Preserving the .EU domain

Daniel Bicho
daniel.bicho@fccn.pt
Arquivo.pt is a research infrastructure for the study of web archived material.
What is .EU domain?

- .EU is the top-level domain for the European Union (EU).
- Launched on 7 December 2005, the domain is available for organisations and residents of EU member states.
- There are more than 3.9 million registered .eu domain names. [https://www.eurid.eu/en/about-us](https://www.eurid.eu/en/about-us)
Why preserve .EU domain?

- An **European Identity** on the Internet for European citizens.
"One very simple, but radical, idea: to democratise Europe."
Why preserve .EU domain?

- Used by the European Union institutions (europa.eu).
European Parliament

http://www.europarl.europa.eu
Why preserve .EU domain?

- European Research and Development projects
Thorax project

http://www.thorax-project.eu/
Neuroscience project

http://www.neurobid.eu/

Summary

Brain disease, encompassing both psychiatric and neurologic disease, represents a major public health problem in Europe and worldwide. It is one of the most prevalent and costly groups of diseases in Europe, with estimated total annual costs (2004 data) amounting to 386 billion Euro or 829 Euro per individual living in Europe (1). Data collected by the World Health Organization (WHO) suggest that brain diseases are responsible for 35% of Europe's total disease burden (2). An analysis of all epidemiological and health economic studies in Europe published by the European Brain Council in 2005 found that across 28 European countries (the European Union plus Iceland, Norway, and Switzerland) with a total population of 466 Million, 127 million (or 27%) of Europeans, children and adults at all stages, are affected by at least one form of brain disease (1). Improved prevention and treatment strategies are paramount in the attempt to reduce the individual and societal burdens associated with cerebral palsy, multiple sclerosis, stroke, schizophrenia and the many other devastating disorders that involve the central nervous system (CNS). In this scenario, prevention and treatment options rely heavily upon optimized pharmacologic access to the CNS. The major barriers between the systemic and CNS compartments are the blood-brain parenchyma barrier (microvessels) and blood-cerebrospinal fluid barrier (choroid plexus). Throughout this project, we refer to both collectively as the blood brain barrier (BBB) unless directing a specific research topic at only one of these interfaces.
Not preserved website

Making SME meat processors more competitive by using automatic in-line fat analysis in meat trimmings

Automatic in-line fat analysis in meat trimmings

The TRIMSCAN solution consists in the development of an in-line fat analyser to scan pork and beef trimmings contained in plastic bins. The solution uses a contactless system to predict the fat content with an expected accuracy below 1.5%.

TRIMSCAN will provide SME meat processors with a practical tool at an affordable cost to make them more competitive by reducing lean meat over-use and to avoid the need of inaccurate and time consuming manual sampling and recipe reformulation.

TRIMSCAN will significantly improve the production of processed meat by homogenising the quality of the final product and by reducing the percentage of rejections due to products that do not meet the legislation and/or customer of specifications.

[Read More]

http://trimscan.eu
Not preserved website


Hrm.

Wayback Machine doesn't have that page archived.
Who preserves the .EU domain?

● National institutions preserve national domains.

● Which institution preserves the .EU domain?

● This can lead to the loss of European Cultural Heritage for future generations.
RESAW

RESAW is an European network that aims to create a Research Infrastructure for the Study of Archived Web Materials established in late 2012.

Arquivo.pt is one of the Web Archiving initiatives participating in RESAW.

International and National Projects.

http://resaw.eu
Archiving .EU project

- A pilot exploratory crawl of the .EU domain.
- The objective was to gain insight about how to preserve the .EU domain.
How a Crawler Works

Crawler is a robot that automatically navigates from page to page.
- Crawler gets a start page (seed URL).
- Extracts addresses available on the page.
- Follows the extracted addresses to new pages.
- Repeat process for all sites.
URL Seeds

- We gathered seeds from several sources:
  - HttpArchive - http://www.httparchive.org
  - Websiteisworth - http://goo.gl/0if6pe
  - Search our Crawl Logs for .eu URL addresses.
  - Google Search Queries for .eu URL addresses.

- Gathered a total of 34 thousand URL seeds.
First crawl Configuration

- 5 links followed from homepage.
- 10,000 files per site.
Information Crawled

- 1 million hosts.
- 250 million files.
- Crawler ran for 23 days.
- 5.8 Terabytes of .eu domain content stored.
Second Crawl Configuration

- 5 links followed from homepage.
- 1,000 files per site.
Information Crawled

- 72 thousand hosts.
- 14 million files.
- Crawler ran for 19 days.
- 3.1 Terabytes of .eu domain content stored.
What we learned

● Redirects to other top level domains are common.
  ○ 9% of the URL seeds were redirected to other top level domains.

● Poorly designed online shops that generate infinite number of URLs. (http://autobazar.eu)
What we learned

● Lots of Web Spam.
  ○ Link farm sites. (http://in-links.eu/)

● Sub-domains that referenced multilingual versions.
Search Prototype

- We developed a URL and Full-Text Search prototype on the crawled content.

