Café com o Arquivo.pt



Dia mundial da preservação digital

5 novembro 2026

Novidades do Arquivo.pt

Ivo.Branco@ai2portugal.pt

FCCN serviços digitais fct | fct Fundação para a Ciência e a Tecnologia | ai2 Agência para a Investigação e Inovação

Novidades

Pesquisa de páginas por texto (reimplementação e duplicação da informação)

Novas funcionalidades de pesquisa

- Linha temporal
- Equilíbrio entre anos
- Pesquisa por língua

Sugestões de pesquisa

Pesquisa avançada

ArchivePageNow: proteção contra utilização abusiva

Melhorias partilhadas com a comunidade arquivos da web (pywb)

RGPD e privacidade

APIs abertas e Docker

Arquivo.pt como plano de contingência

Desafios



Pesquisa de páginas por texto

Duplicámos a informação pesquisável.

- Espaço das coleções indexado em TB: 658 → 1479 (+125%)
- Nº de coleções indexadas: 106 → 194 (+83%)

Reimplementação da tecnologia: Nutchwax → Apache Solr.

Esforço de vários anos.

Hardware

- 2 clusters de servidores
- Cada cluster com 10 servidores
- Cada servidor com 512GB de RAM e 40TB SSD



Quando é que um tema foi importante?

Problema: A pesquisa devolvia uma lista de resultados, mas não mostrava em que anos o tema esteve mais presente na Web.

Novidade: Barras por ano acima do seletor de datas mostram o impacto da **pesquisa ao longo do tempo**.

Impacto normalizado pelo volume do arquivo em cada ano, para comparar anos de forma justa

API: timeline=true; devolve counts (páginas por ano) e impact (valor normalizado)



Equilíbrio entre anos na pesquisa

Problema: O arquivo tem menos conteúdo antes de 2015 do que depois de 2020. Por isso, os resultados das pesquisas vêm sobretudo dos anos mais recentes.

Novidade: Nova opção Equilíbrio de anos na Pesquisa Avançada de Páginas: Desligado (predefinição), 0.25, 0.5, 0.75 e 1.

Faz subir nos resultados as páginas dos anos com menos conteúdo.

É opcional e está desligada por omissão.

Também está disponível na API.

Estado: Experimental. Ainda vamos avaliar o desempenho e a qualidade dos resultados.



Pesquisa por língua

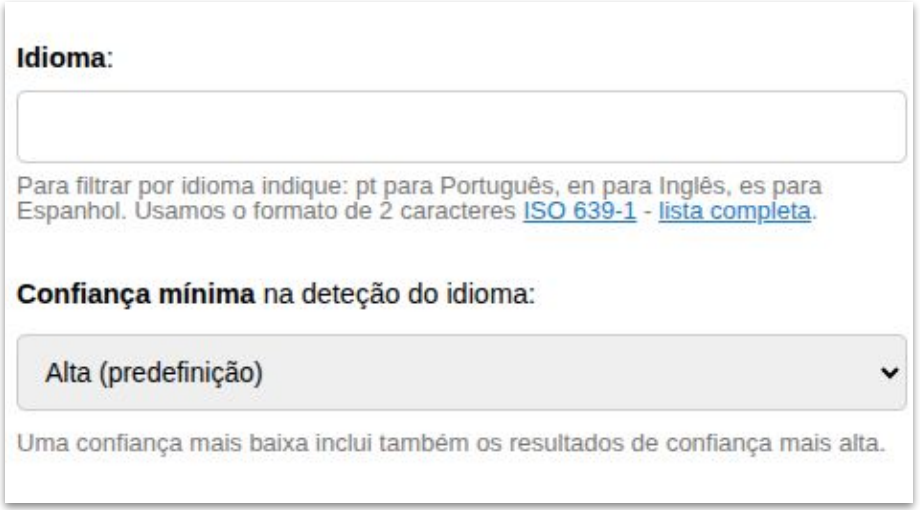
Novo filtro **Língua** na Pesquisa Avançada de Páginas

(ex.: pt, en, es)

Confiança mínima na deteção da língua:

