

Preserving Websites Of Research & Development Projects

Daniel Bicho
daniel.bicho@fccn.pt

Daniel Gomes
daniel.gomes@fccn.pt

Project arcomem: ARchive COmmunities MEMories



ABOUT ARCOMEM ▼

TECH DEMOS

SYSTEM DEMOS

TRAINING

OPEN SOURCE

VIDEOS

PUBLICATIONS &

ABOUT ARCOMEM

ARCOMEM is about memory institutions like archives, museums, and libraries in the age of the Social Web. Memory institutions are more important now than ever: as we face greater economic and environmental challenges we need our understanding of the past to help us navigate to a sustainable future. This is a core function of democracies, but this function faces stiff new challenges in face of the Social Web, and of the radical changes in information creation, communication and citizen involvement that currently characterise our information society (e.g., there are now more social network hits than Google searches). Social media are becoming more and more pervasive in all areas of life. In the UK, for example, it is now not unknown for a government minister to answer a parliamentary question using Twitter, and this material is both ephemeral and highly contextualised, making it increasingly difficult for a political archivist to decide what to preserve.

<http://www.arcomem.eu/>



Information about scientific events



ABOUT ARCOMEM ▼ TECH DEMOS SYSTEM DEMOS TRAINING OPEN SOURCE VIDEOS PUBLICATIONS

IPRES 2013

iPRES2013

10th International Conference on
Preservation of Digital Object

1st International Workshop on Archiving Community Memories

6 September 2013, Lisbon, Portugal

www.arcomem.eu | twitter.com/arcomem | [#arcomem](https://twitter.com/arcomem)

Dissemination and training materials



[ABOUT ARCOMEM](#) ▾ [TECH DEMOS](#) [SYSTEM DEMOS](#) **[TRAINING](#)** [OPEN SOURCE](#) [VIDEOS](#) [PUBLICATIONS](#)

TRAINING

To explain the ARCOMEM project to detail we have created online training material so that the end-users of the ARCOMEM system can train themselves in understanding and using the AROMEM system.

ARCOMEM MADE EASY

As a beginner you can start with having a look at the "ARCOMEM made Easy" presentation. This presentation shows you a general overview on how the ARCOMEM system has been set up. If you want to dig deeper you can follow the links to the Powerpoint presentations on the different modules of the ARCOMEM system.

The ARCOMEM system

Demonstrations about the project



ABOUT ARCOMEM ▾ TECH DEMOS **SYSTEM DEMOS** TRAINING OPEN SOURCE VIDEOS PUBLICATIONS ▾

SYSTEM DEMOS

Memory institutions like archives, museums, and libraries are more important now than ever. But our understanding of the past faces stiff new challenges in face of the Social Web. History is not only written in books anymore. The radical changes in information creation, communication and citizen involvement that currently characterise our information society (e.g., there are now more social network hits than Google searches) need a [new approach on preservation](#). Social media are becoming more and more pervasive in all areas of life. But how do we separate the wheat from the chaff in Social media, how is it being preserved and will it still be available twenty years from now?

ARCOMEM will enable political archivists and journalists to preserve Social media around certain entities, topics or events. To give you a quick insight [here is a short movie](#) on how journalism can benefit from Social media archiving.

We would like to share our work with you so please come and have a look at our system demos! We have installed

Content is not available anymore



File not found.

<http://www.arcomem.eu/>

Waste of significant investments

European Union FP7 Work Programme invested **59 million euros in R&D projects.**

Part of this funding was **spent** developing R&D projects websites.

Loss of **Knowledge**

R&D project sites publish important scientific outputs (e.g. data sets, tools).

Work as aggregators of project outputs (ex. news and events).

How to preserve R&D websites?

Web Archives **need to identify project URLs** to preserve R&D project websites.

Where to get project URLs?



CORDIS - EU research projects under FP7 (2007-2013)

Publisher

Publications Office »

Description

This dataset contains projects funded by the European Union under the seventh framework programme for research and technological development (FP7) from 2007 to 2013. Grant information is provided for each project, including reference, acronym, dates, funding, programmes, participant countries, subjects and objectives. You can also find a separate file with organisation information (project participants, coordinators) and another with results in brief written by CORDIS for selected FP7 projects.

