Opportunities and challenges in collecting and studying national webs

Daniel Gomes
Collecting the national Web of Portugal

**Trimestral broad** crawls: .PT + user suggestions

**Daily selective** 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 collection to preserve R&D project websites

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

Over 25 000 projects 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>Millions of files collected</th>
<th>Data volume (TB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>November, December 2014</td>
<td>129</td>
<td>5.8 TB</td>
</tr>
<tr>
<td>January 2016</td>
<td>61</td>
<td>3.1 TB</td>
</tr>
<tr>
<td>June, July 2017</td>
<td>105</td>
<td>11 TB</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>295</strong></td>
<td><strong>20 TB</strong></td>
</tr>
</tbody>
</table>

Searchable and accessible at: arquivo.pt/resawdev

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

Only 29% of the URLs were under the .EU domain.
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

arquivo.pt/awards

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