Alta (predefinição), Média ou Baixa

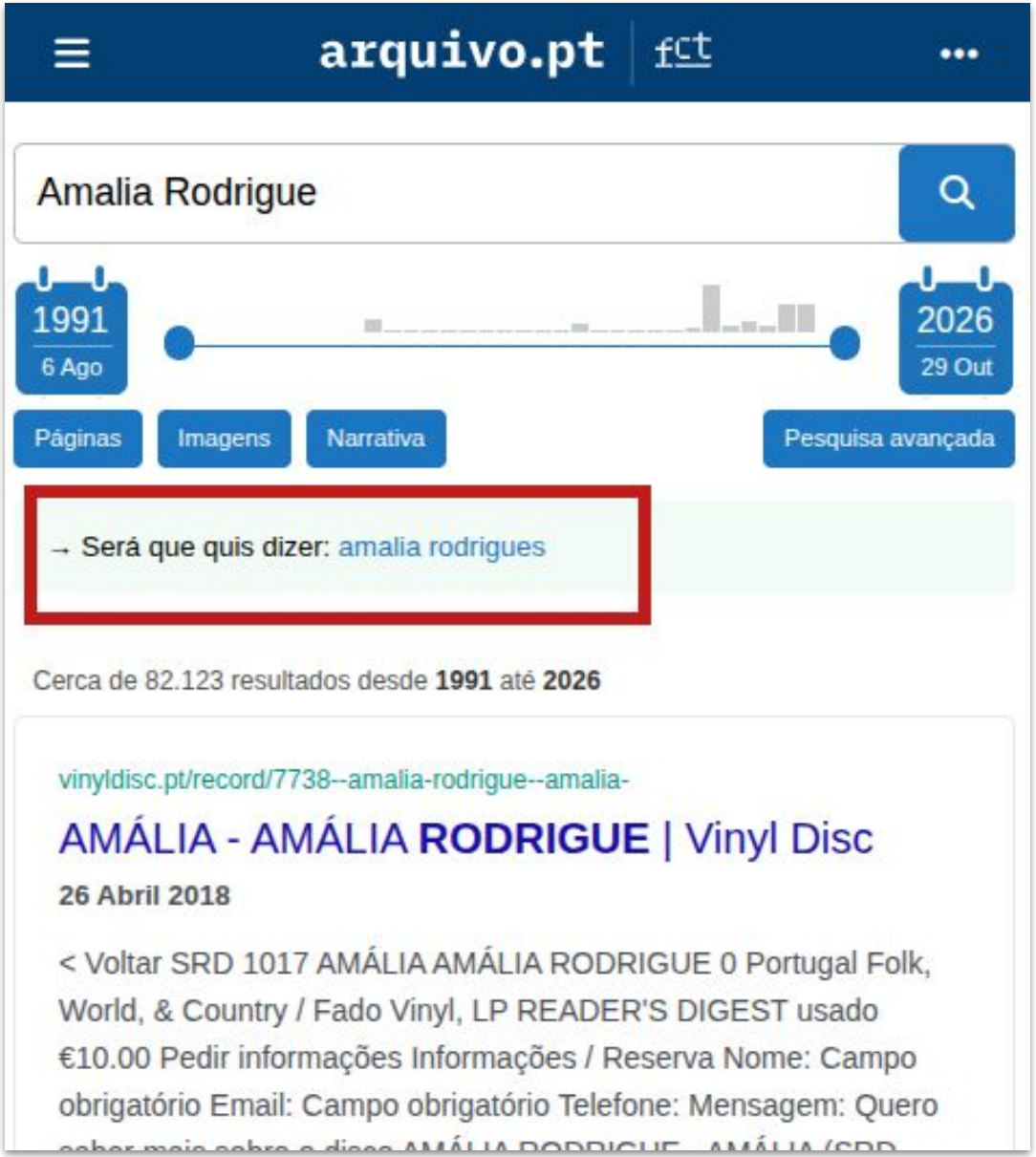
Também na caixa de pesquisa (language:pt) e na API





Sugestões de pesquisa

- Corrige erros de escrita:
  - **Lizboa → Lisboa**
  - **Amalia Rodrigue → Amalia Rodrigues**
- Na pesquisa de páginas e de imagens
- Sem dependência de serviços externos
  - usa os dados do Arquivo.pt
  - através do Apache Solr



