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

WEZARD WEather haZARDs for aeronautics

- Home
- The Project
- The Consortium
- The Advisory Board
- Deliverables
- Documentation
- Publications
- Events
- Related projects
- Contact

Funded by: European Commission

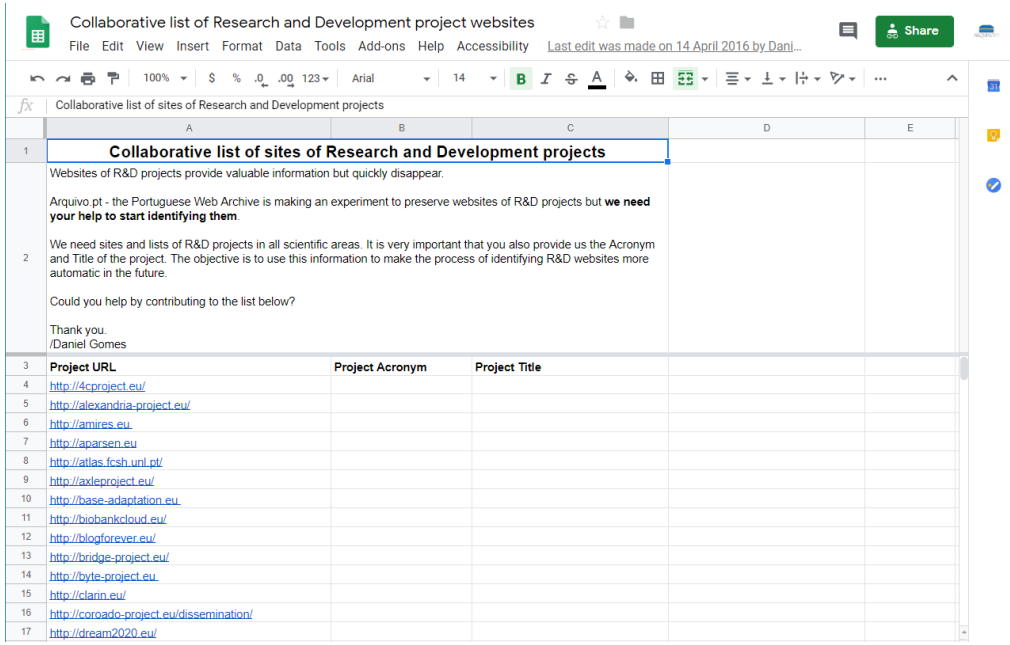
WEZARD Home 15 Feb 2013

Overview

The European WEZARD project (acronym standing for Weather Hazards for Aeronautics) aims at preparing the future research community in the area of air transport system robustness when it is faced with weather hazards. Its precise objectives are to provide: (i) an interdisciplinary and cross-sector network comprising relevant experts; (ii) a state-of-the-art review of the on-going research actions; (iii) an analysis which will identify the shortcomings, areas for improvements and the type of activity needed to limit the effects of disruptive events; (iv) a set of recommendations and a roadmap validated by the main stakeholders of the aeronautics community.

The WEZARD consortium consists of 3 airframers, 2 engine manufacturers, 1 system supplier, 1 network of meteorological offices, 4 research centers, 1 provider of test facilities and 1 civil aviation authority over 2 years. An Advisory Board gathering a panel of international experts in relevant domains has been set up to provide advice on the vision, priorities and directions proposed by the project.

The project runs for 24 months from July 2011 until June 2013.



Collaborative list of Research and Development project websites

File Edit View Insert Format Data Tools Add-ons Help Accessibility Last edit was made on 14 April 2016 by Dani...

Share

Collaborative list of sites of Research and Development projects

Collaborative list of sites of Research and Development projects

Websites of R&D projects provide valuable information but quickly disappear.

Arquivo.pt - the Portuguese Web Archive is making an experiment to preserve websites of R&D projects but **we need your help to start identifying them**.

We need sites and lists of R&D projects in all scientific areas. It is very important that you also provide us the Acronym and Title of the project. The objective is to use this information to make the process of identifying R&D websites more automatic in the future.

Could you help by contributing to the list below?

