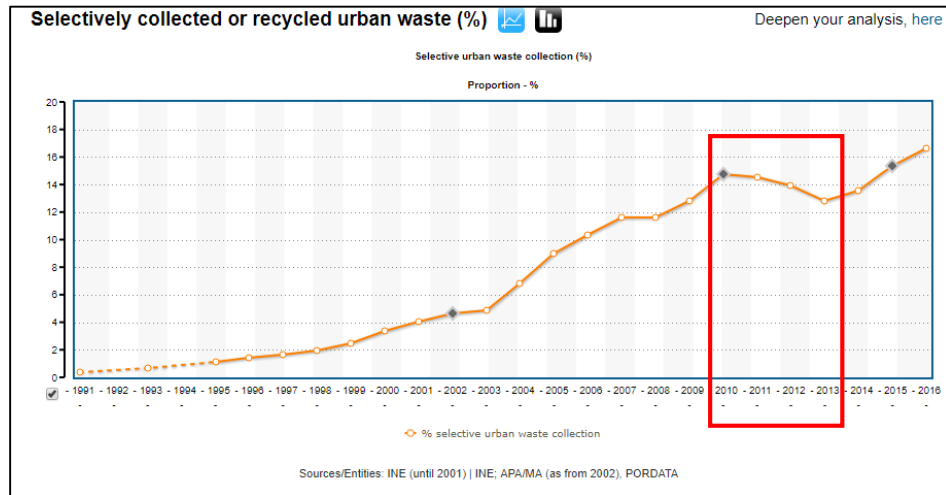# Management: documentation/audit of finished project

"Everything was on the website"

Maintain Portfolio/CV

"My best work was a website that no longer exists…"

60% of the websites disappear just after 2 years.

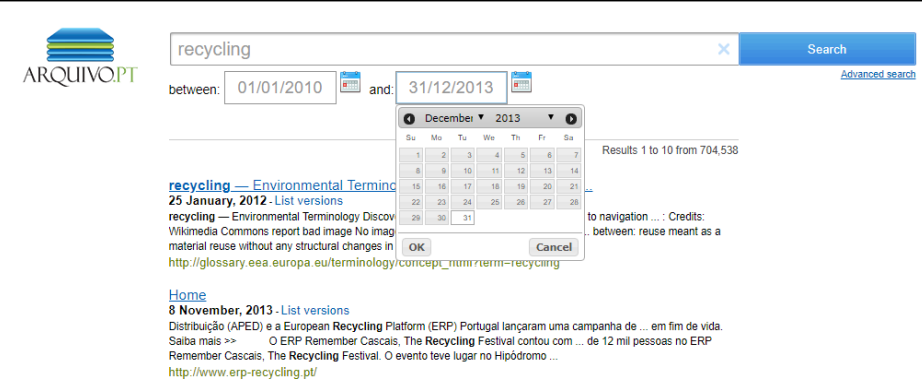**Server not found**

Firefox can't find the server at order.waytekwire.com.

- Check the address for typing errors such as **ww**.example.com instead of **www**.example.com
- If you are unable to load any pages, check your computer's network connection.
- If your computer or network is protected by a firewall or proxy, make sure that Firefox is permitted to access the Web.