Reference data (countries, funding schemes/types of action, subjects (SIC codes)) can be found in this dataset: <https://data.europa.eu/euodp/en/data/dataset/cordisref-data>

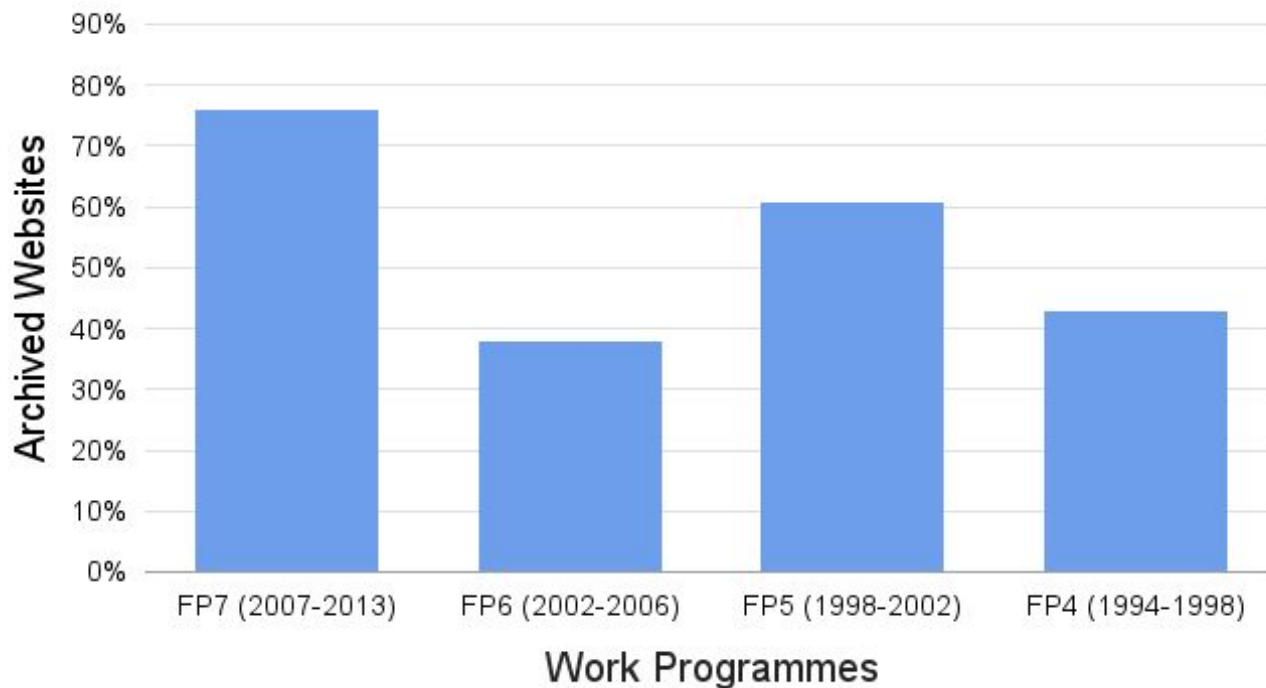
<https://open-data.europa.eu/>

