Arquivo.pt APIs: enabling automatic analytics over historical web data

Fernando Melo < fernando.melo@fccn.pt >
João Nobre < joao.nobre@fccn.pt >

API – Application Programming Interface

Fast development of new services and applications

Automatic access

Easy integration

Eliminates the need to understand the core code

What are Arquivo’s APIs for?

What are Arquivo's APIs for?

Image Search

What are Arquivo's APIs for?

Aggregator of information about politicians

Memento API

(URL Search)

Memento API

http://timetravel.mementoweb.org/guide/api/

"Time Travel helps you find and view versions of web pages that existed at some time in the past."

Memento API

time travel

http://nytimes.com

2002-02-22 | 12:19:37 | Find | Reconstruct

Find Mementos in Internet Archive, Archive-It, British Library, archive.today, GitHub and many more!

Memento API

URL Search

Multiple Web Archives

http://arquivo.pt/apis/memento

OpenSearch API

(URL and Full-text Search)

OpenSearch - Definition

http://www.opensearch.org/Home

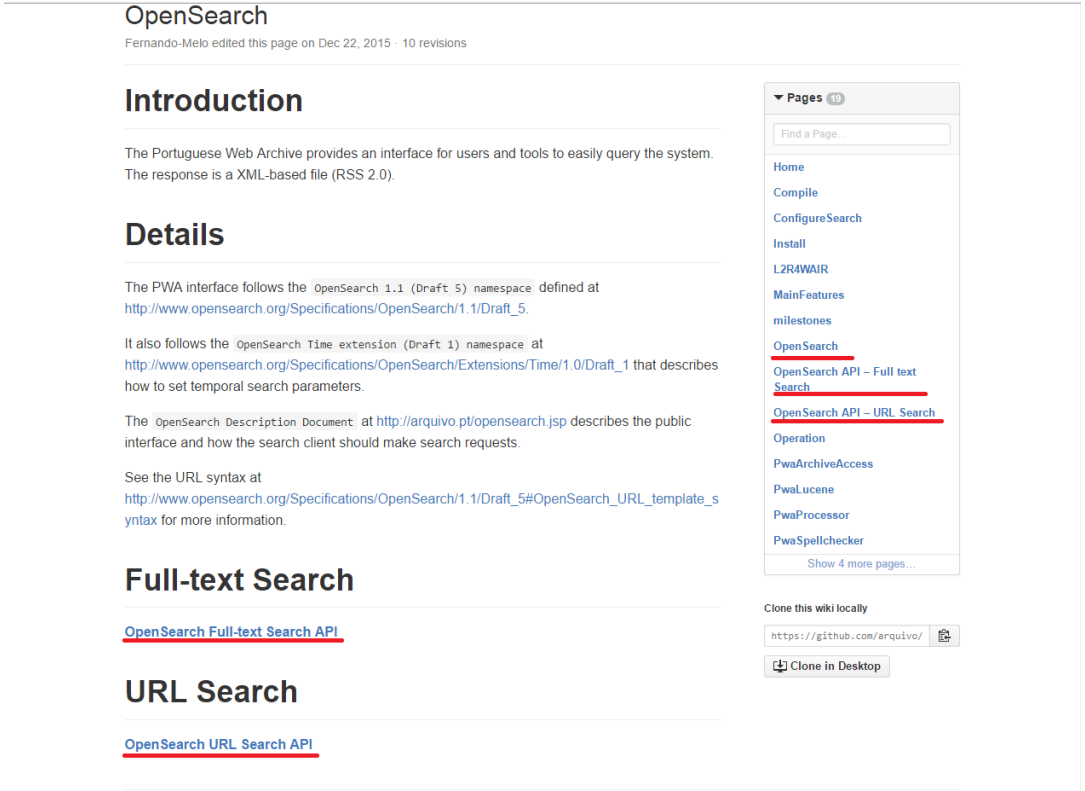
"OpenSearch is a collection of simple formats for the sharing of search results."

OpenSearch based Arquivo.pt API

Extended to support temporal searches in the Portuguese web archive

OpenSearch – Arquivo.pt Wiki

http://arquivo.pt/apis/opensearch



OpenSearch

Fernando-Melo edited this page on Dec 22, 2015 · 10 revisions

Introduction

The Portuguese Web Archive provides an interface for users and tools to easily query the system. The response is a XML-based file (RSS 2.0).

