How Arquivo.pt is Preserving Scientific Research Project Websites and Promoting Data Reuse

Ricardo Basílio, Arquivo.pt web curator

Agenda

- What is Arquivo.pt
- Websites of European R&D projects
- How researchers use Arquivo.pt
- Collaborations and tools

Arquivo.pt – the Portuguese Web Archive

Arquivo.pt's headquarters in Lisbon. FCCN, Foundation for Science and Technology digital services unit.

https://arquivo.pt

Arquivo.pt – the Portuguese web archive

- 1.6 Petabytes
- 73 million websites in Portuguese and many other languages

Guided tour of the data center, our “Cloud,” by Daniel Gomes, founder of Arquivo.pt (2019)

Know more:
https://sobre.arquivo.pt/en/about/press/the-portuguese-web-archive-in-numbers/

Arquivo.pt – the Portuguese web archive

- Search by URL, Text, and Image
- Access via API for automatic processing
- Special reuse for research

FAIR

Menu

arquivo.pt | fct

dcc.ac.uk

Search

1991 6 Aug

2026 15 Feb

Pages | Images | Narrative

Advanced search

About 441 results from **1991** to **2026**

← Table List →

2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026
								23 Nov		8 Jan	13 Jun	8 Apr	1 Jan	6 Feb	12 Jan	26 Jan	31 Jan	3 Feb		
										14 Mar		18 Jul	2 Jan	6 Feb	8 Mar	1 Feb	31 Jan	3 Feb		
										15 Mar		29 Jul	3 Jan	7 Feb	10 Jun	22 Apr	16 Feb	12 May		
										13 May		30 Jul	4 Jan	8 Feb	1 Jul	22 Apr	15 Mar	12 May		
										19 May		31 Jul	5 Jan	9 Feb	9 Sep					
												1 Aug	6 Jan	10 Feb	21 Sep					
												2 Aug	7 Jan	11 Feb	17 Oct	12 Aug	24 Apr	7 Aug		
												3 Aug	8 Jan	13 Feb		24 Oct	30 Jul	7 Aug		

Example of URL search

Menu

arquivo.pt | fct

Options

idcc conference

Search

1991
6 Aug

2026
6 Feb

Pages | Images | Narrative

Advanced search

About 7.830 results from **1991** to **2026**

www.dcc.ac.uk/events/international-digital-curation-conference-idcc

International Digital Curation **Conference** (**IDCC**) | Digital Curation Centre

7 June 2017

International Digital Curation **Conference** (**IDCC**) | Digital Curation Centre Skip to Navigation Contact us Search this site: Home > Events > International Digital Curation **Conference Idcc** International Digital Curation **Conference** (**IDCC**) Our International Digital Curation **Conference** (**IDCC**) is an ...

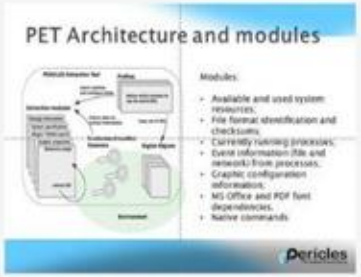
4cproject.eu/news-and-comment/current-news/78-4c-at-the-ninth-international-digital-curation-conference-idcc

4C at the Ninth International Digital Curation **Conference** (**IDCC**) - 4C Project

7 January 2016

4C at the Ninth International Digital Curation **Conference** (**IDCC**) - 4C Project 4C - Collaboration to ... 4C Partners 4C at t Digital Curation **Conference** (**IDCC**) The 4C Project is ... . The **IDCC conference** is a valuable opportunity for the 4C Project and the community to examine how ...