European Open Data Portal Datasets

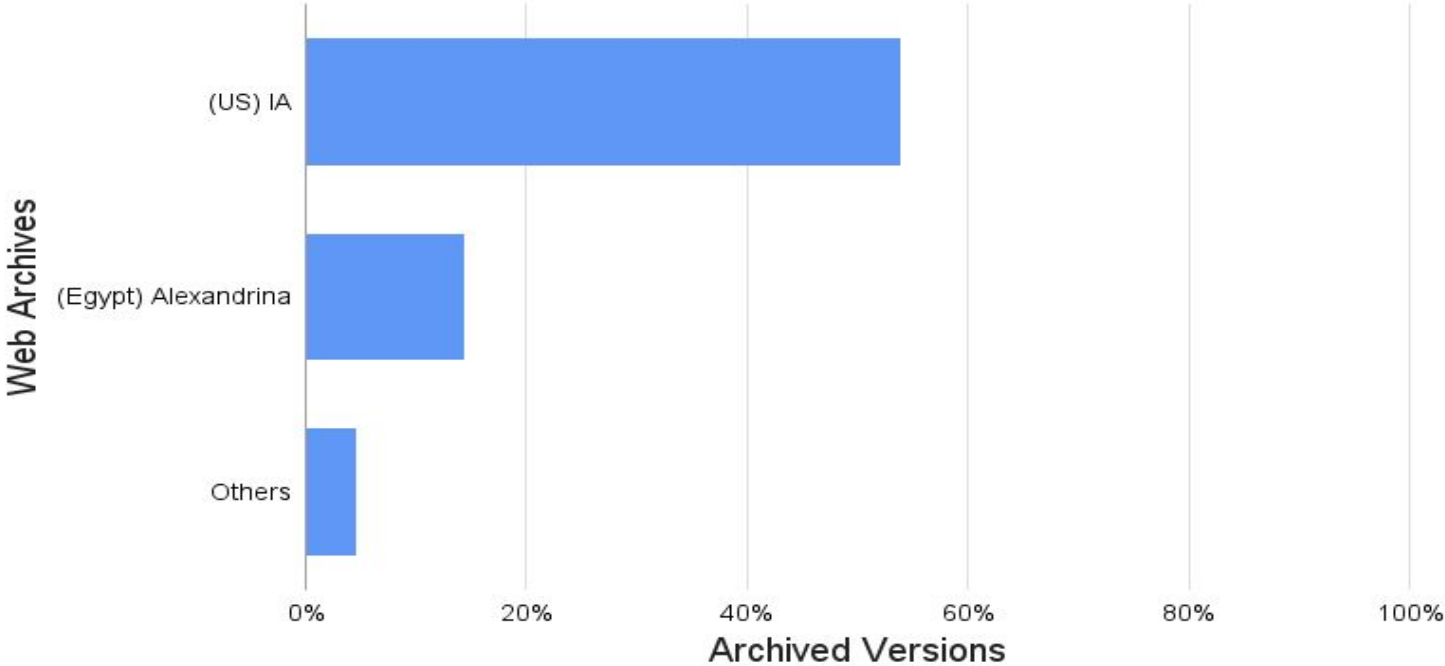
CORDIS EU research projects funded under work programmes (FP4, FP5, FP6, FP7).

Include information like **project URL**, acronym, title, dates, funding, objectives etc.