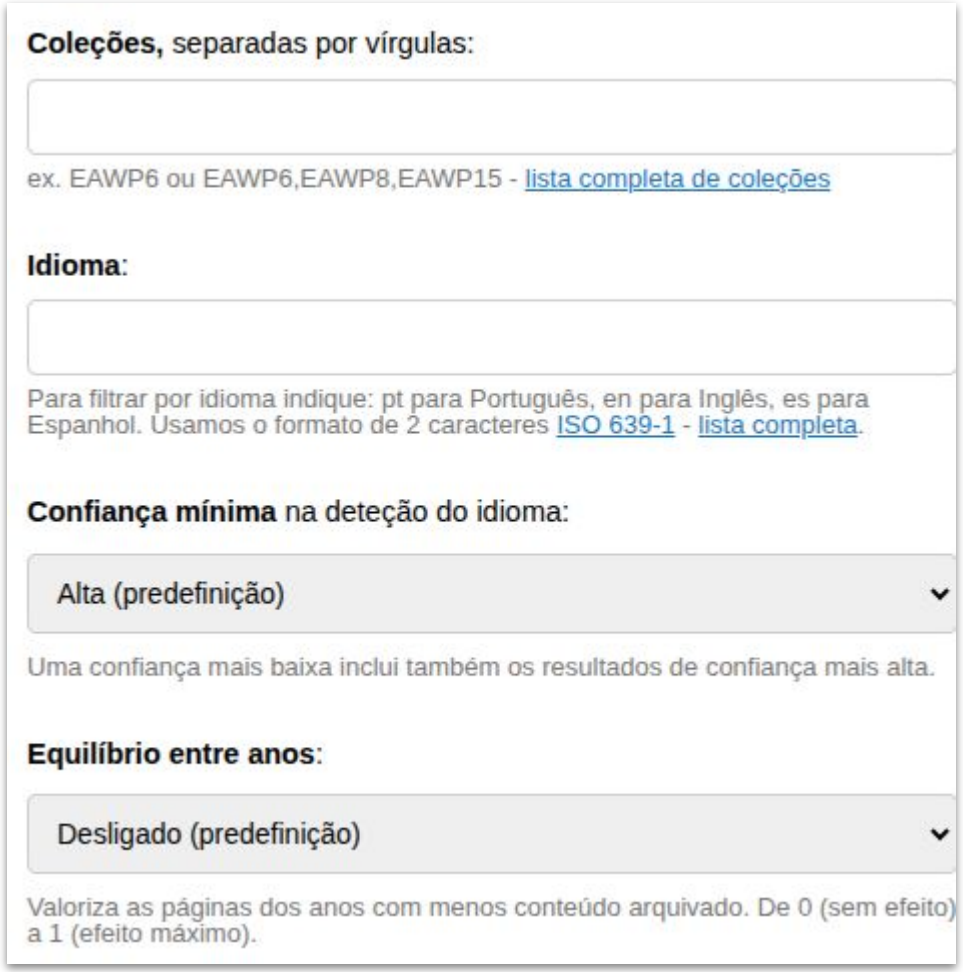
Pesquisa avançada: novos critérios

Pesquisa de páginas por texto

- **Coleção**: pesquisar só numa ou em várias coleções (ex.: EAWP6,EAWP8), nas páginas e nas imagens
- **Língua** e **confiança mínima** da língua detetada
- **Equilíbrio de anos**: dar mais destaque aos anos com menos conteúdo