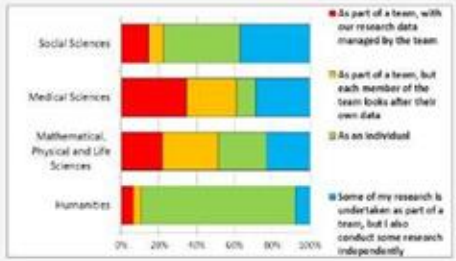
Example of text search



→ http://pericles-project.eu/…
8 January 2016

→ https://connect.geant.org…
26 May 2022

→ http://www.dcc.ac.uk/ne…
18 July 2018



→ http://blogs.it.ox.ac.uk/da…
17 June 2017

→ https://wwwen.uni.lu/snt/…
18 September 2021

→ https://wwwfr.uni.lu/snt/n…
24 August 2022

→ https://www.eudat.eu/eve…
29 December 2020

→ http://www.vre4eic.eu/
9 January 2016

→ http://eudat.eu/events/tra…
14 March 2016

→ https://www.fairsfair.eu/ar…
23 September 2021

→ https://www.dariah.eu/ev…

→ https://www.dcc.ac.uk/ev…

Coalition for Networked Information

→ http://www.bibalex.org/a2…
4 July 2009

→ http://www.dcc.ac.uk/ne…

Example image search

→ https://conservativehome…

In this section

International Digital Curation Conference (IDCC)

Data Management Roadshows

Research Data Management Forum (RDMF)

Workshops

Other DCC events

External events

International Digital Curation Conference (IDCC)

Our **International Digital Curation Conference (IDCC)** is an established annual event with a unique place in the digital curation community, reaching out to individuals, organisations and institutions across all disciplines and domains involved in curating data and providing an opportunity to get together

Testimonials

This is probably the single most important general meeting for those working in the area of data curation.

Frankly one of the most useful conferences I have attended ... I especially appreciated the practical applications that many of the speakers addressed.

a DCC me ideas, some of which we are already putting into action specifically regarding the "Designing Data Management Training"

Example of the replay

```
serviceName:          "Arquivo.pt - the Portuguese web-archive"
linkToService:        "https://arquivo.pt"
▼ response_items:
  ▼ 0:
      title:          "International Digital Curation Conference (IDCC) | Digital Curation Centre"
      originalURL:    "http://www.dcc.ac.uk/events/international-digital-curation-conference-idcc"
      linkToArchive:  "https://arquivo.pt/wayback/20170607214014/http://www.dcc.ac.uk/events/international-digital-curation
      tstamp:         "20170607214014"
      contentLength:  20197
      digest:         "VC57XOV53ZOQFBDOMIKW3CT5QJFCDVKZ"
      mimeType:       "text/html"
      encoding:       "UTF-8"
      date:           "1496871614"
      linkToScreenshot:  "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20170607214014%2Fhttp
                         c.uk%2Fevents%2Finternational-digital-curation-conference-idcc"
      linkToNoFrame:     "https://arquivo.pt/noFrame/replay/20170607214014/http://www.dcc.ac.uk/events/international-digital-c
                         idcc"
      linkToExtractedText:  "https://arquivo.pt/textextracted?m=http%3A%2F%2Fwww.dcc.ac.uk%2Fevents%2Finternational-digital-curat
                         c%2F20170607214014"
      linkToMetadata:    "https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fwww.dcc.ac.uk%2Feven                gital-c
                         idcc%2F20170607214014"
      linkToOriginalFile:  "https://arquivo.pt/noFrame/replay/20170607214014id_/http://www.dcc.ac.uk                -digital
                         ce-idcc"
      snippet:           'International Digital Curation Conference (IDCC) | Digital Curation Centre Skip to Navigation<span c
```


Example of the API output

1

Websites of European R&D projects

Websites of European R&D projects

- Researchers publish supplementary information on projects on the Web
- Web content disappears in a short time

Websites of European R&D projects

- 31 000 projects funded by European Union (H2020*)
- 46% didn’t mentioned a project website or webpage

* Framework Programme, funded by the European Commission (EU)

Websites of European R&D projects

K1 | projectUrl

“projectURL”

