Research the Past Web using Web archives
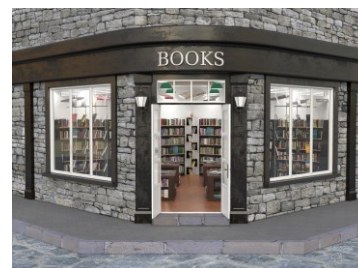
Daniel.Gomes@fccn.pt

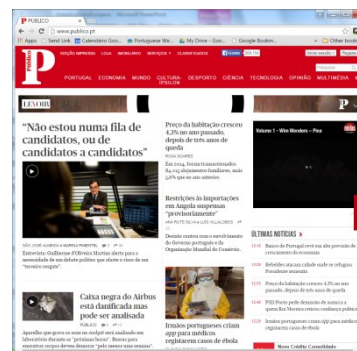# 80%

of web pages
change their content or disappear
within 1 year

Archives are also require to preserve information published online
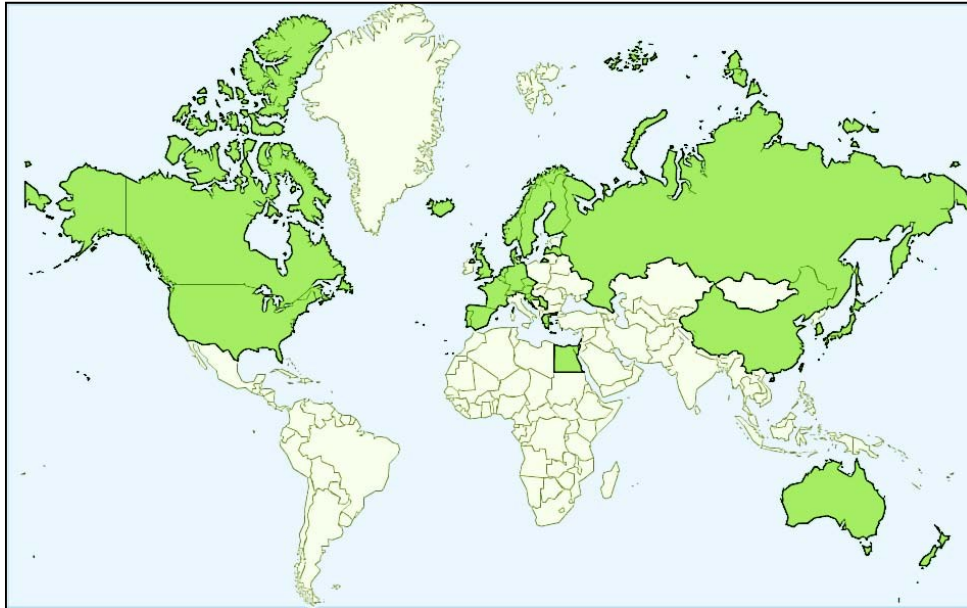
Preserve the Press

Preserve the Web

[Archive.org](): the first web archive

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

93 **web archiving** initiatives around the World (2020)

List of Web archiving initiatives, Wikipedia

Collaborative collections

IIPC members are collaborating to build public web archive collections based on themes or events that align with the following criteria:

- of high interest to IIPC members
- broader than any one member's responsibility or mandate
- of higher value to research due to broader perspective provided by multiple institutions
- transnational in scope.

The project is administered by the IIPC Content Development Group. More information about the project is available on their page.

Collections

🔗 Novel Coronavirus (Covid-19) (2020)
🔗 Climate change (2019)
🔗 Online News Around the World (2018)
🔗 Olympics/Paralympics collections (2010-2018)
🔗 European Refugee Crisis (2015-2016)
🔗 World War One Commemoration (2015-)
🔗 Intergovernmental Organizations (2015-)

56 members, netpreserve.org

Interesting thematic collections

I work at the **Arquivo.pt web archive**

Research infrastructure provided by Foundation for Science and Technology (Portugal)

Preserves **publicly accessible** information related with:

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

Free online services to research the Past Web

Arquivo.pt is used word-wide

Country	Users	% Users
1. 🇵🇹 Portugal	87,532	47.69%
2. 🇺🇸 United States	34,784	18.95%
3. 🇧🇷 Brazil	6,991	3.81%
4. 🇯🇵 Japan	5,306	2.89%
5. 🇬🇧 United Kingdom	4,996	2.72%
6. 🇷🇺 Russia	3,227	1.76%
7. 🇩🇪 Germany	2,649	1.44%
8. 🇨🇦 Canada	2,475	1.35%
9. 🇪🇸 Spain	2,385	1.30%
10. 🇮🇳 India	2,086	1.14%

52% of our users were international in 2019

User Interfaces and documentation also in English

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

The Past Web

Examples of international web-archived pages and collections

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 (N...

Before going to CERN, Tim was a founding director of Image Computer S... consultant in hardware and software system design, real-time communicat... 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.
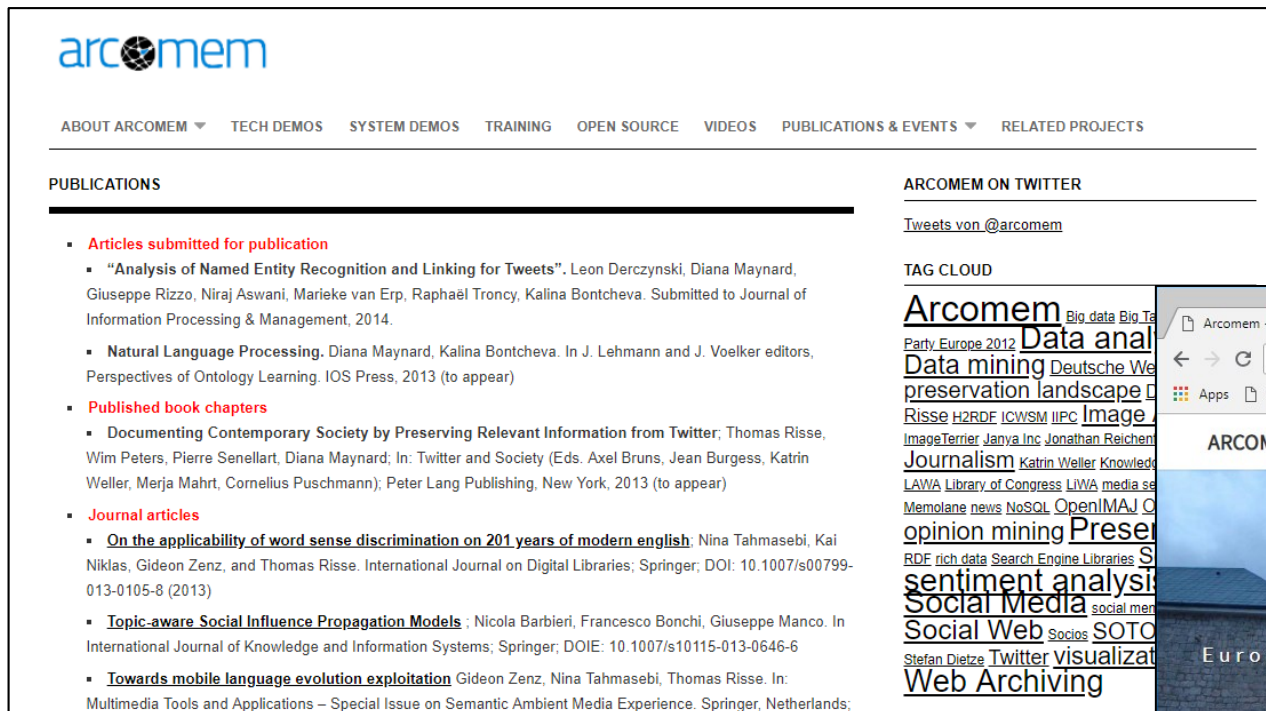
W3C works with the global community to produce specifications and reference software. W3C is funded by industrial members but its products are freely available to all. The Consortium is run by MIT LCS and by INRIA, in collaboration with CERN where the web originated. Please see the list of members to 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

