Arquivo.pt behind the curtains

daniel.gomes@fccn.pt

Fundação para a Ciência e a Tecnologia

Who are we?

Free online service to research the Past Web

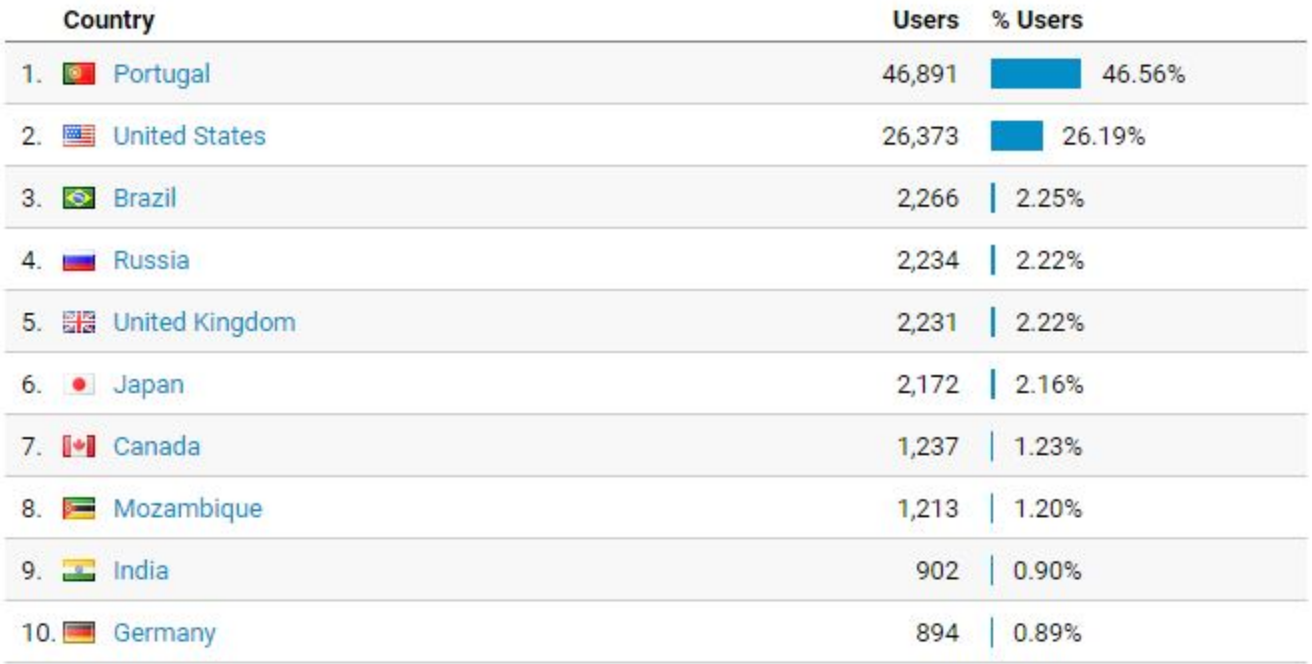
Preserves **publicly accessible** information related with:

- Portugal
- **Research** and **Education** (international)

Governmental service provided by
Foundation for Science and Technology (Portugal)

A digital research infrastructure

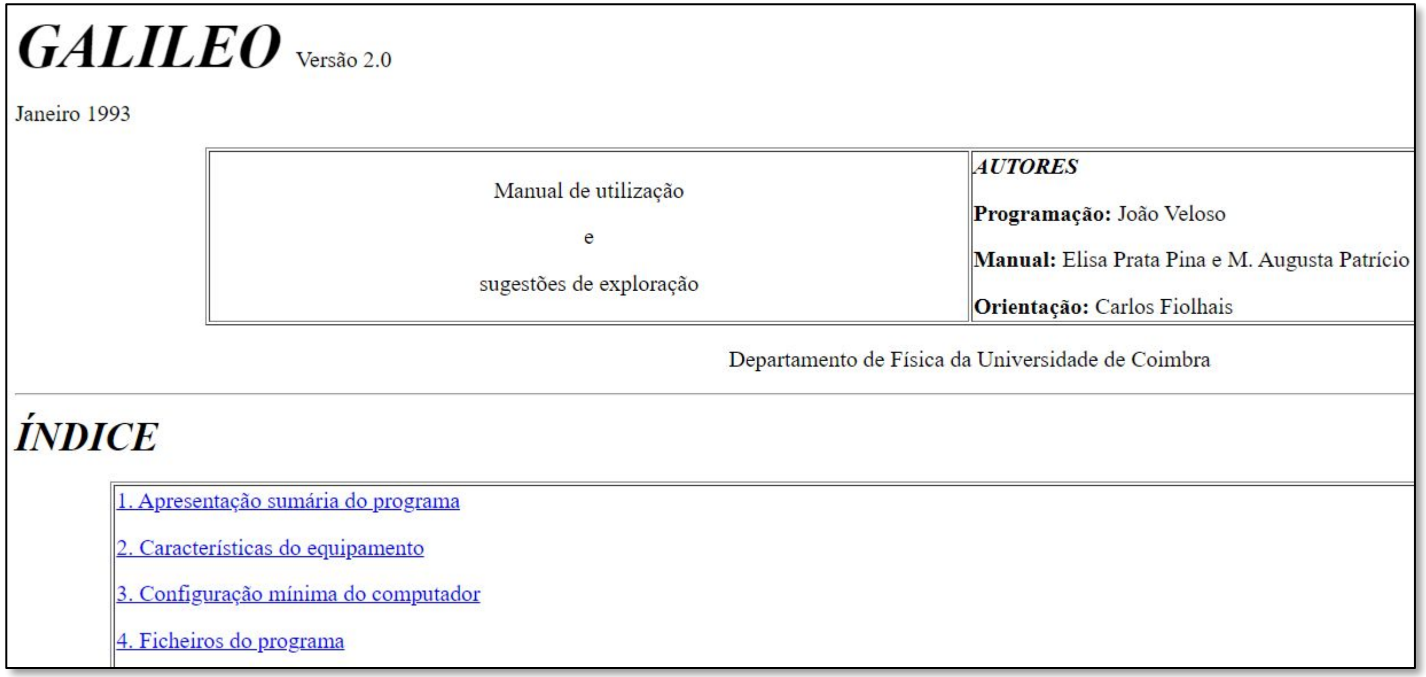
Arquivo.pt is used word-wide



	Country	Users	% Users
1.	Portugal	46,891	46.56%
2.	United States	26,373	26.19%
3.	Brazil	2,266	2.25%
4.	Russia	2,234	2.22%
5.	United Kingdom	2,231	2.22%
6.	Japan	2,172	2.16%
7.	Canada	1,237	1.23%
8.	Mozambique	1,213	1.20%
9.	India	902	0.90%
10.	Germany	894	0.89%

- 53% of users are international
- User Interfaces and documentation also in English
- Combined with Google Translate enables **cross-lingual access** to preserved content

Arquivo.pt preserves national and international historical web content



GALILEO Versão 2.0

Janeiro 1993

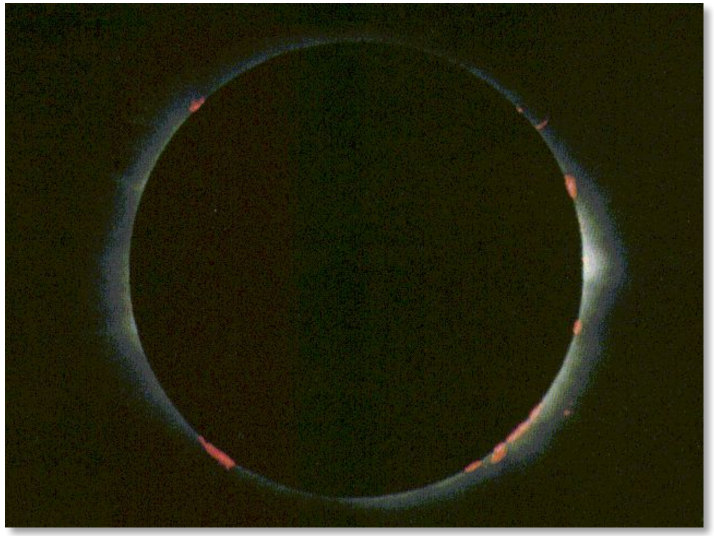
Manual de utilização e sugestões de exploração	***AUTORES*** **Programação:** João Veloso **Manual:** Elisa Prata Pina e M. Augusta Patrício **Orientação:** Carlos Fiolhais

Departamento de Física da Universidade de Coimbra

ÍNDICE

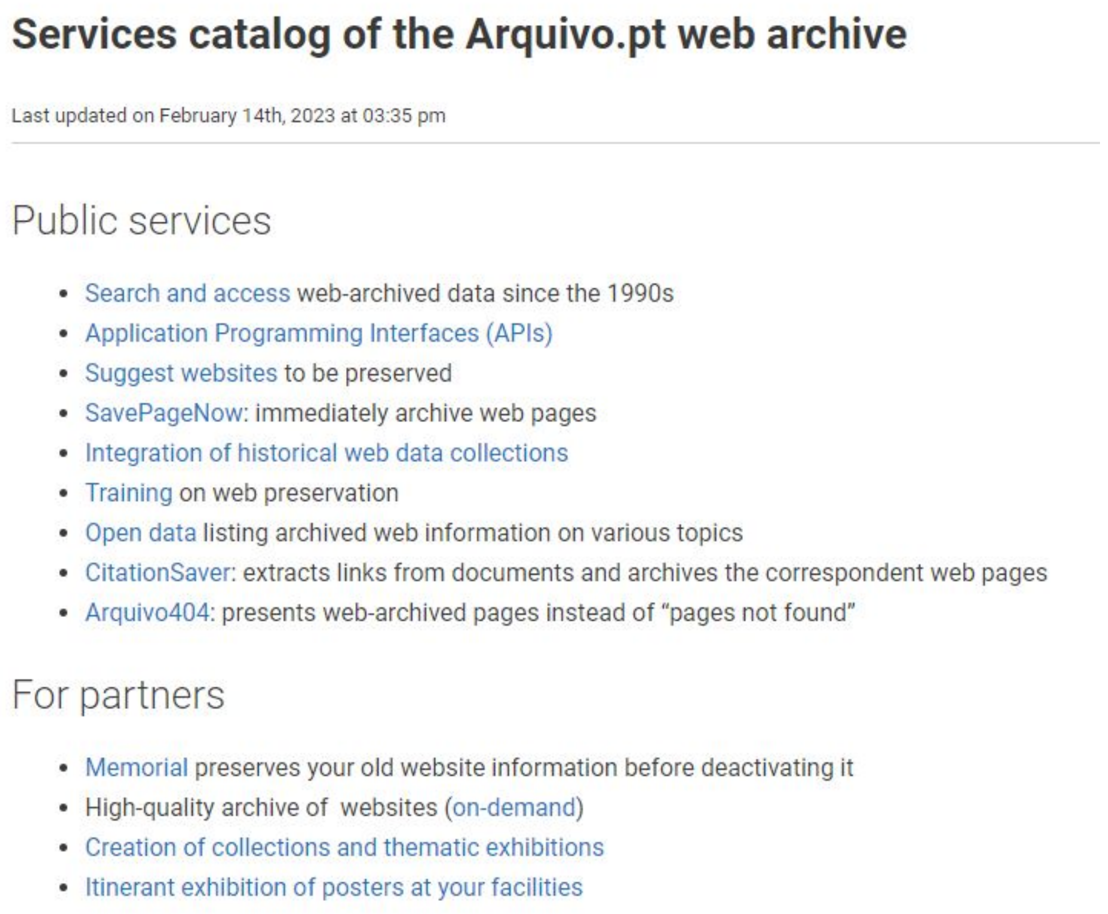
1. Apresentação sumária do programa
2. Características do equipamento
3. Configuração mínima do computador
4. Ficheiros do programa

nautilus.fis.uc.pt- 1993
(oldest page)



spacelink.nasa.gov – 1992
(oldest image)

Arquivo.pt supports 13 services



Services catalog of the Arquivo.pt web archive

Last updated on February 14th, 2023 at 03:35 pm

Public services

- Search and access web-archived data since the 1990s
- Application Programming Interfaces (APIs)
- Suggest websites to be preserved
- SavePageNow: immediately archive web pages
- Integration of historical web data collections
- Training on web preservation
- Open data listing archived web information on various topics
- CitationSaver: extracts links from documents and archives the correspondent web pages
- Arquivo404: presents web-archived pages instead of "pages not found"

For partners

- Memorial preserves your old website information before deactivating it
- High-quality archive of websites (on-demand)
- Creation of collections and thematic exhibitions
- Itinerant exhibition of posters at your facilities

arquivo.pt/catalog

Our **successful** official History!

2007: Start of the Portuguese Web Archive project

2008: 1st crawl of the Portuguese web

2009: Recommendations for publishers

2010: Public prototype of the search and access system

2011: Open source available

2012: Application Programming Interface

2013: Decree-Law mandating our web archiving activities

2017: Training program

2018: Arquivo.pt Award

2019: Arquivo.pt Memorial

2021: Image search service

2022: Arquivo.pt was considered the best Digital Service of 2022!

ARQUIVO.PT

Behind the curtains

The adversities

2007: Start of the Portuguese Web Archive project

9/2013: Total collapse due to hardware malfunction

- Irrecoverable data loss of 17% (17 TB)
- Crawling interruptions
- Suspension of search service

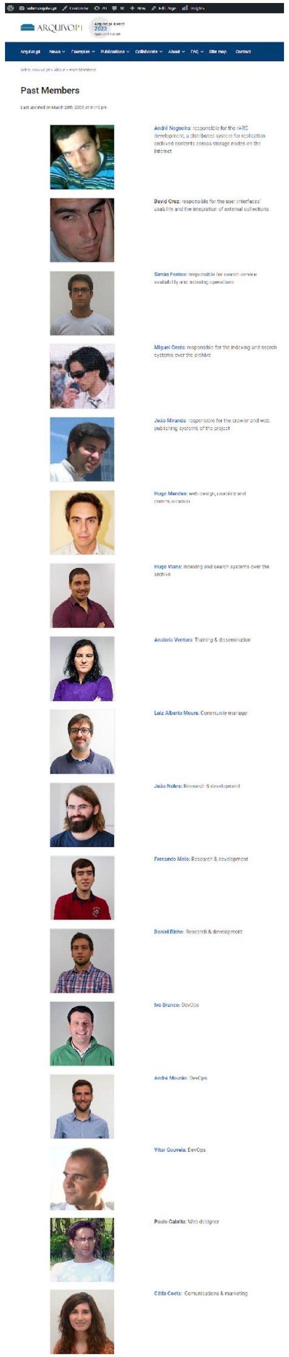
2014 - 2016: Recovery and improving robustness

- Ground-zero recovery
- 9,2 hours of downtime in 1 year

2022: Service availability of 99.894%

High staff rotation: 23 members in 15 years with average team size of 4

- On average: each 8 months, one member leaves



Past Members

Last updated on March 28th, 2023 at [illegible] pm

André Nogueira: responsible for the rARC development, a distributed system for replication archived contents across storage nodes on the Internet

David Cruz: responsible for the user interfaces' usability and the integration of external collections

Simão Fontes: responsible for search service availability and indexing operations

Miguel Costa: responsible for the indexing and search systems over the archive

João Miranda: responsible for the crawler and web publishing systems of the project

Hugo Mendes: web design, usability and communication

Hugo Viana: indexing and search systems over the archive

Anabela Ventura: Training & dissemination

Luiz Alberto Moura: Community manager

João Nobre: Research & development

Fernando Melo: Research & development

Daniel Bicho: Research & development

Ivo Branco: DevOps

André Mourão: DevOps

Vitor Gouveia: DevOps

Paulo Cabrita: Web designer

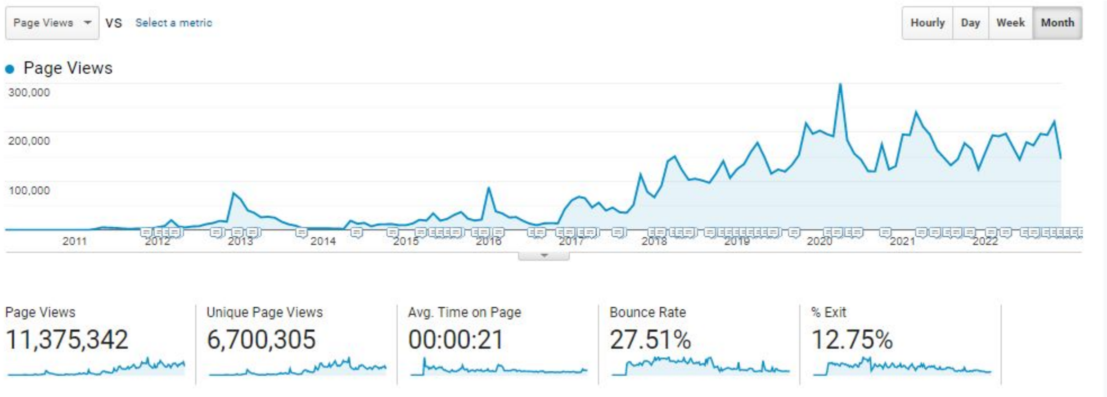
Cátia Costa: Comunications & marketing

Workload (only human users)

	Average month	Total 2022
Users	10 453	125 439
Sessions	14 373	172 478
PageViews	188 253	2 259 039

Exciting figures for an archive **but** humble for a web service.