	G	H	I	J	K	L
1	frameworkProgra	title	startDate	endDate	projectUrl	object
2	FP7	ALFRED - Personal Interactive Assistant for Independent Living and Active Ageing	01/10/2013	30/09/2016	http://alfred.eu/	***Pe
3	FP7	Microbial Biomarker Records in Tibetan Peats: Monsoon Variability and its Impact o	01/10/2015	30/09/2016	http://www.thefreedictionary.com/Tibetan-	It is cr
4	FP7	Post-glacial recolonisation and Holocene anthropization impact on populations of sl	05/05/2014	04/05/2016	http://www.thefreedictionary.com/Insectiv	At the
5	FP7	Molecular Mechanisms Employed by the Newly Assigned RNA-binding Protein FASTK	01/01/2015	31/12/2016		Posttr
6	FP7	Identifying the targets and mechanism of action of the SUMO targeted ubiquitin liga	01/07/2014	19/08/2017	http://www.dundee.ac.uk/research/media/	The
7	FP7	Archaeological Investigatio				
8	FP7	Nano -structural and -dyna				
9	FP7	Synthesis and Biological Eva				
10	FP7	Microbially catalyzed electr				

I	J	K	L
startDate	endDate	projectUrl	object
01/10/2013	30/09/2016	http://alfred.eu/	***Pe
01/10/2015	30/09/2016	http://www.thefreedictionary.com/Tibetan-	It is cr
05/05/2014	04/05/2016	http://www.thefreedictionary.com/Insectiv	At the
01/01/2015	31/12/2016		Posttr

Websites of European R&D projects

Arquivo.pt identified URLs related to R&D projects

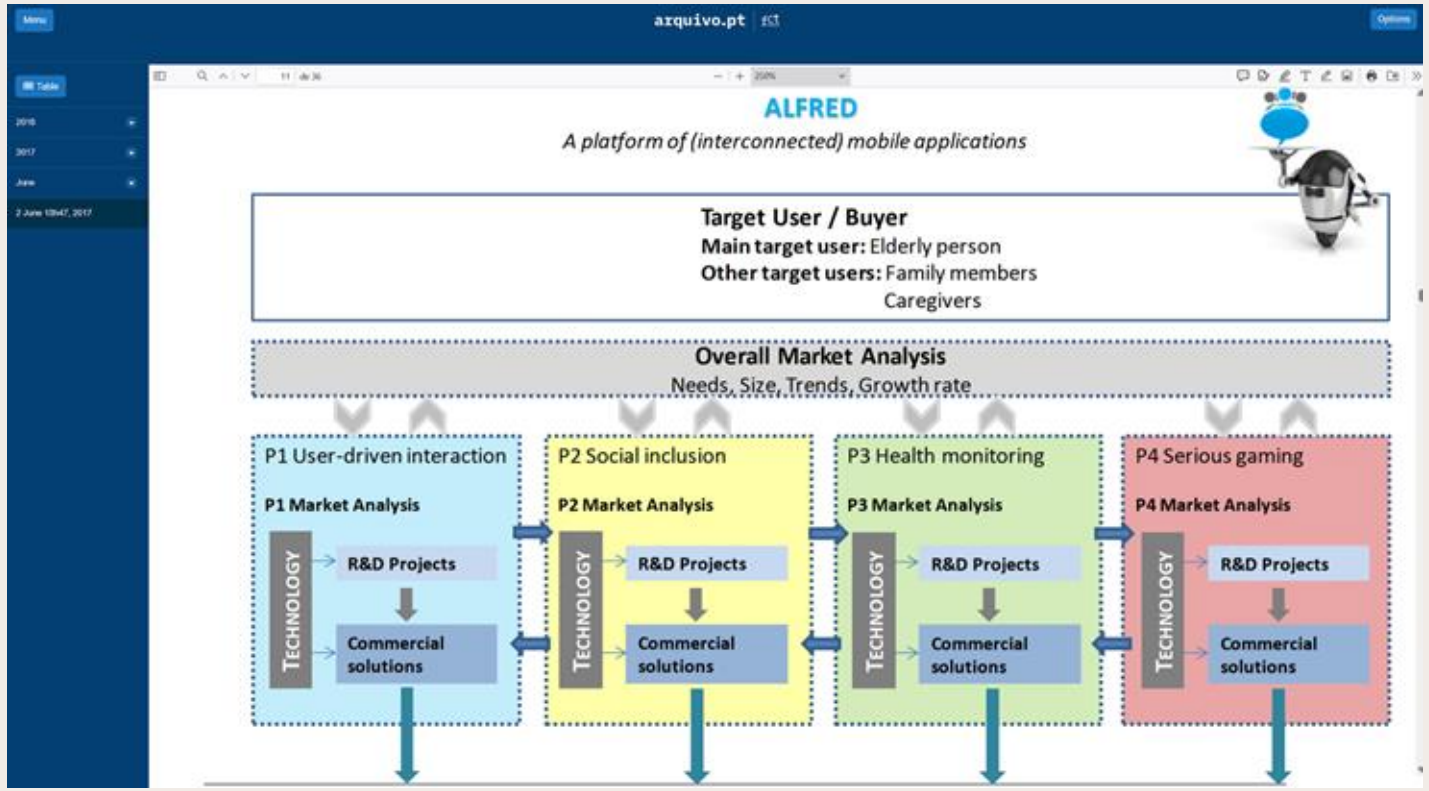
- acronym + title of the project
- acronym + title of the project –Cordis +words