Details

The PWA interface follows the `OpenSearch 1.1 (Draft 5) namespace` defined at http://www.opensearch.org/Specifications/OpenSearch/1.1/Draft_5.

It also follows the `OpenSearch Time extension (Draft 1) namespace` at http://www.opensearch.org/Specifications/OpenSearch/Extensions/Time/1.0/Draft_1 that describes how to set temporal search parameters.

The `OpenSearch Description Document` at http://arquivo.pt/opensearch.jsp describes the public interface and how the search client should make search requests.

See the URL syntax at http://www.opensearch.org/Specifications/OpenSearch/1.1/Draft_5#OpenSearch_URL_template_syntax for more information.

Full-text Search

OpenSearch Full-text Search API

URL Search

OpenSearch URL Search API

Pages 19

Find a Page...

Home
Compile
ConfigureSearch
Install
L2R4WAIR
MainFeatures
milestones
OpenSearch
OpenSearch API – Full text Search
OpenSearch API – URL Search
Operation
PwaArchiveAccess
PwaLucene
PwaProcessor
PwaSpellchecker

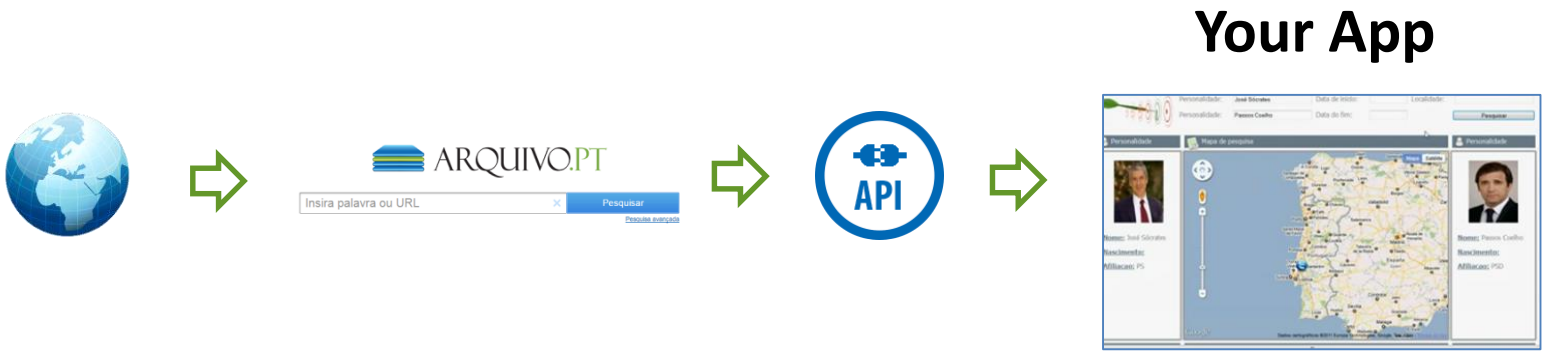
Show 4 more pages...

Clone this wiki locally

https://github.com/arquivo/

Clone in Desktop

Arquivo.pt OpenSearch API:



Full-text search

URL search

OpenSearch – Full-text search

Full-text search **request**

http://arquivo.pt/opensearch?query=**euro 2004**

Default parameters:

start=0 (starting result)
hitsPerSite=2 (max number of results from the same site)
hitsPerPage=10 (number of results requested)

Full-text search request - **next 10 results**

http://arquivo.pt/opensearch?query=**euro 2004 &start=10**

start=10 (next 10 results)

Full-text search request - **2000 results**

http://arquivo.pt/opensearch?query=**euro 2004 &start=0&hitsPerPage=2000**

Full-text search request – **search an expression**

http://arquivo.pt/opensearch?query=**“euro 2004”**

Search results with the expression: “**euro 2004**”

Full-text search request results – **excluding a word**

http://arquivo.pt/opensearch?query=**euro 2004+-coin**

Search results
containing words **euro , 2004**
but **without** the word **coin**