Steady growth since 2010



Page Views 11,375,342

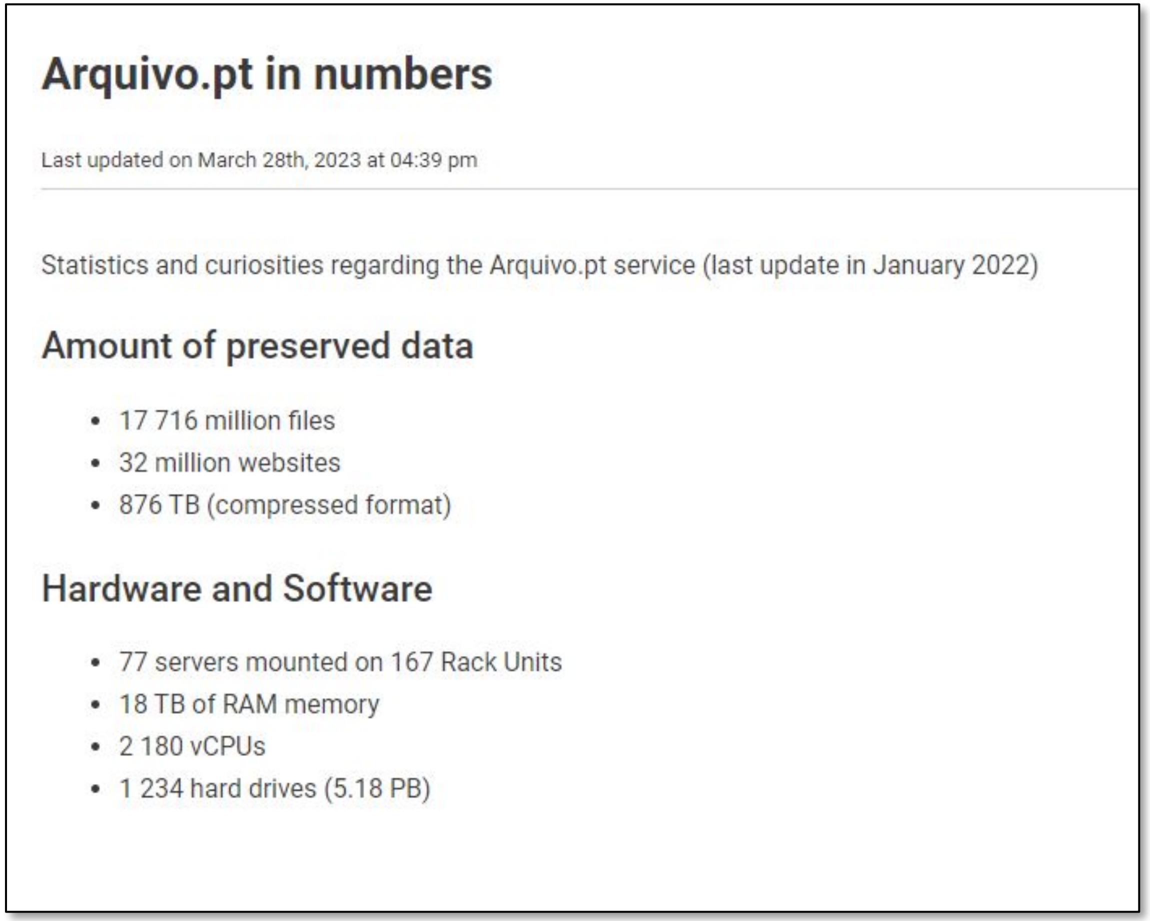
Unique Page Views 6,700,305

Avg. Time on Page 00:00:21

Bounce Rate 27.51%

% Exit 12.75%

Arquivo.pt is a medium-size web archive



Arquivo.pt in numbers

Last updated on March 28th, 2023 at 04:39 pm

Statistics and curiosities regarding the Arquivo.pt service (last update in January 2022)

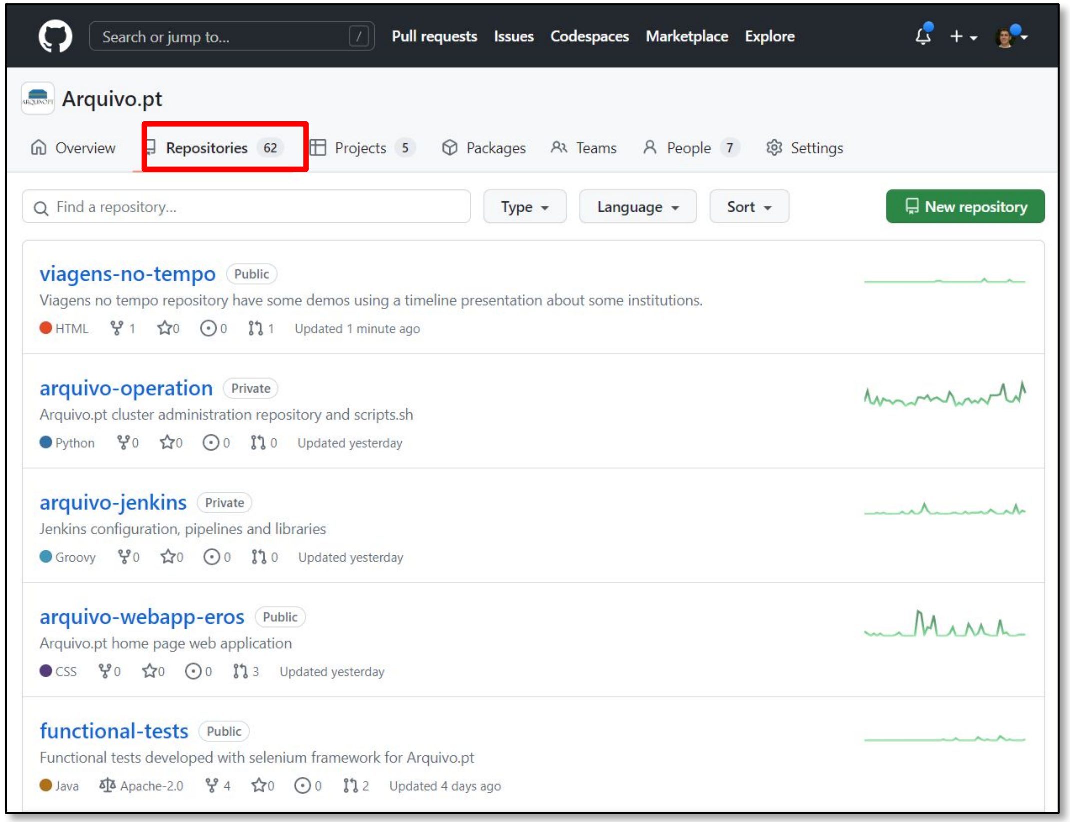
Amount of preserved data

- 17 716 million files
- 32 million websites
- 876 TB (compressed format)

Hardware and Software

- 77 servers mounted on 167 Rack Units
- 18 TB of RAM memory
- 2 180 vCPUs
- 1 234 hard drives (5.18 PB)

Supported by a lot of software



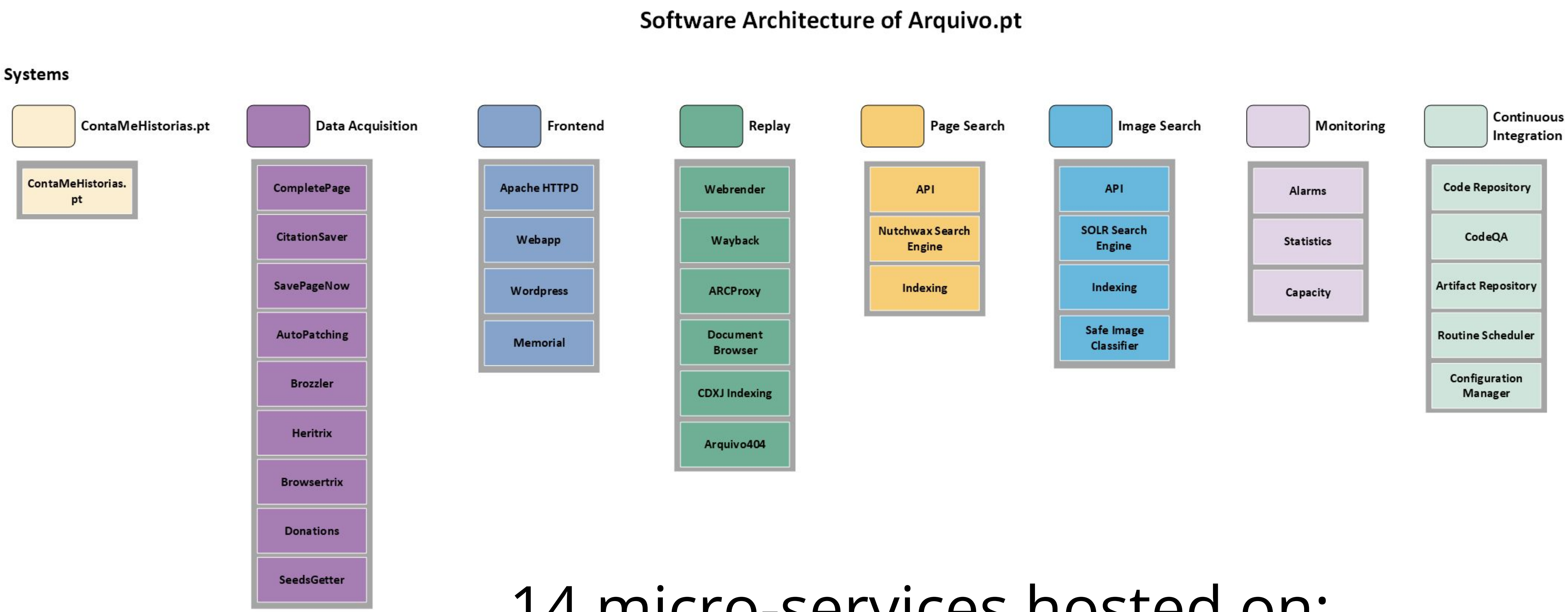
viagens-no-tempo Public
Viagens no tempo repository have some demos using a timeline presentation about some institutions.
HTML · 1 · 0 · 0 · 1 · Updated 1 minute ago

arquivo-operation Private
Arquivo.pt cluster administration repository and scripts.sh
Python · 0 · 0 · 0 · 0 · Updated yesterday

arquivo-jenkins Private
Jenkins configuration, pipelines and libraries
Groovy · 0 · 0 · 0 · 0 · Updated yesterday

arquivo-webapp-eros Public
Arquivo.pt home page web application
CSS · 0 · 0 · 0 · 3 · Updated yesterday

functional-tests Public
Functional tests developed with selenium framework for Arquivo.pt
Java · Apache-2.0 · 4 · 0 · 0 · 2 · Updated 4 days ago



Software Architecture of Arquivo.pt

Systems

- ContaMeHistorias.pt
  - ContaMeHistorias.pt
- Data Acquisition
  - CompletePage
  - CitationSaver
  - SavePageNow
  - AutoPatching
  - Brozzler
  - Heritrix
  - Browsertrix
  - Donations
  - SeedsGetter
- Frontend
  - Apache HTTPD
  - Webapp
  - Wordpress
  - Memorial
- Replay
  - Webrender
  - Wayback
  - ARCProxy
  - Document Browser
  - CDXJ Indexing
  - Arquivo404
- Page Search
  - API
  - Nutchwax Search Engine
  - Indexing
- Image Search
  - API
  - SOLR Search Engine
  - Indexing
  - Safe Image Classifier
- Monitoring
  - Alarms
  - Statistics
  - Capacity
- Continuous Integration
  - Code Repository
  - CodeQA
  - Artifact Repository
  - Routine Scheduler
  - Configuration Manager

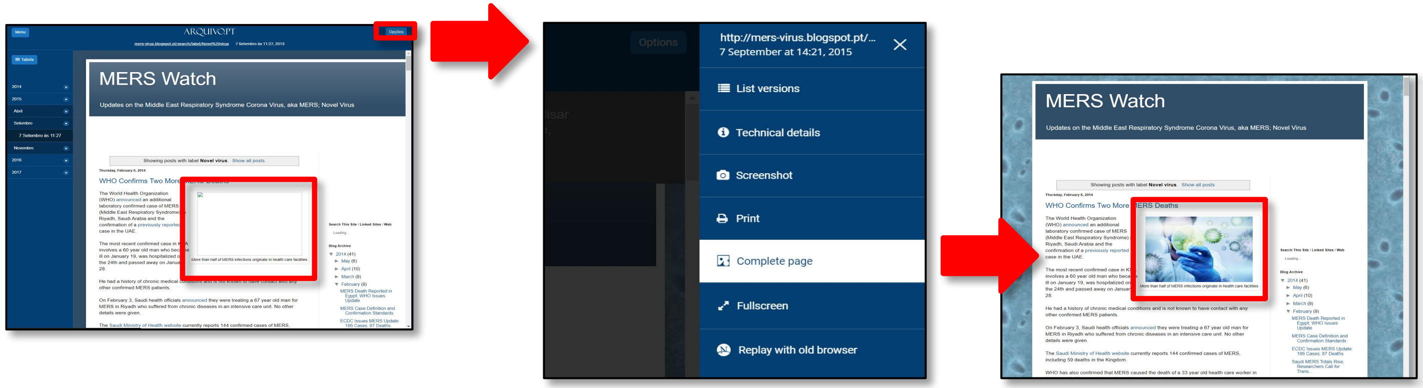
14 micro-services hosted on:

- 8 systems
- 35 components
- 88 servers

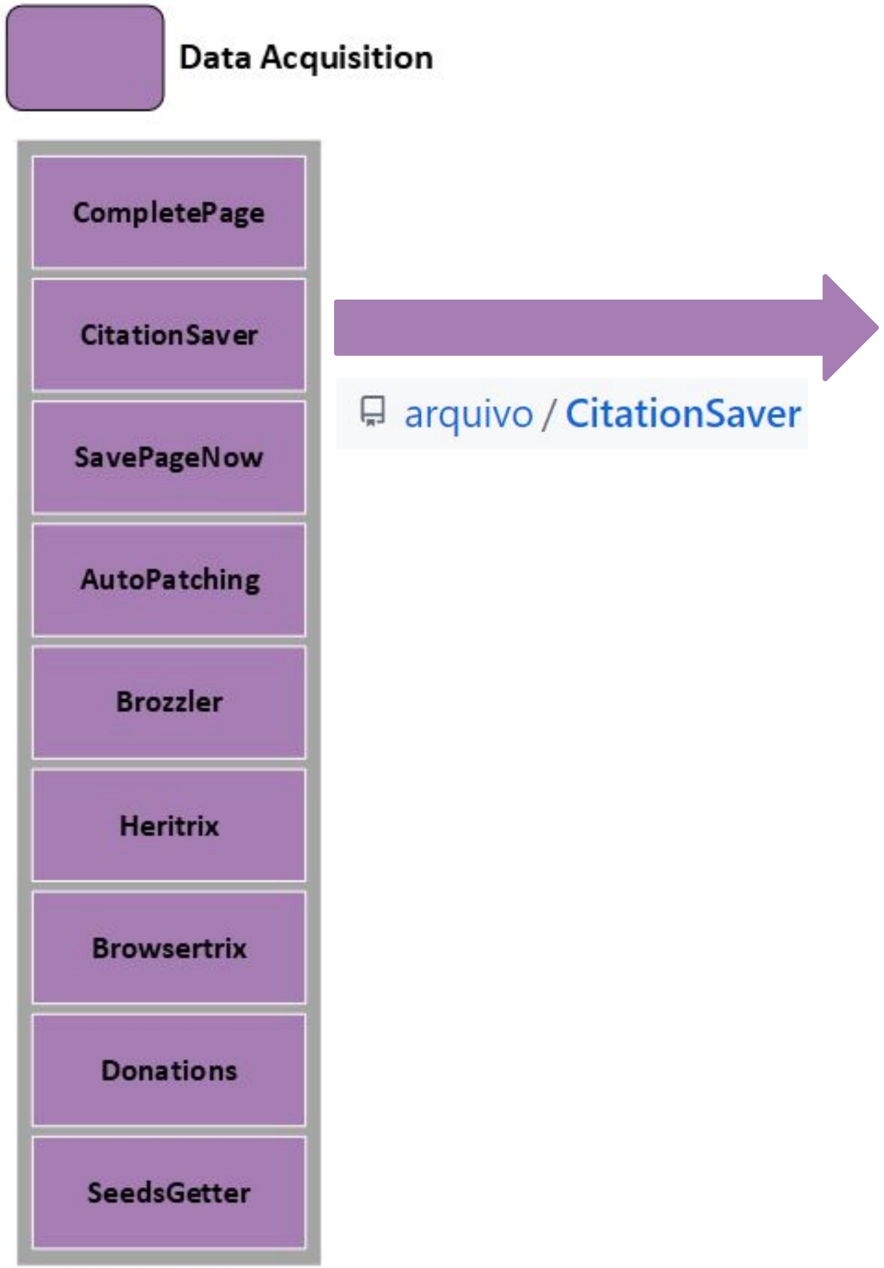
Data Acquisition

- CompletePage
- CitationSaver
- SavePageNow
- AutoPatching
- Brozzler
- Heritrix
- Browsertrix
- Donations
- SeedsGetter

pywb


ARQUIVO.PT
Options
MERS Watch
Updates on the Middle East Respiratory Syndrome Corona Virus, aka MERS; Novel Virus
WHO Confirms Two More MERS Deaths
http://mers-virus.blogspot.pt/...
7 September at 14:21, 2015
List versions
Technical details
Screenshot
Print
Complete page
Fullscreen
Replay with old browser