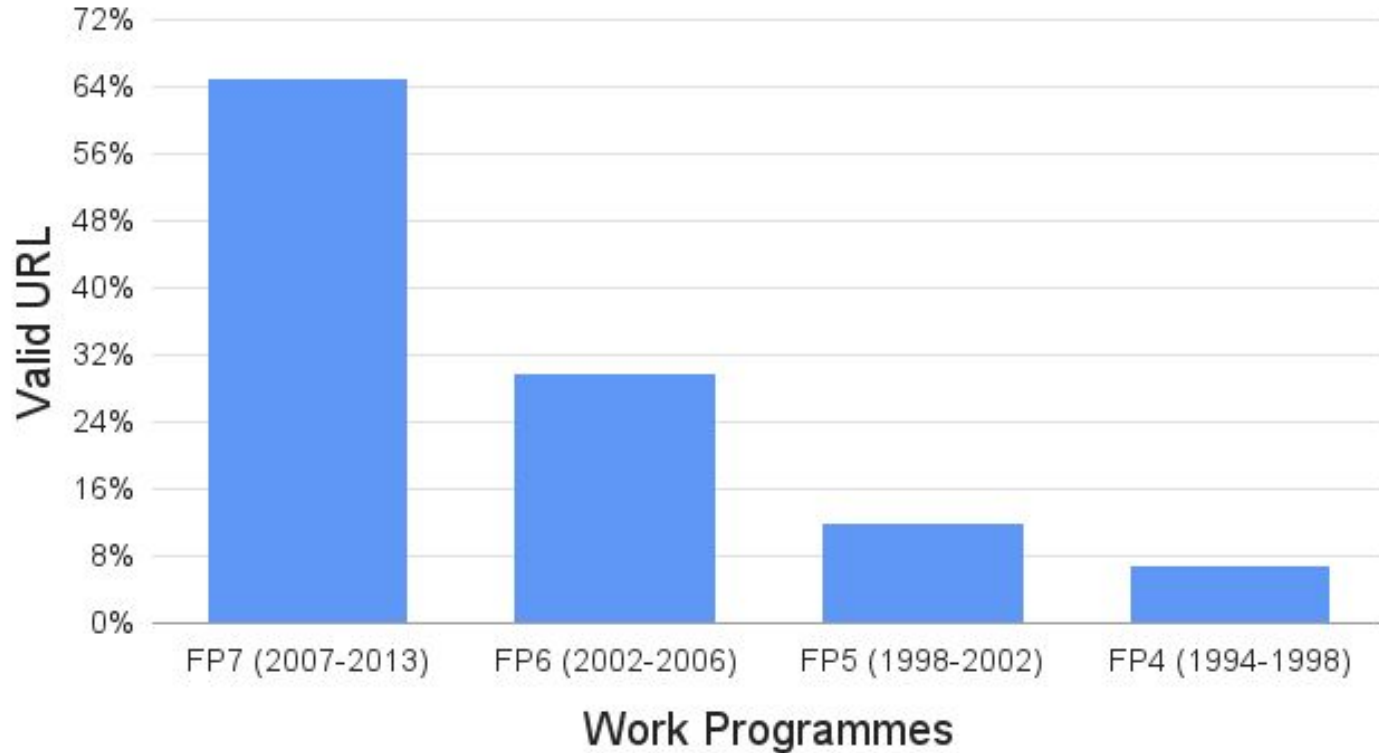
Web-archived project URLs since FP4



Project URLs of EU-funded research are mostly being preserved outside Europe



Valid project URLs in 2016



CORDIS datasets provide incomplete information

25 000 projects were funded by the FP7 work programme.

Only **8%** have an associated **project URL**.

Missing information regarding **project URL**

acronym	title	projectUrl
ALFRED	ALFRED - Person	
TIBETMETH	Microbial Biomark	
SMALL_MAM_RECOL	Post-glacial recol	
MOMEFAST	Molecular Mecha	
RNF4 IN THE DDR	Identifying the ta	
ARIEL	Archaeological In	
NANODYNATCELLVATION	Nano -structural &	
TBKO	Synthesis and Bi	
	Microbially cataly	
MICROBIOELECTROSYN	cathode in bioele	
THINFACE	Thin-film Hybrid I	
IMPACTS	The impact of the	
FLEXISTAT	Production Flexib	
3FLEX	Depth enabled wo	
GUIDENANO	Assessment and	
AZNETAC	A zebrafish mode	
PLASMANANOSMART	Plasma- and elec	
RASMIM	Reactivity of Alu	
FAMILIESANDSOCIETIES	Changing families	http://www.familiesandsocieties.eu/

Problem

R&D websites provide **important** information but quickly **disappear**.

Funded projects datasets are **incomplete**.

How to automatically identify project URLs to be preserved (with limited resources)?

Proposed approach

Search Engines already index the Web

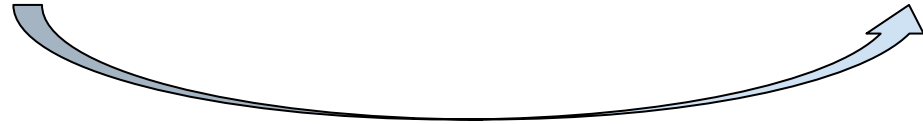
Open Data Portal provides meta-data about R&D projects