Pesquisa de imagens

- **Número de resultados** na pesquisa de imagens: até 200



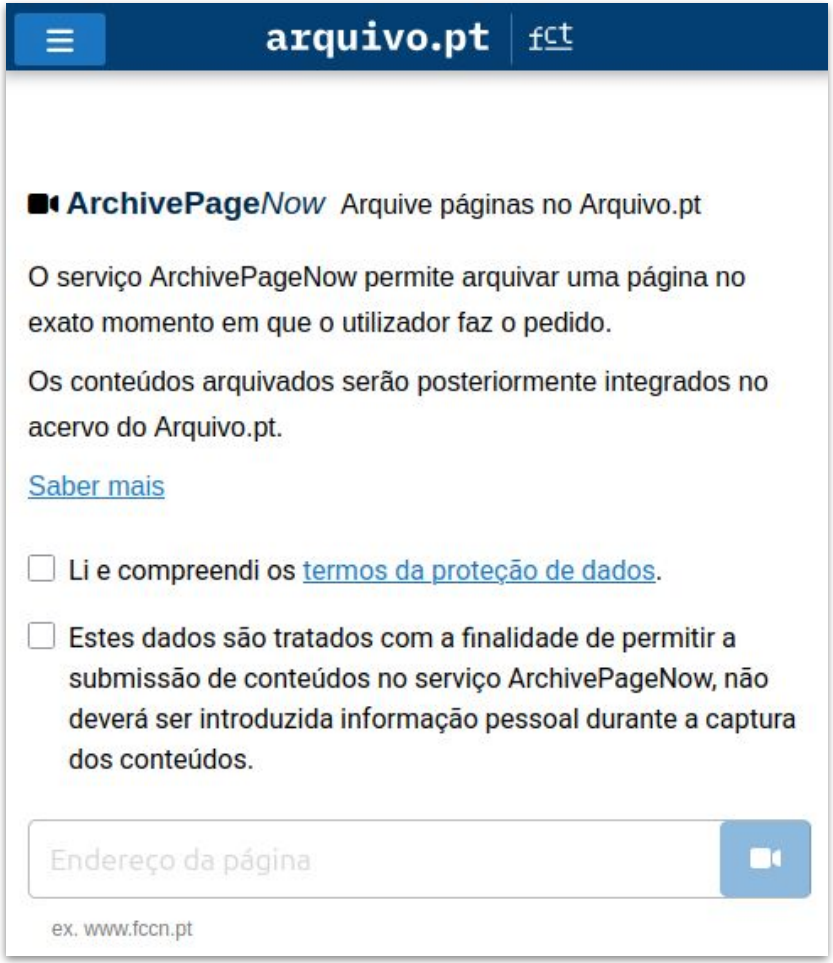
ArchivePageNow: proteção contra abusos

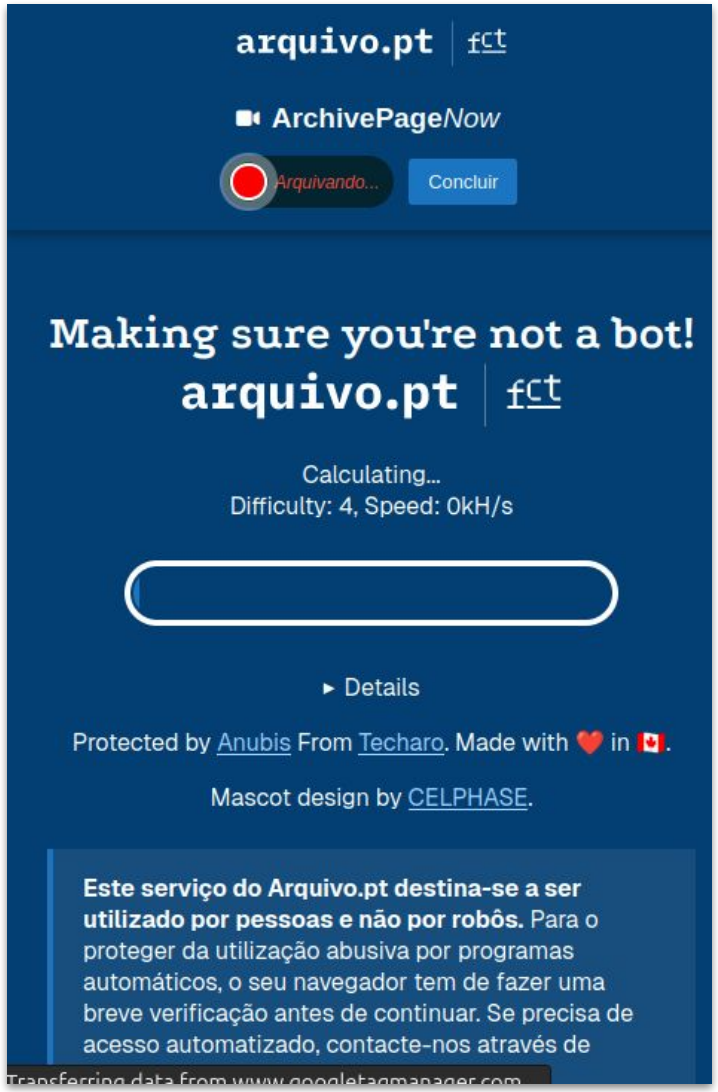
O serviço foi usado por agentes de IA automáticos para fazer pedidos a sites governamentais.