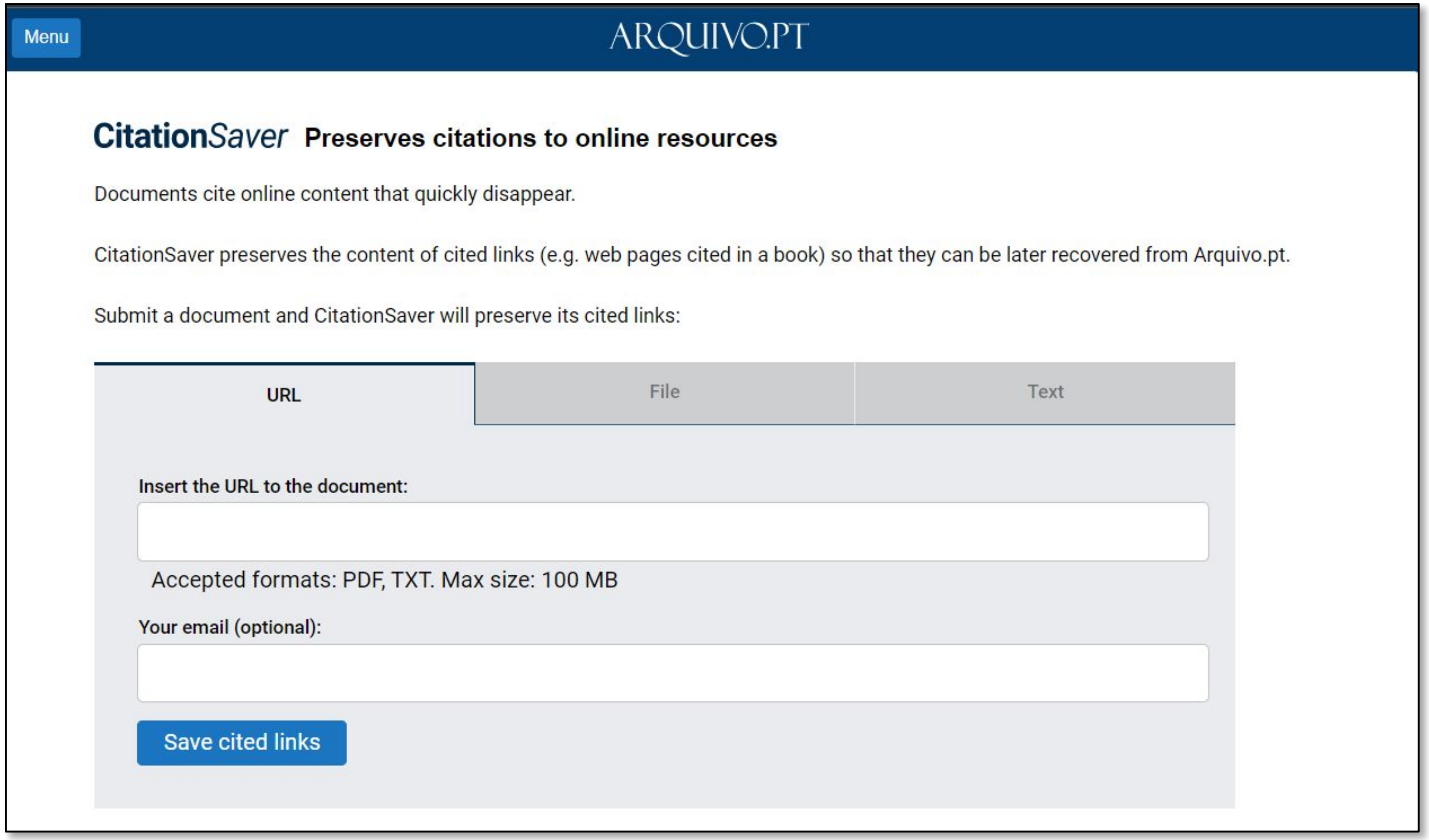
Missing files obtained from the live-web/archives are integrated to improve the quality of the archived page



Data Acquisition

- CompletePage
- CitationSaver
- SavePageNow
- AutoPatching
- Brozzler
- Heritrix
- Browsertrix
- Donations
- SeedsGetter

arquivo / CitationSaver



CitationSaver **Preserves citations to online resources**

Documents cite online content that quickly disappear.

CitationSaver preserves the content of cited links (e.g. web pages cited in a book) so that they can be later recovered from Arquivo.pt.

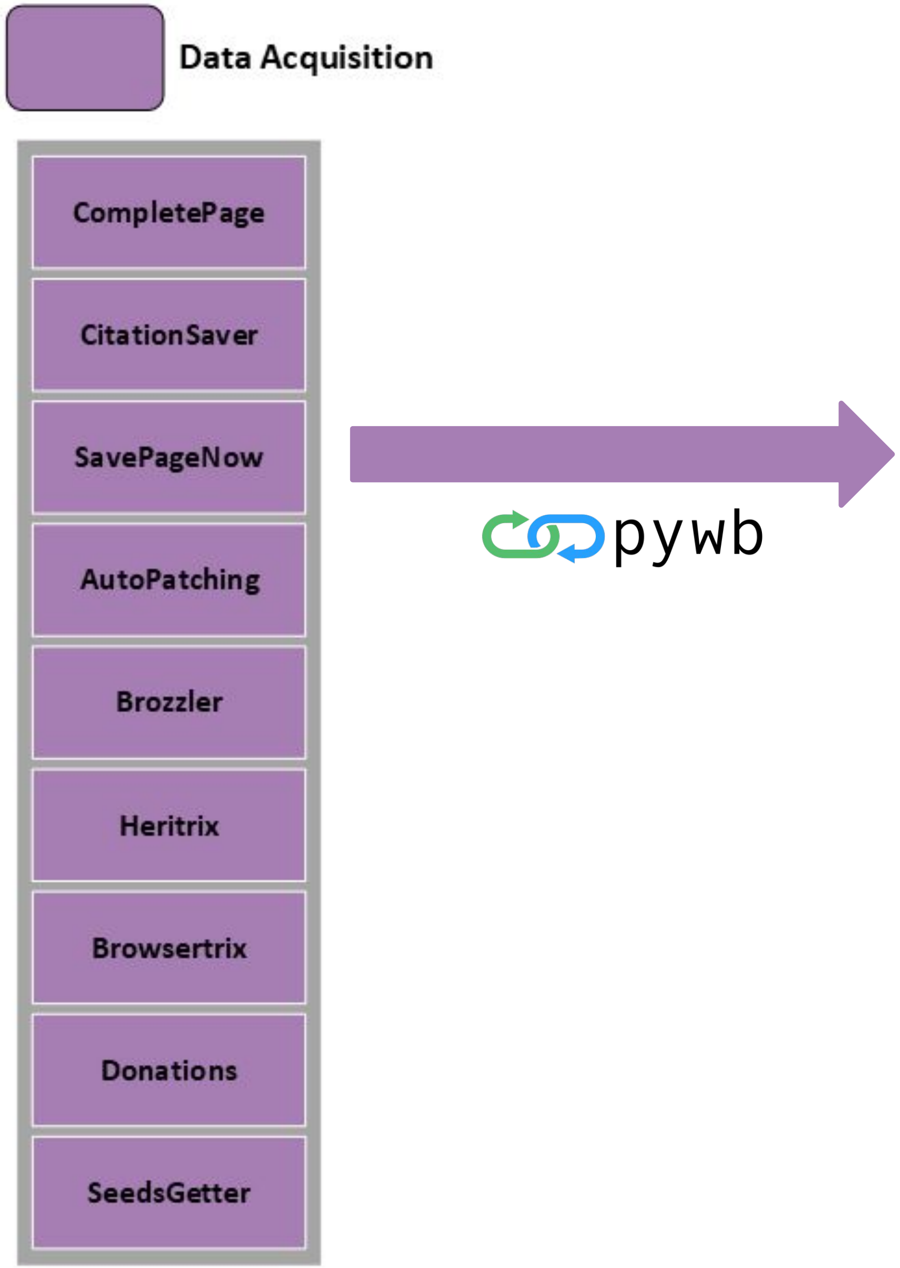
Submit a document and CitationSaver will preserve its cited links:

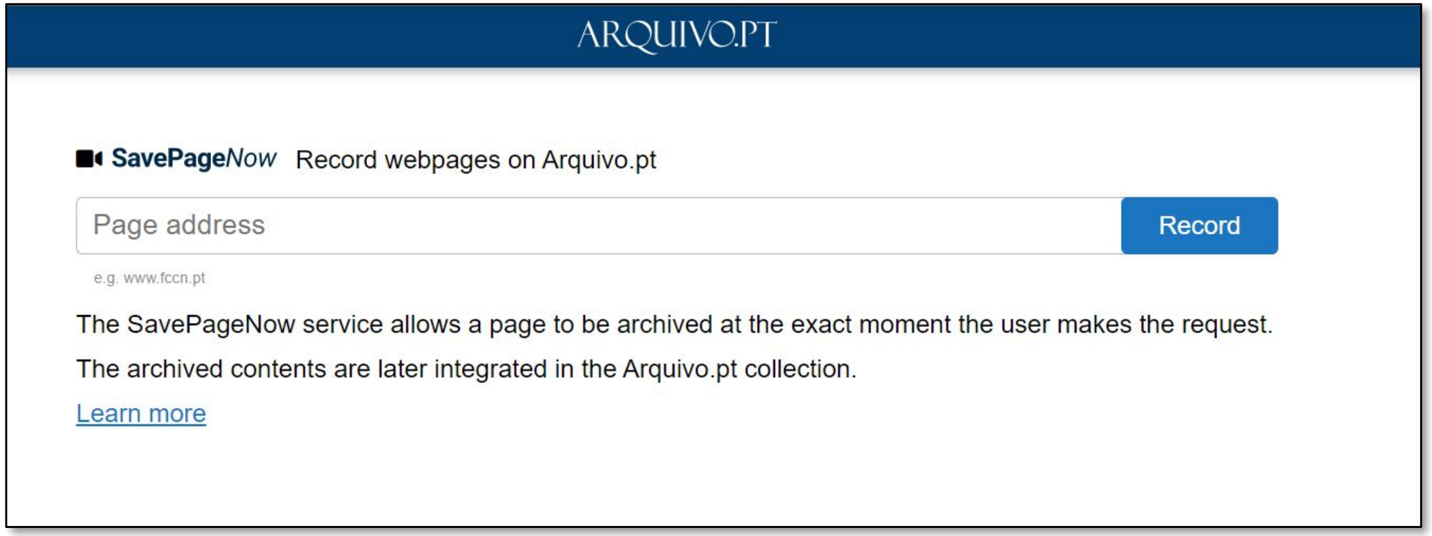
Insert the URL to the document:

Accepted formats: PDF, TXT. Max size: 100 MB

Your email (optional):

arquivo.pt/citationsaver





SavePage*Now* Record webpages on Arquivo.pt

Page address [Record]

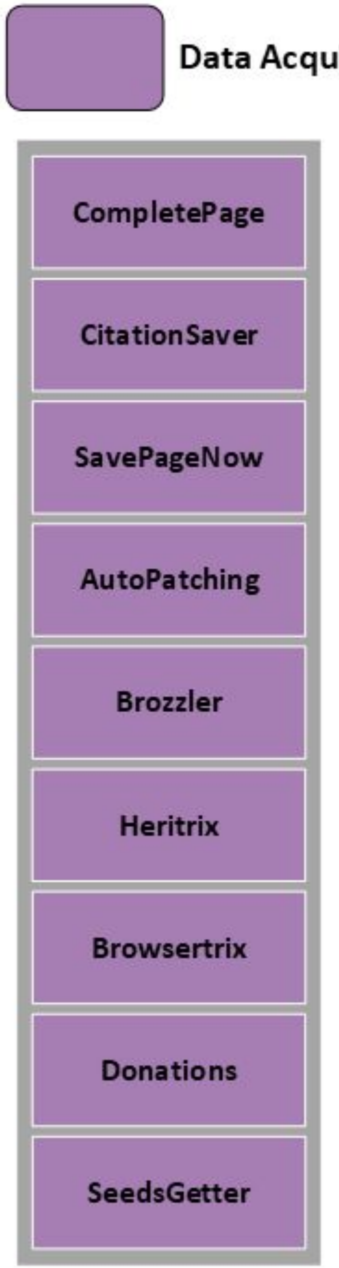
e.g. www.fccn.pt

The SavePageNow service allows a page to be archived at the exact moment the user makes the request.

The archived contents are later integrated in the Arquivo.pt collection.

Learn more

arquivo.pt/savepagenow



Data Acquisition

CompletePage

CitationSaver

SavePageNow

AutoPatching

Brozzler

Heritrix

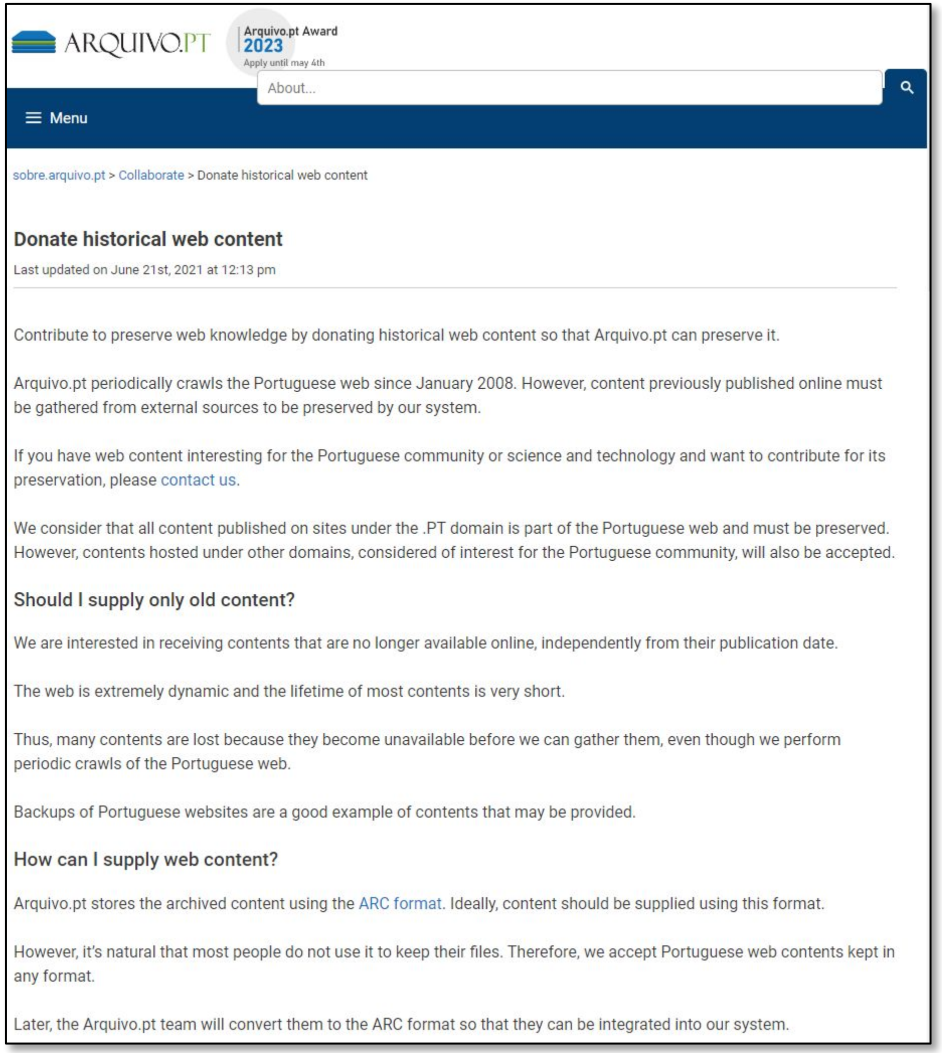
Browsertrix

Donations

SeedsGetter

ARQUIVO.PT

Libraries to convert files to WARC format



sobre.arquivo.pt > Collaborate > Donate historical web content

Donate historical web content

Last updated on June 21st, 2021 at 12:13 pm

Contribute to preserve web knowledge by donating historical web content so that Arquivo.pt can preserve it.

Arquivo.pt periodically crawls the Portuguese web since January 2008. However, content previously published online must be gathered from external sources to be preserved by our system.

If you have web content interesting for the Portuguese community or science and technology and want to contribute for its preservation, please contact us.

We consider that all content published on sites under the .PT domain is part of the Portuguese web and must be preserved. However, contents hosted under other domains, considered of interest for the Portuguese community, will also be accepted.

Should I supply only old content?

We are interested in receiving contents that are no longer available online, independently from their publication date.

The web is extremely dynamic and the lifetime of most contents is very short.

Thus, many contents are lost because they become unavailable before we can gather them, even though we perform periodic crawls of the Portuguese web.

Backups of Portuguese websites are a good example of contents that may be provided.

How can I supply web content?

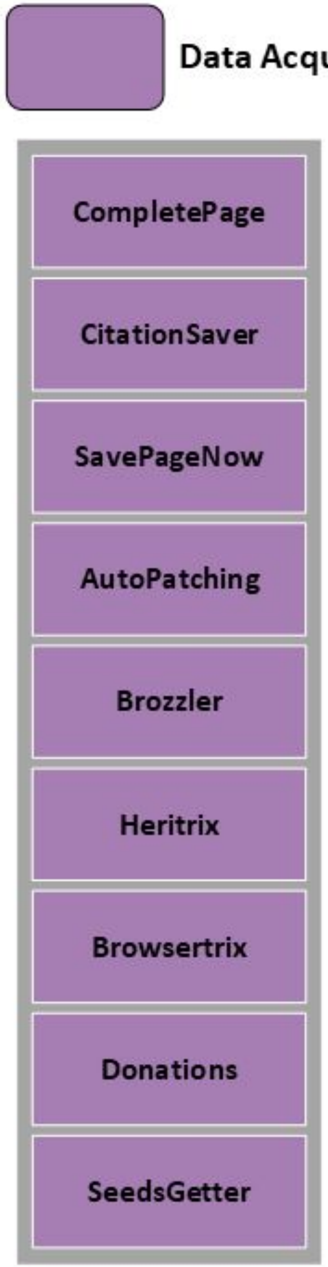
Arquivo.pt stores the archived content using the ARC format. Ideally, content should be supplied using this format.

However, it's natural that most people do not use it to keep their files. Therefore, we accept Portuguese web contents kept in any format.

Later, the Arquivo.pt team will convert them to the ARC format so that they can be integrated into our system.

