Opportunities and challenges in collecting and studying national webs

Daniel Gomes
Collecting the national Web of Portugal

Broad crawls: .PT + user suggestions
Daily crawls: 361 selected websites
Special crawls: events such as elections
High-quality crawls: on-demand
An attempt to archive the .EU domain
Tried to perform a collaborative about R&D projects.

Google Sheet to gather websites about R&D and development projects.

Over 25 000 project on the EU database.
Archive the .EU domain: a “brute-force” attempt to preserve R&D websites

3 crawls of .EU domain

<table>
<thead>
<tr>
<th>Time of crawl</th>
<th># of Files collected</th>
<th>Data volume (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21/11/2014 to 2014/12/16</td>
<td>129 793 987</td>
<td>5.8</td>
</tr>
<tr>
<td>2016/01/07 to 2016/01/26</td>
<td>61 863 684</td>
<td>3.1</td>
</tr>
<tr>
<td>02/06/2017 to 10/07/2017</td>
<td>105 823 552</td>
<td>11</td>
</tr>
</tbody>
</table>

Table 1. Comprehensive .eu crawls 2014-2017

Searchable and accessible at: https://arquivo.pt/resawdev

Main problem: web spam

Main problem: web spam
Automatic selection and preservation of websites related to R&D projects

thorax-project.eu, 2014
Studying past webs
Training courses on web preservation and research

New ways of searching the past

Any Internet user

Publishing preservable information on the web

Web authors

Automatic processing of information preserved from the Web

Developers

arquivo.pt/training
Investiga XXI (Research XXI)

Communication Studies
Transformations of the Websites of Portuguese Newspapers

Information Science
FCSH on the Web: virtual exhibition

Social Sciences
Straight-Edge in the Lisbon metropolitan area

All videos, presentations, reports at: arquivo.pt/research
Any subject
Arquivo.pt as main source of information
Submissions in Portuguese

[link to information]
arquivo.pt/prizes

1st place: 10 000 €
2nd place: 3 000 €
3rd place: 2 000 €