[Try Again]
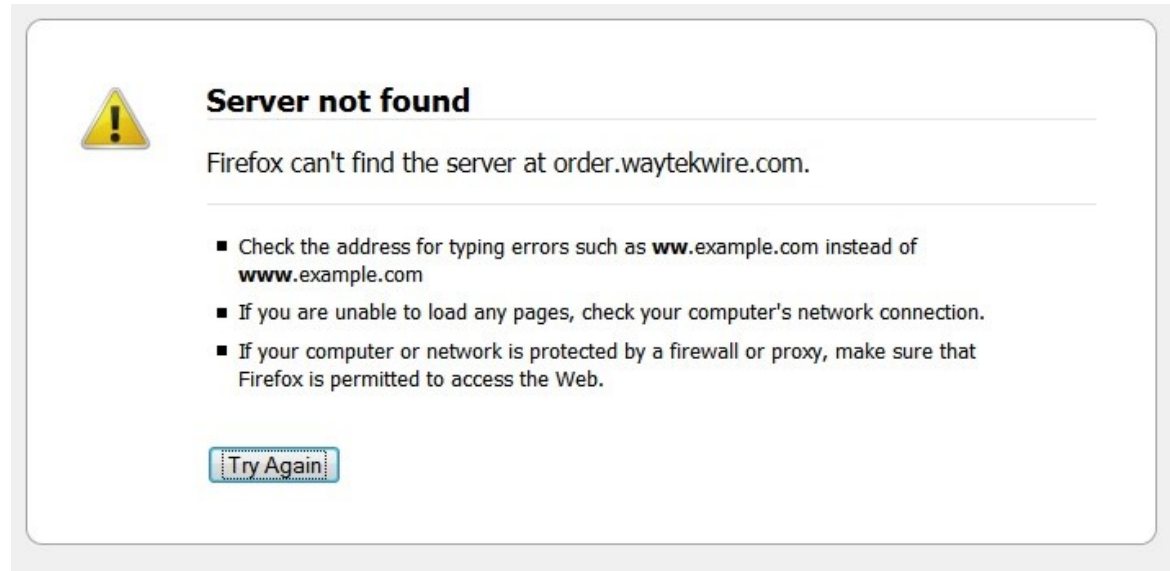
Support fast qualitative analysis for historical research

PORDATA: **quantitative** analysis

Statistical event:
Recycling decrease 2010-2013

Why?

Arquivo.pt: **qualitative** analysis

Search for news related to recycling between 2010-2013

Search the Past Web

Anyone can do it.

URL Search:
Internet Archive Wayback Machine

Archive.org

World wide content

URL search

New York Times news on
February 14th, 2009

New York Times (2005)

New York Times on 9/11, 2001

New York Times on the early days of the Web

URL search across several web archives

Memento Time Travel Portal

TimeTravel.mementoweb.org

Limitation of URL search

Users **do not know** the
exact URL of the **web page of the Past** that contained the information
that they search for.