Full-text search request - **search by mimetype pdf**

http://arquivo.pt/opensearch?query=**euro 2004+ type=application/pdf**

Search results containing the words **euro** , **2004** *in* **.pdf**

Full-text search **response –results summary**

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<rss xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/">
<channel>
    <title>PWA Search Engine</title>
    <description>PWA search results for query: euro 2004</description>
    <link>http://archive.pt</link>
    <opensearch:totalResults>18185160</opensearch:totalResults>
    <opensearch:startIndex>0</opensearch:startIndex>
    <opensearch:itemsPerPage>10</opensearch:itemsPerPage>
    <opensearch:Query role="request" searchTerms="euro 2004" startPage="1"/>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
</channel>
</rss>
```

Full-text search **response –result items**

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<rss xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/">
<channel>
    <title>PWA Search Engine</title>
    <description>PWA search results for query: euro 2004</description>
    <link>http://archive.pt</link>
    <opensearch:totalResults>18185160</opensearch:totalResults>
    <opensearch:startIndex>0</opensearch:startIndex>
    <opensearch:itemsPerPage>10</opensearch:itemsPerPage>
    <opensearch:Query role="request" searchTerms="euro 2004" startPage="1"/>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
    <item>...</item>
</channel>
</rss>
```

Full-text search **response item**

```
<item>
    <title>Euro 2004 - Estádios do Euro 2004</title>
    <source url="http://helder2004.no.sapo.pt/">
        Original URL of Euro 2004 - Estádios do Euro 2004
    </source>
    <link>http://arquivo.pt/wayback/20131106141720/http://helder2004.no.sapo.pt/</link>
    <pwa:id>2506798</pwa:id>
    <pwa:index>24</pwa:index>
    <pwa:arcname>IAH-20131105224049-01265-p13.arquivo.pt</pwa:arcname>
    <pwa:arcoffset>86869029</pwa:arcoffset>
    <pwa:digest>7bb196359aef86a45767a34b21a78a97</pwa:digest>
    <pwa:tstamp>20131105224456000</pwa:tstamp>
    <pwa:contentLength>1718</pwa:contentLength>
    <pwa:primaryType>text</pwa:primaryType>
    <pwa:subType>html</pwa:subType>
</item>
```

OpenSearch –URL search

URL search **request example**

http://arquivo.pt/opensearch?query=
date:19960101000000-20150101000000+
exacturlexpand: http://www.nytimes.com/
&**hitsPerSite**=10000&**waybackQuery**=true
&**start**=0&**hitsPerPage**=10000

URL search request – **date options**

http://arquivo.pt/opensearch?query=

date: Search versions of a URL between 2 dates.

OR

closestdate: For each collection, lists the version of a URL whose date is closest to the requested date.

URL search request – **date examples**

http://arquivo.pt/opensearch?query=

date:19960101000000-20151022163016

OR

closestdate:20140101010101

URL search request - **example**

http://arquivo.pt/opensearch?query=
closestdate:19960101000000+
exacturlexpand:http://nytimes.com/
&**hitsPerSite**=10000&**waybackQuery**=true
&**start**=0&**hitsPerPage**=10000