- Full-Text Search makes more easy the research for crawled contents on .EU domain.

- Publicly accessible through: http://arquivo.pt/resawdev
Search Prototype

Search from the past

Know the project

http://arquivo.pt/resawdev
Search Prototype

European Research Projects

between: 01/01/1996 and: 31/12/2015

ERC: European Research Council | ERC: European Research Coun…
21 November, 2014 - other dates
ERC: European Research Council | ERC: European Research Council Jump to Navigation Advanced Search | Glossary / A-Z | Sitemap Home Funding and Grants Projects and Results Media and Events About ERC ... ERC Grants Open also to non-European researchers ERC News 17.11.14 ERC President speaks to research ...
http://erc.europa.eu/

Research Projects / University of Ostrava
22 November, 2014 - other dates
Research Projects / University of Ostrava Home UO Portal FAQ Contacts UO Premises Site Map ... Art. Activity Main Areas of Research Research Projects Seventh Framework Programme Faculties and ... Institute for Research and Applications of Fuzzy Modeling European Research Institute for Social Work ...

European Commission
21 November, 2014 - other dates
European Commission Skip to main content Language български (bg) čeština (cs) dansk (da) Deutsch ... European Commission News Highlight 01/11/2014 The Juncker Commission A strong and experienced team ...
http://ec.europa.eu/index_en.htm

http://arquivo.pt/resawdev
Research Data


- Talk with us if you need more information for your research.
Future/Current Work

- Preservation of R&D projects.
  - they publish valuable scientific outputs.
  - they vanish shortly after the project funding ends.
  - constitute a trans-national, multi-lingual and cross-field set of historical web data for researchers (e.g. social scientists).
Open Data Portal

https://open-data.europa.eu/
<table>
<thead>
<tr>
<th>#</th>
<th>C</th>
<th>H</th>
<th>K</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>acronym</td>
<td>Development of an automated classification system for polished stone slabs and tiles</td>
<td>projectUrl</td>
<td>limitation in quality control performance represent</td>
</tr>
<tr>
<td>2</td>
<td>PROVIPS</td>
<td>ABC GATES FOR EUROPE</td>
<td><a href="http://abc4eu.com/">http://abc4eu.com/</a></td>
<td>During the last years, many ABC Gates have been</td>
</tr>
<tr>
<td>3</td>
<td>ACOBAR</td>
<td>Acoustic Technology for observing the interior of the Arctic Ocean</td>
<td><a href="http://acobar.nerc.no/">http://acobar.nerc.no/</a></td>
<td>ACOBAR will develop an observing system for t</td>
</tr>
<tr>
<td>5</td>
<td>ADAPO</td>
<td>Development of an automatic process of in-house collection, storage and application of adaptive bacteria culture for fish farms</td>
<td><a href="http://adapond.eii.ee/">http://adapond.eii.ee/</a></td>
<td>Evidence from recirculation systems shows tha</td>
</tr>
<tr>
<td>6</td>
<td>NANOFOODS</td>
<td>Development of foods containing nanoencapsulated ingredient</td>
<td><a href="http://allegati.unina.it/ricerca/progi">http://allegati.unina.it/ricerca/progi</a></td>
<td>The preservation of bioactive food ingredients</td>
</tr>
<tr>
<td>7</td>
<td>ANTICARB</td>
<td>Monoclonal ANTIBody-targeted CARBon nanobues against cancer</td>
<td><a href="http://anticarb.org/">http://anticarb.org/</a></td>
<td>ANTICARB attempts to exploit the advantages</td>
</tr>
<tr>
<td>8</td>
<td>APARET</td>
<td>African Programme for Advanced Research Epidemiology Training</td>
<td><a href="http://aparet.org/index.php?option">http://aparet.org/index.php?option</a></td>
<td>The APARET fellowship programme will catalys</td>
</tr>
<tr>
<td>9</td>
<td>AQUAVALENS</td>
<td>Protecting the health of Europeans by improving methods for the detection of pathogens in drinking water and water used in food preparation</td>
<td><a href="http://aquavalens.org/">http://aquavalens.org/</a></td>
<td>The AQUAVALENS consortium has brought toget</td>
</tr>
<tr>
<td>10</td>
<td>ARCH</td>
<td>Agenda for research on chernobyl health (ARCH): to develop of a strategic plan for research on the health consequences of radiation from the chernobyl accident</td>
<td><a href="http://arch.iarc.fr">http://arch.iarc.fr</a></td>
<td>The Chernobyl accident led to the most serious</td>
</tr>
</tbody>
</table>
Recovering project URLs

- 25,608 R&D projects funded by FP7 listed by the EU Open Data Portal.

- Only 7.9% had an associated project URL.

- Researching techniques to automatically identify this project URLs.