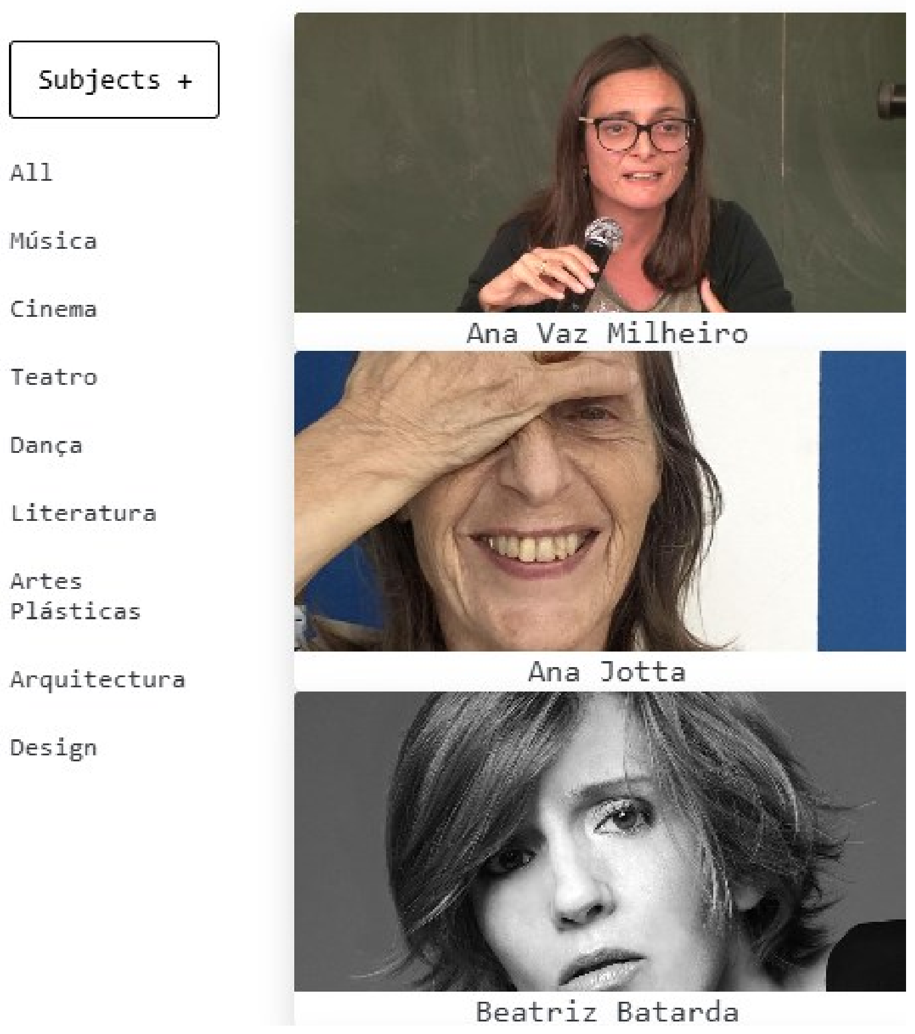
Arquivo.pt API and Bulk access usage

When collecting large amounts of data for archiving purposes it is important to ensure that the data is accessible and searchable so that researchers can find the useful data they need. For this reason, in 2018 Arquivo.pt implemented Application Program Interface (API) access, which allowed for automatic access to the information and for microservices to be built on top of it. Consequently, now close to half of the web traffic to Arquivo.pt is made through API requests. Nowadays there are four different APIs which cater to different needs of the community, they have supported several projects which wouldn't have been possible without them. Arquivo.pt also promotes the usage of its data through the Arquivo.pt Awards, a yearly event meant to reward any innovative works which use Arquivo.pt as its primary source of data.

Text Search API

Implemented in 2018, it's Arquivo.pt oldest API. It allows users to search the archived pages for text content, similar to what a conventional search engine might do over the live web. It's usage has been growing steadily and since 2023 it represented about 1/3 of all requests made to arquivo.pt.