Websites that document R&D projects funded by the European Union

arcomem.eu, 2014
@Internet Archive

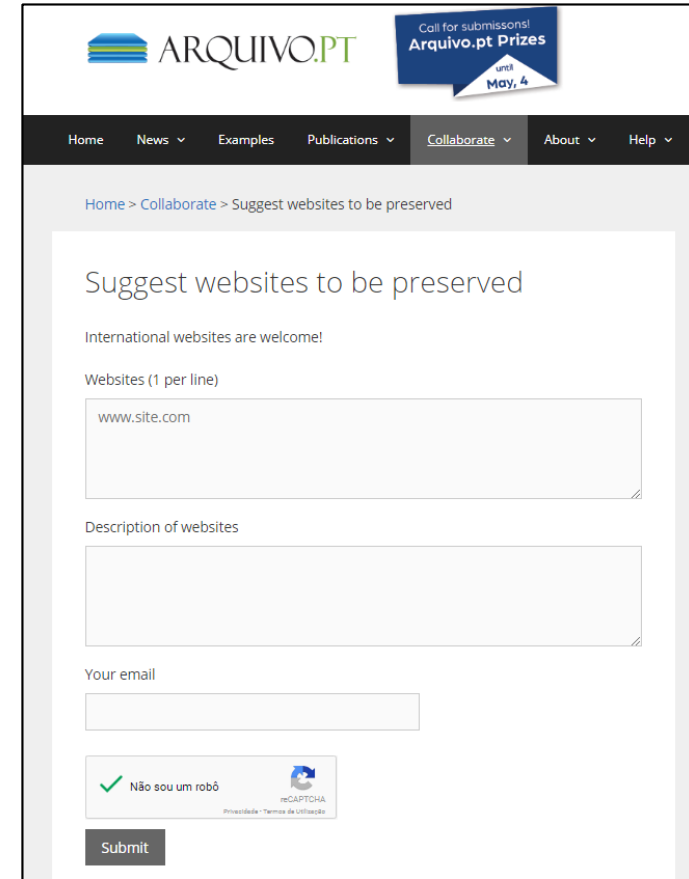# Cross-lingual collaborative web collections (e.g. European Elections 2019)

Suggest **any website** before it vanishes and you miss it!

Any domain

Any language

Broad interest for research and education

Save your research outputs!

arquivo.pt/suggest

Use cases for a web archive

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

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

P.A. Ioannidis*,†‡

» Abstract
Full Text
Full Text (PDF)

...nd Epidemiology,
...reece. E-mail:

+ Classifications
+ Services

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   Issue

Article first published online: 3 NOV 2006

DOI: 10.1002/asi.20513

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

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

14

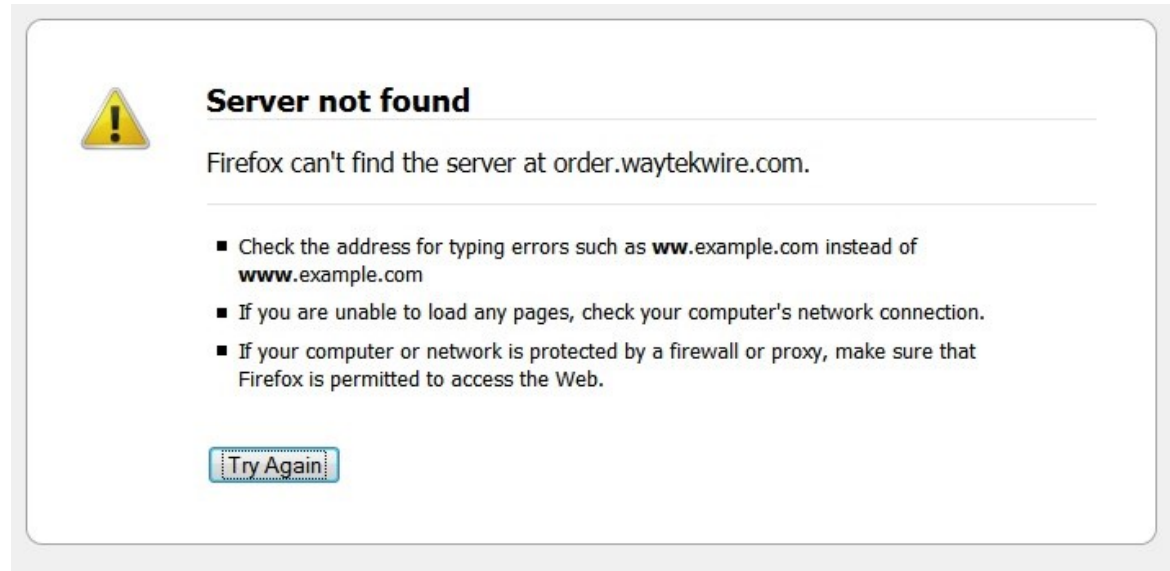# Management: documentation/audit of finished project

"Everything was on the website"