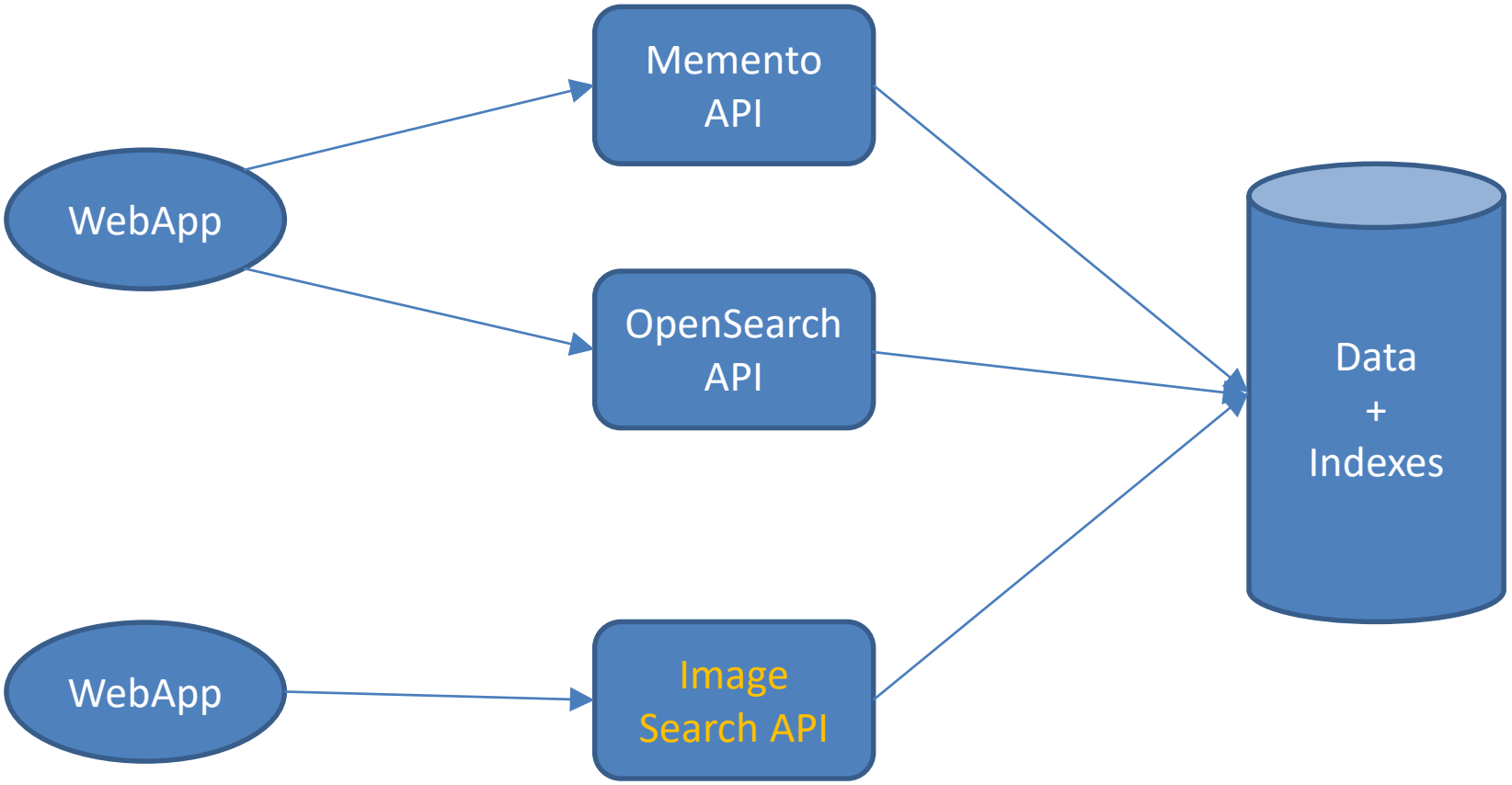
Image Search API

Arquivo.pt – under development

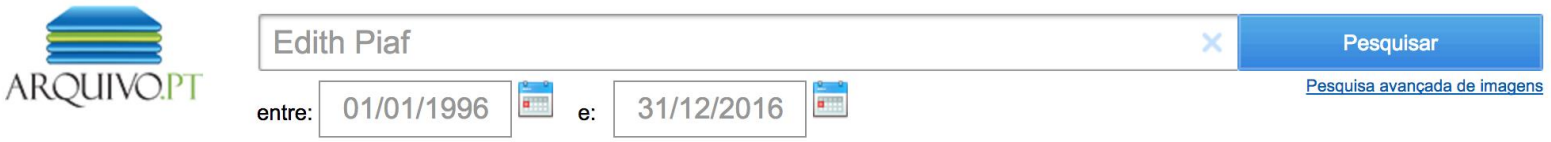
João Nobre

joao.nobre@fccn.pt

Search Image API

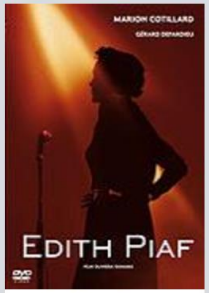


What is search image API for?



