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
However, this valuable information quickly vanishes
80% of web pages change their content or disappear within 1 year.
Archives are also required 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”.
List of Web archiving initiatives

This page contains a list of Web archiving initiatives worldwide. For easier reading, the information is divided into three tables: web archiving initiatives, archived data, and access methods.

This Wikipedia page was originally generated from the results obtained for the research paper A survey on web archiving initiatives published by the Arquivo.pt team (the Portuguese Web-archiving team).

Web archiving initiatives

<table>
<thead>
<tr>
<th>Name</th>
<th>Country</th>
<th>Creation Year</th>
<th>Technologies</th>
<th>Number of Employees</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australians Web Archive</td>
<td>Australia</td>
<td>1996</td>
<td>PANDORA Digital Archiving System (PANDAS), NLA Trove, HTTrack</td>
<td>10</td>
<td>Full-time</td>
</tr>
</tbody>
</table>

The PANDORA Archive which takes a selective approach is a collaborative program of 11 agencies that provide an estimate of average monthly staffing equivalent to around 10 FTE. IT support provided by the National Library of Australia; 0.25 person-month. Whole .au domain harvests have been conducted annually since 2005 in collaboration with the Internet...
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

<table>
<thead>
<tr>
<th>Country</th>
<th>Users</th>
<th>% Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Portugal</td>
<td>45,891</td>
<td>46.56%</td>
</tr>
<tr>
<td>2. United States</td>
<td>25,373</td>
<td>26.19%</td>
</tr>
<tr>
<td>3. Brazil</td>
<td>2,266</td>
<td>2.25%</td>
</tr>
<tr>
<td>4. Russia</td>
<td>2,234</td>
<td>2.22%</td>
</tr>
<tr>
<td>5. United Kingdom</td>
<td>2,231</td>
<td>2.22%</td>
</tr>
<tr>
<td>6. Japan</td>
<td>2,172</td>
<td>2.16%</td>
</tr>
<tr>
<td>7. Canada</td>
<td>1,237</td>
<td>1.23%</td>
</tr>
<tr>
<td>8. Mozambique</td>
<td>1,213</td>
<td>1.20%</td>
</tr>
<tr>
<td>9. India</td>
<td>902</td>
<td>0.90%</td>
</tr>
<tr>
<td>10. Germany</td>
<td>894</td>
<td>0.89%</td>
</tr>
</tbody>
</table>

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 companies with the mission to realize the full potential of the Web. He is a Principal Research Scientist in 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 Ltd., a consultant in hardware and software system design, real-time communications processing; and a principal engineer with Plessey Telecommunications, in Oxford University.

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

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

www.w3.org, 1996
@Arquivo.pt
No Google, no Facebook, you had to navigate through the links
Web-born information **NOT** collected online: web pages from CD-ROMs (digital archeology)


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

arepublicano.blogspot.com, 2010
@Arquivo.pt
Open-access research data and publications
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.
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.
Support fast qualitative analysis for historical research

PORyDA: 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

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

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
Op-Ed: 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
Plan to Transform Brooklyn Waterfront Advances at City Hall
By DIANE CARDWELL  6:28 PM ET
Officials agreed on a plan that would transform the north Brooklyn waterfront into a neighborhood of residential towers.

African Strain of Polio Virus Hits Indonesia
By DONALD G. MCNEIL JR.  9:17 PM ET
The case of polio indicates that an outbreak spreading from Nigeria has reached the world's fourth-most-populous country.
World Trade Center Toppled in Attack

Thousands Fear Dead in New York and Washington
By JAMES BARRON
In what appeared to be coordinated attacks, hijacked jetliners crashed into both towers of the World Trade Center and the Pentagon Tuesday morning.

In Washington, an Attack on a Symbol of American Power
By DAVID STOUT
Shortly after the attacks in New York, another aircraft crashed into the Pentagon, the massive building that is the very symbol of the American defense establishment.

New York Times on the early days of the Web

Welcome! First-time users register here or take a QuickTour of the site.
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.


Which was the address of the web page?
Keyword search related to home pages of websites

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
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.

In Portugal

Michael Jackson performed only once in Portugal.

The show was the 26th of September of 1992 at the Estádio José Alvalade in the capital Lisbon. The singer only spoke to the audience twice, saying "I love you" ("I love you") the start and "peace" ("Peace") in farewell.
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](http://TimeTravel.memento.org)
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)