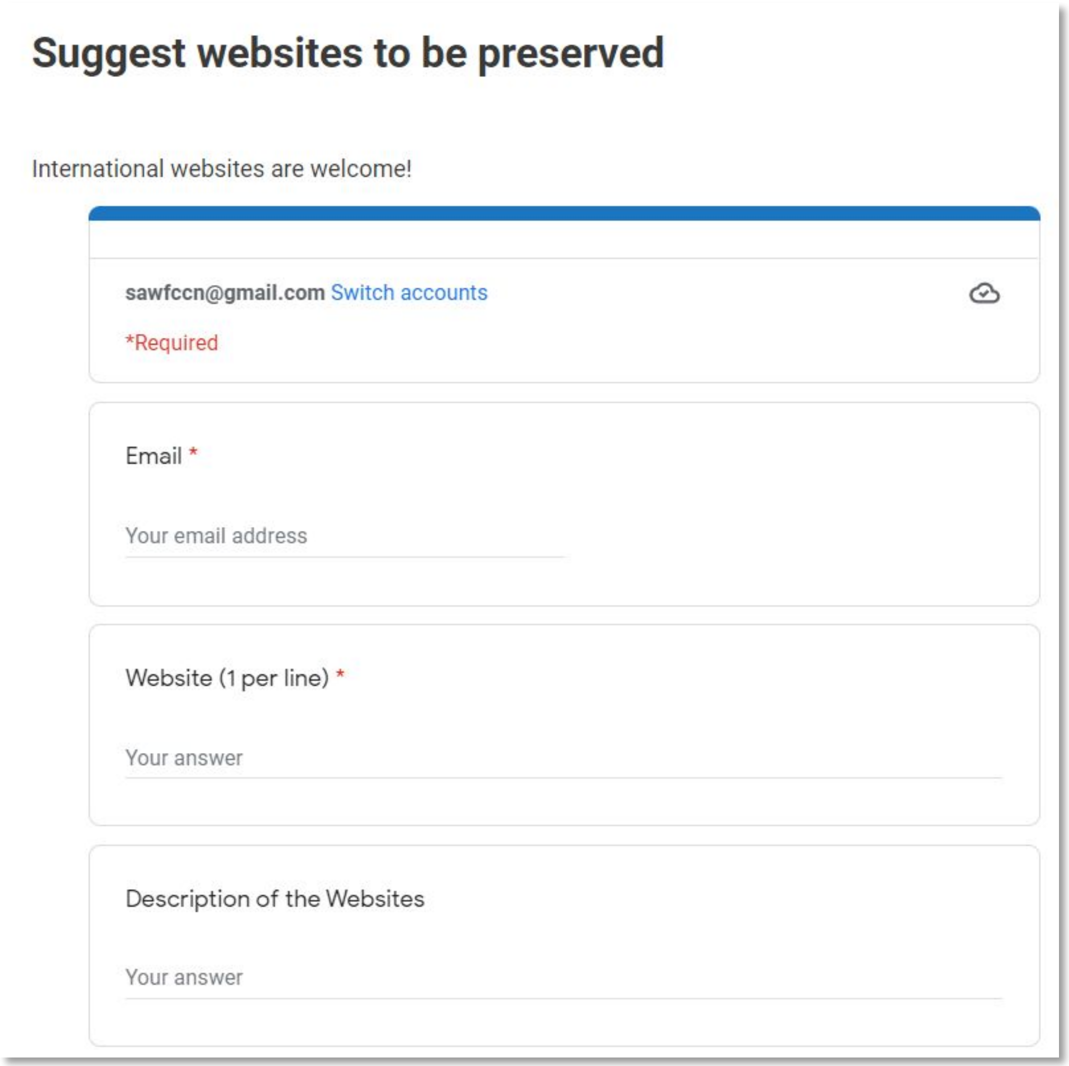
arquivo.pt/donate

arquivo.pt/suggest



ARQUIVO.PT

Queries to external APIs and crawl log analysis



Suggest websites to be preserved

International websites are welcome!

sawfccn@gmail.com Switch accounts

*Required

Email *

Your email address

Website (1 per line) *

Your answer

Description of the Websites

Your answer

🟪 **Data Acquisition**

CompletePage	pywb
CitationSaver	ARQUIVO.PT arquivo / CitationSaver
SavePageNow	pywb
AutoPatching	pywb
Brozzler	internetarchive / brozzler
Heritrix	internetarchive / heritrix3
Browsertrix	Browsertrix
Donations	ARQUIVO.PT Libraries to convert files to WARC format
SeedsGetter	ARQUIVO.PT Queries to external APIs and crawl log analysis

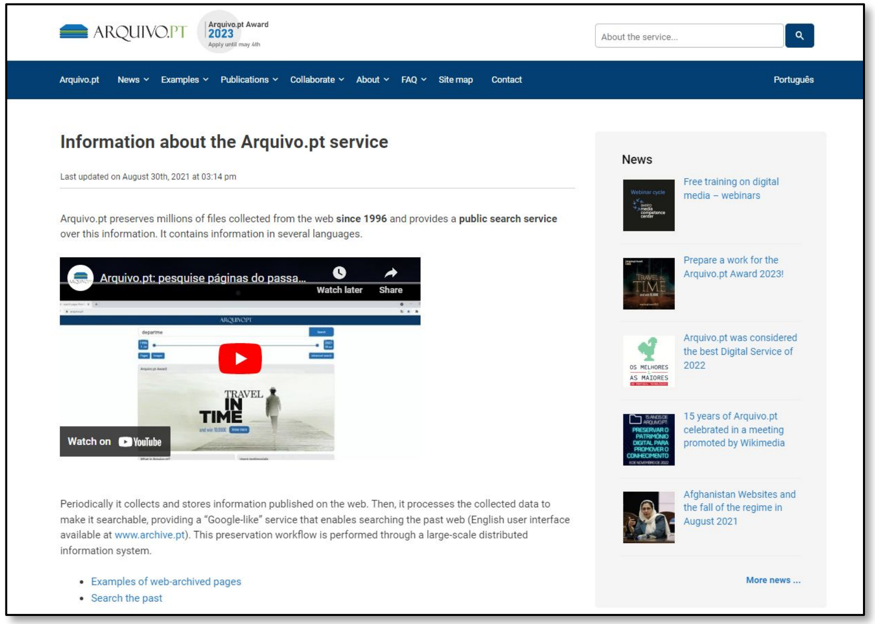
Frontend

Apache HTTPD

Webapp

Wordpress

Memorial


ARQUIVO.PT Arquivo.pt Award 2023 Apply until may 4th

About the service...

Arquivo.pt · News · Examples · Publications · Collaborate · About · FAQ · Site map · Contact · Português

Information about the Arquivo.pt service

Last updated on August 30th, 2021 at 03:14 pm

Arquivo.pt preserves millions of files collected from the web **since 1996** and provides a **public search service** over this information. It contains information in several languages.

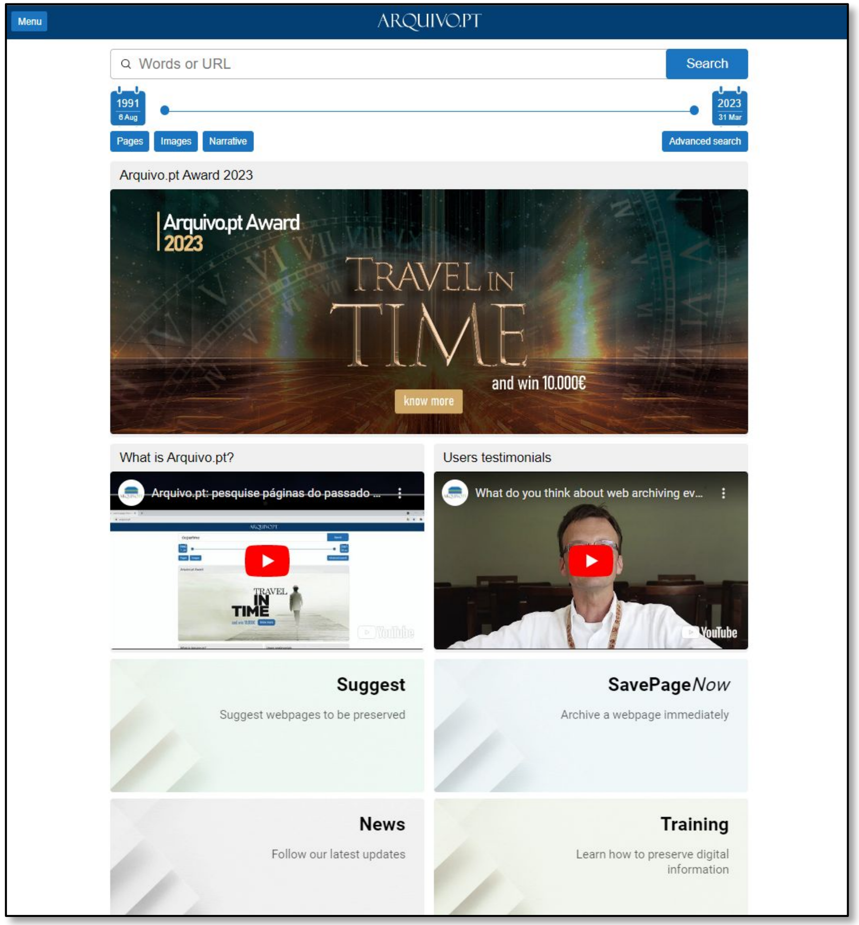
Periodically it collects and stores information published on the web. Then, it processes the collected data to make it searchable, providing a "Google-like" service that enables searching the past web (English user interface available at www.archive.pt). This preservation workflow is performed through a large-scale distributed information system.

- Examples of web-archived pages
- Search the past

News

- Free training on digital media – webinars
- Prepare a work for the Arquivo.pt Award 2023!
- Arquivo.pt was considered the best Digital Service of 2022
- 15 years of Arquivo.pt celebrated in a meeting promoted by Wikimedia
- Afghanistan Websites and the fall of the regime in August 2021

More news ...


Menu ARQUIVO.PT

Words or URL · Search

1991 8 Aug — 2023 31 Mar

Pages · Images · Narrative · Advanced search

Arquivo.pt Award 2023

Arquivo.pt Award 2023
TRAVEL IN TIME
and win 10.000€
know more

What is Arquivo.pt?

Arquivo.pt: pesquise páginas do passado ...

Users testimonials

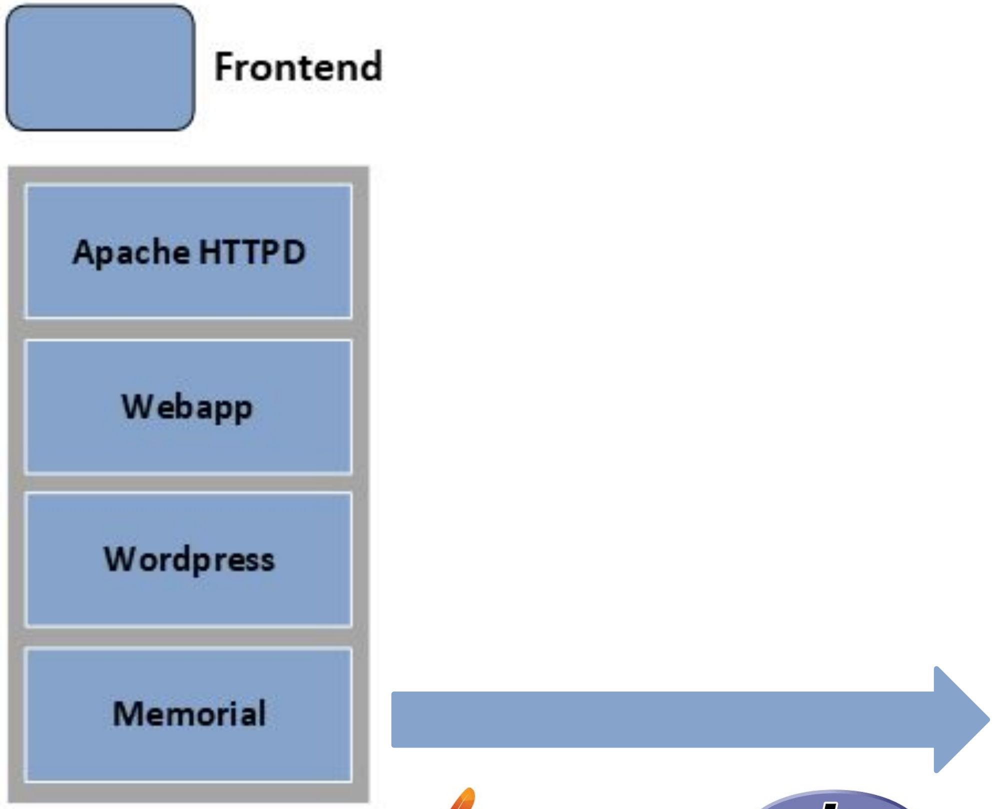
What do you think about web archiving ev...

Suggest
Suggest webpages to be preserved

SavePage*Now*
Archive a webpage immediately

News
Follow our latest updates

Training
Learn how to preserve digital information



Frontend

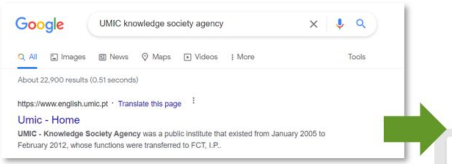
Apache HTTPD

Webapp

Wordpress

Memorial

Don’t kill your historical website!
Preserve it in the **Arquivo.pt Memorial**: arquivo.pt/memorial

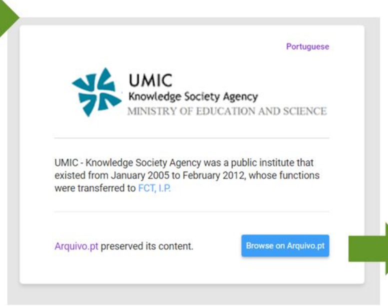


About 22,900 results (0.51 seconds)

https://www.english.umic.pt · Translate this page

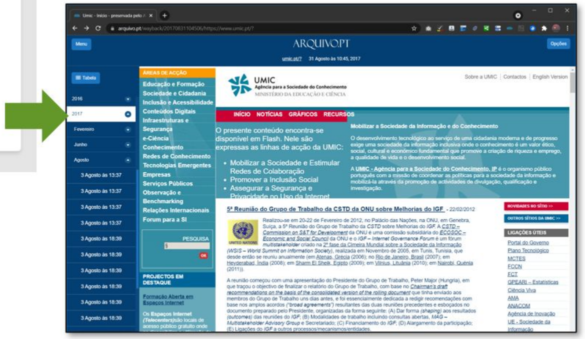
Umic - Home

UMIC - Knowledge Society Agency was a public institute that existed from January 2005 to February 2012, whose functions were transferred to FCT, I.P..



UMIC - Knowledge Society Agency was a public institute that existed from January 2005 to February 2012, whose functions were transferred to FCT, I.P.

Arquivo.pt preserved its content.



O presente conteúdo encontra-se disponível em Flash. Nele são expressas as linhas de acção da UMIC:

- Mobilizar a Sociedade e Estimular Redes de Colaboração
- Promover a Inclusão Social
- Assegurar a Segurança e Privacidade no Uso da Internet

Mobilizar a Sociedade da Informação e do Conhecimento

O desenvolvimento tecnológico ao serviço de uma cidadania moderna e de progresso exige uma sociedade da informação inclusiva onde o conhecimento é um valor ético, social, cultural e económico fundamental que promove a criação de riqueza e emprego, a qualidade de vida e o desenvolvimento social.

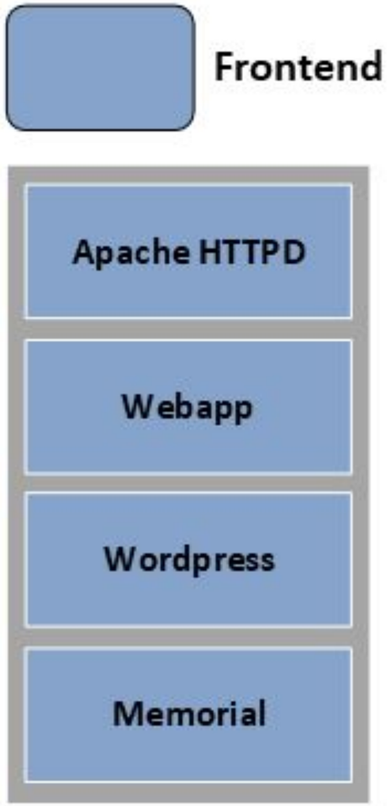
A **UMIC - Agência para a Sociedade do Conhecimento, IP** é o organismo público português com a missão de coordenar as políticas para a sociedade da informação e mobilizá-la através da promoção de actividades de divulgação, qualificação e investigação.

5ª Reunião do Grupo de Trabalho da CSTD da ONU sobre Melhorias do IGF - 22/02/2012

Realizou-se em 20-22 de Fevereiro de 2012, no Palácio das Nações, na ONU, em Genebra, Suíça, a 5ª Reunião do Grupo de Trabalho da *CSTD* sobre Melhorias do *IGF*. A *CSTD – Commission on S&T for Development* da ONU é uma comissão subsidiária do *ECOSOC – Economic and Social Council* da ONU e o *IGF – Internet Governance Forum* é um fórum *multistakeholder* criado na 2ª fase da Cimeira Mundial sobre a Sociedade da Informação (*WSIS – World Summit on Information Society*), realizada em Novembro de 2005, em Tunis, Tunísia, que desde então se reuniu anualmente (em Atenas, Grécia (2006); no Rio de Janeiro, Brasil (2007); em Heyderabad, Índia (2008); em Sharm El Sheik, Egipto (2009); em Vilnius, Lituânia (2010); em Nairobi, Quénia (2011)).

A reunião começou com uma apresentação do Presidente do Grupo de Trabalho, Peter Major (Hungria), em que traçou o objectivo de finalizar o relatório do Grupo de Trabalho, com base no *Chairman's draft recommendations on the basis of the consolidated version of the rolling document* que tinha enviado aos membros do Grupo de Trabalho uns dias antes, e foi essencialmente dedicada a redigir recomendações com base nos amplos acordos (*"broad agreements"*) resultantes das duas reuniões precedentes e esboçados no documento preparado pelo Presidente, organizadas da forma seguinte: (A) Dar forma (*shaping*) aos resultados (*outcomes*) das reuniões do *IGF*; (B) Modalidades de trabalho incluindo consultas abertas, *MAG – Multistakeholder Advisory Group* e Secretariado; (C) Financiamento do *IGF*; (D) Alargamento da participação; (E) Ligações do *IGF* a outros processos/mecanismos/entidades.

