

Constructing and sharing historical web link graphs from web archives

vasco.rato@fccn.pt

Arquivo.pt text search is old...

- Textsearch uses NutchWAX which hasn't had updates since 2010
- It's old enough that its project page is no longer live



Arquivo.pt text search is old...

- But you can still access it using arquivo.pt ;)



The screenshot shows the Arquivo.pt interface for the NutchWAX project. The top navigation bar includes a 'Menu' button, the 'arquivo.pt' logo, and an 'Opções' button. The breadcrumb trail reads 'archive-access.sourceforge.net/projects/nutchwax' and the timestamp is '18 Janeiro às 18h44, 2023'. A left sidebar contains a 'Tabela' button and a year-based navigation menu with options for 2009, 2011, 2019, 2020, 2021, 2022, 2023, and Janeiro. The main content area features the NutchWAX logo, a 'Last Published: 08 Mar 2009' notice, and a navigation menu with links to Sourceforge, Heritrix, Archive Access, Internet Archive, and Home. The page content is divided into sections: 'Introduction' and 'Project Sponsors'. The 'Introduction' section describes NutchWAX as a search tool for web archive collections, mentioning its use of Nutch and Web Archive eExtensions (WAX). The 'Project Sponsors' section shows the logo for the International Internet Preservation Consortium (IIPC).

Menu

arquivo.pt | fct

Opções

archive-access.sourceforge.net/projects/nutchwax 18 Janeiro às 18h44, 2023

Tabela

2009

2011

2019

2020

2021

2022

2023

Janeiro

Last Published: 08 Mar 2009

Sourceforge | Heritrix | Archive Access | Internet Archive | Home

NutchWAX

Home

- Downloads
- Getting Started
- Building from Source
- User Query-time Help
- Regression Test Suite
- Wayback-NutchWAX
- Praxis
- FAQ

Project Documentation

- Project Information
- Project Reports

built by:

maven

Introduction

NutchWAX ("Nutch + Web Archive eExtensions") searches web archive collections. The Web Archive eExtensions (WAX) include adaptation of the Nutch fetcher step to go against web archives rather than crawl the open net -- adaptation currently does Internet Archive ARC files only -- and plugins to add extra fields to the index that return an Archive Records' location in the repository, its collection name, etc.

Project Sponsors

IIPC

The International Internet Preservation Consortium (IIPC) is a consortium of

Solr based text search

- We already had a Solr based solution for image search
- Experimented with Solrwayback
- ...But ended up adapting our image search solution for text searching



Constructing the dataset

- New text search engine required reindexing our whole archive
- Used the opportunity to also generate link graph!
- From WARCs we extract outlinks directly
- After all outlinks are extracted, we convert them to inlinks

Constructing the dataset

- Converting outlinks into inlinks

pt,fct)/

- Doc #1 (url: "https://www.fct.pt")
 - Captured: "2024-03-01 09:00:00"
 - Outlinks:
 - (pt.fct/,"Home")

We find **fct.pt** archived at **2024-03-01 09:00:00**

Caso concreto: Links dataset

- Converting outlinks into inlinks

pt,fct)/

- Doc #1 (url: "https://www.fct.pt")
 - Captured: "2024-03-01 09:00:00"
 - Outlinks:
 - (pt.fct/,"Home")
- Inlink #1
 - Captured: "2024-03-01 09:00:00"
 - Source: pt,fct)/
 - Anchor: "Home"

Caso concreto: Links dataset

- Converting outlinks into inlinks

pt,fct)/

- Doc #1 (url: "https://www.fct.pt")
 - Captured: "2024-03-01 09:00:00"
 - Outlinks:
 - (pt.fct/,"Home")
- Inlink #1
 - Captured: "2024-03-01 09:00:00"
 - Source: pt,fct)/
 - Anchor: "Home"

pt,fccn)/

- Doc #1 (url: "https://fccn.pt")
 - Captured: "2024-03-01 10:00:00"
 - Outlinks:
 - (pt.fct/,"Fundação Ciência Tec.")
 - (pt.fccn/quem-somos,"Quem somos")