Combine open data sets with search engines to identify R&D project URLs

Automatic workflow to identify R&D project URLs

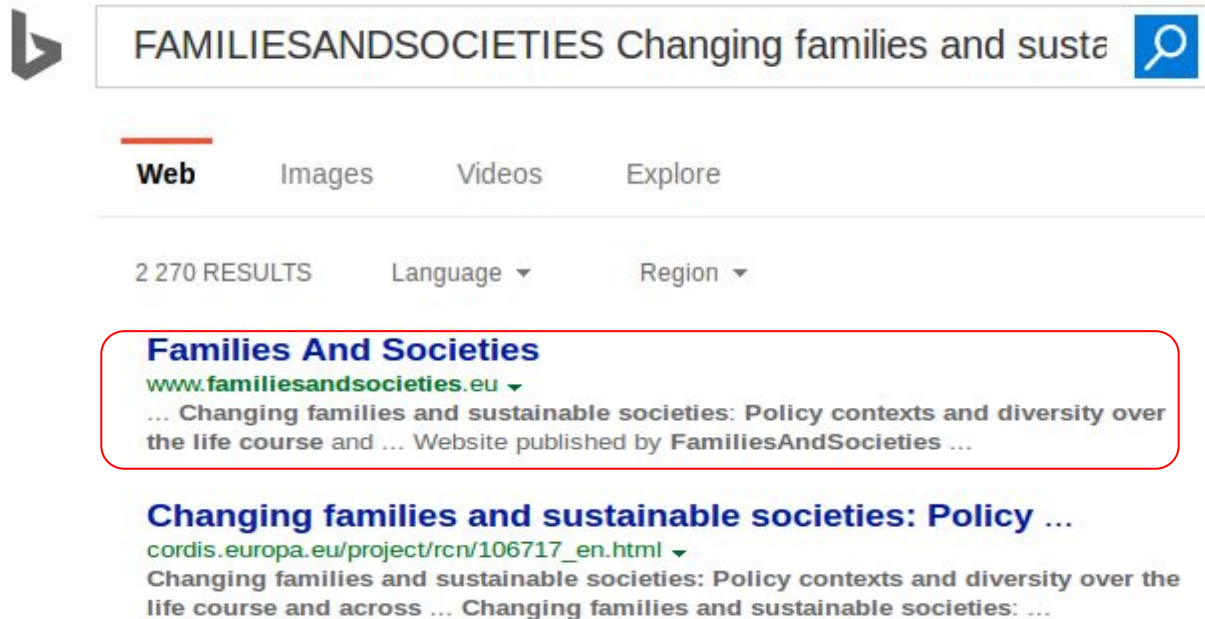
ACRONYM	TITLE	PROJECT URL
DIP3	The 3Ps of Distributed Information delivery ...	missing

ACRONYM	TITLE	PROJECT URL
DIP3	The 3Ps of Distributed Information delivery ...	www.dip-3.eu



Identify project URL through
Bing Web Search API

Web search by project Acronym + Title



The screenshot shows a search engine interface. At the top left is a blue logo resembling a stylized 'b'. To its right is a search bar containing the text 'FAMILIESANDSOCITIES Changing families and susta' and a magnifying glass icon. Below the search bar are four tabs: 'Web' (highlighted with a red underline), 'Images', 'Videos', and 'Explore'. Underneath the tabs, it displays '2 270 RESULTS' followed by 'Language' and 'Region' dropdown menus. The first search result is enclosed in a red rounded rectangle and consists of the title 'Families And Societies' in bold blue text, the URL 'www.familiesandsocieties.eu' with a dropdown arrow, and a snippet: '... Changing families and sustainable societies: Policy contexts and diversity over the life course and ... Website published by FamiliesAndSocieties ...'. Below this, a second result is visible with the title 'Changing families and sustainable societies: Policy ...' in bold blue text, the URL 'cordis.europa.eu/project/rcn/106717_en.html' with a dropdown arrow, and a snippet: 'Changing families and sustainable societies: Policy contexts and diversity over the life course and across ... Changing families and sustainable societies: ...'.

Experiment

Evaluate heuristics to **maximize** the performance of the automatic identification of R&D project URLs.

+Acronym +Title

+Acronym +Title -Cordis

+Acronym +Title -Cordis -EC

+Acronym +Title -Cordis -EC +CommonTerms

Evaluating the heuristics

A test collection was built based on the FP7 dataset by **manually** validating each R&D project URL.

The project URLs returned by each heuristic were compared to the test collection to measure if they **matched**.

Heuristics Performance (F-measure)

Heuristics	Top 1 Results	Top 10 Results
+Acronym +Title	44%	12%
+Acronym +Title -Cordis	45%	11%
+Acronym +Title -Cordis -EC	47%	11%
+Acronym +Title -Cordis -EC +project	48%	11%

*F-measure is a combination of **recall** and **precision**.*

Preserving R&D projects websites

Applied heuristic with best performance to the 23 588 projects that were **missing the project URL**.

Identified **20 429 new project URLs**.