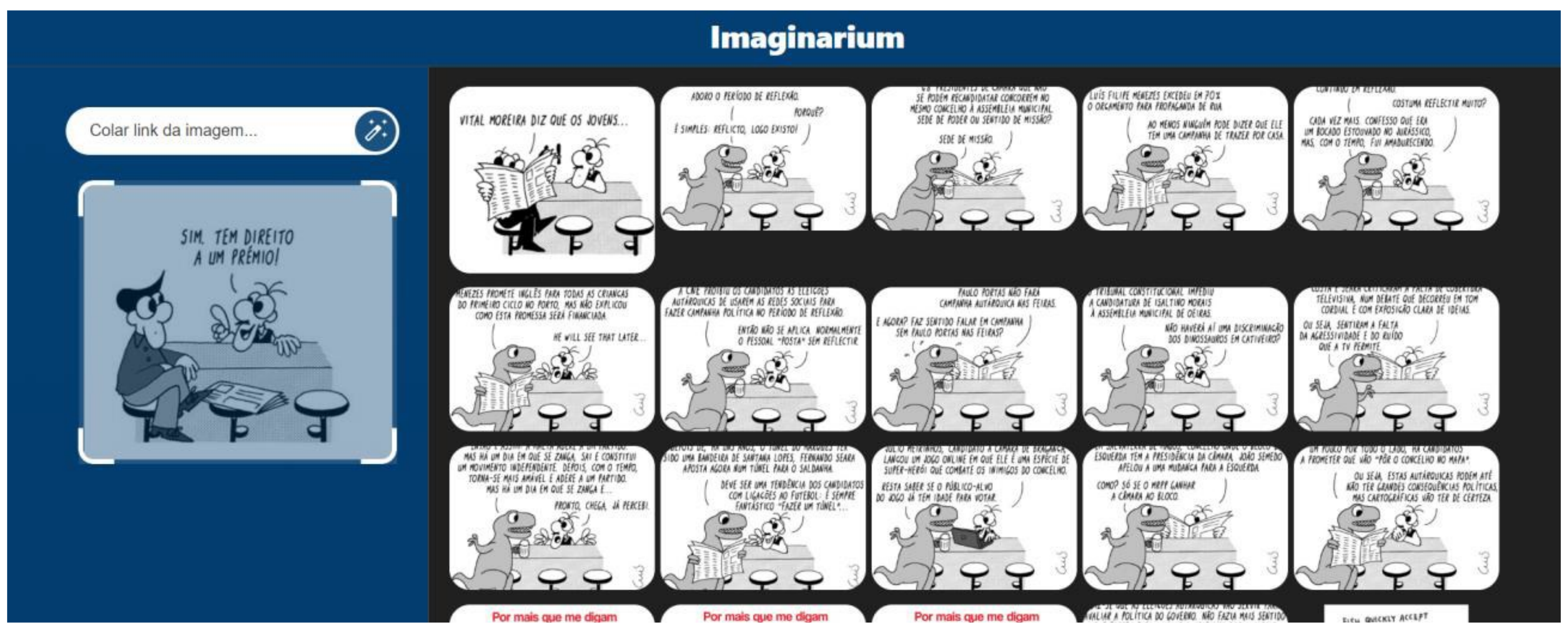
An example of a project that used the text search API is the 2nd place winner of the 2023 Arquivo.pt Awards, "Representatividade das mulheres artistas na imprensa nacional". This work is a study on the representativity of artist women in the Portuguese press, it was made by Cláudia Sevivas and Miguel Boavida, both from Universidade Europeia. On the project's website you can find information about Portuguese women artists and explore news articles that mention them over the years.