Fizemos uma auditoria de segurança ao serviço.

Novas proteções:

- verificação anti-robôs (Anubis)
- limites de utilização
- aceitação obrigatória dos termos de proteção de dados

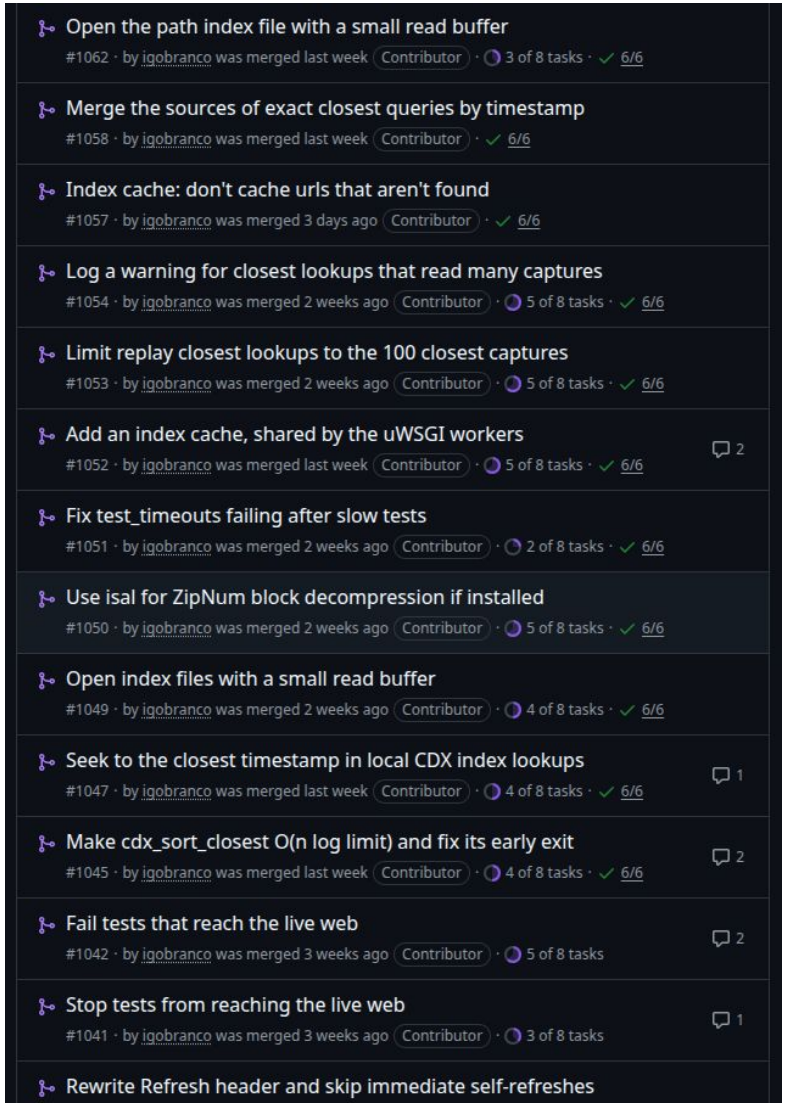




Melhorias partilhadas com a comunidade

No replay (componente: pywb) partilhado melhorias com a comunidade de arquivos da web.