Maintain Portfolio/CV

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

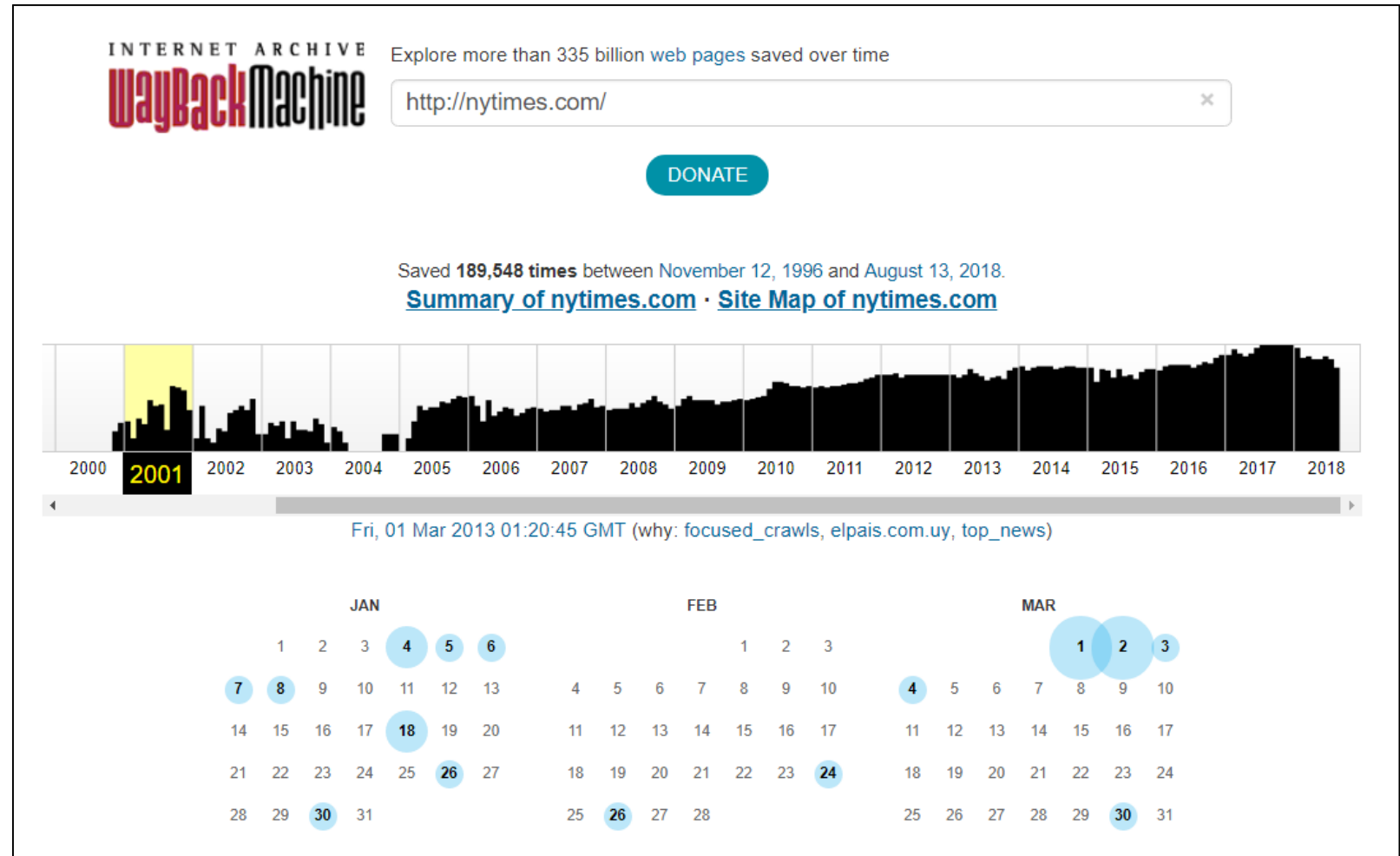60% of the websites disappear just after 2 years.

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

The New York Times

Saturday, February 14, 2009   Last Update: 1:11 AM ET

Search | Try Our **EXTRA** Home Page | ▸ Get Home Delivery | Personalize Your Weather

JOBS
REAL ESTATE
AUTOS
ALL CLASSIFIEDS

WORLD
U.S.
POLITICS
N.Y./REGION
BUSINESS
TECHNOLOGY
SPORTS
SCIENCE
HEALTH
OPINION
ARTS
Books
Movies
Music

Stimulus Plan Sets New Limits on Executive Pay

By EDMUND L. ANDREWS and ERIC DASH

A provision buried in the stimulus bill passed by Congress would impose restrictions on executive bonuses at companies reaping bailout money.

Stimulus Plan Receives Final Approval in Congress

Stephen Farrell/The New York Times

Iraqi Militants Become Patrons of the Arts
By STEVEN LEE MYERS

OPINION »

TALK SHOW
When Updike and Cheever Came to Visit
Dick Cavett recalls the rare thrill of having two great writers as guests on the same show.

· Collins: Stock a Cabinet
  ☞ Comments
· Editorial: California Prisons
· Op-Ed: No Price on Love
· Op-Ed: Baby Bailout
· Op-Ed: Stagflation
· Opinionator: Renditions

MOVIES »

Lesser-Known Mobsters
"Gomorrah" takes an unsentimental look at Italian organized crime.

BUSINESS »

A Stress Test for the Latest Bailout Plan
The stress test finds out how crippled big banks would be

New York Times on 9/11, 2001

New York Times on the early days of the Web, 1996

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?
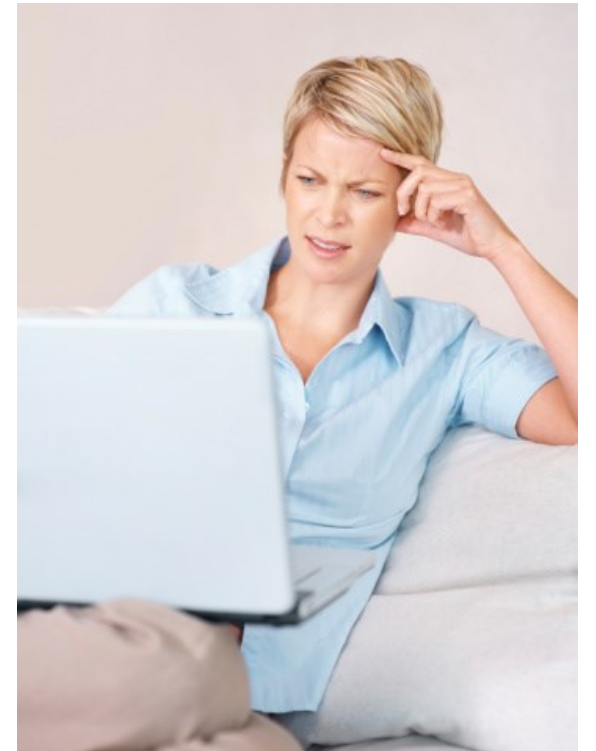
Keyword search related to home pages of websites

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.

Search **images** from the past
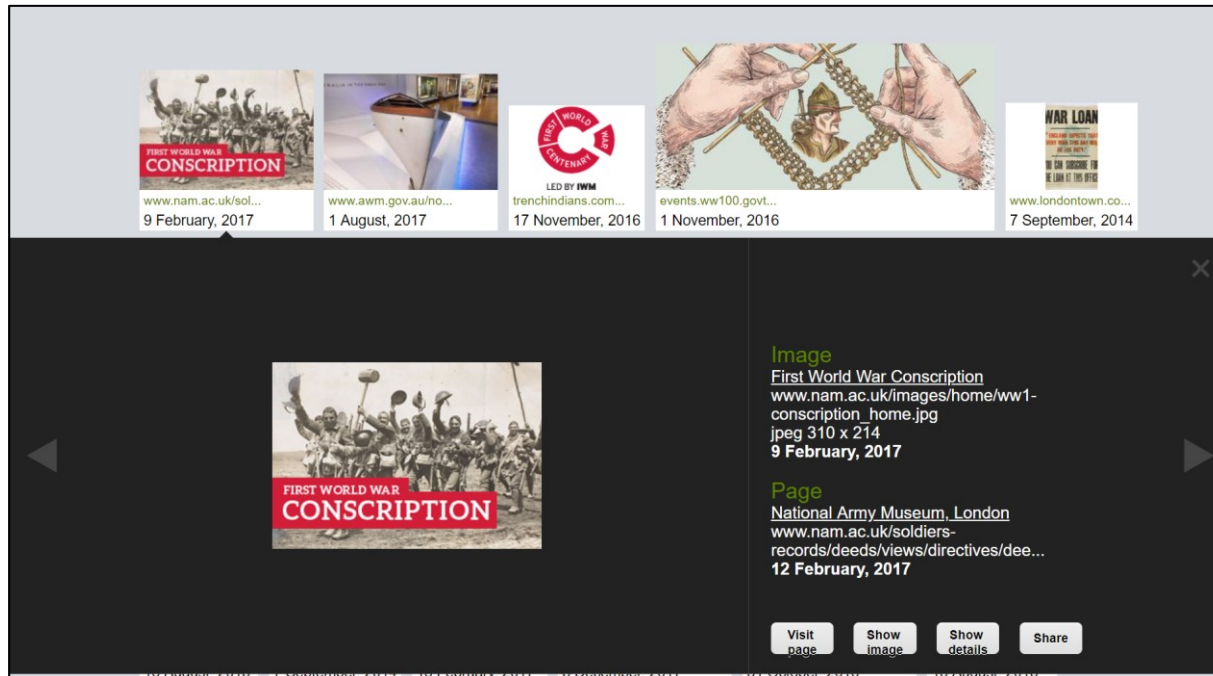
Image **source** and details