We find **fccn.pt** archived at **2024-03-01 10:00:00**

Caso concreto: Links dataset

- Converting outlinks into inlinks

pt,fct)/

- Doc #1 (url: "https://www.fct.pt")
 - Captured: "2024-03-01 09:00:00"
 - Outlinks:
 - (pt.fct/,"Home")
- Inlink #1
 - Captured: "2024-03-01 09:00:00"
 - Source: pt,fct)/
 - Anchor: "Home"

- Inlink #2
 - Captured: "2024-03-01 10:00:00"
 - Source: pt,fccn)/
 - Anchor: "Fundação Ciência Tec."

pt,fccn)/

- Doc #1 (url: "https://fccn.pt")
 - Captured: "2024-03-01 10:00:00"
 - Outlinks:
 - (pt.fct/,"Fundação Ciência Tec.")
 - (pt.fccn/quem-somos,"Quem somos")

pt,fccn)/quem-somos

- Inlink #1
 - Captured: "2024-03-01 10:00:00"
 - Source: pt,fccn)/
 - Anchor: "Quem somos"

```
{
  "url": "(pt,fct,)/apoios/cooptrans/eranets/jpco_fund/index.phtml.en",
  "count": 6,
  "countInternal": 5,
  "countExternal": 1,
  "captureDate": "2021-10-27T02:19:35",
  "inlinks": [
    {
      "date": "2021-10-10T22:45:16",
      "source": "(pt,fct,)/apoios/cooptrans/eranets/index.phtml.en",
      "anchor": "JPCOFUND2"
    },
    {
      "date": "2021-10-10T22:23:51",
      "source": "(pt,fct,)/concurso/index.phtml.en",
      "anchor": "Joint Transnational Call 2021 of the ERA-NET Cofund JPcofuND 2 is open"
    },
    {
      "date": "2021-10-10T21:46:19",
      "source": "(pt,fct,)/apoios/cooptrans/eranets/jpco_fund/index.phtml.pt",
      "anchor": "PT | EN"
    },
    {
      "date": "2021-10-10T23:49:46",
      "source": "(pt,fct,)/calendario/index.phtml.en?mes=3&ano=2021",
      "anchor": "Joint Transnational Call 2021 of the ERA-NET Cofund JPcofuND 2 is open"
    },
    {
      "date": "2021-10-10T22:45:11",
      "source": "(pt,fct,)/apoios/cooptrans/eranets/",
      "anchor": "JPCOFUND2"
    },
    {
      "date": "2021-11-15T03:39:01",
      "source": "(eu,biomark-uc,)/projects-oligofit",
      "anchor": "https://www.fct.pt/apoios/cooptrans/eranets/jpco_fund/index.phtml.en"
    }
  ]
}
```

Inlink dataset overview






The Arquivo.pt link dataset combines three distinct web collections:

- **PWA9609** (1996-2009): 89 million pages that capture the initial evolution of the Internet, centered on the .pt domain. This historical collection provides insights into early linking patterns on the Web.
- **AWP38** (Oct-Nov 2021): 44 million pages that offer a contemporary portrait of Web connectivity, with emphasis on the .pt domain, but including broader Internet content.
- **FAWP47** (Oct-Dec 2021): 8 million pages from daily captures of .pt domain content, designed to track short-term changes in linking patterns.

Sharing the Data

- Available for download at arquivo.pt/datasets/linkgraphs

Index of /datasets/linkgraphs

 [ICO]	<u>Name</u>	<u>Last modified</u>	<u>Size</u>	<u>Description</u>
 [PARENTDIR]	Parent Directory		-	
 [DIR]	AWP38/	2025-02-05 18:24	-	
 [DIR]	FAWP47/	2025-02-05 18:24	-	
 [DIR]	PWA9609/	2025-02-13 17:28	-	

Thank you!

contacto@arquivo.pt