Os **Espaços Internet** *(Telecenters)* são locais de acesso público gratuito onde



Frontend

Apache HTTPD

Webapp

Wordpress

Memorial

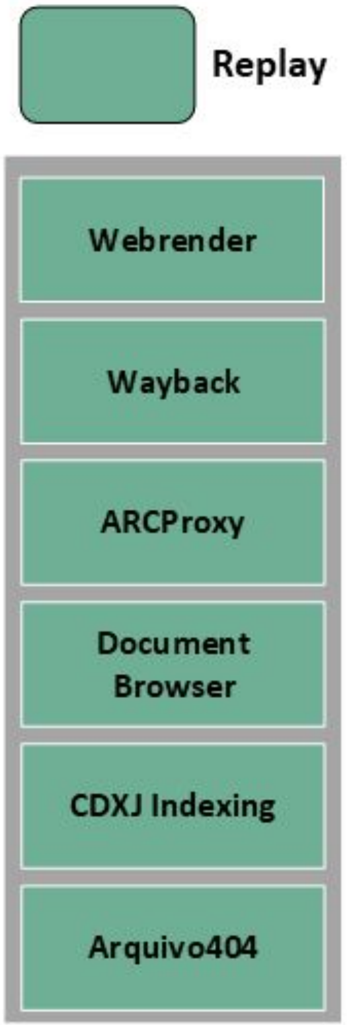
APACHE HTTP SERVER PROJECT

node JS

WordPress

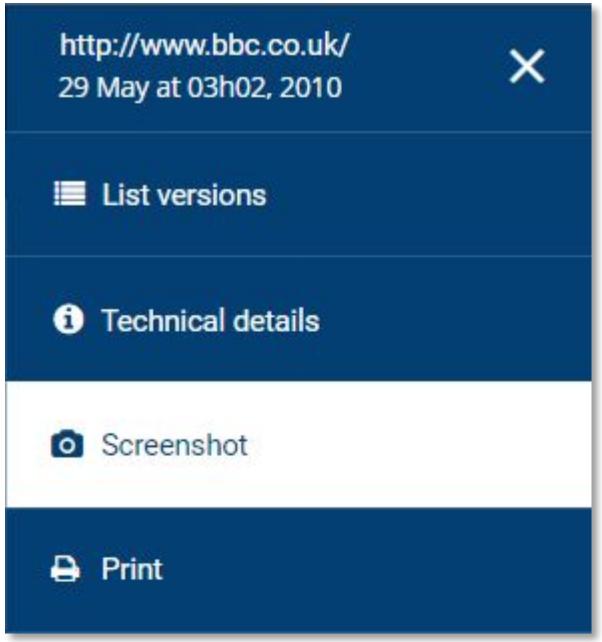
APACHE HTTP SERVER PROJECT

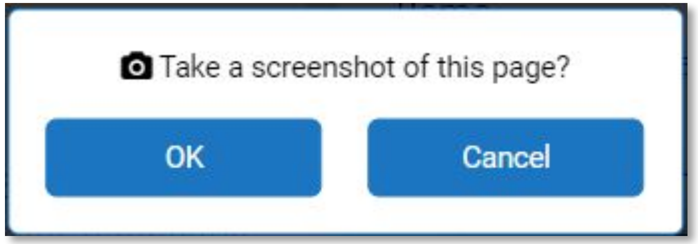
php

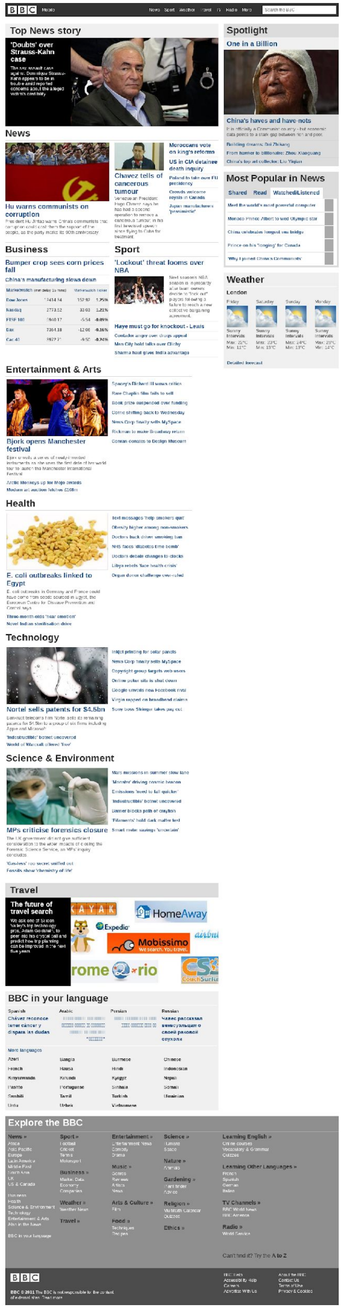


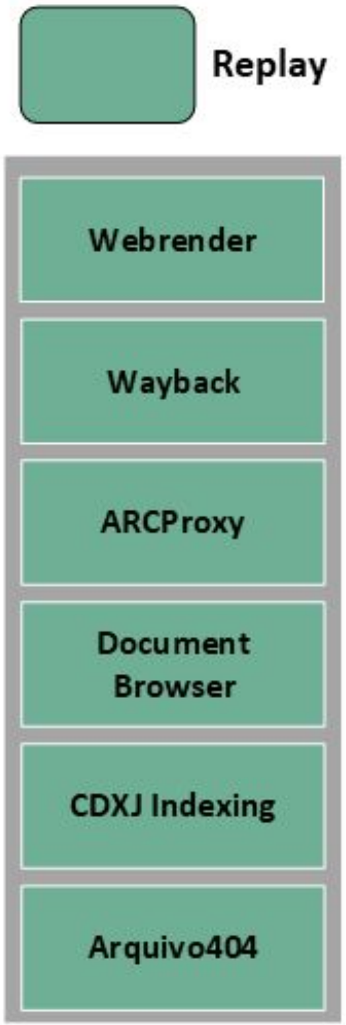
node JS

Puppeteer

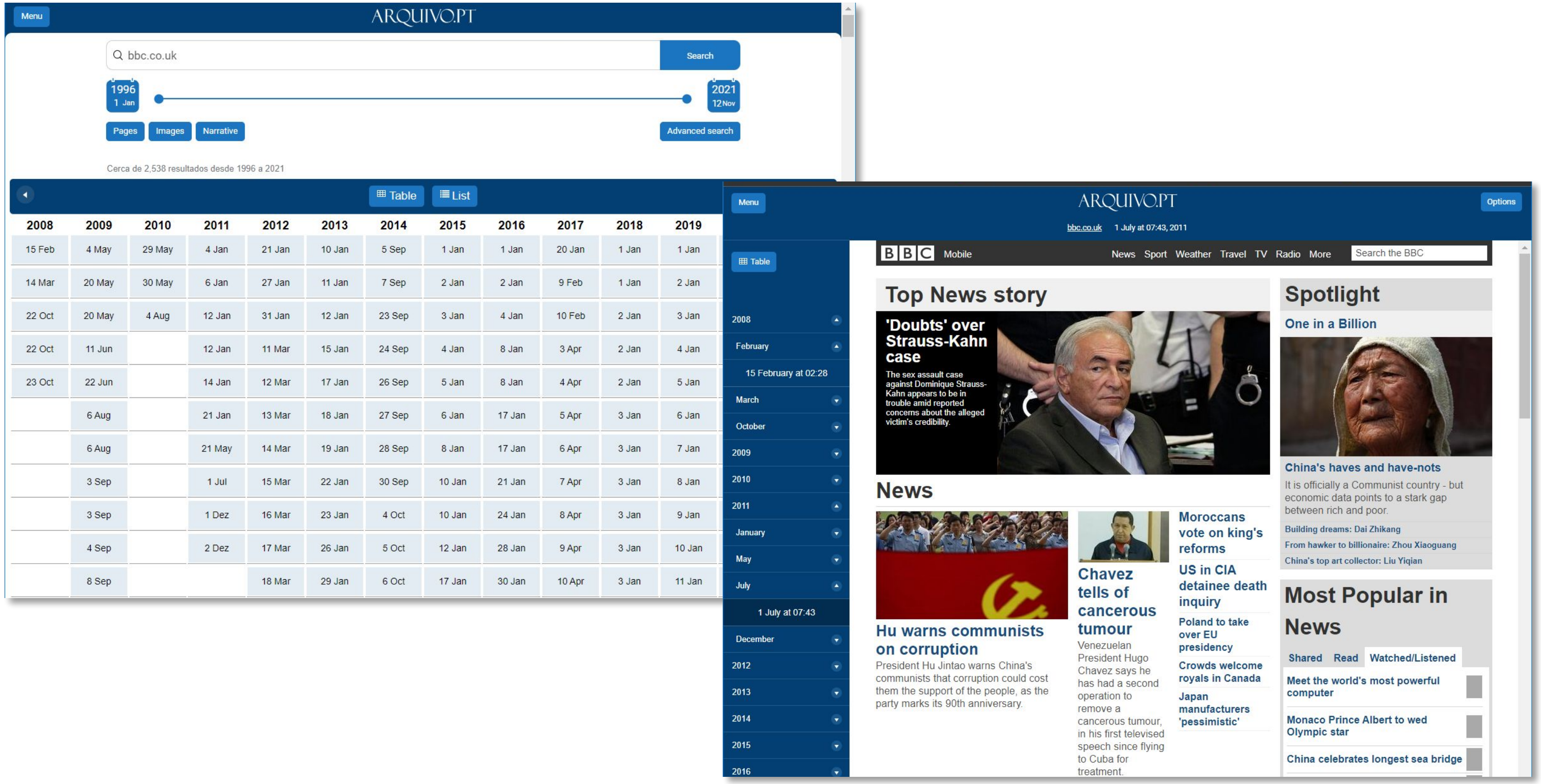

http://www.bbc.co.uk/
29 May at 03h02, 2010


Take a screenshot of this page?





pywb



Cerca de 2,538 resultados desde 1996 a 2021

2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
15 Feb	4 May	29 May	4 Jan	21 Jan	10 Jan	5 Sep	1 Jan	1 Jan	20 Jan	1 Jan	1 Jan
14 Mar	20 May	30 May	6 Jan	27 Jan	11 Jan	7 Sep	2 Jan	2 Jan	9 Feb	1 Jan	2 Jan
22 Oct	20 May	4 Aug	12 Jan	31 Jan	12 Jan	23 Sep	3 Jan	4 Jan	10 Feb	2 Jan	3 Jan
22 Oct	11 Jun		12 Jan	11 Mar	15 Jan	24 Sep	4 Jan	8 Jan	3 Apr	2 Jan	4 Jan
23 Oct	22 Jun		14 Jan	12 Mar	17 Jan	26 Sep	5 Jan	8 Jan	4 Apr	2 Jan	5 Jan
	6 Aug		21 Jan	13 Mar	18 Jan	27 Sep	6 Jan	17 Jan	5 Apr	3 Jan	6 Jan
	6 Aug		21 May	14 Mar	19 Jan	28 Sep	8 Jan	17 Jan	6 Apr	3 Jan	7 Jan
	3 Sep		1 Jul	15 Mar	22 Jan	30 Sep	10 Jan	21 Jan	7 Apr	3 Jan	8 Jan
	3 Sep		1 Dez	16 Mar	23 Jan	4 Oct	10 Jan	24 Jan	8 Apr	3 Jan	9 Jan
	4 Sep		2 Dez	17 Mar	26 Jan	5 Oct	12 Jan	28 Jan	9 Apr	3 Jan	10 Jan
	8 Sep			18 Mar	29 Jan	6 Oct	17 Jan	30 Jan	10 Apr	3 Jan	11 Jan

Top News story

'Doubts' over Strauss-Kahn case

The sex assault case against Dominique Strauss-Kahn appears to be in trouble amid reported concerns about the alleged victim's credibility.

News

Hu warns communists on corruption

President Hu Jintao warns China's communists that corruption could cost them the support of the people, as the party marks its 90th anniversary.

Chavez tells of cancerous tumour

Venezuelan President Hugo Chavez says he has had a second operation to remove a cancerous tumour, in his first televised speech since flying to Cuba for treatment.

Moroccans vote on king's reforms

US in CIA detainee death inquiry

Poland to take over EU presidency

Crowds welcome royals in Canada

Japan manufacturers 'pessimistic'

Spotlight

One in a Billion

China's haves and have-nots

It is officially a Communist country - but economic data points to a stark gap between rich and poor.

Building dreams: Dai Zhikang

From hawker to billionaire: Zhou Xiaoguang

China's top art collector: Liu Yiqian

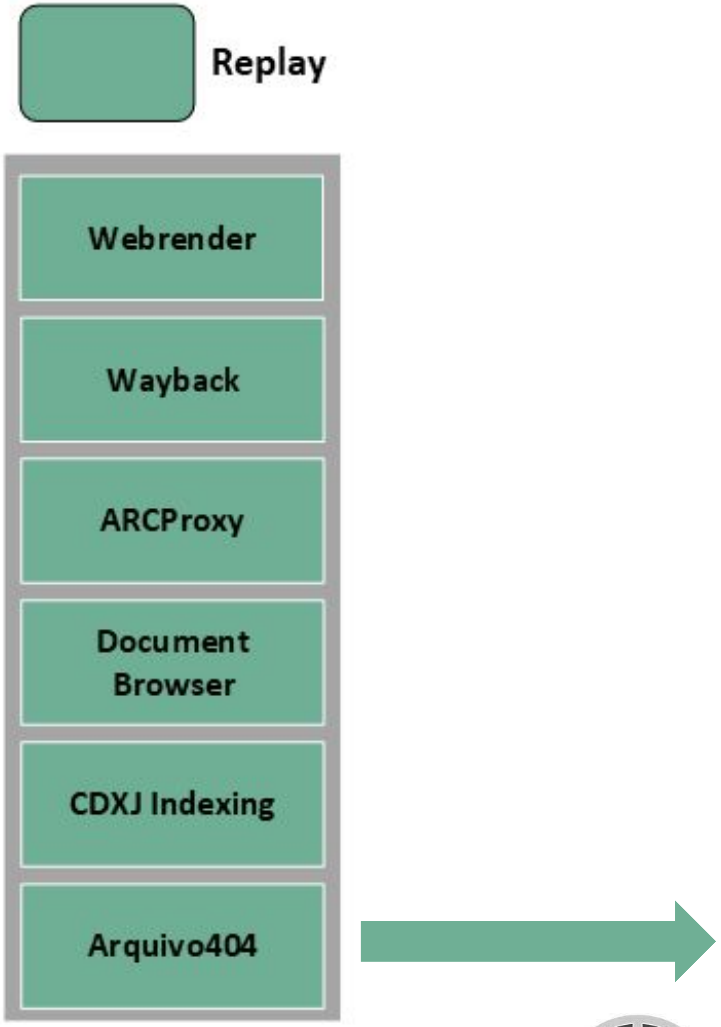
Most Popular in News

Shared Read Watched/Listened

Meet the world's most powerful computer

Monaco Prince Albert to wed Olympic star

China celebrates longest sea bridge

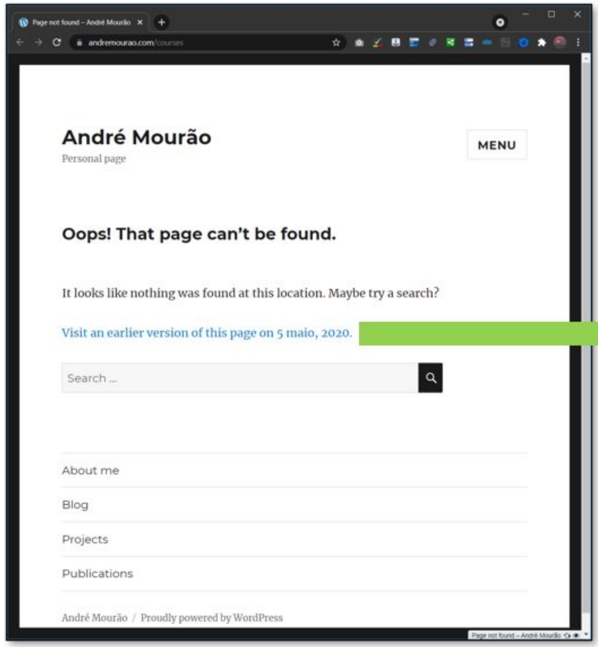


JavaScript

memento

arquivo.pt/arquivo404 is powered by Memento API

presents web-archived pages instead of “pages not found”



Page not found – André Mourão

andremourao.com/courses

André Mourão

Personal page

MENU

Oops! That page can’t be found.

It looks like nothing was found at this location. Maybe try a search?

Visit an earlier version of this page on 5 maio, 2020.

Search ...