Application **P**rograming **I**nterfaces to develop new applications

Arquivo.pt APIs

Your new app
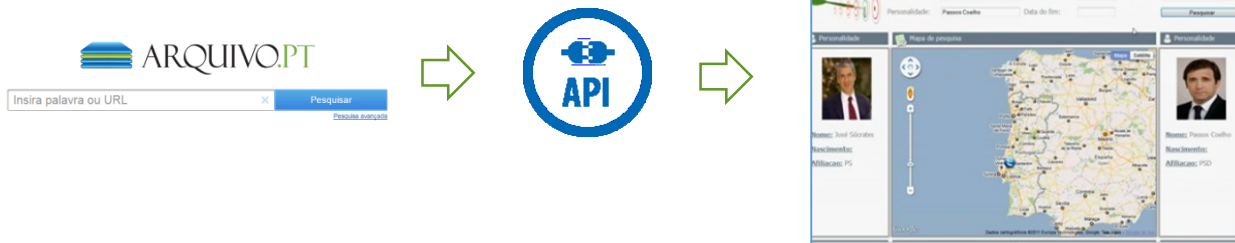
arquivo.pt/api

Custom search engine through "collection" API parameter over 3 crawls of .EU domain (2015-16-17)

arquivo.pt/resawdev/

Innovative service based on Arquivo.pt API: "Tell Me Stories"

Automatically generates narratives about any subject based on online news from the Past.

Winners of the Arquivo.pt Award 2018.
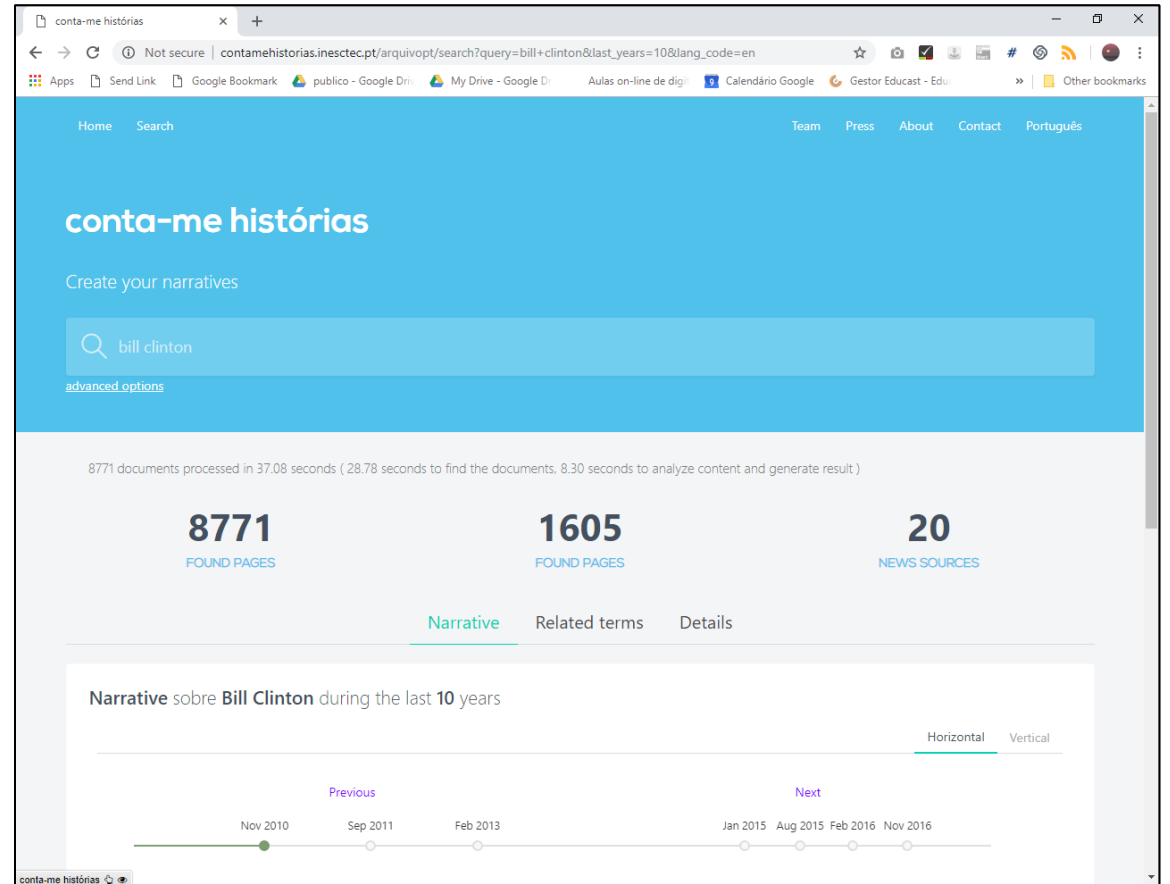
ContaMeHistorias.pt (with English UI)
TellMeStories.pt

