

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



WEZARD WEather hAZARDs for aeronautics

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

Funded by:
European Commission
SEVENTH FRAMEWORK PROGRAMME

WEZARD Home 15 Feb 2013

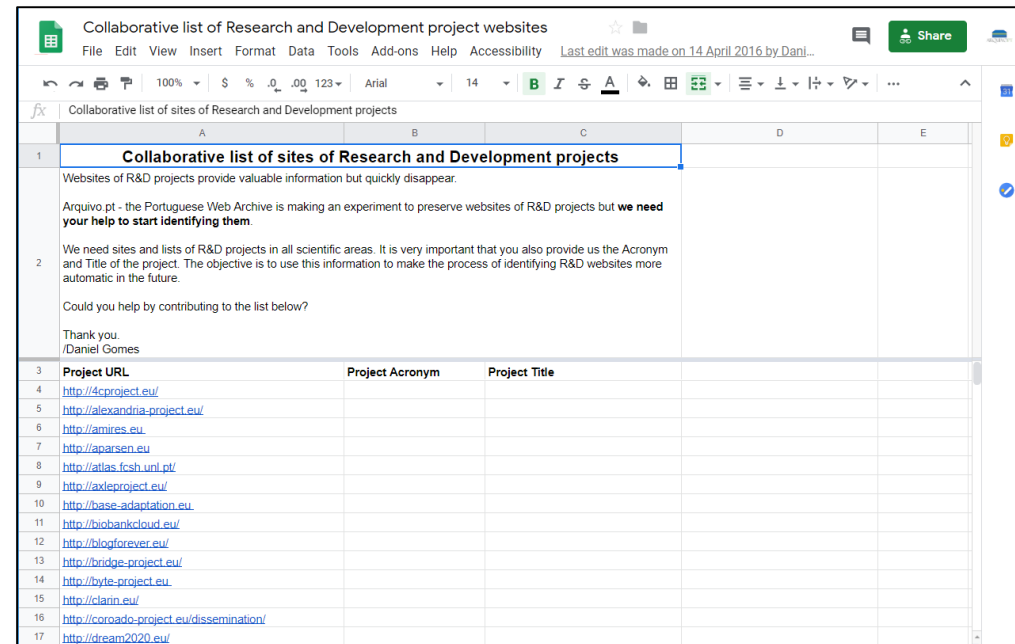
Overview

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...

Collaborative list of sites of Research and Development projects

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 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**



RESEARCH
.EU

[Advanced search](#)

Search pages from the past
[Meet the service](#)

Automatic selection and preservation of websites related to R&D projects

The screenshot shows the THORAX project website. At the top, the URL is www.thorax-project.eu/ and the date is 14 Março, 2016 às 10:08 GMT. The main header features the THORAX logo and the tagline "Thoracic injury assessment for improved vehicle safety" next to the European Union flag. A navigation bar includes links for HOME, SEARCH, SITEMAP, and LOGIN. The main content area is divided into several sections: a "THORAX Final Workshop" announcement for April 25, 2013; a "Co-funded under 7th FP" section with the Seventh Framework Programme logo; a "Facts & Figures" section detailing road accident statistics; a "Project summary" section describing thoracic injuries; an "Events" section with a "View all events" link; and a "News" section with several articles listed.

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/straightedgen

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 €