About me

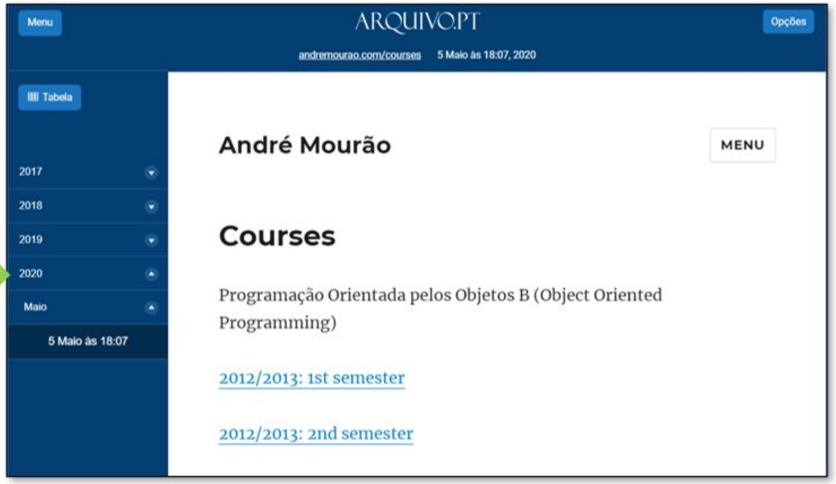
Blog

Projects

Publications

André Mourão / Proudly powered by WordPress

Page not found at live website



Menu ARQUIVO.PT Opções

andremourao.com/courses 5 Maio às 18:07, 2020

Tabela

2017

2018

2019

2020

Maio

5 Maio às 18:07

André Mourão

MENU

Courses

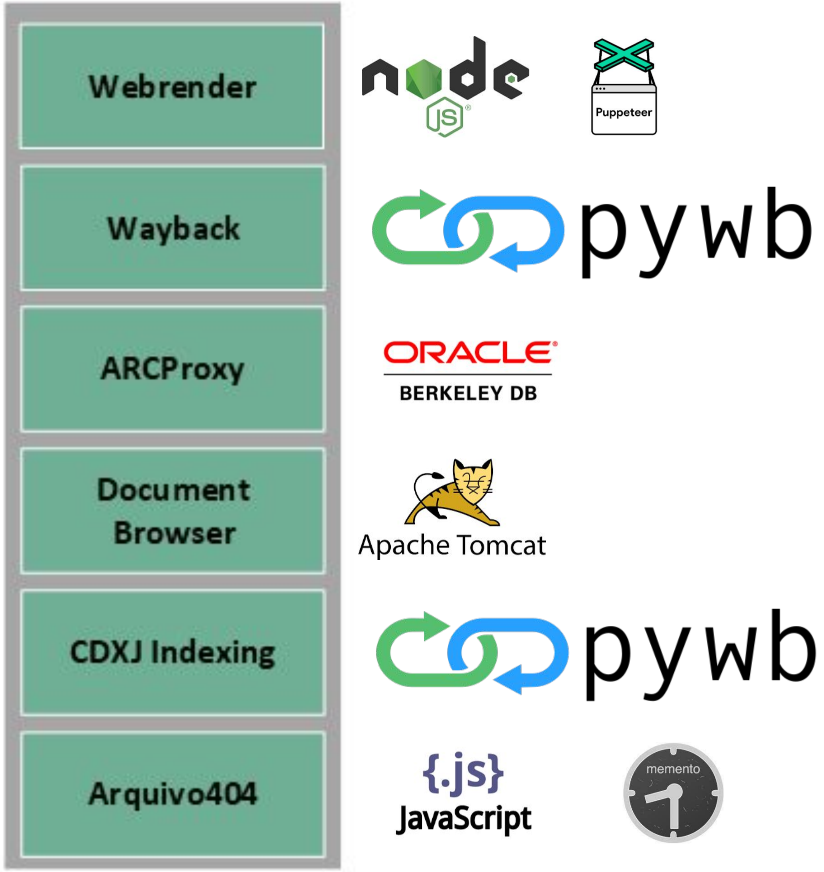
Programação Orientada pelos Objetos B (Object Oriented Programming)

2012/2013: 1st semester

2012/2013: 2nd semester

Page available at Arquivo.pt

🟩 Replay



Webrender

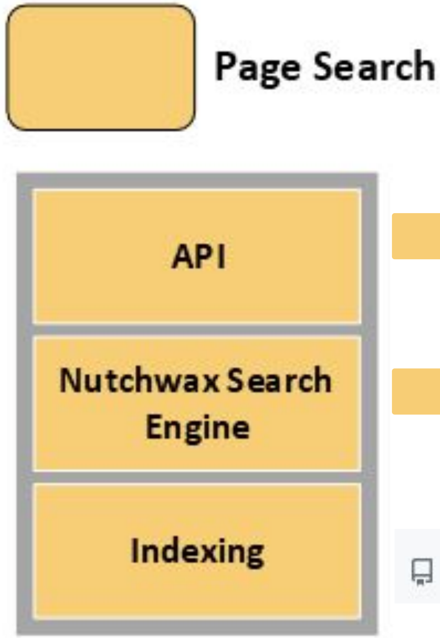
Wayback

ARCProxy

Document Browser

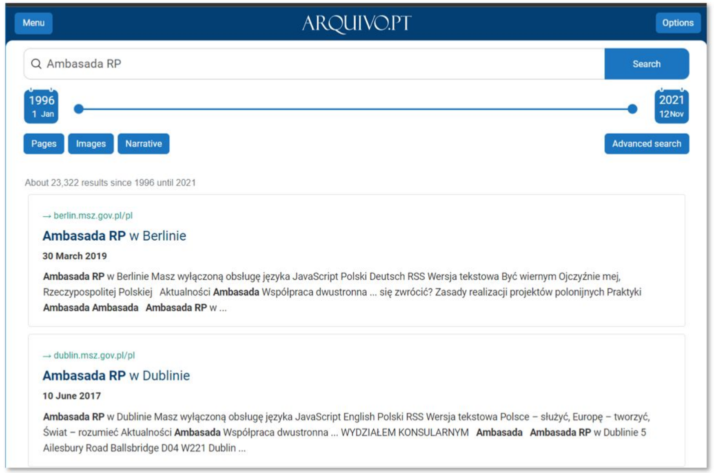
CDXJ Indexing

Arquivo404



arquivo / pwa-technologies

Search **texts** from the past in any language



About 23,322 results since 1996 until 2021

→ berlin.msz.gov.pl/pl

Ambasada RP w Berlinie

30 March 2019

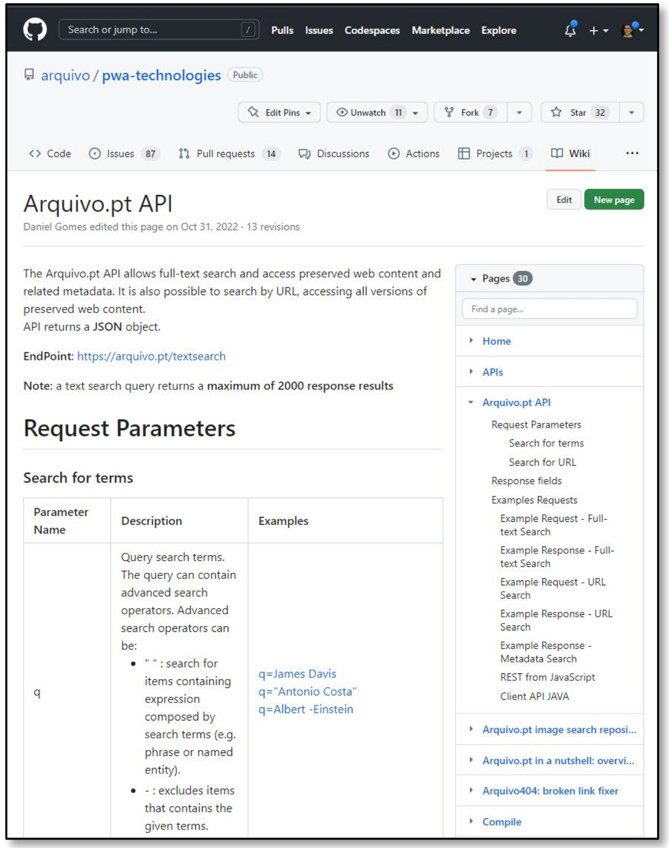
Ambasada RP w Berlinie Masz wyłączoną obsługę języka JavaScript Polski Deutsch RSS Wersja tekstowa Być wiernym Ojczyźnie mej, Rzeczypospolitej Polskiej Aktualności **Ambasada** Współpraca dwustronna ... się zwrócić? Zasady realizacji projektów polonijnych Praktyki **Ambasada Ambasada Ambasada RP** w ...

→ dublin.msz.gov.pl/pl

Ambasada RP w Dublinie

10 June 2017

Ambasada RP w Dublinie Masz wyłączoną obsługę języka JavaScript English Polski RSS Wersja tekstowa Polsce – służyć, Europę – tworzyć, Świat – rozumieć Aktualności **Ambasada** Współpraca dwustronna ... WYDZIAŁEM KONSULARNYM **Ambasada Ambasada RP** w Dublinie 5 Ailesbury Road Ballsbridge D04 W221 Dublin ...



Arquivo.pt API

Daniel Gomes edited this page on Oct 31, 2022 · 13 revisions

The Arquivo.pt API allows full-text search and access preserved web content and related metadata. It is also possible to search by URL, accessing all versions of preserved web content.
API returns a **JSON** object.

EndPoint: https://arquivo.pt/textsearch

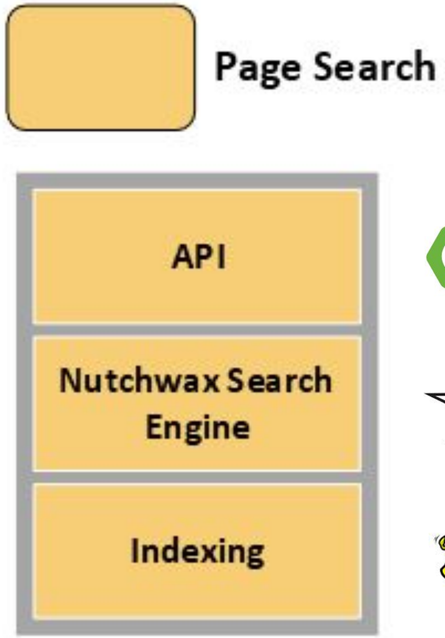
Note: a text search query returns a **maximum of 2000 response results**

Request Parameters

Search for terms

Parameter Name	Description	Examples

q	Query search terms. The query can contain advanced search operators. Advanced search operators can be:
	• " " : search for items containing expression composed by search terms (e.g. phrase or named entity).
	• - : excludes items that contains the given terms.
		q=James Davis
		q="Antonio Costa"
		q=Albert -Einstein



spring boot {JSON}

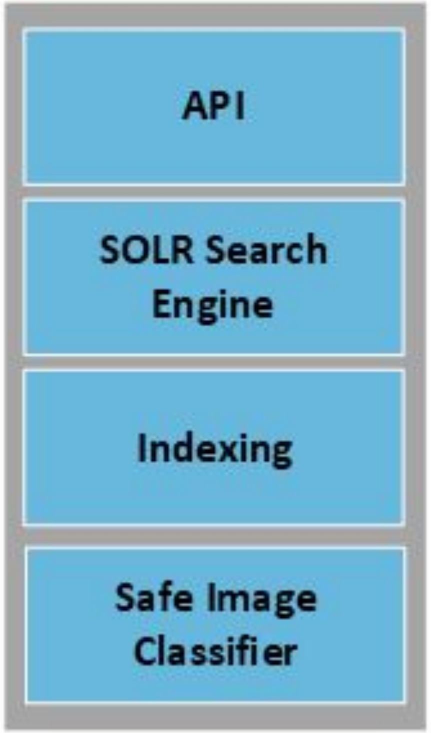
nutch

arquivo / pwa-technologies

hadoop

arquivo / pwa-technologies

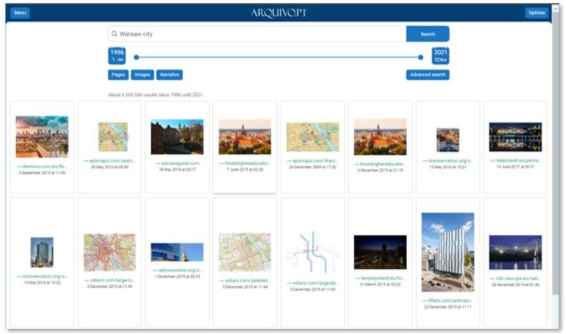
Image Search

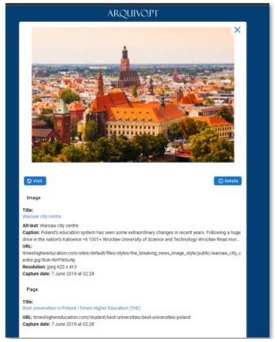


Apache Solr {JSON}

Apache Solr

Search **images** from the past




ARQUIVO.PT

Visit

Details

Image

Title: Warsaw city centre

Alt text: Warsaw city centre

Caption: Poland's education system has seen some extraordinary changes in recent years. Following a huge drive in the nation's Katowice =6 1001+ Wrocław University of Science and Technology Wrocław Read mor...

URL: timeshighereducation.com/sites/default/files/styles/the_breaking_news_image_style/public/warsaw_city_centre.jpg?itok=NYFWGvNj

Resolution: jpeg 620 x 413

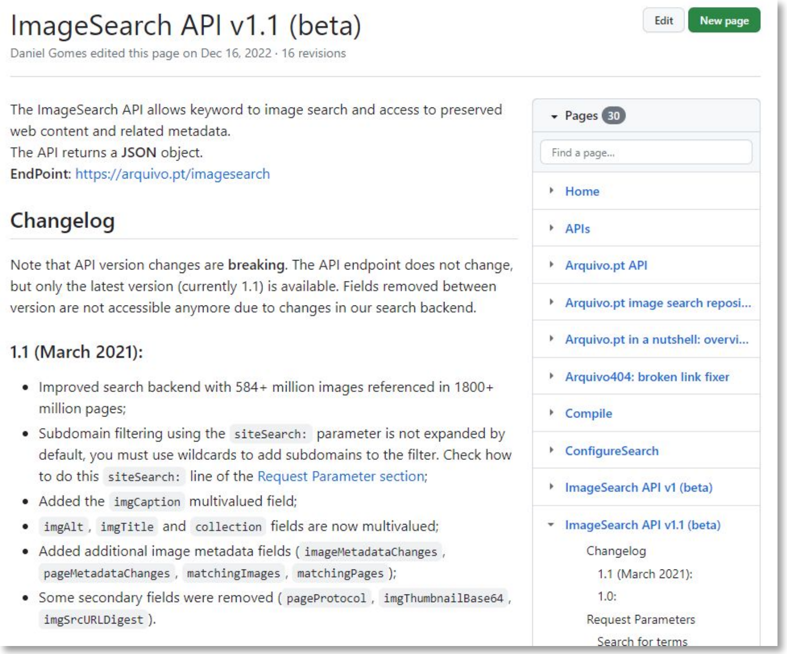
Capture date: 7 June 2019 at 02:28

Page

Title: Best universities in Poland | Times Higher Education (THE)

URL: timeshighereducation.com/student/best-universities/best-universities-poland

Capture date: 7 June 2019 at 02:28


ImageSearch API v1.1 (beta)

Daniel Gomes edited this page on Dec 16, 2022 · 16 revisions

Edit | New page

The ImageSearch API allows keyword to image search and access to preserved web content and related metadata.
The API returns a **JSON** object.
EndPoint: https://arquivo.pt/imagesearch

Changelog

Note that API version changes are **breaking**. The API endpoint does not change, but only the latest version (currently 1.1) is available. Fields removed between version are not accessible anymore due to changes in our search backend.

1.1 (March 2021):

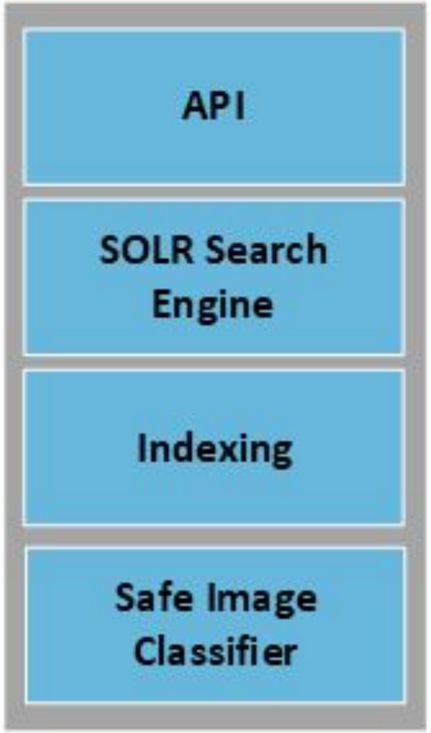
- Improved search backend with 584+ million images referenced in 1800+ million pages;
- Subdomain filtering using the `siteSearch:` parameter is not expanded by default, you must use wildcards to add subdomains to the filter. Check how to do this `siteSearch:` line of the Request Parameter section;
- Added the `imgCaption` multivalued field;
- `imgAlt`, `imgTitle` and `collection` fields are now multivalued;
- Added additional image metadata fields (`imageMetadataChanges`, `pageMetadataChanges`, `matchingImages`, `matchingPages`);
- Some secondary fields were removed (`pageProtocol`, `imgThumbnailBase64`, `imgSrcURLDigest`).

Pages 30

Find a page...