ARQUIVO.PT

Arquivo.pt Award
2020
Apply until May 4th

- Any subject

- Arquivo.pt as main source of information

- Submissions in Portuguese

- arquivo.pt/awards

1st place: 10 000 €

2nd place: 3 000 €

3rd place: 2 000 €

ArchivesUnleashed.org:
automatic processing and analytics
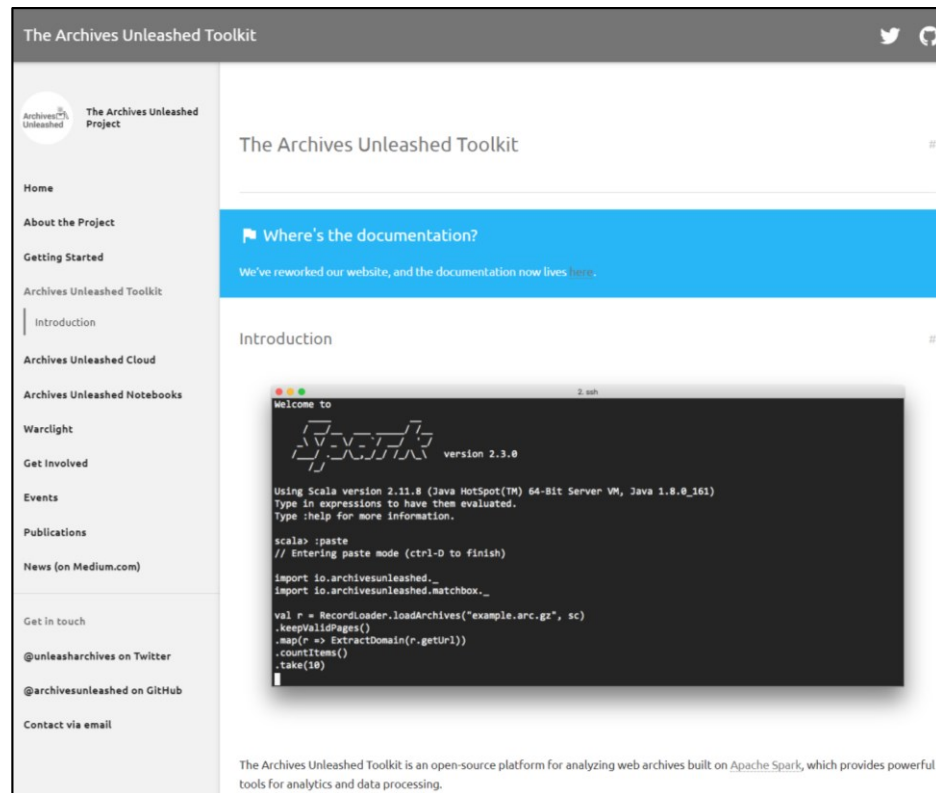
Cloud or Toolkit
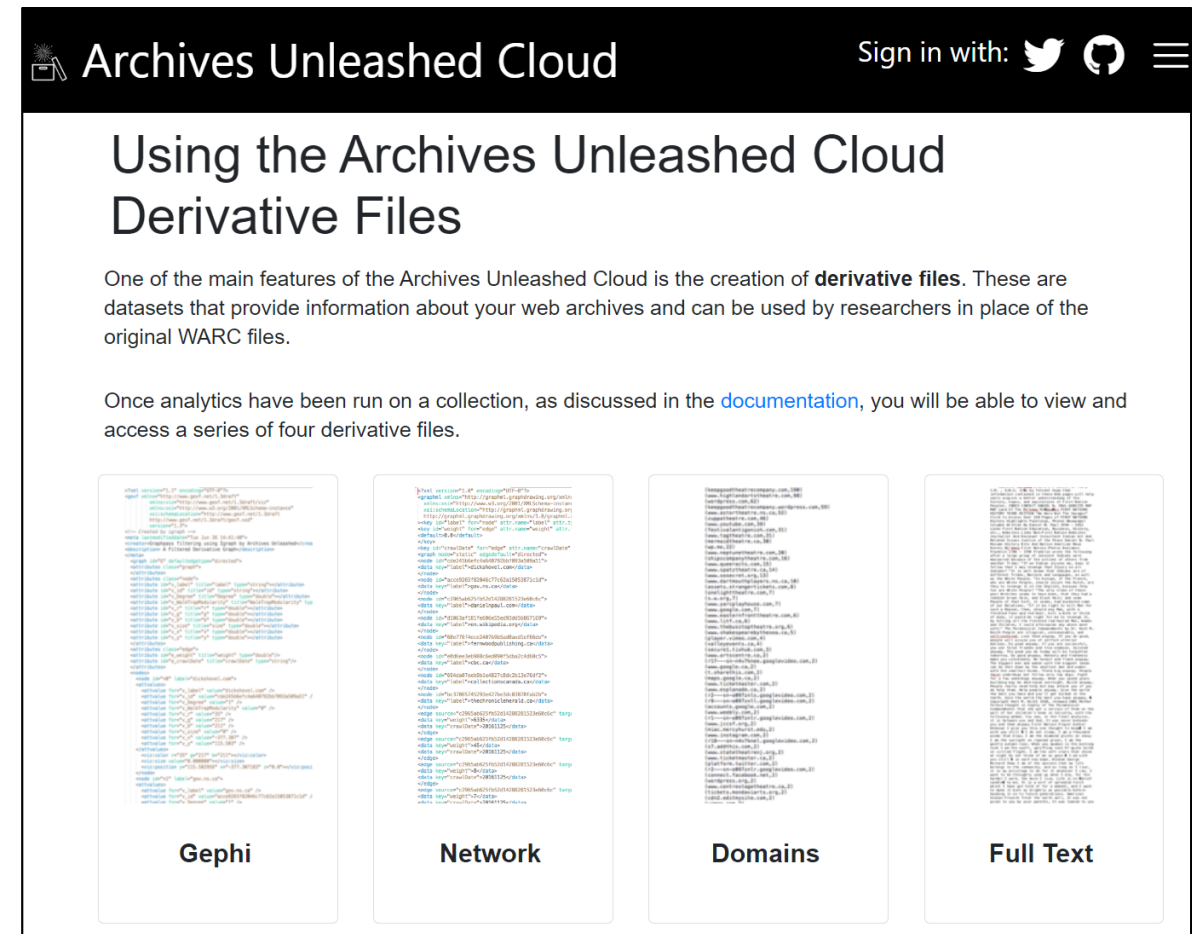
Generates derived results