Search for: Michael Jackson concert before 1998.
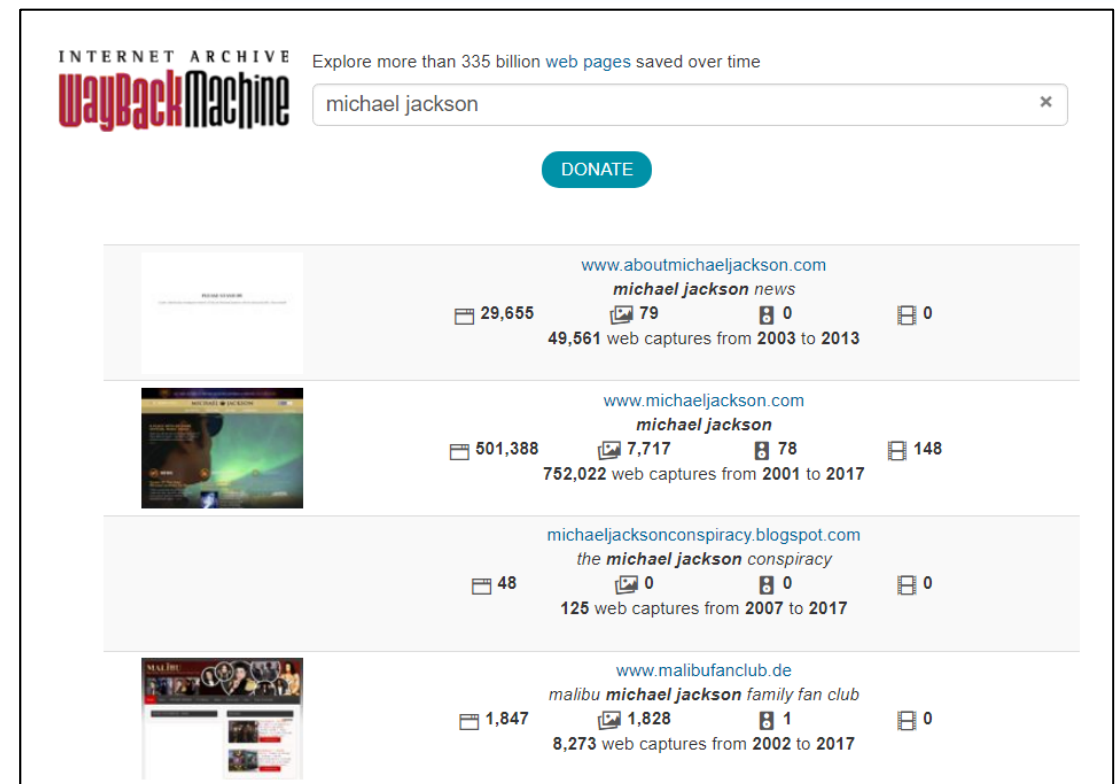
Which **was** the address of the web page?