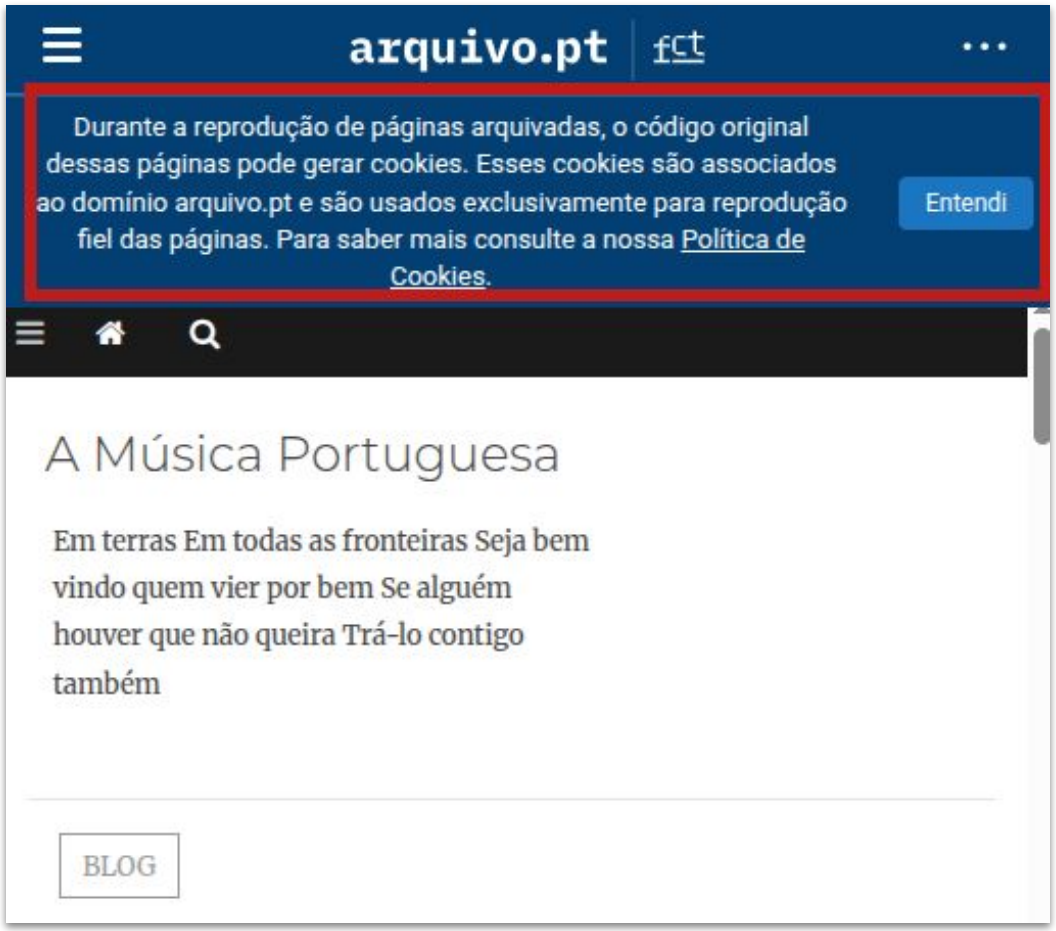
- **Segurança**: controlo dos endereços que o ArchivePageNow pode visitar
- **Desempenho**: replay mais rápido em páginas com muitas versões arquivadas
- **Qualidade**: páginas com redirecionamento Refresh deixam de saltar para a Web atual



github.com/webrecorder/pywb

RGPD e privacidade

- **Consentimento explícito** obrigatório no ArchivePageNow e no CitationSaver
- **Aviso de cookies** ao consultar páginas arquivadas, com uma nova secção na Política de Cookies