Websites of European R&D projects

Example, how to identify contents about the “ALFRED” project

- ALFRED + “ALFRED – Personal interactive assistant”
- ALFRED + “ALFRED – Personal interactive assistant” –Cordis +project

Websites of European R&D projects



Example of a page related to ALFRED project preserved at Arquivo.pt

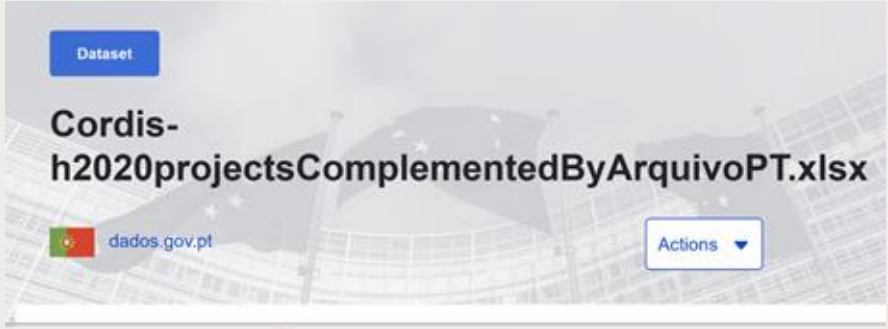
Websites of European R&D projects

Example of the same page on the live web - ALFRED project report

Websites of European R&D projects

Results of H2020 Web preservation

- 17 terabytes at Arquivo.pt - EAWP35
- URL's list available at the Cordis Open Data Portal



Collaboration with the Publications Office of the EU

2

How researchers use Arquivo.pt

How researchers use Arquivo.pt

2 Examples:

- A study about immigration
- LLM

https://arquivo.pt/awards

How researchers use Arquivo.pt

Example 1: Narrative Monitor

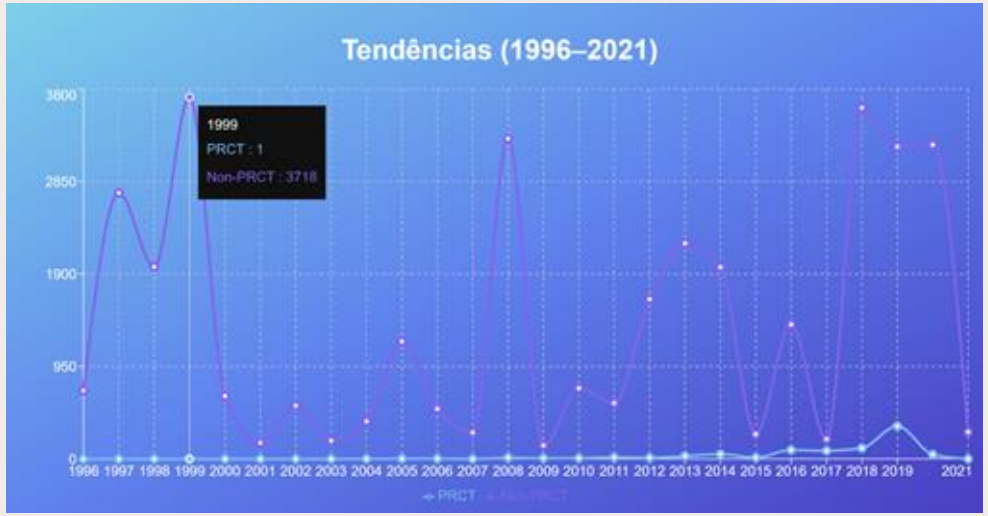
- Processed archived web pages
- Trends in the Portuguese Press about Ethnic Replacement Conspiracy Theories