Thank you.
/Daniel Gomes

Project URL	Project Acronym	Project Title
http://4cproject.eu/		
http://alexandria-project.eu/		
http://amires.eu		
http://aparsen.eu		
http://atlas.fcsh.unl.pt/		
http://axleproject.eu/		
http://base-adaptation.eu		
http://biobankcloud.eu/		
http://blogforever.eu/		
http://bridge-project.eu/		
http://byte-project.eu		
http://clarin.eu/		
http://coroado-project.eu/dissemination/		
http://dream2020.eu/		

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

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

Table 1. Comprehensive .eu crawls 2014-2017

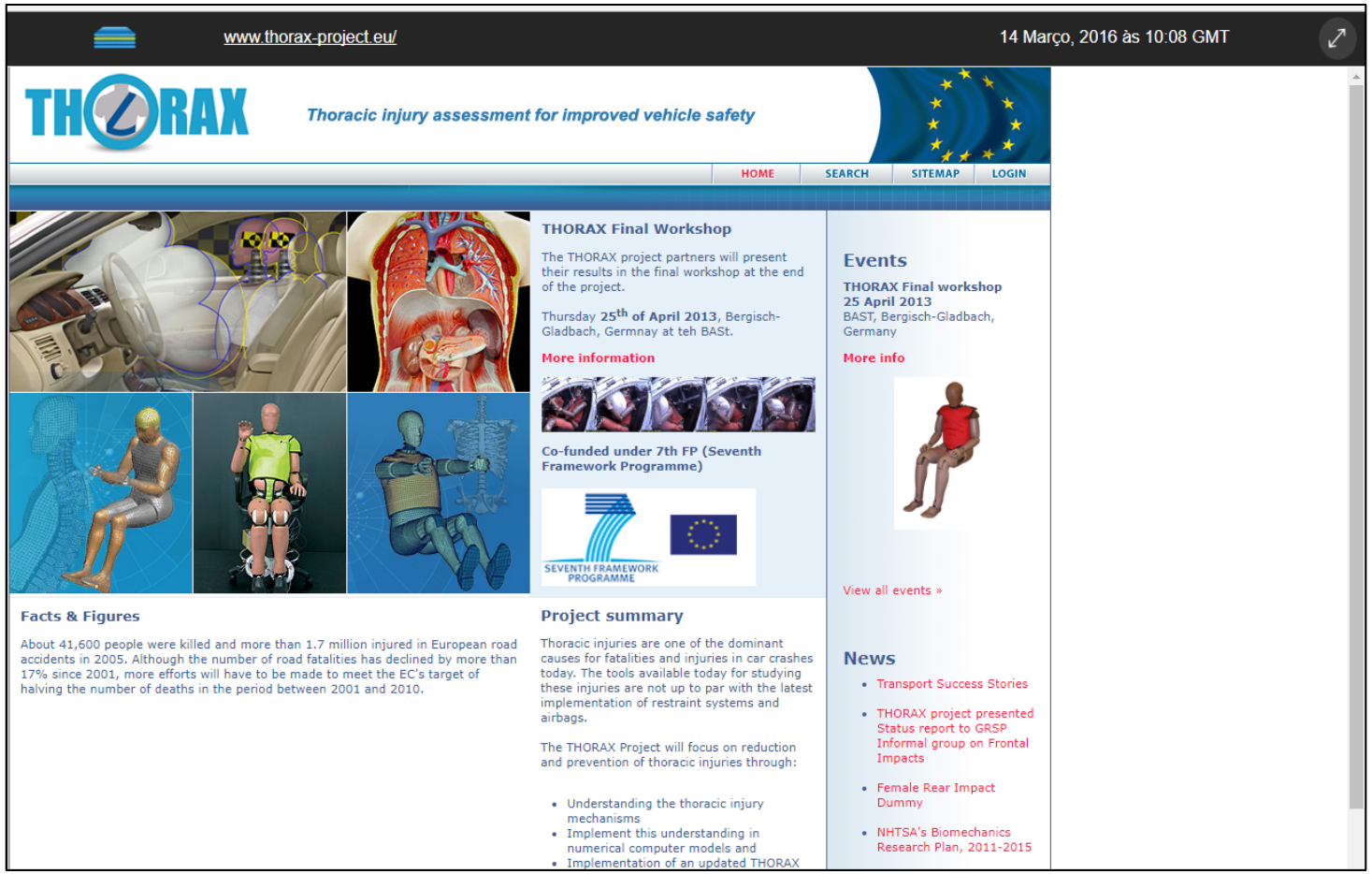
Main problem: **web spam**

Searchable and accessible at:

https://arquivo.pt/resawdev

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

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/prizes

1st place: 10 000 €

2nd place: 3 000 €

3rd place: 2 000 €