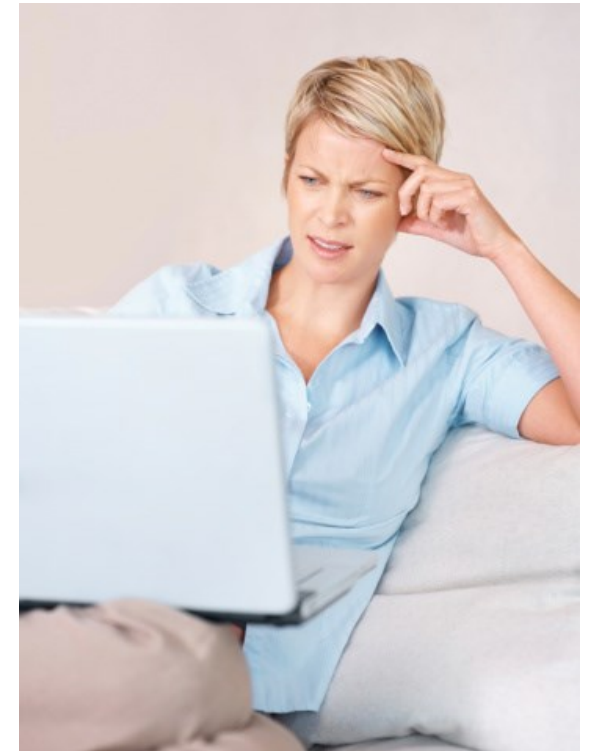
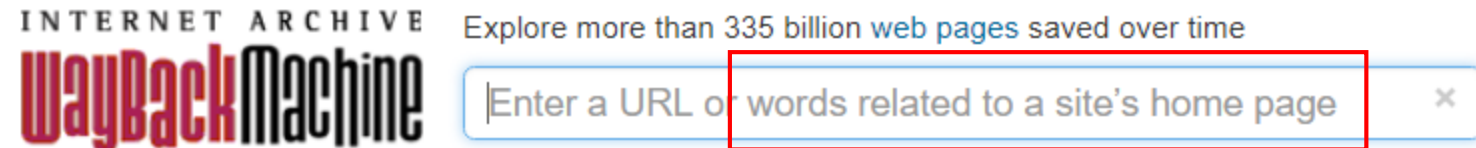# Keyword search related to home pages of websites
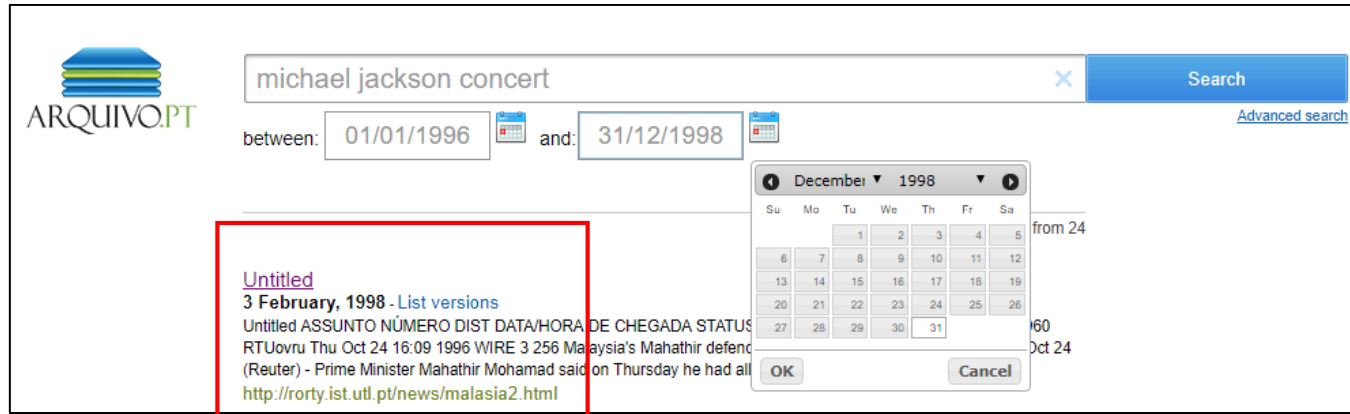