https://arquivo.pt/winners 2025

How researchers use Arquivo.pt

Example 1: Narrative Monitor

- Processed archived web pages
- Trends in the Portuguese Press about Ethnic Replacement Conspiracy Theories



https://arquivo.pt/winners 2025

How researchers use Arquivo.pt

Example 1: Narrative Monitor

Researchers from Évora University, Portugal

3rd place at the 2025 Arquivopt Award

https://arquivo.pt/winners 2025

How researchers use Arquivo.pt

Example 2:

- GlórIA LLM – A Generative and Open Large Language Model for Portuguese
- 1.4 Million archived news from Arquivo.pt

Know more: https://sobre.arquivo.pt/en/artificial-intelligence-processes-data-from-arquivo-pt/

3

Collaborations and tools

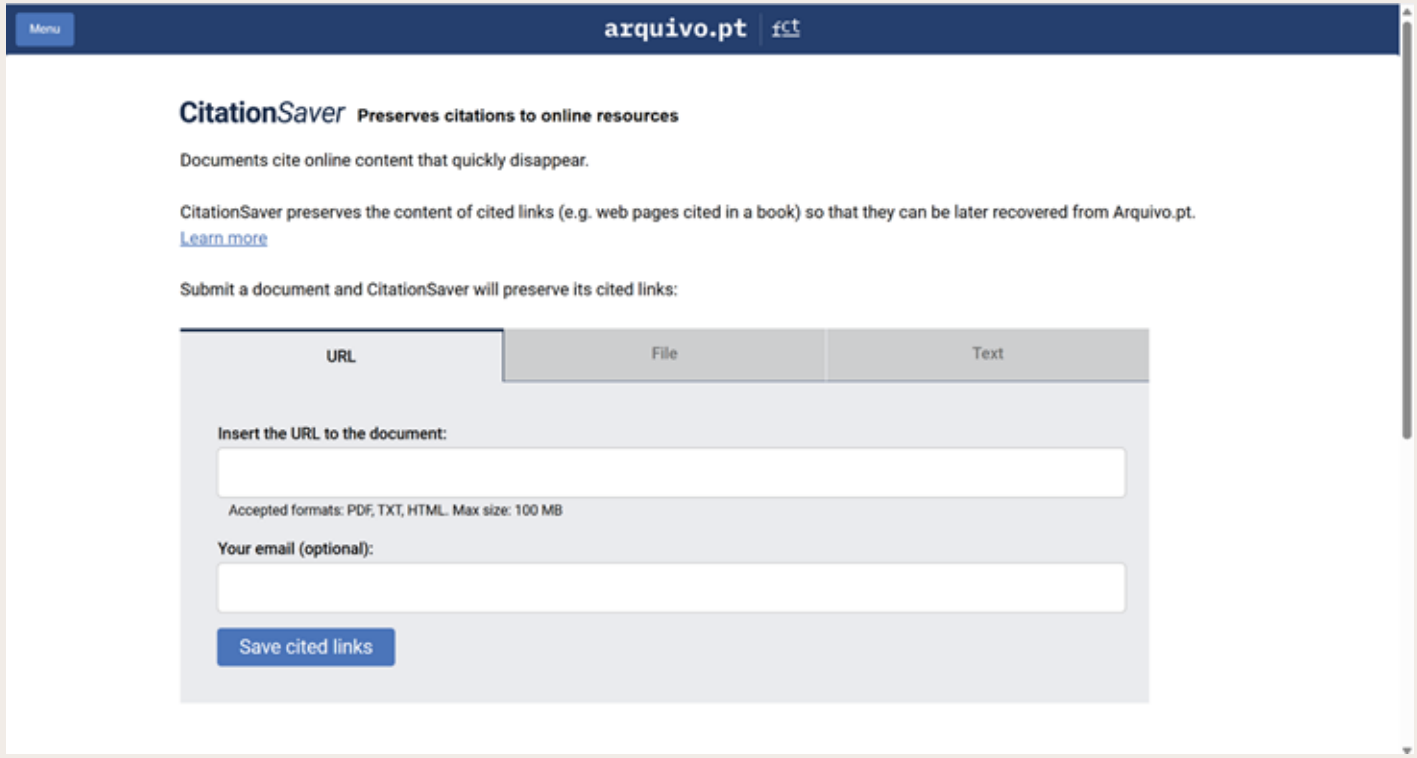
Collaborations and tools

- Extracting **URLs** from PDFs of scientific publications and CVs
- **Motivating** researchers to preserve web content related to projects

Collaborations and tools

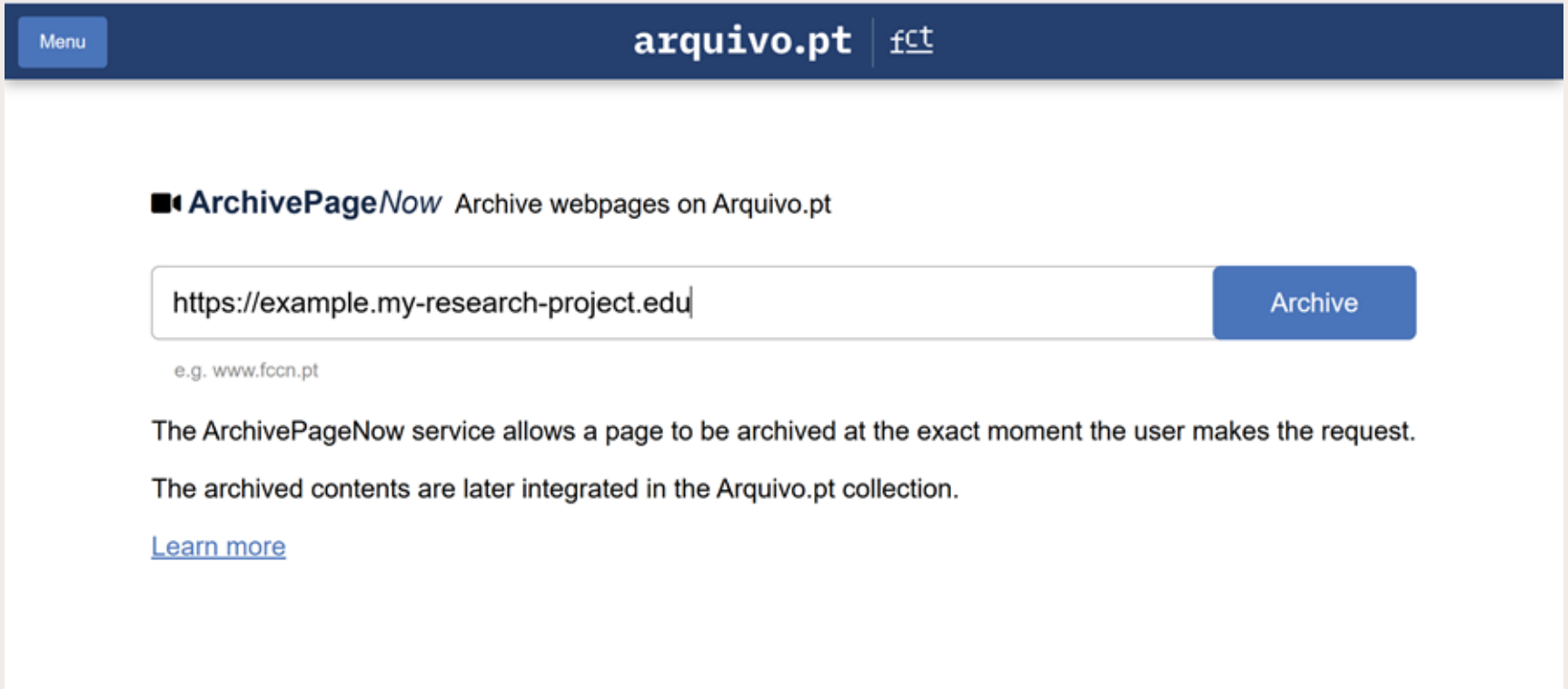
- CitationSaver - preserves citations to online resources
- ArchivePageNow - archives immediately any web content
- Memorial - maintains accessible old websites after being shut down