marius4.com.sapo....
14 Março, 2008

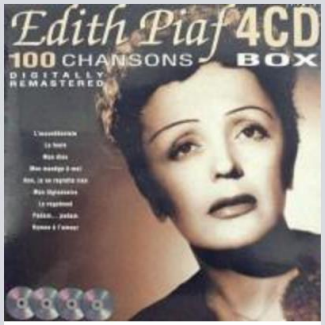
diekopfhoerer.eu/...
22 Novembro, 2014



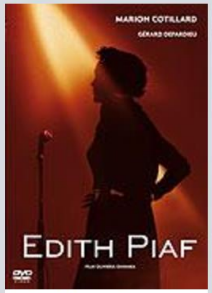
www.kinotip.eu/fi...
26 Novembro, 2014

www.tsf.pt/Pagina...
13 Outubro, 2013

www.tsf.pt/Pagina...
13 Outubro, 2013



warezclub.eu/musi...
25 Novembro, 2014



online-filmy.vset...
25 Novembro, 2014

secure.getgo.de/g...
16 Março, 2008

www.zippyshare.eu...
23 Novembro, 2014

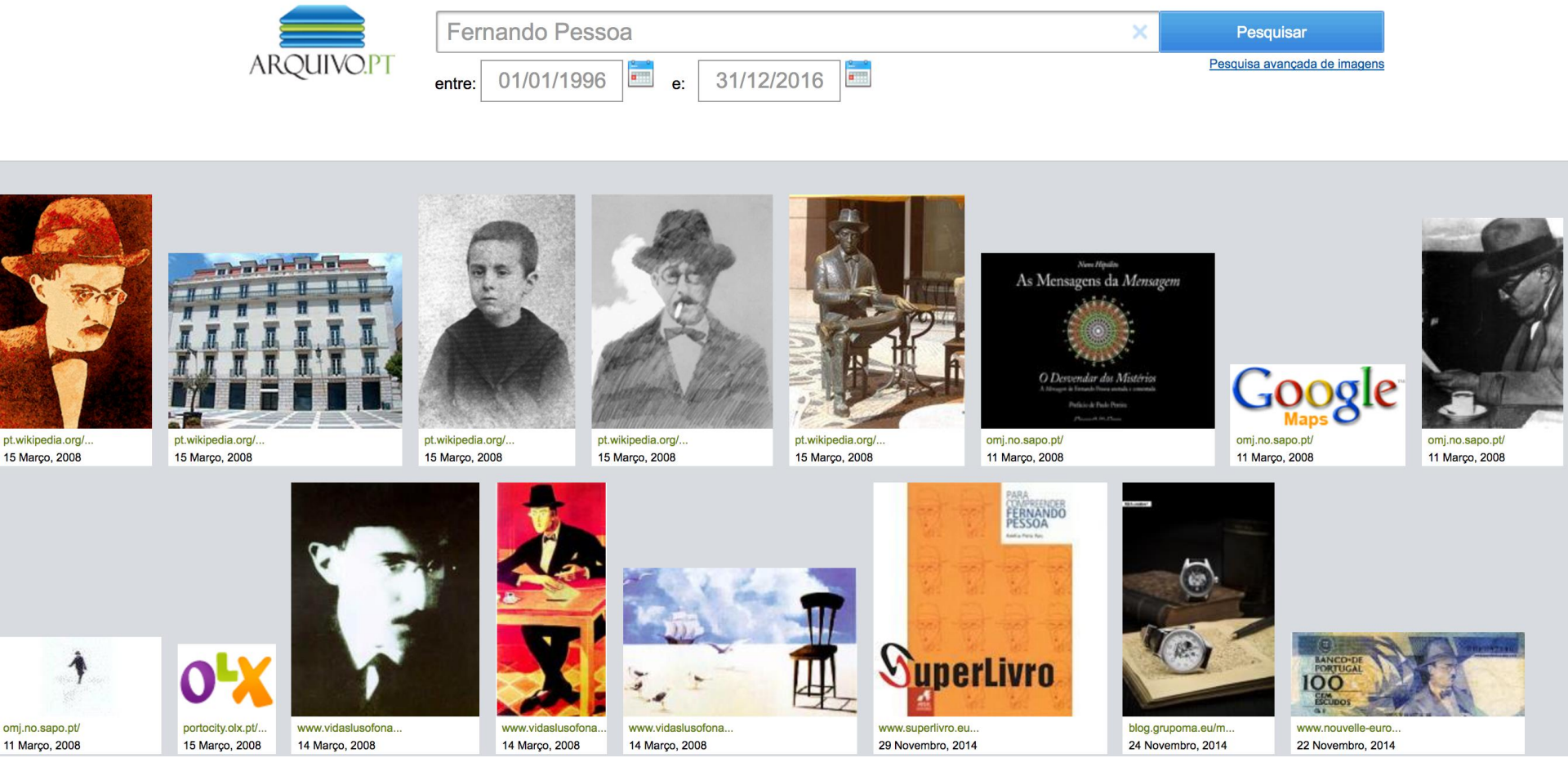
www.zippyshare.eu...
23 Novembro, 2014

www.bibi-piaf.com/
17 Março, 2008

www.bibi-piaf.com/
17 Março, 2008

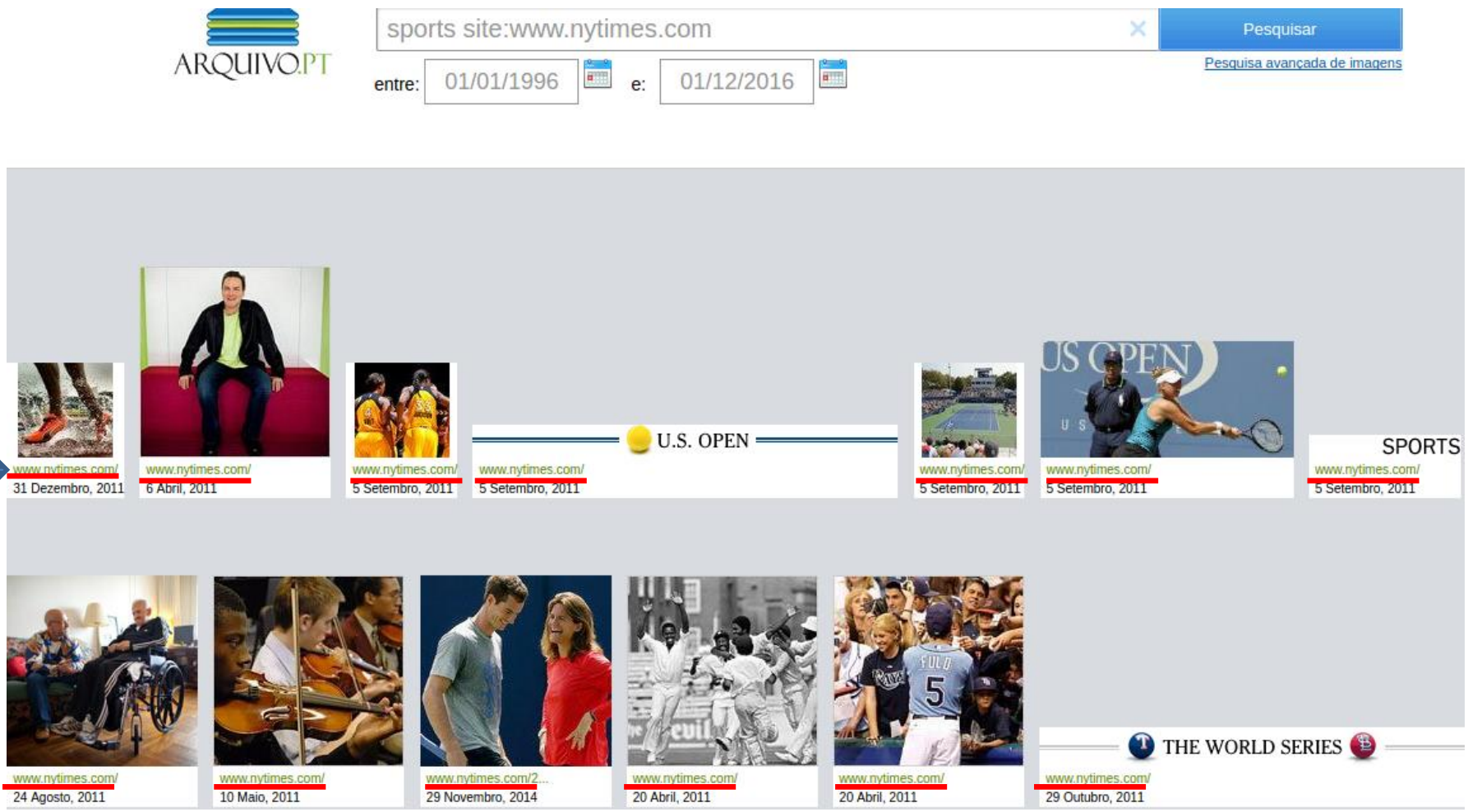
Search Image API - How to use ?