Access the Past Web

Access web pages from the Past

Navigate along **past versions**
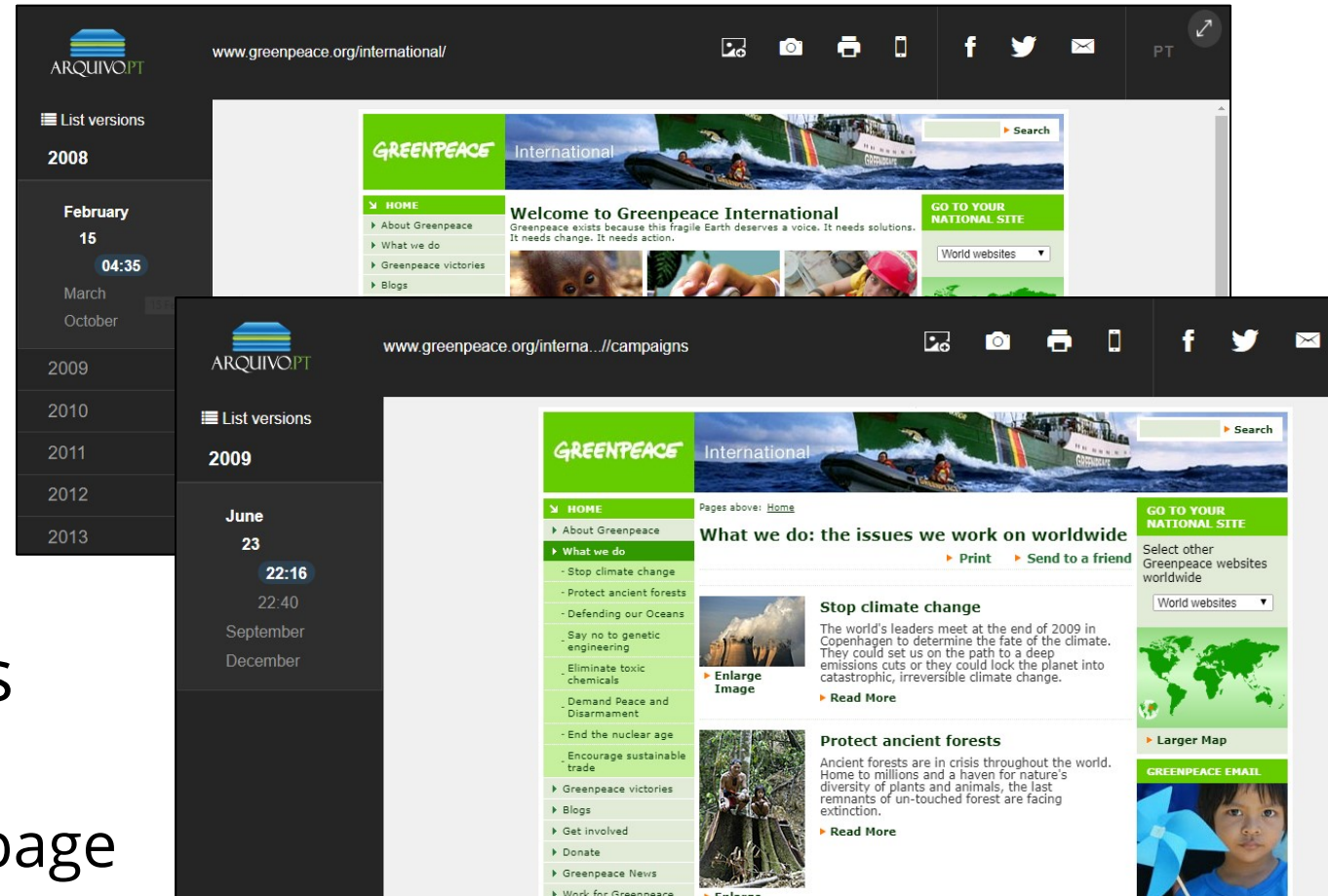
Browse links across pages
– Destiny version with the
closest date to the origin page

Navigate **between web archives**
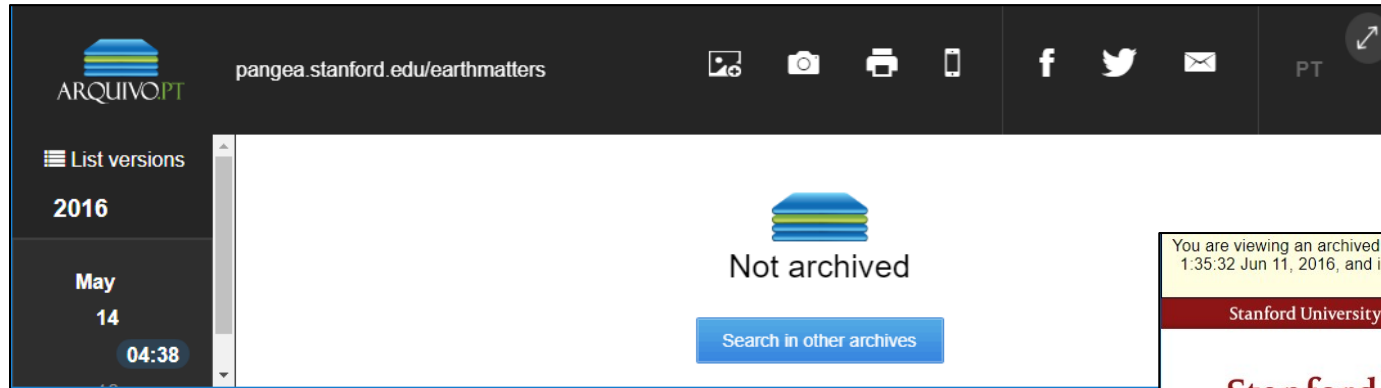
Powered by:
TimeTravel.memento.org