Collaborations and tools



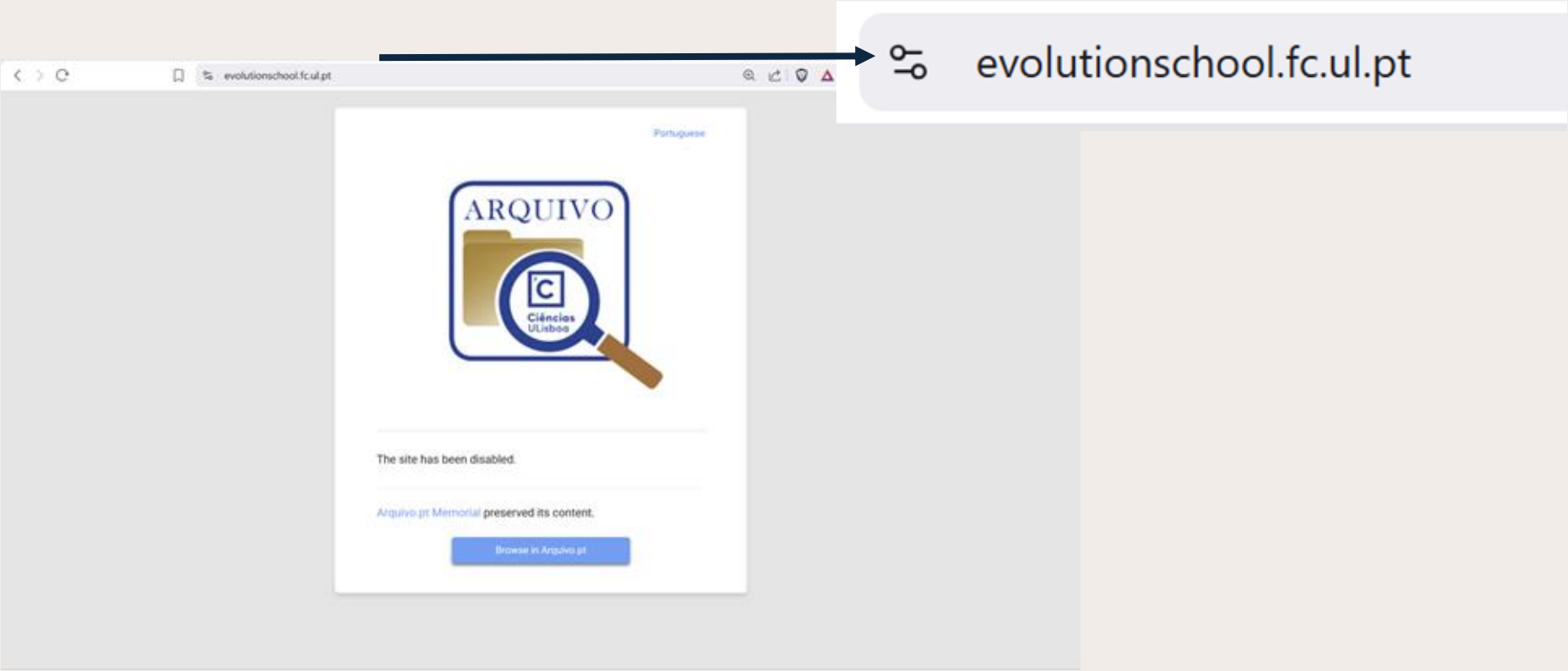
https://arquivo.pt/citationsaver

Collaborations and tools



https://arquivo.pt/archivepagenow

Collaborations and tools



Memorial redirects the DNS to the preserved version (Know more at arquivo.pt/memorial)

Final thoughts

- Use web archives to save the results of your work
- Include web archived materials in the research process
- Use Arquivo.pt and give us your feedback on how useful the service is for the open science ecosystem

Thank you!

Email:
contacto@arquivo.pt

Mailing list
arquivo.pt/subscribe