- Home
- APIs
- Arquivo.pt API
- Arquivo.pt image search reposi...
- Arquivo.pt in a nutshell: overvi...
- Arquivo404: broken link fixer
- Compile
- ConfigureSearch
- ImageSearch API v1 (beta)
- ImageSearch API v1.1 (beta)
  - Changelog
    - 1.1 (March 2021):
    - 1.0:
  - Request Parameters
    - Search for terms

Image Search



Apache Solr {JSON}

Apache Solr

hadoop

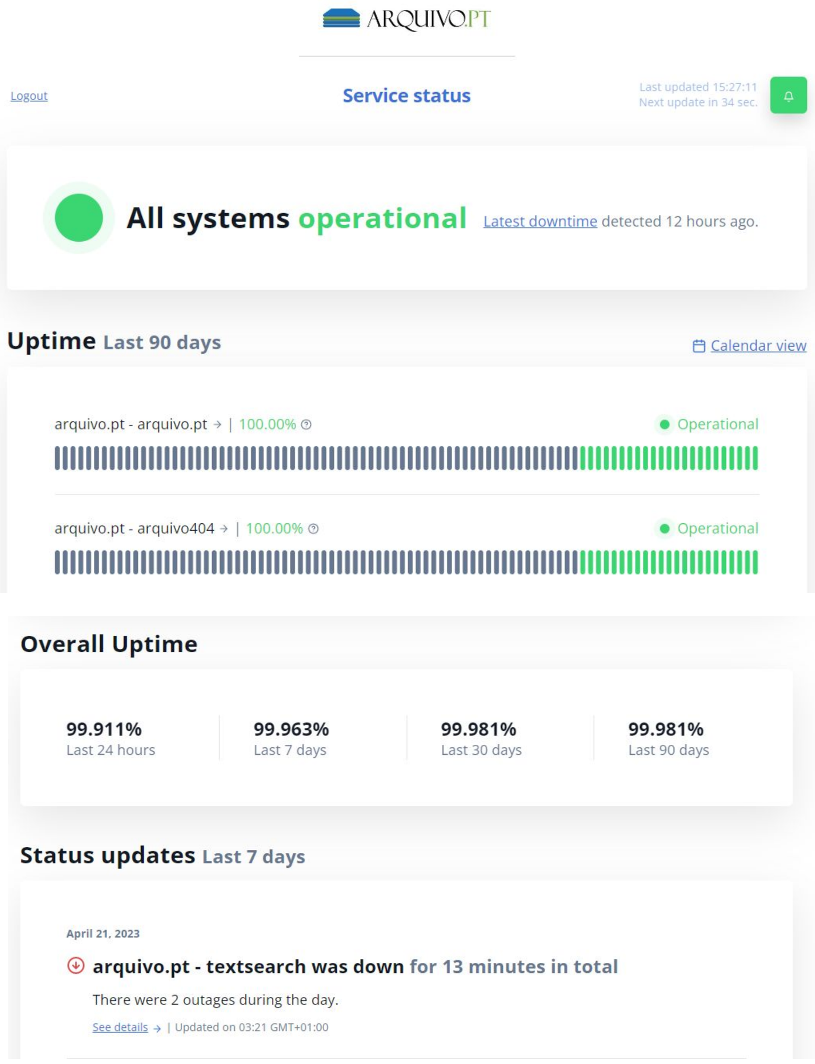
arquivo / image-gpu-classifier

Monitoring

Alarms
Statistics
Capacity

UptimeRobot

ICINGA


ARQUIVO.PT

Logout | **Service status** | Last updated 15:27:11 Next update in 34 sec.

All systems operational Latest downtime detected 12 hours ago.

Uptime Last 90 days — Calendar view

arquivo.pt - arquivo.pt → | 100.00% — ● Operational

arquivo.pt - arquivo404 → | 100.00% — ● Operational

Overall Uptime

99.911%	99.963%	99.981%	99.981%
Last 24 hours	Last 7 days	Last 30 days	Last 90 days

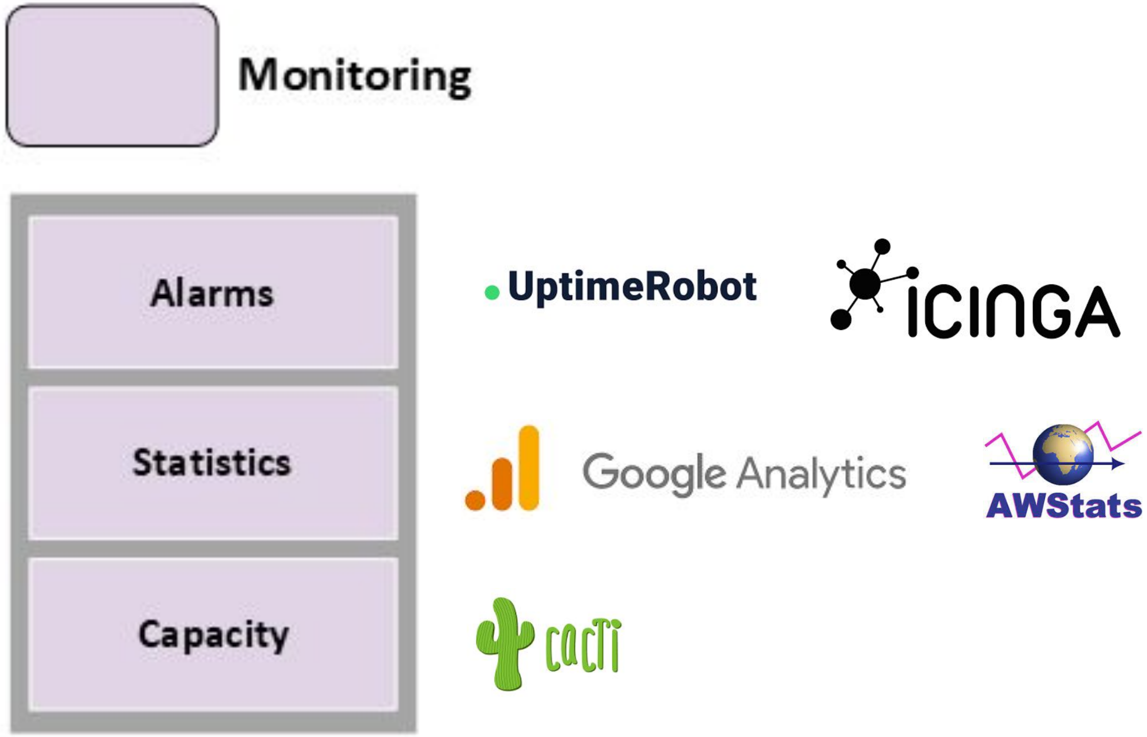
Status updates Last 7 days

April 21, 2023

arquivo.pt - textsearch was down for 13 minutes in total

There were 2 outages during the day.

See details → | Updated on 03:21 GMT+01:00



Monitoring

Alarms

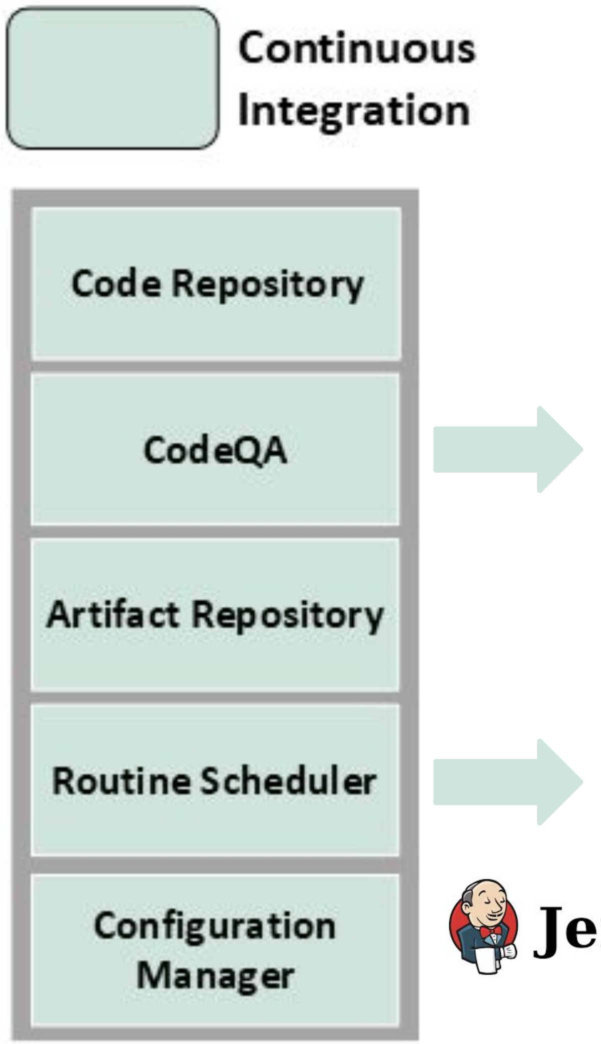
Statistics

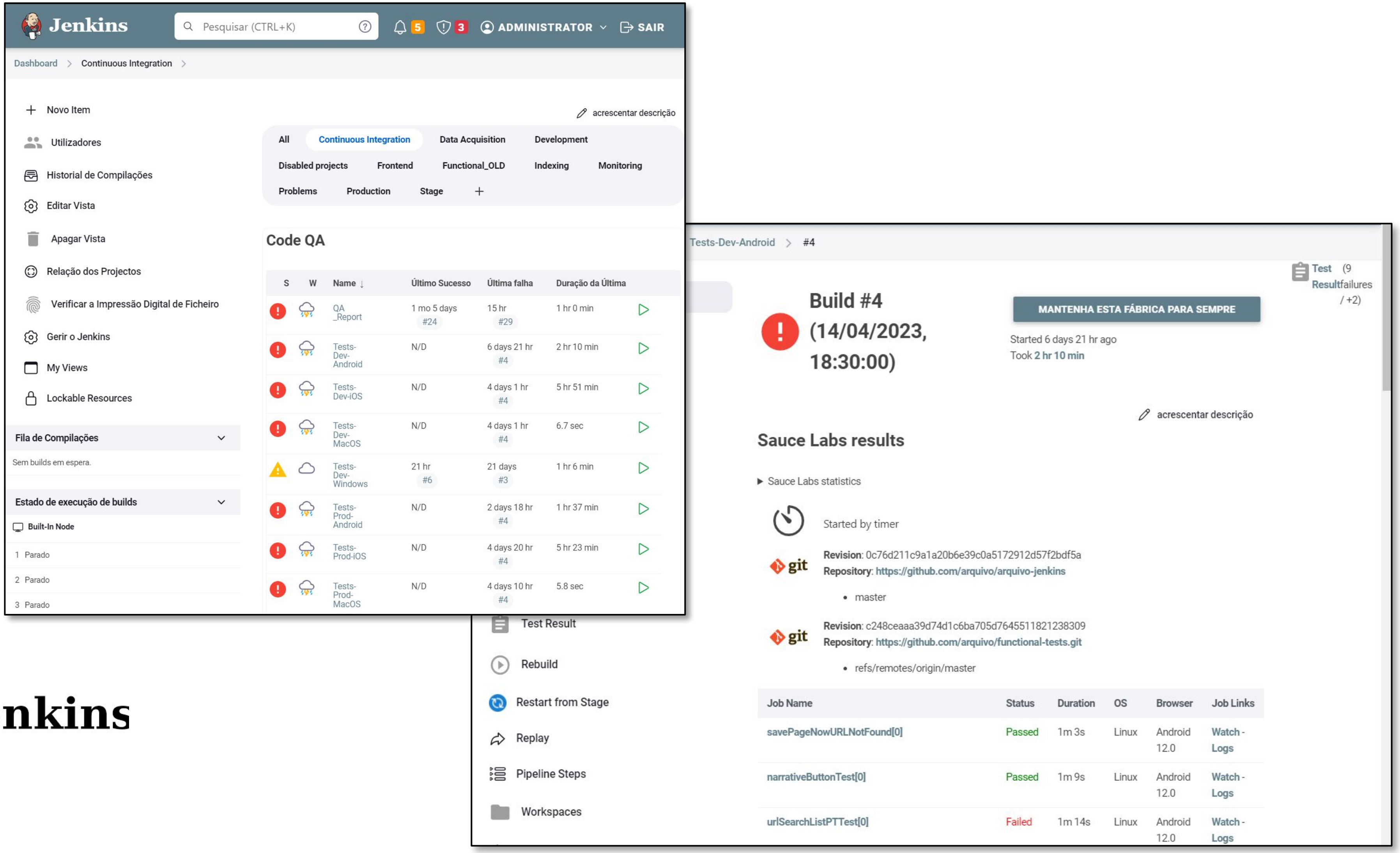
Capacity

494 internal monitors/alarms
140 external monitors/alarms

sonarcloud

SAUCELABS





Dashboard > Continuous Integration >

- Novo Item
- Utilizadores
- Historial de Compilações
- Editar Vista
- Apagar Vista
- Relação dos Projectos
- Verificar a Impressão Digital de Ficheiro
- Gerir o Jenkins
- My Views
- Lockable Resources

Fila de Compilações

Sem builds em espera.

Estado de execução de builds

Built-In Node

1 Parado

2 Parado

3 Parado

acrescentar descrição

All | Continuous Integration | Data Acquisition | Development | Disabled projects | Frontend | Functional_OLD | Indexing | Monitoring | Problems | Production | Stage | +

Code QA

S	W	Name ↓	Último Sucesso	Última falha	Duração da Última	
!		QA _Report	1 mo 5 days #24	15 hr #29	1 hr 0 min	▷
!		Tests-Dev-Android	N/D	6 days 21 hr #4	2 hr 10 min	▷
!		Tests-Dev-iOS	N/D	4 days 1 hr #4	5 hr 51 min	▷
!		Tests-Dev-MacOS	N/D	4 days 1 hr #4	6.7 sec	▷
⚠		Tests-Dev-Windows	21 hr #6	21 days #3	1 hr 6 min	▷
!		Tests-Prod-Android	N/D	2 days 18 hr #4	1 hr 37 min	▷
!		Tests-Prod-iOS	N/D	4 days 20 hr #4	5 hr 23 min	▷
!		Tests-Prod-MacOS	N/D	4 days 10 hr #4	5.8 sec	▷

Tests-Dev-Android > #4

Test Result (9 failures / +2)

Build #4 (14/04/2023, 18:30:00)

MANTENHA ESTA FÁBRICA PARA SEMPRE

Started 6 days 21 hr ago

Took 2 hr 10 min

acrescentar descrição

Sauce Labs results

▶ Sauce Labs statistics

Started by timer

Revision: 0c76d211c9a1a20b6e39c0a5172912d57f2bdf5a
Repository: https://github.com/arquivo/arquivo-jenkins

- master

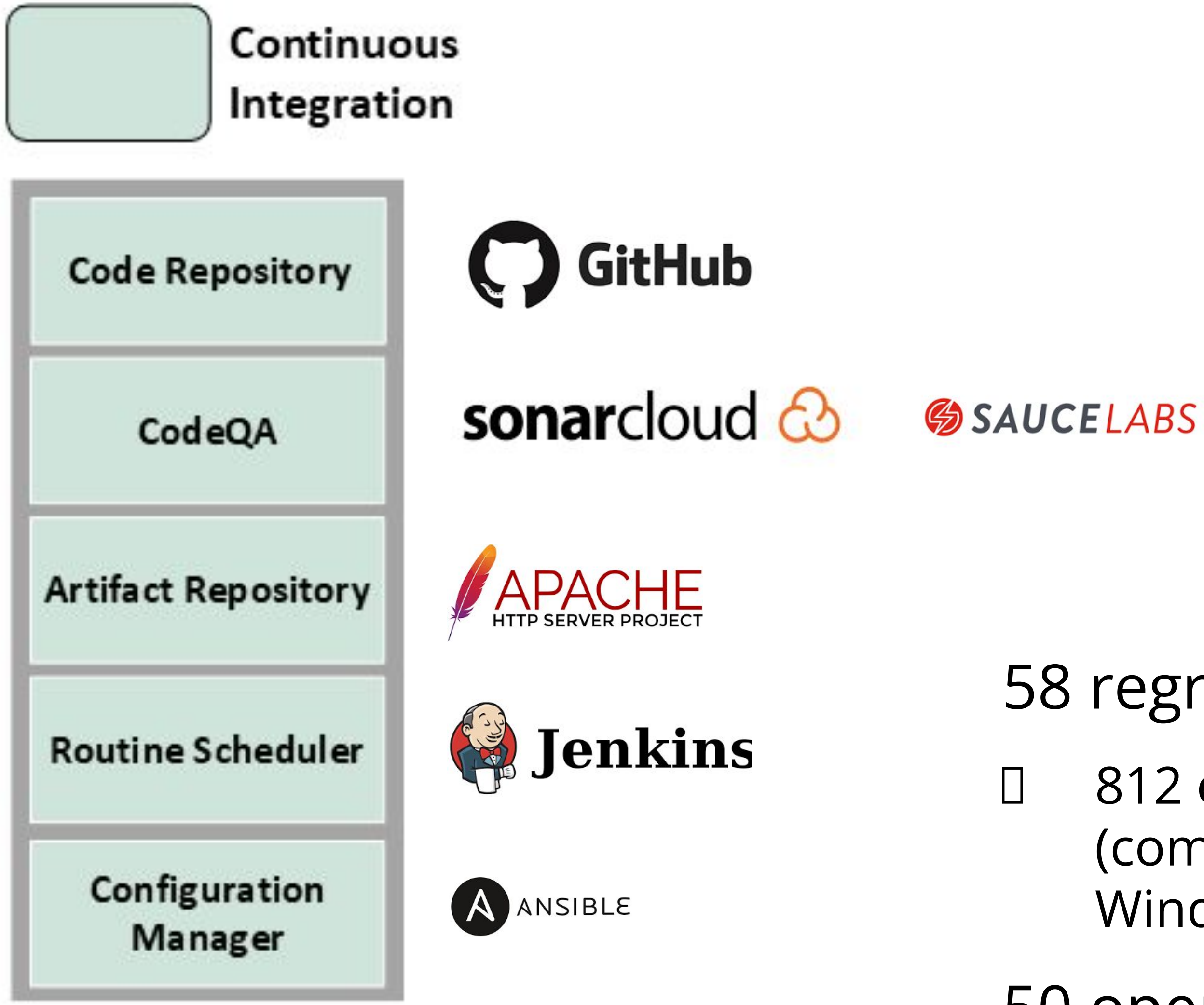
Revision: c248ceaaa39d74d1c6ba705d7645511821238309
Repository: https://github.com/arquivo/functional-tests.git

- refs/remotes/origin/master

- Test Result
- Rebuild
- Restart from Stage
- Replay
- Pipeline Steps
- Workspaces

Job Name	Status	Duration	OS	Browser	Job Links
savePageNowURLNotFound[0]	Passed	1m 3s	Linux	Android 12.0	Watch - Logs
narrativeButtonTest[0]	Passed	1m 9s	Linux	Android 12.0	Watch - Logs
urlSearchListPTTest[0]	Failed	1m 14s	Linux	Android 12.0	Watch - Logs

Jenkins



Continuous Integration

Code Repository

CodeQA

Artifact Repository

Routine Scheduler

Configuration Manager

58 regression tests

- 812 environments tested on Sauce labs (combinations of Firefox, Chrome and Safari on Windows, MacOS, IOS and Android)