Before: 23 588 missing
URLs

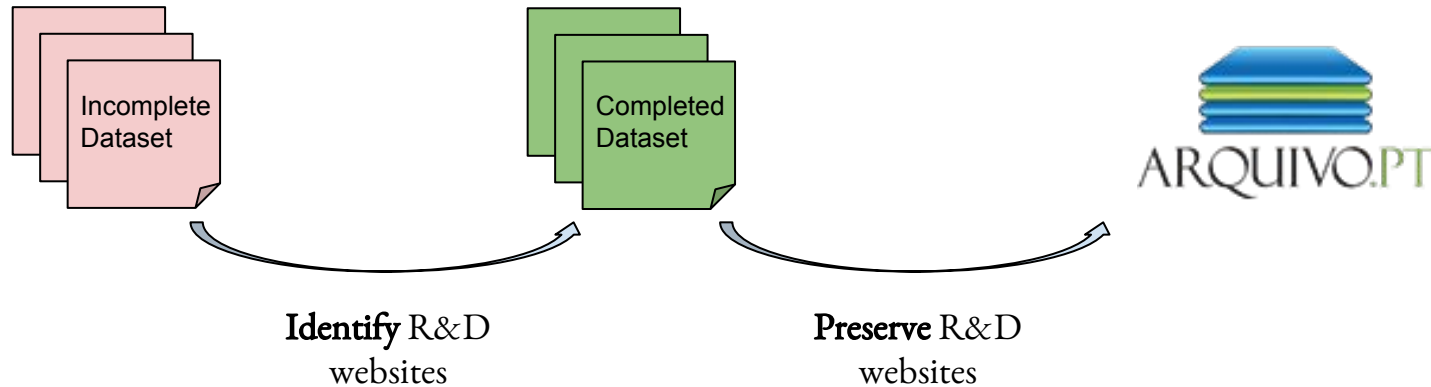
acronym	title	projectUrl
ALFRED	ALFRED - Person	
TIBETMETH	Microbial Biomar	
SMALL_MAM_RECOL	Post-glacial reco	
MOMEFAST	Molecular Mecha	
RNF4 IN THE DDR	Identifying the ta	
ARIEL	Archaeological r	
NANODYNATCELLVATION	Nano -structura	
TBKO	Synthesis and B	
MICROBIOELECTROSYN	Microbially cataly	
THINFACE	Thin-film Hybrid	
IMPACTS	The impact of the	
FLEXISTAT	Production Flexit	
3FLEX	Depth enabled w	
GUIDENANO	Assessment and	
AZNETAC	A zebrafish mode	
PLASMANANOSMART	Plasma- and elec	
RASMIM	Reactivity of Alu	
FAMILIESANDSOCIETIES	Changing familie	http://www.familiesandsocieties.eu/

acronym	title	projectUrl
ALFRED	ALFRED - Person	http://alfred.eu/
TIBETMETH	Microbial Biomar	http://www.thefreedictionary.com/Tib
SMALL_MAM_RECOL	Post-glacial reco	http://www.thefreedictionary.com/Ins
MOMEFAST	Molecular Mecha	
RNF4 IN THE DDR	Identifying the ta	http://www.dundee.ac.uk/research/rn
ARIEL	Archaeological r	http://ucy.ac.cy/ariel/
NANODYNATCELLVATION	Nano -structura	
TBKO	Synthesis and Bio	
MICROBIOELECTROSYN	al systems	
THINFACE	Thin-film Hybrid	http://www.nanogune.eu/projects/th
IMPACTS	The impact of the	http://www.sciencedirect.com/science
FLEXISTAT	Production Flexit	http://www.flexistat.eu/
3FLEX	Depth enabled w	http://www.hhi.fraunhofer.de/depart
GUIDENANO	Assessment and	http://www.guidenano.eu/News
AZNETAC	A zebrafish mode	
PLASMANANOSMART	Plasma- and elec	http://portal.tpu.ru/departments/cent
RASMIM	Reactivity of Alu	https://www.highbeam.com/doc/1G1
FAMILIESANDSOCIETIES	Changing familie	http://www.familiesandsocieties.eu/

After: 20 429 new
URLs

Crawled the identified R&D project URLs to be preserved.

Nr. Project URL Seeds	20 429
Stored Content (compressed)	1.4 TB



Conclusions

R&D websites are **important** but are quickly **disappearing**.

European Datasets about funded R&D projects are **incomplete**. Only **8%** have a project URL associated.

54% of the EU-funded R&D project URLs are being web-archived. Mostly **outside the European Union**.

Conclusions

Automatic heuristics to identify R&D project URLs.

All outputs of this study are available in open access

- Test collection
- Completed CORDIS data sets with new project URLs
- Extensive technical report

<https://github.com/arquivo/Research-Websites-Preservation>

Thank You