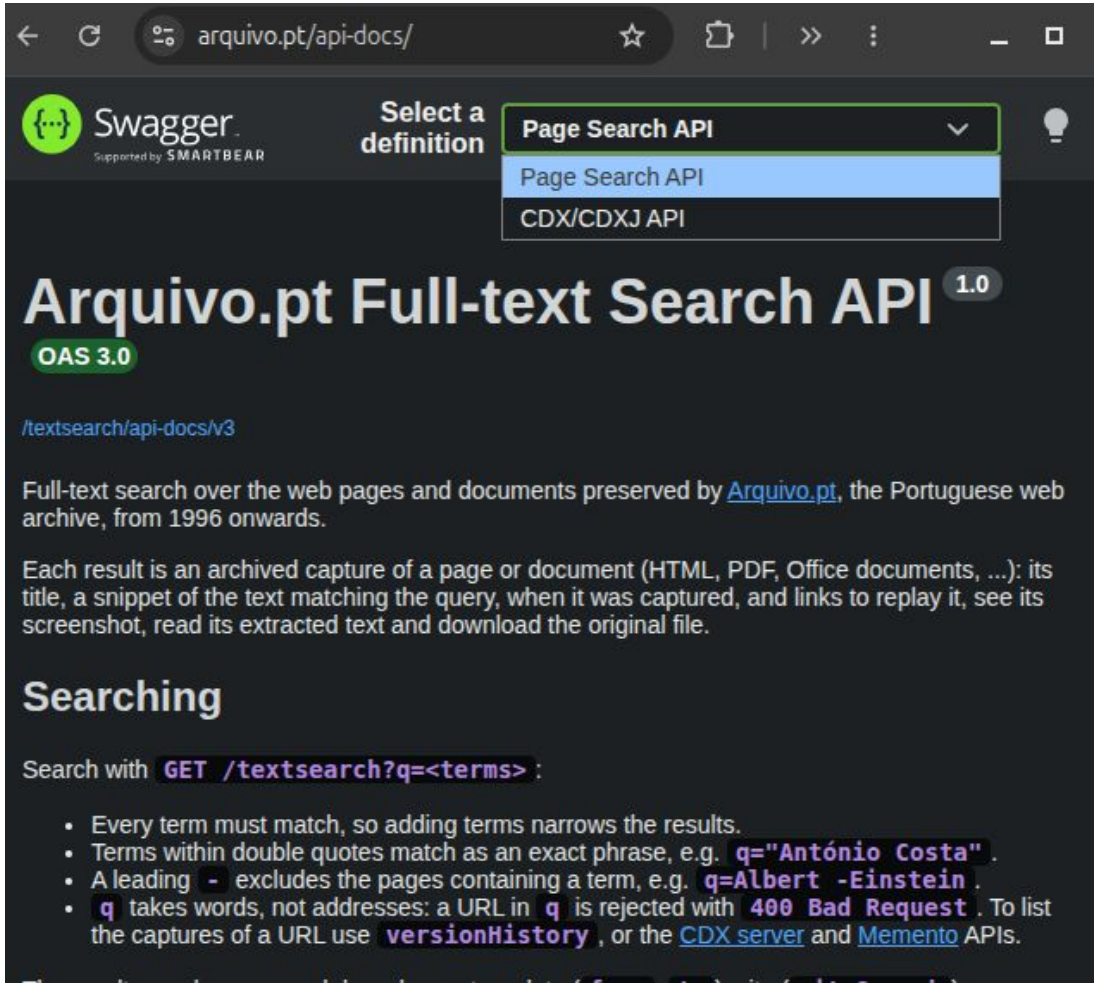
Tecnologia: APIs abertas e Docker

Documentação das APIs (OpenAPI v3 / Swagger):

- Pesquisa de páginas por texto: /textsearch
- Pesquisa por URL (CDX): /wayback/cdx
- Pesquisa de imagens: brevemente

Docker

- A maioria dos serviços corre agora em contentores, o que facilita a operação e a inovação



