

How Arquivo.pt is preserving scientific research project websites and inspiring the future of research

Abstract

This presentation shows how Arquivo.pt is preserving websites related to scientific research projects and how web information from the past is becoming a source of data for research projects and the development of Artificial Intelligence applications. It is argued that the preserved contents of the web are also data and, as such, should be included in issues related to curation.

Arquivo.pt (<https://arquivo.pt>) is the Portuguese web archive, a government service around science, research, and innovation, which provides web content from the 1990s to the present day. All its data (1.4 Petabyte) is open and accessible to both researchers and ordinary citizens, both nationally and internationally, via the web. Content can be searched by text, image or URL. It also has APIs for automatic processing.

First, this presentation starts by briefly describing how Arquivo.pt identified and collected the websites of European research projects, from FP4 to FP7 and H2020. In 2021, the European Union's Open Data Portal published a data set from the Community Research and Development Information Service (CORDIS) that documents H2020 research projects. However, from the 31 129 projects listed, only 46% presented a project URL. Arquivo.pt developed a low-cost methodology that automatically identifies URLs related to R&D projects to be preserved. Fifty-two million files related to scientific projects were preserved, resulting in 17 terabytes of information. Since then, every year, Arquivo.pt has been collecting information project websites and science-related news in the media. To raise awareness among researchers and curators, it collaborated with the PTCRIS.pt program by sending an email to the principal investigator and providing a training program on web archiving and the processing of archived content through APIs.

Secondly, this presentation mentions two cases of use of Arquivo.pt to illustrate how researchers are beginning to use preserved web data in the context of

The 20th International Digital Curation Conference takes place on 16-18 February 2026 in Zagreb, Croatia.

URL: <https://www.dcc.ac.uk/events/idcc26>

Copyright rests with the authors. This work is released under a Creative Commons Attribution License, version 4.0. For details please see <https://creativecommons.org/licenses/by/4.0/>.



research and experimentation. One example, among dozens of others from the annual Arquivo.pt Award competition, is a work on the theme of immigration called “Narrative Monitoring.” The other case, called “GlorIA,” in the field of Artificial Intelligence, uses data from Arquivo.pt to create a Large Language Model (LLM) for European Portuguese.

In conclusion, this presentation emphasizes the accessibility of Arquivo.pt and the fact that anyone can find content in many languages, including the 24 languages of the European Union. With these two good reasons, it invites curators and researchers to access Arquivo.pt and to critically identify issues that still need to be resolved, both in curating and using data from the web of the past.