Complete page with resources from external web archives

Powered by: TimeTravel.memento.org
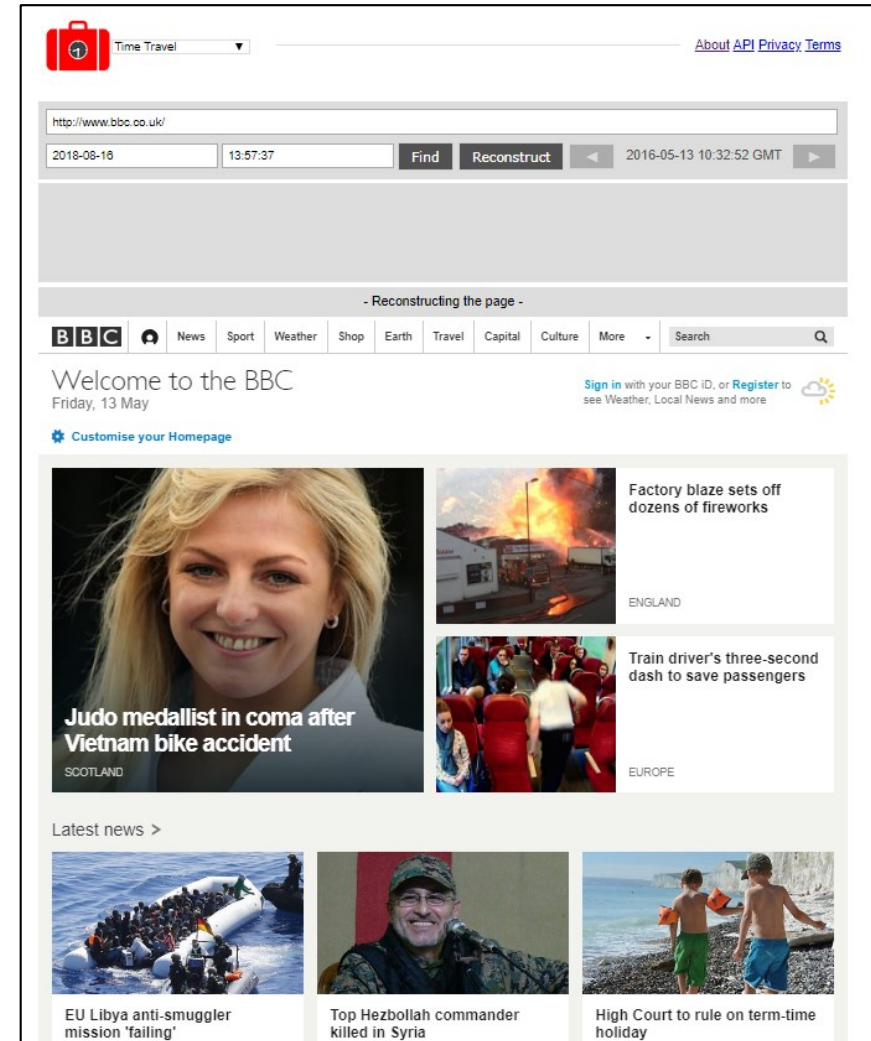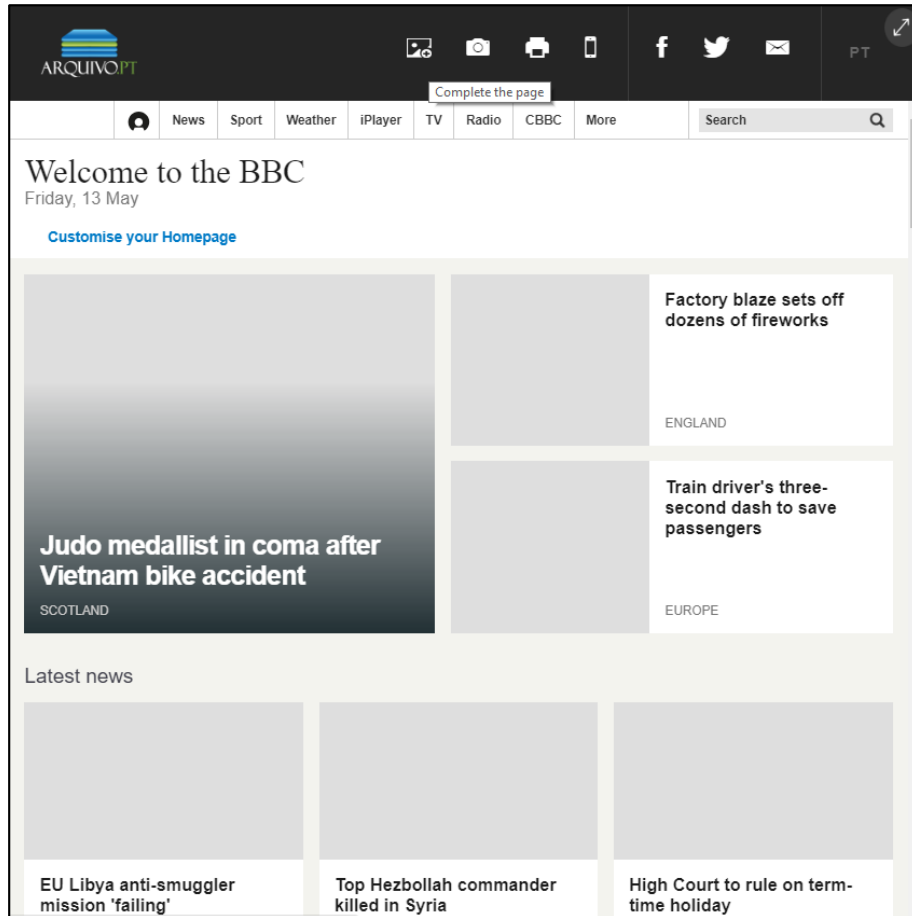
Save **full page** instead of partial "Print Screen"

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

Keep local copies
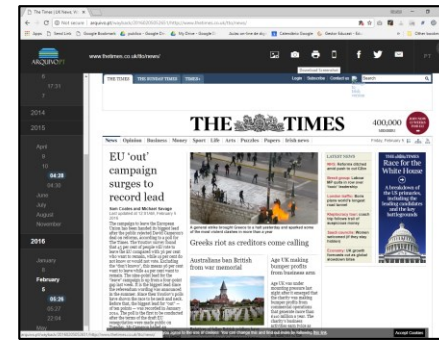
Access to **technical details**
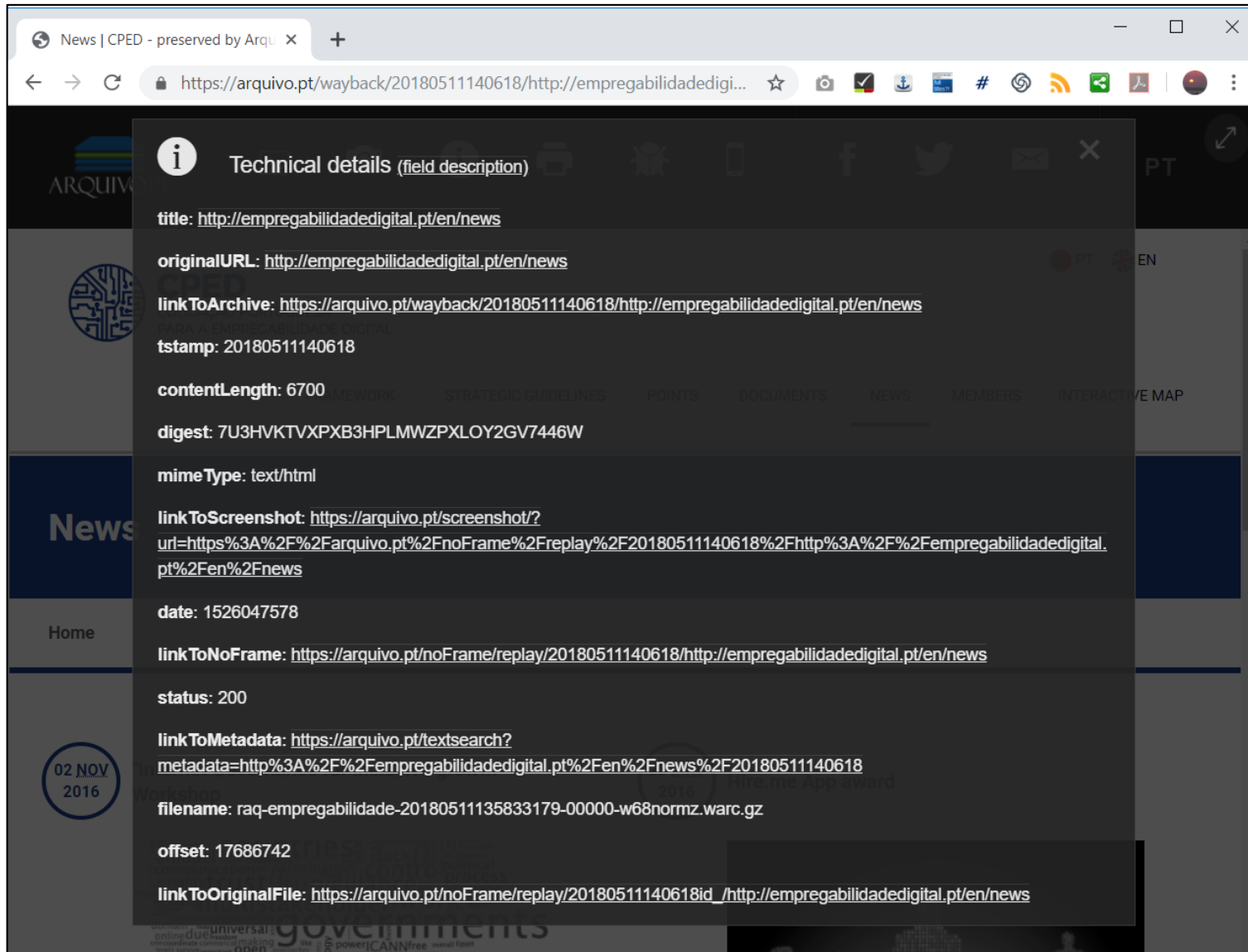
Useful for debug