Archive.org

Keywords from the links to the homepages of websites.

Search results **do not include:**

- Home pages that do not receive links

- Internal webpages (not home pages)

- Cannot define time span for search

Full-text search over all pages with Arquivo.pt
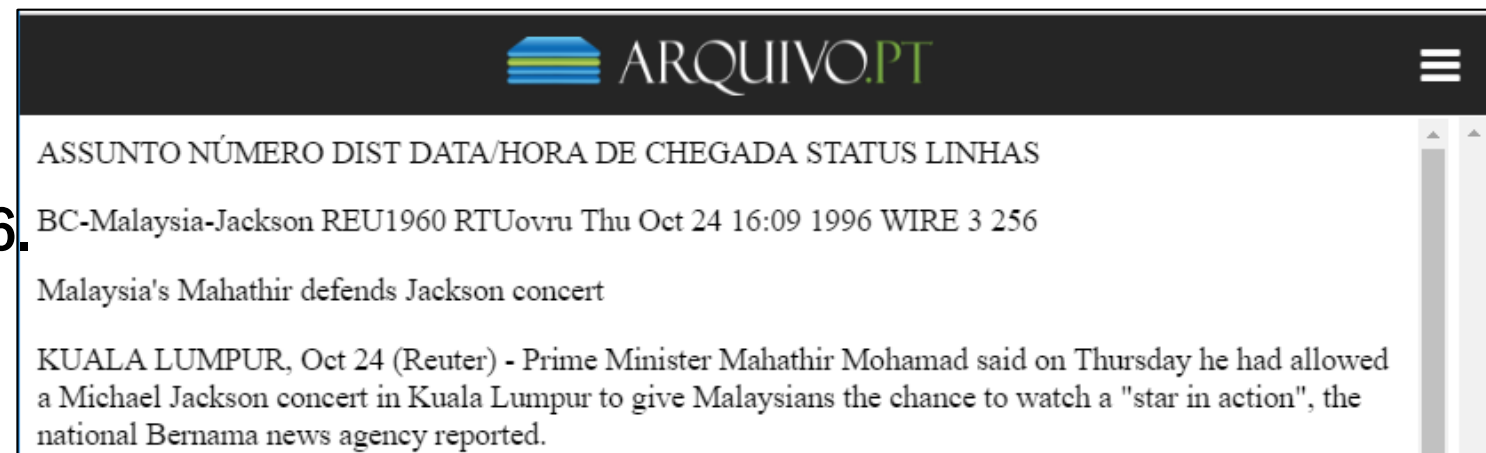
Not site home page

"Untitled"

There was a controversial Michael Jackson concert in Malaysia in 1996.

Advanced search:
did Michael Jackson ever came
to Portugal?

Results in Portuguese can be automatically translated to answer our question.

Did Michael Jackson ever came to Portugal? Yes, in 1992.

Access the Past Web

Access web pages from the Past

Navigate along **past versions**

Browse links across pages
- – Destiny version with the closest date to the origin page