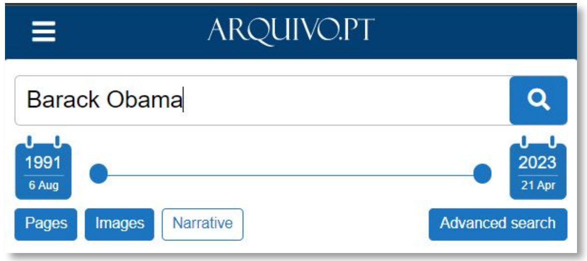
50 operational routines (Jenkins)

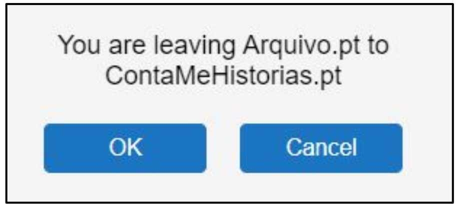
50 orchestration playbooks (Ansible)

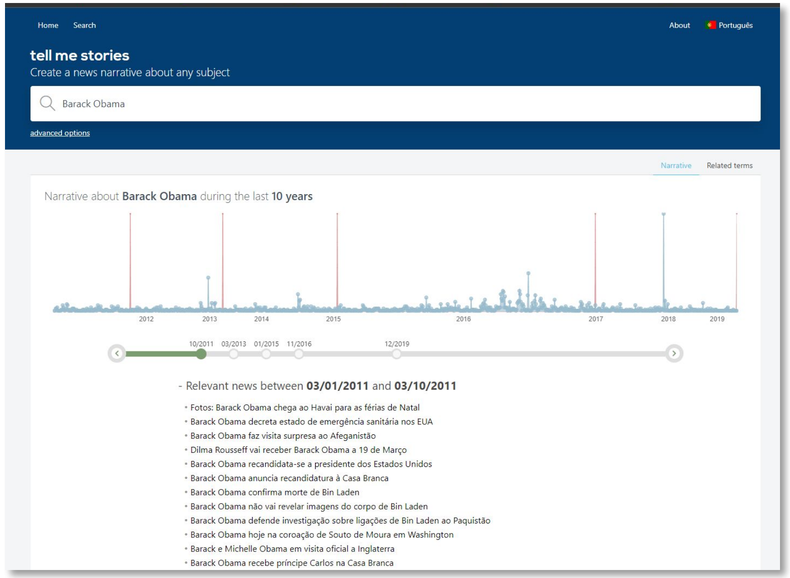
Third party service: “Tell me stories”

ContaMeHistorias.pt

ContaMeHistorias.pt




You are leaving Arquivo.pt to ContaMeHistorias.pt



Narrative about **Barack Obama** during the last **10 years**

- Relevant news between **03/01/2011** and **03/10/2011**

- Fotos: Barack Obama chega ao Havai para as férias de Natal
- Barack Obama decreta estado de emergência sanitária nos EUA
- Barack Obama faz visita surpresa ao Afeganistão
- Dilma Rousseff vai receber Barack Obama a 19 de Março
- Barack Obama recandidata-se a presidente dos Estados Unidos
- Barack Obama anuncia recandidatura à Casa Branca
- Barack Obama confirma morte de Bin Laden
- Barack Obama não vai revelar imagens do corpo de Bin Laden
- Barack Obama defende investigação sobre ligações de Bin Laden ao Paquistão
- Barack Obama hoje na coroação de Souto de Moura em Washington
- Barack e Michelle Obama em visita oficial a Inglaterra
- Barack Obama recebe príncipe Carlos na Casa Branca

Narrative: new Arquivo.pt function for free!

Automatically generates **narratives** about any subject based on online news from the Past.

Winners of the Arquivo.pt Award 2018.

- Advantage of exposing public APIs!

Communication and documentation tools

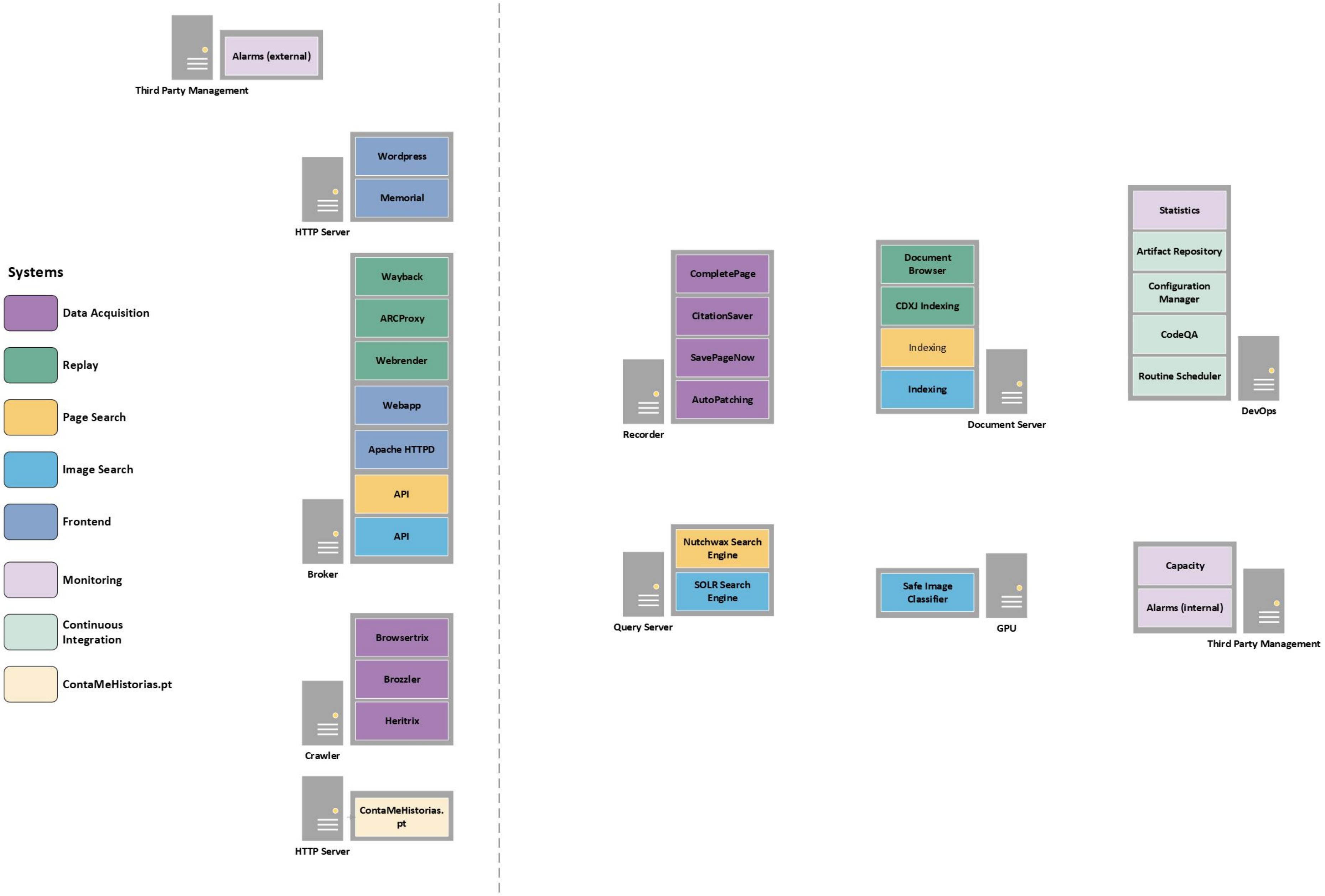
More tools, less communication

There are over 25 official communication tools used within our organization.

We use 5 tools in Arquivo.pt and it is more than enough

- GitHub
- Email
- Cell phone
- WhatsApp
- Google Drive

Hardware Architecture of Arquivo.pt



Alarms (external)
Third Party Management

Wordpress
Memorial
HTTP Server

Wayback
ARCProxy
Webrender
Webapp
Apache HTTPD
API
API
Broker

Browsertrix
Brozzler
Heritrix
Crawler

ContaMeHistorias.pt
HTTP Server

CompletePage
CitationSaver
SavePageNow
AutoPatching
Recorder

Document Browser
CDXJ Indexing
Indexing
Indexing
Document Server

Statistics
Artifact Repository
Configuration Manager
CodeQA
Routine Scheduler
DevOps

Nutchwax Search Engine
SOLR Search Engine
Query Server

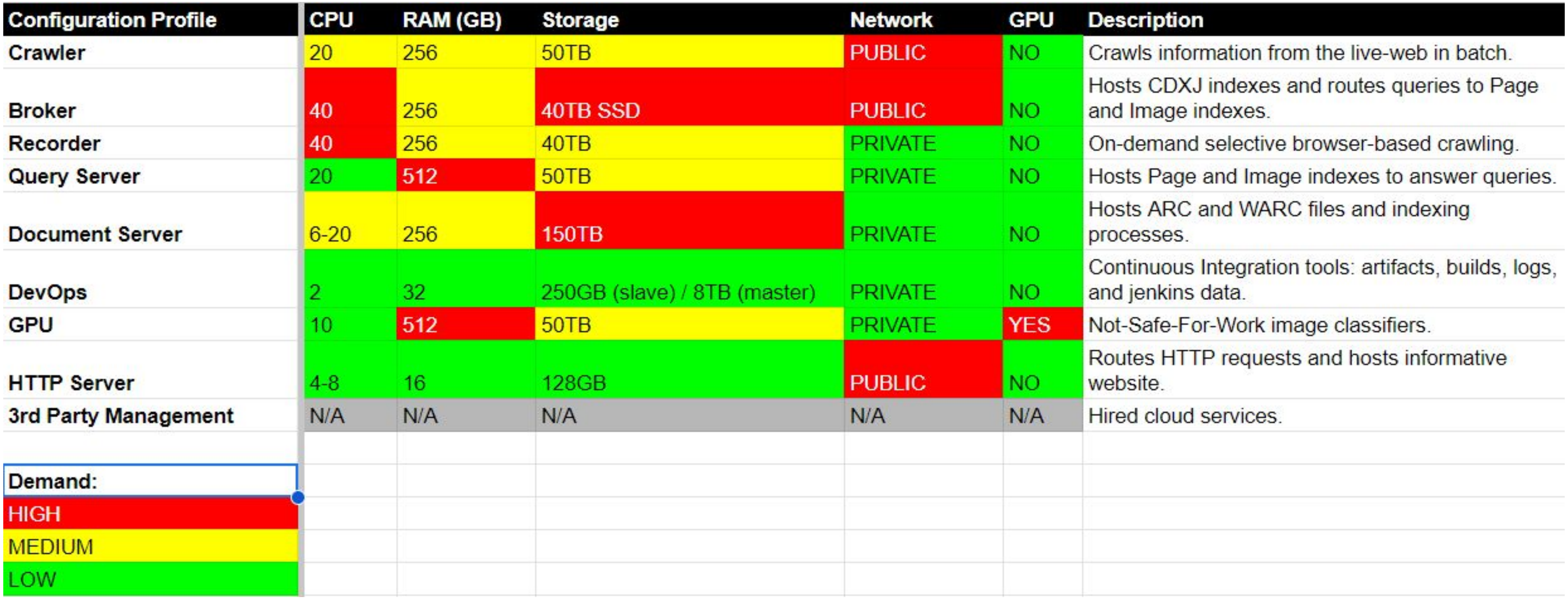
Safe Image Classifier
GPU

Capacity
Alarms (internal)
Third Party Management

Systems

- 🟪 Data Acquisition
- 🟩 Replay
- 🟨 Page Search
- 🟦 Image Search
- 🟦 Frontend
- 🟪 Monitoring
- 🟩 Continuous Integration
- 🟨 ContaMeHistorias.pt

9 hardware profiles for servers to facilitate acquisitions



Configuration Profile	CPU	RAM (GB)	Storage	Network	GPU	Description
Crawler	20	256	50TB	PUBLIC	NO	Crawls information from the live-web in batch.
Broker	40	256	40TB SSD	PUBLIC	NO	Hosts CDXJ indexes and routes queries to Page and Image indexes.
Recorder	40	256	40TB	PRIVATE	NO	On-demand selective browser-based crawling.
Query Server	20	512	50TB	PRIVATE	NO	Hosts Page and Image indexes to answer queries.
Document Server	6-20	256	150TB	PRIVATE	NO	Hosts ARC and WARC files and indexing processes.
DevOps	2	32	250GB (slave) / 8TB (master)	PRIVATE	NO	Continuous Integration tools: artifacts, builds, logs, and jenkins data.
GPU	10	512	50TB	PRIVATE	YES	Not-Safe-For-Work image classifiers.
HTTP Server	4-8	16	128GB	PUBLIC	NO	Routes HTTP requests and hosts informative website.
3rd Party Management	N/A	N/A	N/A	N/A	N/A	Hired cloud services.

Demand:

HIGH

MEDIUM

LOW

Own data-center shared with other public administration services

77 servers mounted on 167 Rack Units

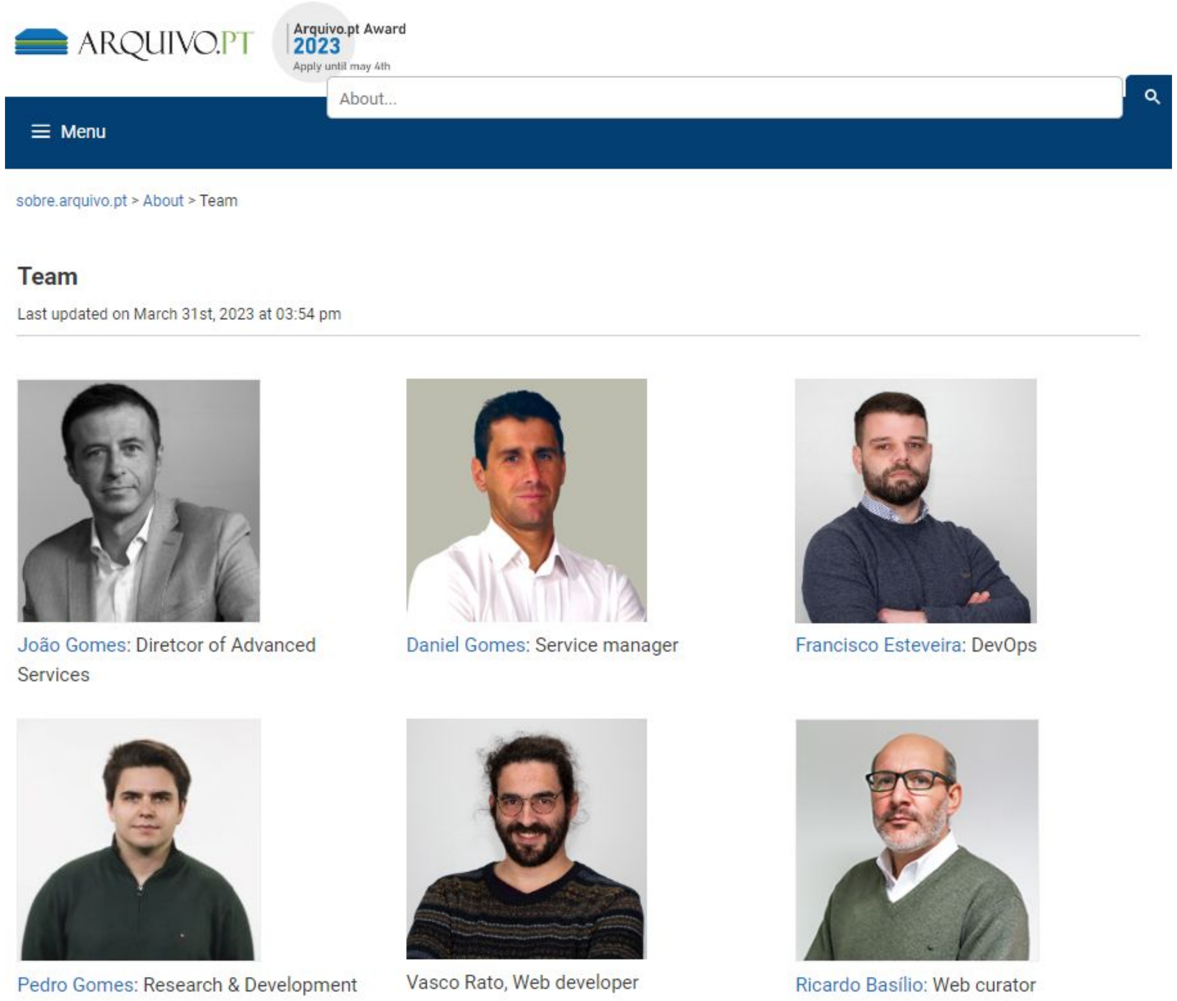
 7.4 meters in linear space

18 TB of RAM memory

2 180 vCPUs

1 234 hard drives (5.18 PB)

The people behind the computers



sobre.arquivo.pt > About > Team

Team

Last updated on March 31st, 2023 at 03:54 pm

João Gomes: Diretcor of Advanced Services

Daniel Gomes: Service manager

Francisco Esteveira: DevOps

Pedro Gomes: Research & Development

Vasco Rato, Web developer

Ricardo Basílio: Web curator

Suggested publications

- Web archives as research infrastructure for digital societies: the case study of Arquivo.pt (2022), Daniel Gomes
- The Anatomy of a Web Archive Image Search Engine (2022), André Mourão
- Information Search in Web Archives (2014), Miguel Costa
- Learning Temporal-Dependent Ranking Models (2014), Costa et al.

□ arquivo.pt/publications

Recommendations to manage a web archive

No part-time web archivists

- Web archiving is complex, requires training and full dedication
- Document thoroughly, prepare the leave when you hire someone

Listen to your users

- Don’t guess what they need, hire User Experience researchers to find it

Use existing tools and micro-services

- Webrecorder.net, Archive-it
- Micro-services facilitate technology replacements

Design-to-fail architecture

- Service must resist to any component fault
- Test, test, test: maintaining is harder than developing