A Memory Infrastructure
By Daniel Gomes

Daniel Gomes
daniel.gomes@fccn.pt
Fact:
The Web has been replacing printed media
However, all these valuable information quickly disappears
80%
Disappear or change within 1 year
The right to be forgotten is naturally granted on the Web
The right to be remembered on the Web is a challenge.
77 web archiving initiatives across the world to preserve Humanity's online memory
Most web archives are part of the IIPC
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 in 2012.
Archive.pt in a nutshell

➔ A free preservation service provided to web authors
  ◆ Only preserves publicly available information
  ◆ Mainly related to the Portuguese community
  ◆ Respects imposed exclusions (robotstxt.org)
  ◆ Opt-out option for authors

➔ A past-web search engine available to web users
  ◆ A complement for live-web search engines
  ◆ A “Google” for the past!
1 700 million archived files since 1996 (57 TB) ...and growing
How can anyone find web archived information?
Using **URL search** like the Internet Archive Wayback Machine
Problem with URL search

➔ Users do not know the URL from the past that contained the information that they need.
Archive.pt also provides full-text search over 1.2 billion web files archived since 1996.
Archive.pt is used by international users

- Archived web pages written in several languages
- Includes all Portuguese speaking domains (.AO, .MZ, .CV) except Brazil
- 22% of the Sessions were not from Portugal
Web archives.
For Who? For What?

➔ Journalist revisits past news
➔ Historian analyses digital documents
➔ Professor recovers presentation
➔ Manager documents project
➔ Web user recovers page from broken link
➔ ...
➔ All of us.
Access mechanisms for automatically processing web-archived data
OpenSearch programming interface access to extend functionality
OpenSearch to develop innovative services over web-archived information

Web Applications course to develop information aggregators about:

- Politicians (2011)
- Football (2012)
Big Data analytics over web-archived data through our large-scale processing platform (Hadoop cluster)
Hadoop cluster used to automatically measure the accessibility of the Web for people with disabilities

In Rui Lopes, Daniel Gomes, Luís Carriço, Web Not For All: A Large Scale Study of Web Accessibility, 2010.
Which kind of information can we find on Archive.pt?

Everything.

Examples of web-archived pages
The first Portuguese web page (90's)
International events

2011 Egyptian revolution: a web-based revolution, a web-archived revolution
Regional events

Results for the 2001 Portuguese elections
Open access Science: publications, test collections, data sets, tools. (1998-2010)

Linguateca project: resources for the computational processing of the Portuguese language
Previous governmental portals (2011)

New government erased previous content
The official gazette of the Republic of Portugal is published exclusively online since 2006 (e.g. DRE.pt contains all the Laws)

➔ 4 copies exclusively prepared and printed for preservation
➔ Printed documents are usually digitized and put online for preservation!

Define o Novo Regime de Publicação Exclusivamente Electrónica do Diário da Assembleia da República e Novas Regras para o Uso de Novas Tecnologias de Informação e Comunicação no Trabalho Parlamentar

Resolução da Assembleia da República n.º 68/2003, de 7 de Agosto

A Assembleia da República resolve, nos termos do n.º 5 do artigo 166.º da Constituição, o seguinte:

Artigo 1.º

Diário da Assembleia da República

1 - A partir de 15 de Setembro de 2003, a 1.ª série do Diário da Assembleia da República passa a ser exclusivamente disseminada em formato electrónico através do portal da Assembleia da República na Internet.

2 - A 2.ª série do Diário da Assembleia da República deixará igualmente de ser publicada em suporte tradicional, devendo ser adoptadas todas as medidas necessárias para que a respectiva publicação electrónica integral ocorra no mais curto prazo.

3 - A edição electrónica do Diário da Assembleia da República faz fé plena e a publicação dos actos através dela realizada vale para todos os efeitos legais e regimentais, devendo ser utilizado mecanismo que assinale, quando apropriado, a respectiva data e hora de colocação em leitura pública.

4 - Os serviços preparam, editam e depositam na Biblioteca da Assembleia da República e na Biblioteca Nacional quatro exemplares de uma versão impressa das duas séries do Diário, preparada unicamente para tal efeito.
Web archived DRE.pt since 2004
Printed publications are also web archived

→ Newspaper front pages on 24 May 2009
Old photos from the pre-Digital Era that were published online (1966)
Historical documents with added value (meta-data)

Blog post with digitization of the republican newspaper (1915)
Even insignificant individual information. Teenager blogged: “math test went bad”
What if this teenager will become a...

➔ Famous singer?
   (who used to be a "nerd")

➔ Physics Nobel?
   (that was not that good at math)
Alert!
People are not preserving their digital data

Web archives may be the only source of memories to many people
Collaborate with Archive.pt

We need your collaboration!
For non-techies: suggest the archive of interesting sites

arquivo.pt/sugerir
For techies: All our software is freely available and you can improve it

pwa-technologies at code.google.com
For web authors: follow recommendations to publish preservable sites

Recommendations for authors to enable web archiving

This page presents recommendations to create web pages that can be efficiently archived and accessed across time. It is impossible to preserve the web without the authors collaboration.

Last update on September 16, 2010.

A web archive works similarly to a web search engine like Google. Therefore, the presented recommendations contribute to create contents that can be correctly processed or both these types of systems.

The following recommendations were classified in fundamental and advisable. They were written with the purpose of helping any author. Please contact us if you have any doubt.

Web site organization

- One link for each content (fundamental)
- Crawler-friendly homepage (fundamental)
- Maintain address to the same content across time (advisable)
- Robots Exclusion Protocol to indicate access restrictions (advisable)

arquivo.pt/recommitations
For everybody: disseminate Archive.pt

arquivo.pt/news
Final message

Web archives are crucial infrastructures for Humanity's Memory
Thank you!

daniel.gomes@fccn.pt