Navigate **between web archives**

Powered by:
TimeTravel.memento.org

Complete page with resources from external web archives
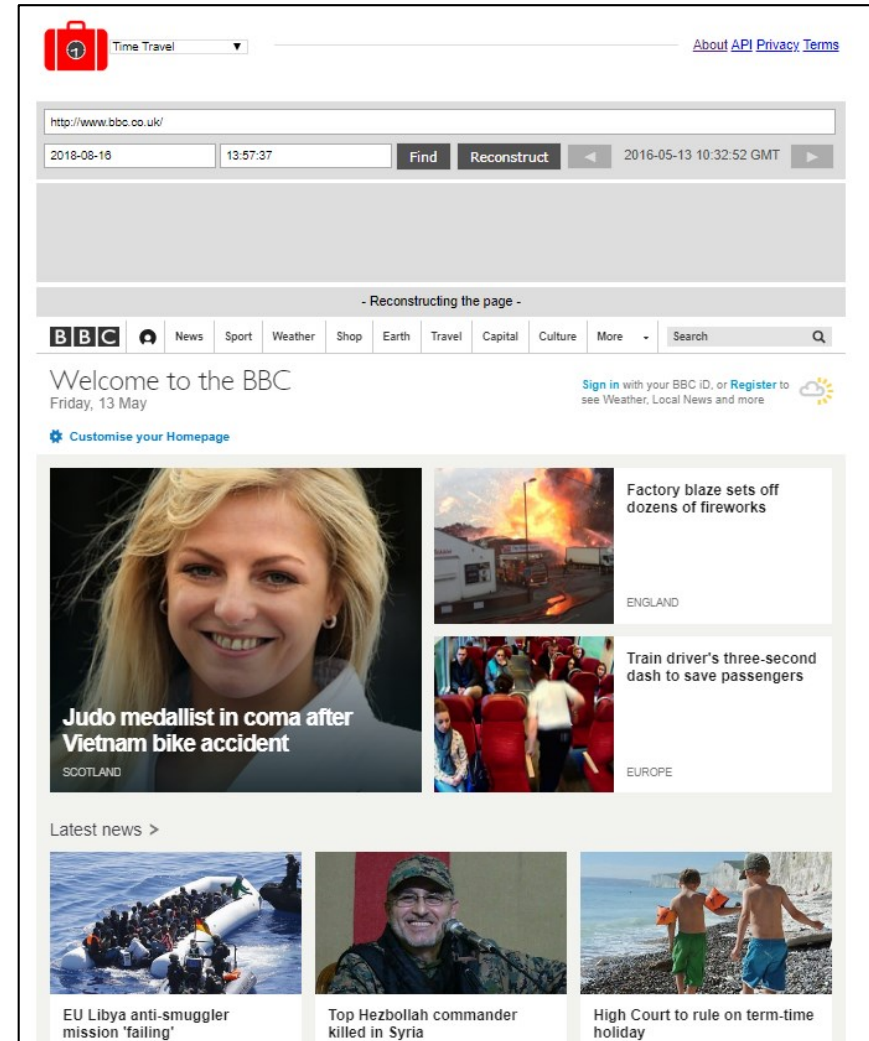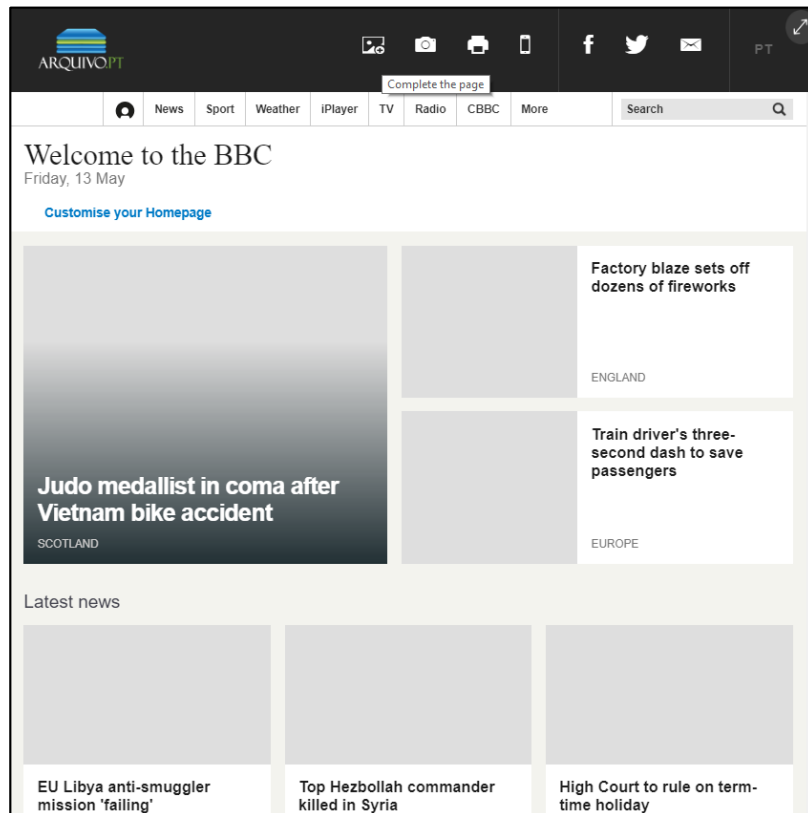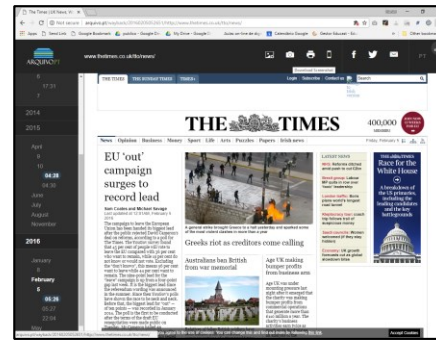
Save **full page** instead of partial "Print Screen"