arquivo.pt/api

arquivo.pt/api-docs

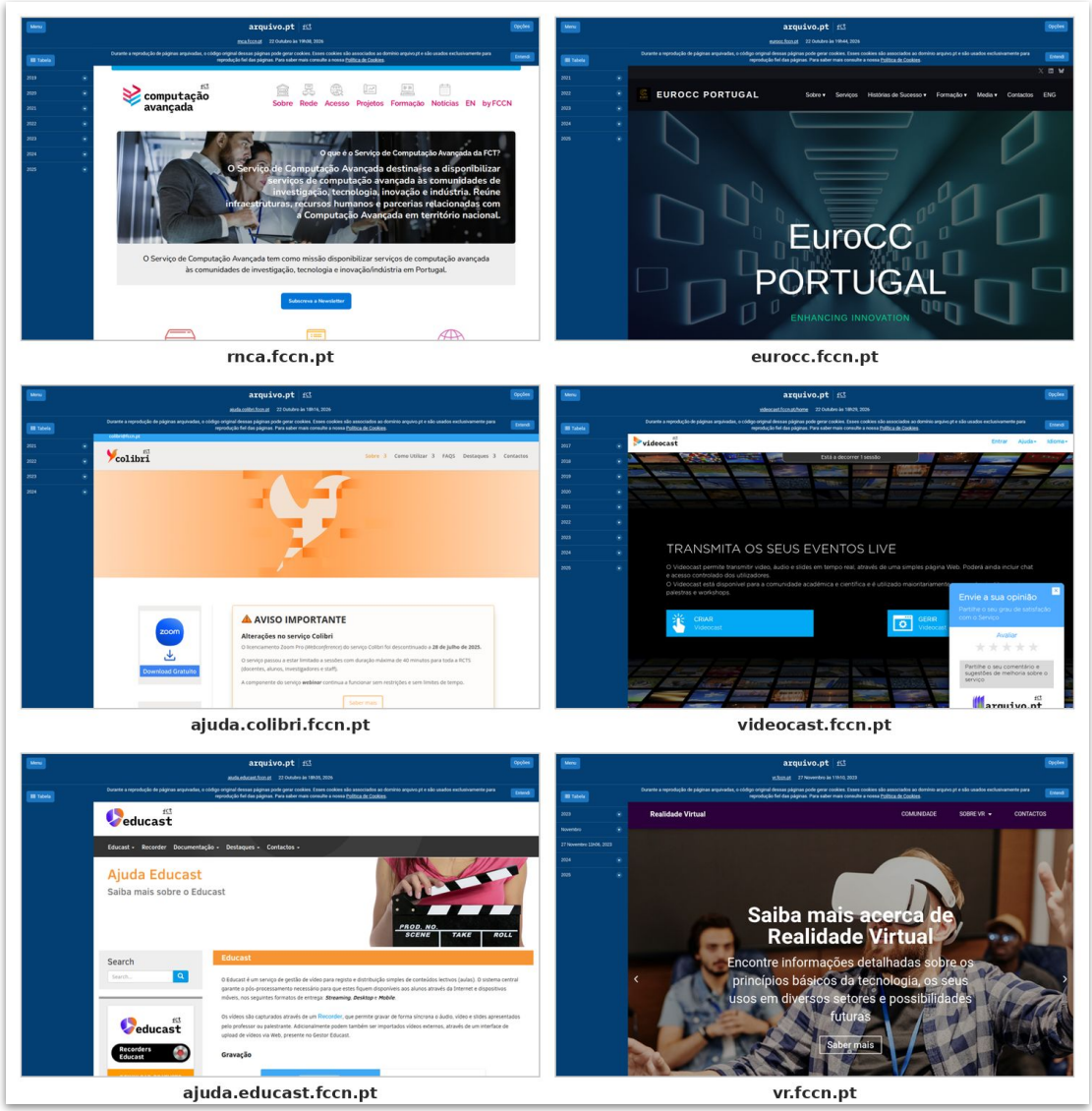
Arquivo.pt como plano de contingência

Problema: em outubro de 2026 foi divulgada uma vulnerabilidade crítica no PHP (PHP2Shell) que permitia executar código remotamente nos servidores.

10 sites afetados e redirecionados temporariamente para o Arquivo.pt

Resposta: em menos de 24 horas recolhemos os sites afetados e passámos a encaminhar o tráfego para a versão arquivada no Arquivo.pt.

Novidade: o novo sobre.arquivo.pt é um site estático feito com Astro, sem PHP e sem base de dados. Há menos componentes que podem ser atacados.



Desafios: manter o Arquivo.pt aberto a todos

- Uso massivo do Arquivo.pt por agentes de IA e robôs.
- Objetivo: continuar a ser um **serviço público e aberto**, **sem obrigar a autenticação** para consultar o arquivo.
- **Fair admission**: dividir a capacidade de forma justa entre utilizadores. Fase experimental até ao fim de 2026.





Obrigado

contacto@arquivo.pt