Advanced users

– E.g. researchers

[Oldweb.today](https://oldweb.today): geocities.com, 1996

1997 browser:
Netscape Navigator v.3.04, Macintosh
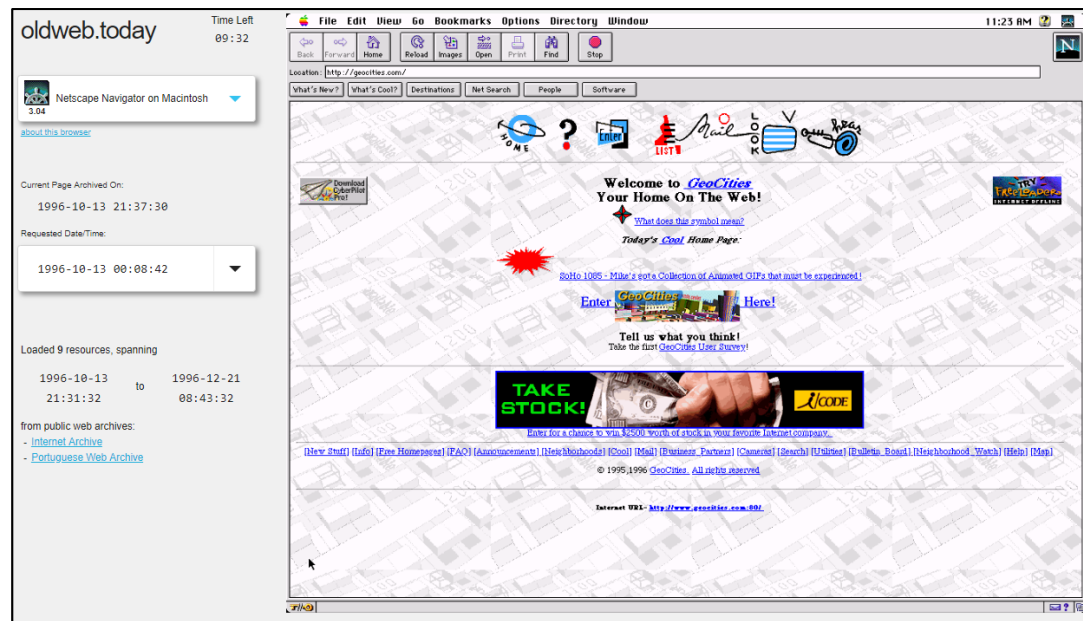
Modern browser:
Google Chrome, Windows 10
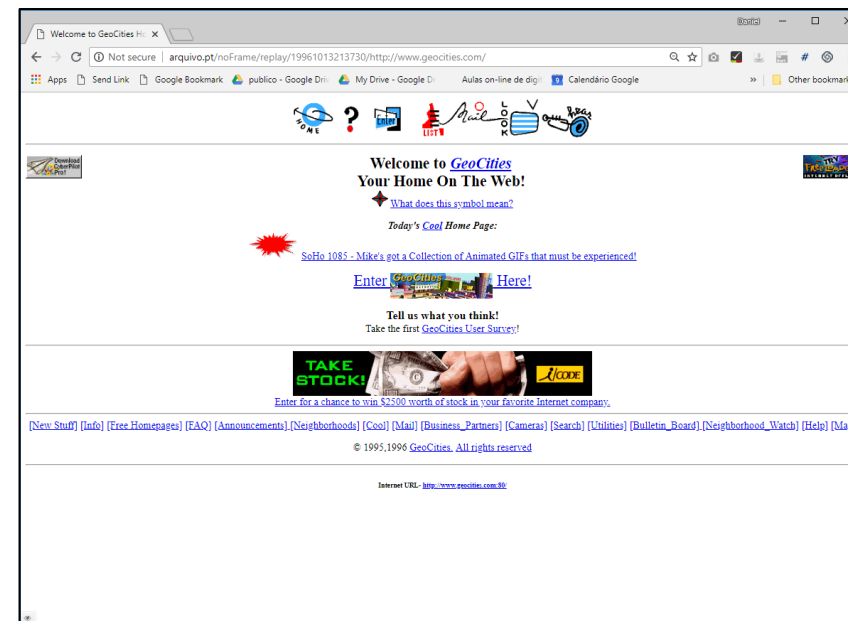
Training and research

Training courses on web preservation and research

New ways of searching the past (module A)

- **Any Internet user**

Publishing preservable information on the web (B)

- Web authors

Automatic processing of information preserved from the Web (C)

- Developers, **Researchers**

Do-it-yourself web archiving (D)

- **Researchers**, Librarians, Archivists

arquivo.pt/training

Investiga XXI (Research XXI)

Communication Studies

Information Science

Social Sciences

Transformations of the Websites of Portuguese Newspapers

FCSH on the Web: virtual exhibition

Straight-Edge in the Lisbon metropolitan area

Short link to this page: arquivo.pt/newspapers

Short link to this page: arquivo.pt/fcshontheweb

Short link to this page: arquivo.pt/straightedgeen

All videos, presentations, reports at: arquivo.pt/research

Web archives sinergies with CLEOPATRA

ESR 1. Fact extraction and cross-lingual alignment:
cross-lingual collections (e.g. IIPC Olympics/Paralympics collections, Arquivo.pt EU Elections 2019).

ESR 3. Crowd quality and training in hybrid multilingual information processing and analytics:
generate test collections and evaluate quality of web archive tools.

ESR 5. Fact validation across multilingual text corpora:
web archives as additional source of historical information for fact checking.

ESR 7. Relations of textual and visual information & ESR 8. Contextualisation of images in multilingual sources:
web archive image search.

ESR 9. National and transnational media coverage of European parliamentary elections, 2004-2014:
Arquivo.pt EU Elections 2019.

ESR 10. Nationalism, internationalism and sporting identity: the London and Rio Olympics/Paralympics:
IIPC Olympics/Paralympics collections, Arquivo.pt daily crawls of online news.

ESR 11. Information propagation with barriers:
study how information propagates across time on the Web, IIPC European Refugee Crisis.

ESR 12. Cross-lingual news reporting bias: IIPC Online News Around the World collection.

ESR 13. Multilingual Wikipedia as 'first draft of history': final layout of Wikipedia pages along time (not database dump).

ESR 14. NLP for under-resourced languages & ESR 15 Cross-lingual sentiment detection:
under-resourced languages in web archives.

Search and Access the Past Web

Search the Present -> Web Search Engines

Search the Past   -> Web Archives!

Useful research tools
- – Archive.org: URL search
- – TimeTravel.mementoweb.org: search across web archives
- – Arquivo.pt: full text search
- – ArchivesUnleashed.org: data analytics
- – Oldweb.today: access using old browsers

Keep in touch!

Subscribe our channels:

➢ arquivo.pt/subscribe: mailing list in English

➢ arquivo.pt/inscrever: mailing list in Portuguese

➢ arquivo.pt/news: social networks and videos