Embed in documentation (e.g. thesis, papers)

Keep local copies

Oldweb.today: geocities.com, 1996

1997 browser:
Netscape Navigator v.3.04, Macintosh
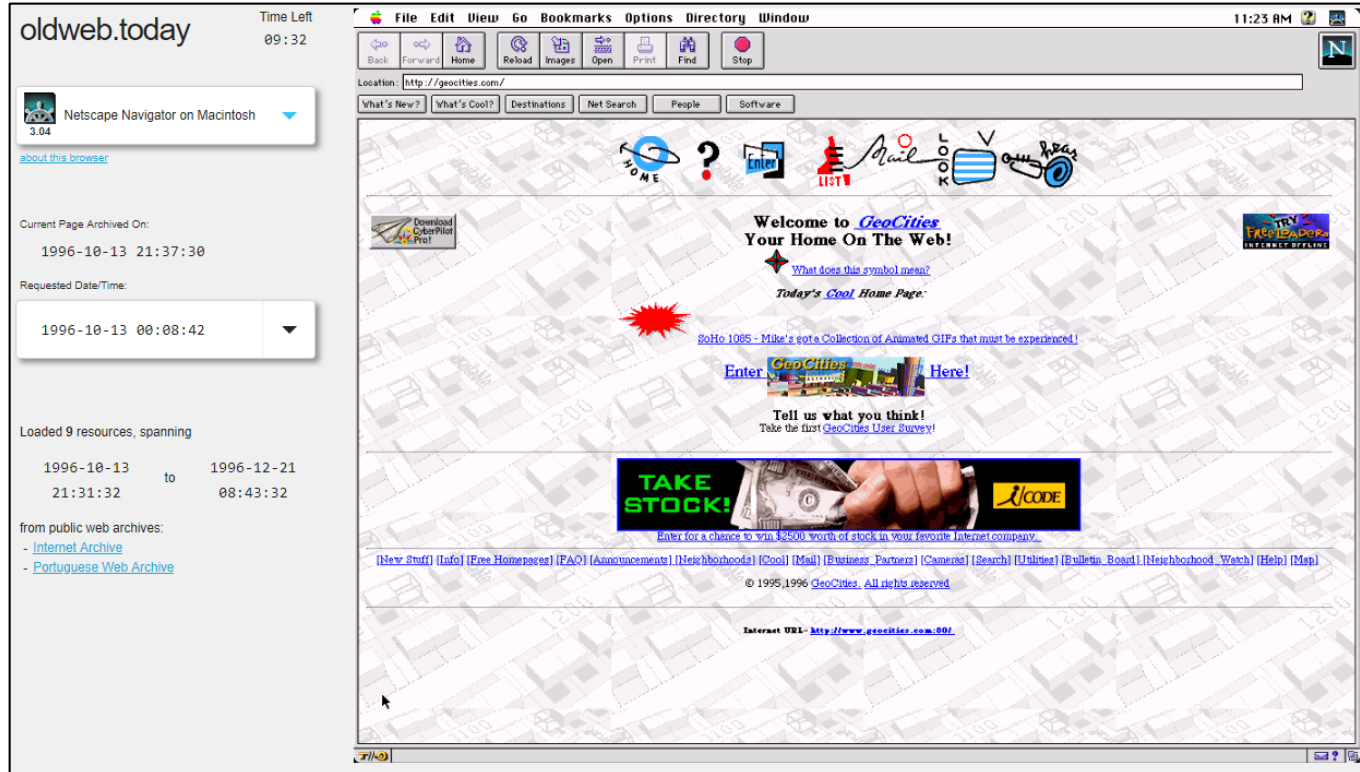
2018 browser:
Google Chrome v.68, Windows 10
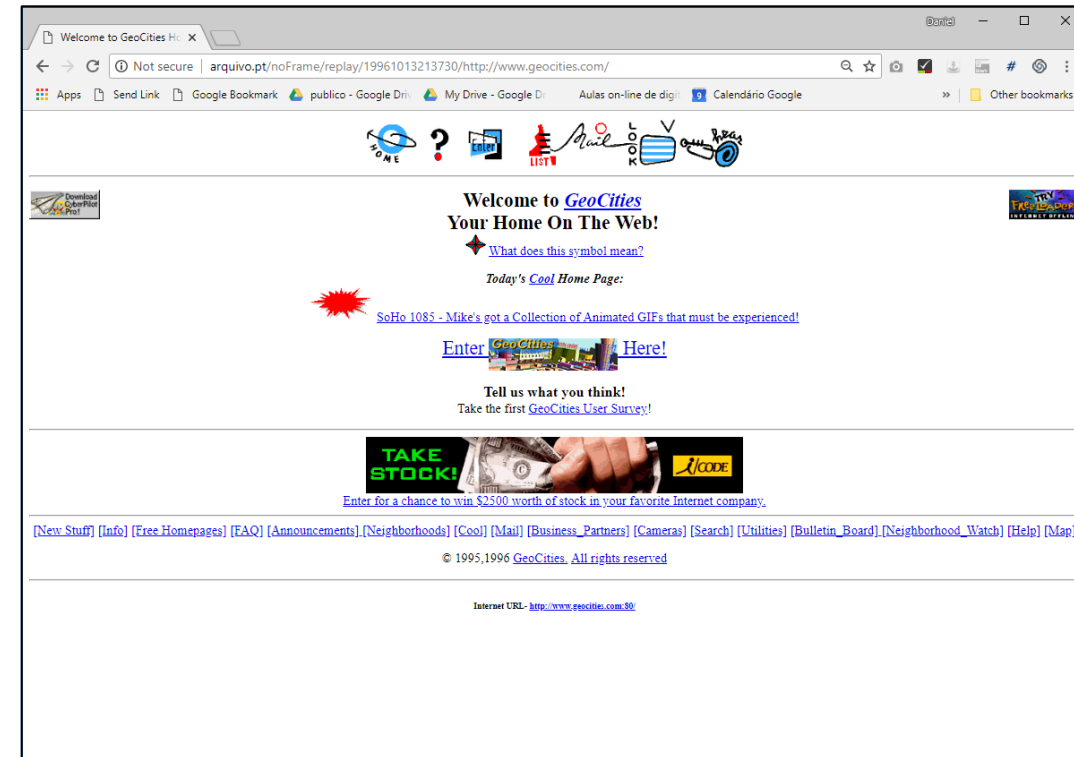
Search and Access the Past Web

Search the Present -> Web Search Engines

Search the Past   -> Web Archives!

Free online services

- Archive.org: URL search

- TimeTravel.mementoweb.org: search across web archives

- Arquivo.pt: full text search

- Oldweb.today: access using old browsers

Tutorial outline

Research the Past Web using Web archives

1. **Search and access**
   – The Past Web: examples and use cases
   – Public online services

2. **Publish and preserve**
   – Recommendations to publish preservable information
   – How to preserve information collected from the Web

3. **Automatic processing**
   – Protocols
   – Application Programming Interfaces (API)

arquivo.pt/training