Opportunities and challenges in collecting and studying national webs

Daniel Gomes

Collecting the national Web of Portugal

Trimestral broad crawls: .PT + user suggestions

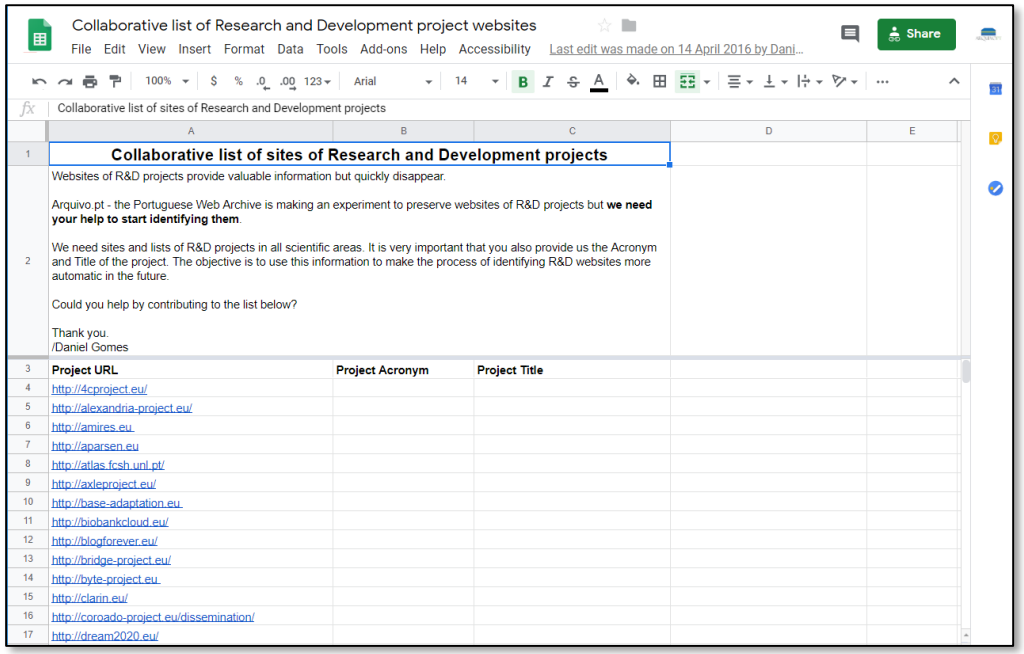
Daily selective crawls: 361 selected websites

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

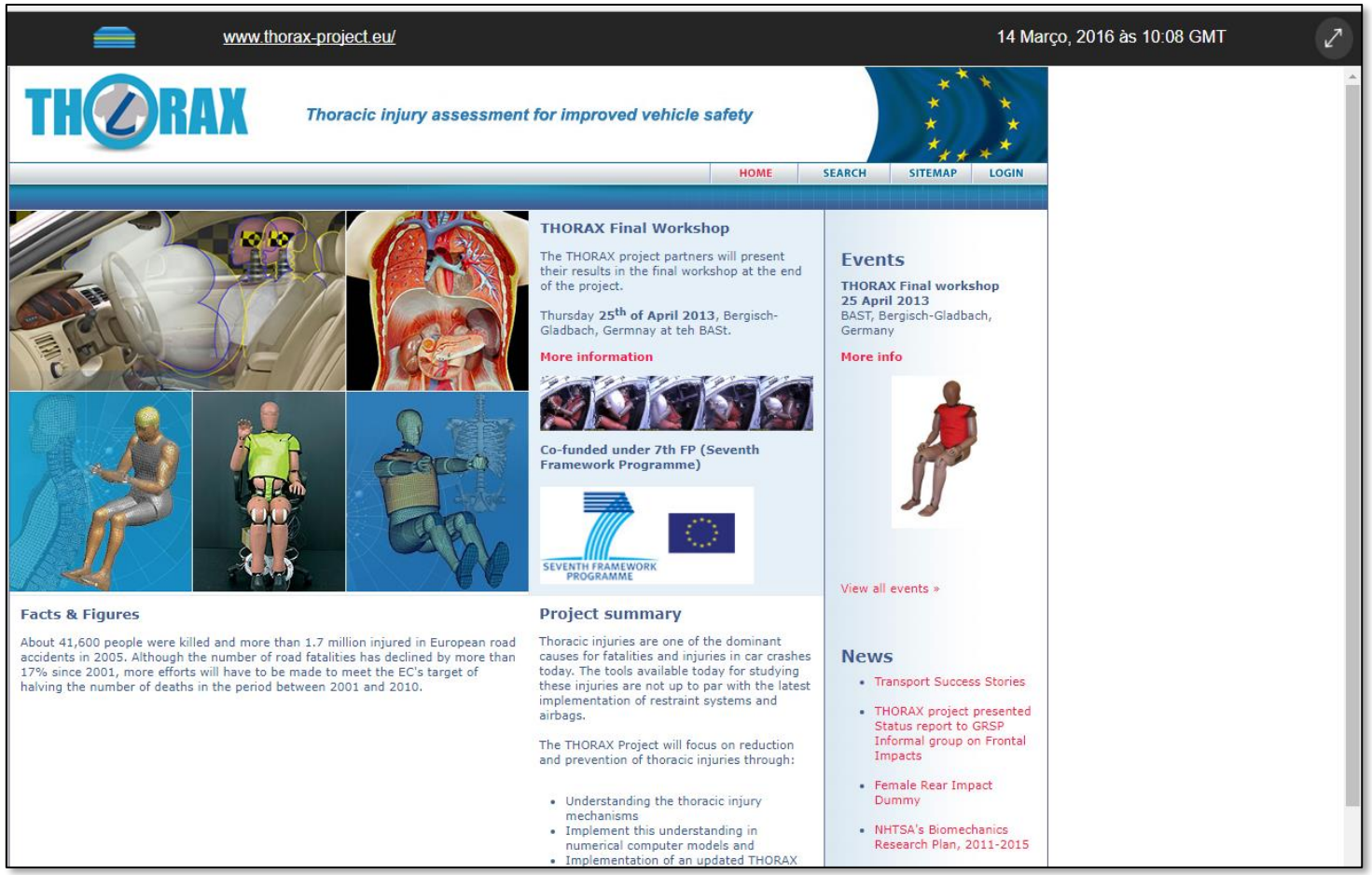
3 crawls of .EU domain

Time of crawl	Millions of files collected	Data volume (TB)
November, December 2014	129	5.8 TB
January 2016	61	3.1 TB
June, July 2017	105	11 TB
Total	**295**	**20 TB**

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

Short link to this page: arquivo.pt/newspapers

Information Science

FCSH on the Web: virtual exhibition

Short link to this page: arquivo.pt/fcshontheweb

Social Sciences

Straight-Edge in the Lisbon metropolitan area

Short link to this page: arquivo.pt/straightedgeen

All videos, presentations, reports at: arquivo.pt/research

ARQUIVO.PT

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 €