existo.pt

Image Search API

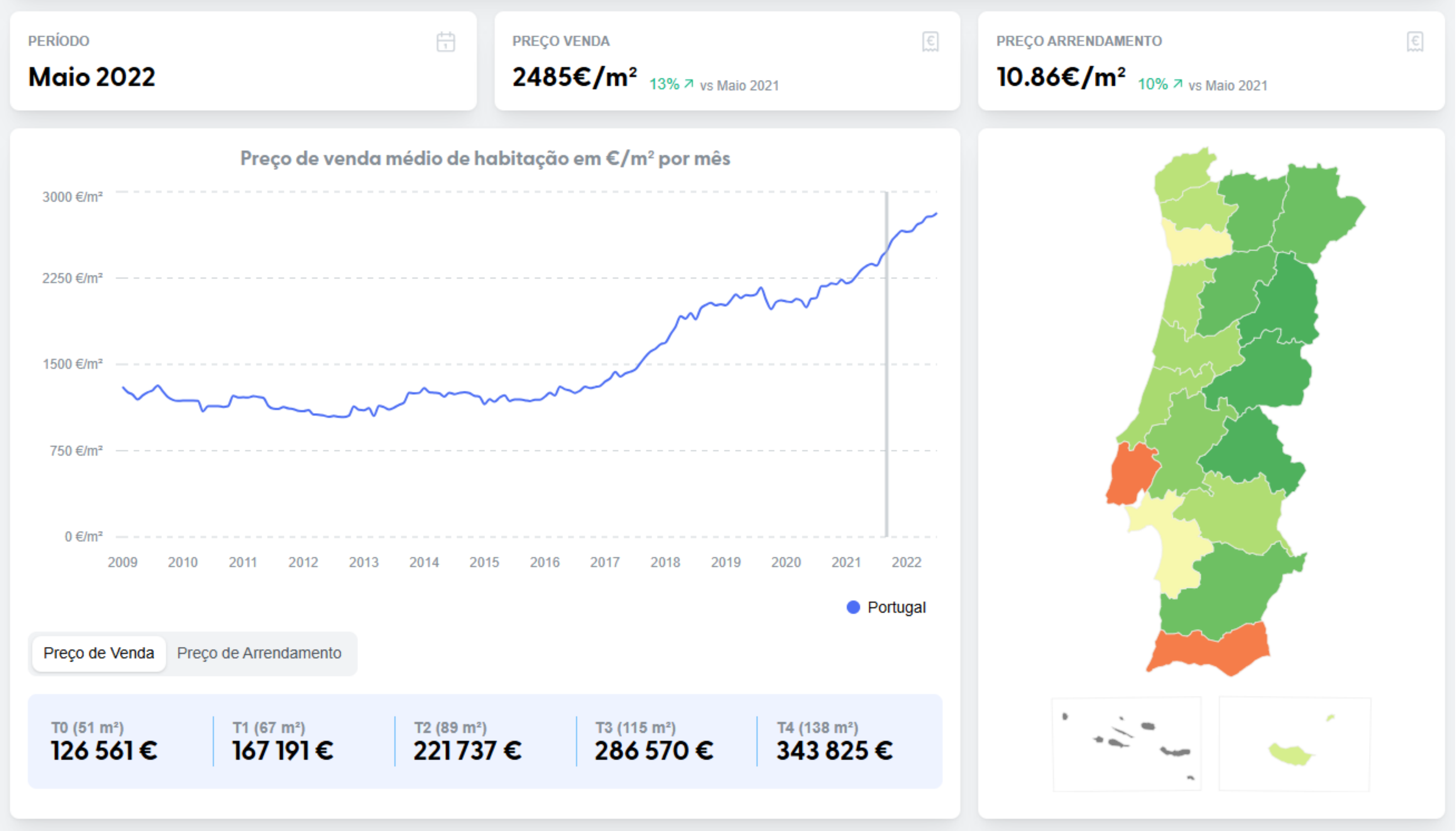
Arquivo.pt was the first web archive to implement an image search functionality, accessible from both the arquivo.pt website and the image search API. Using the image search API, projects such as Imaginarium were built. Developed by Diogo Santos from Universidade Nova de Lisboa, Imaginarium is an image search engine which uses other images rather than search terms to search for similar pictures. This project had an honorary mention at the 2023 Arquivo.pt Awards.



imaginarium.pages.dev

CDX Server API

Our CDX server API gives users automatic access to list, sort, and filter preserved pages from a given URL. Using this API users can, for example, from a selected list of domains get all archived pages from those domains over time and analyze their evolution. This was the approach taken by Habitação.NET, a study made by Diogo Gonçalves which aimed to study the real estate market evolution and tendencies over the past 15 years. It was awarded the second place at the 2024 Arquivo.pt Awards.



habitacao.net

Memento API

The Memento API is an International standard that allows interoperability with other web archives. At Arquivo.pt this API supports services such as Complete*Page*, a function that allows the user to recover missing elements in web-archived pages from other web archives or the live web.

Bulk Access to CDXJ files

To cope with the increasing demand for data for research, in 2023 Arquivo.pt began making all its CDXJ index files publicly available for download. This resulted in an over sixtyfold increase of bandwidth of data downloaded for research. One notable use case was the support of GlórIA, an LLM for European Portuguese with 35 billion tokens. The project was led by Prof. David Semedo, together with Ricardo Lopes and Prof. João Magalhães from Universidade Nova de Lisboa.

Find more

The projects displayed here are just a small sample of the works that have been made using Arquivo.pt's data using APIs. Find more examples by consulting Arquivo.pt Award's history at:

arquivo.pt/awards

Contact us

If you wish to know more about our APIs or the Portuguese web archive, visit us at arquivo.pt or contact us directly via:

contacto@arquivo.pt

arquivo.pt/api

fct arquivo.pt
FCCN serviços digitais fct
fct Fundação para a Ciência e a Tecnologia