Keyword on based image search



Search Image API – Advanced Search Operators

site: site-only search



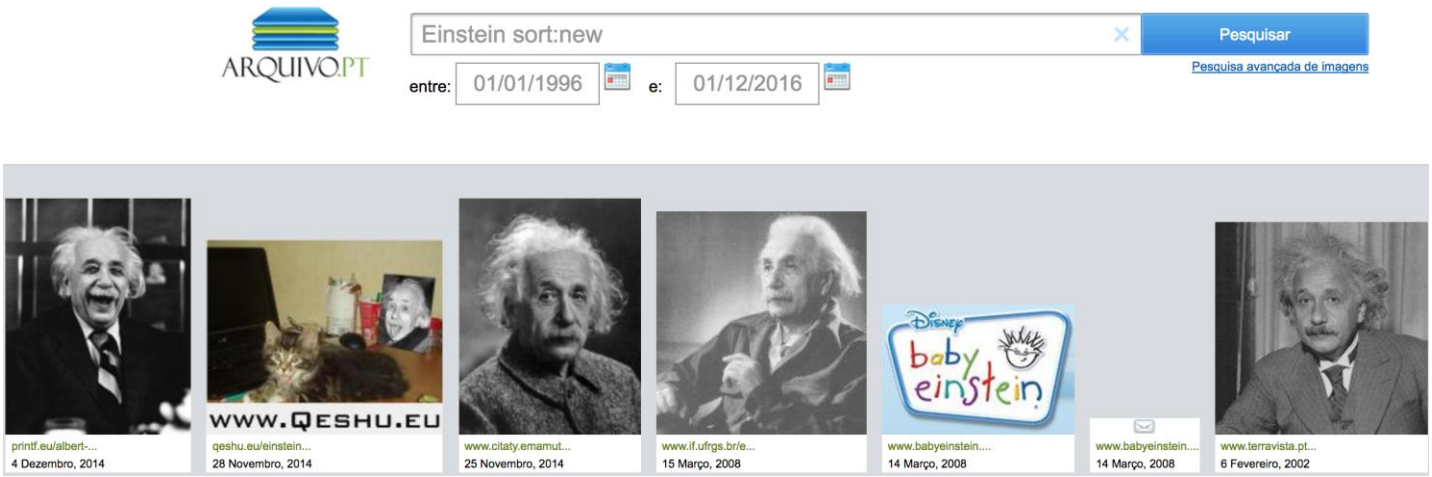
Search Image API – Advanced Search Operators

***sort**:* return results ordered by:

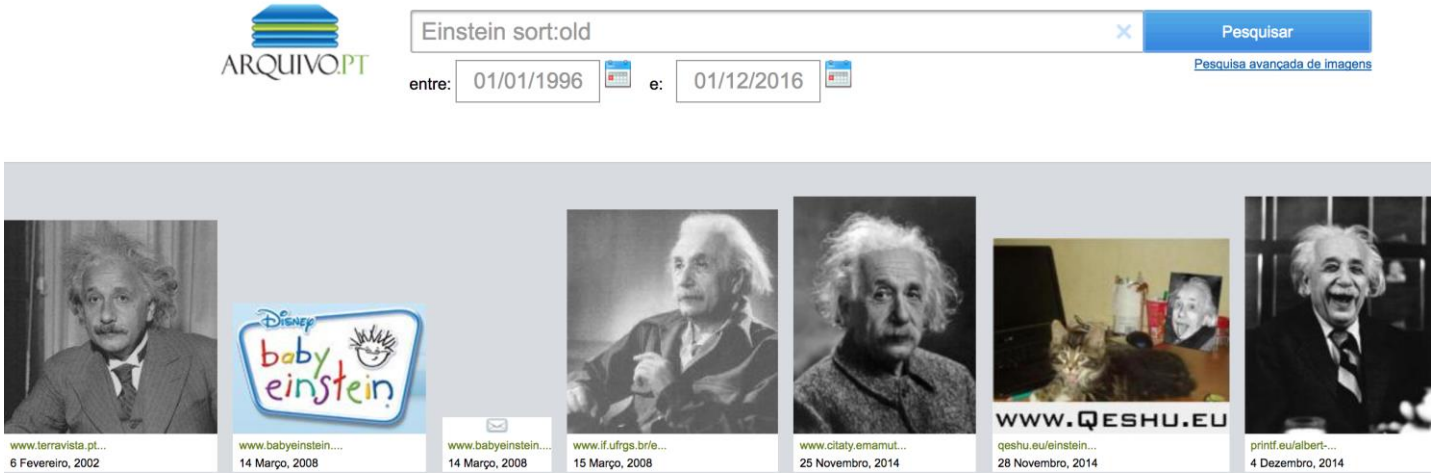
- relevance (default): value assigned based on their attributes
- new: from the most recent to the oldest
- old: from the oldest to the most recent

Search Image API – Advanced Search Operators

Sorted by **newest**

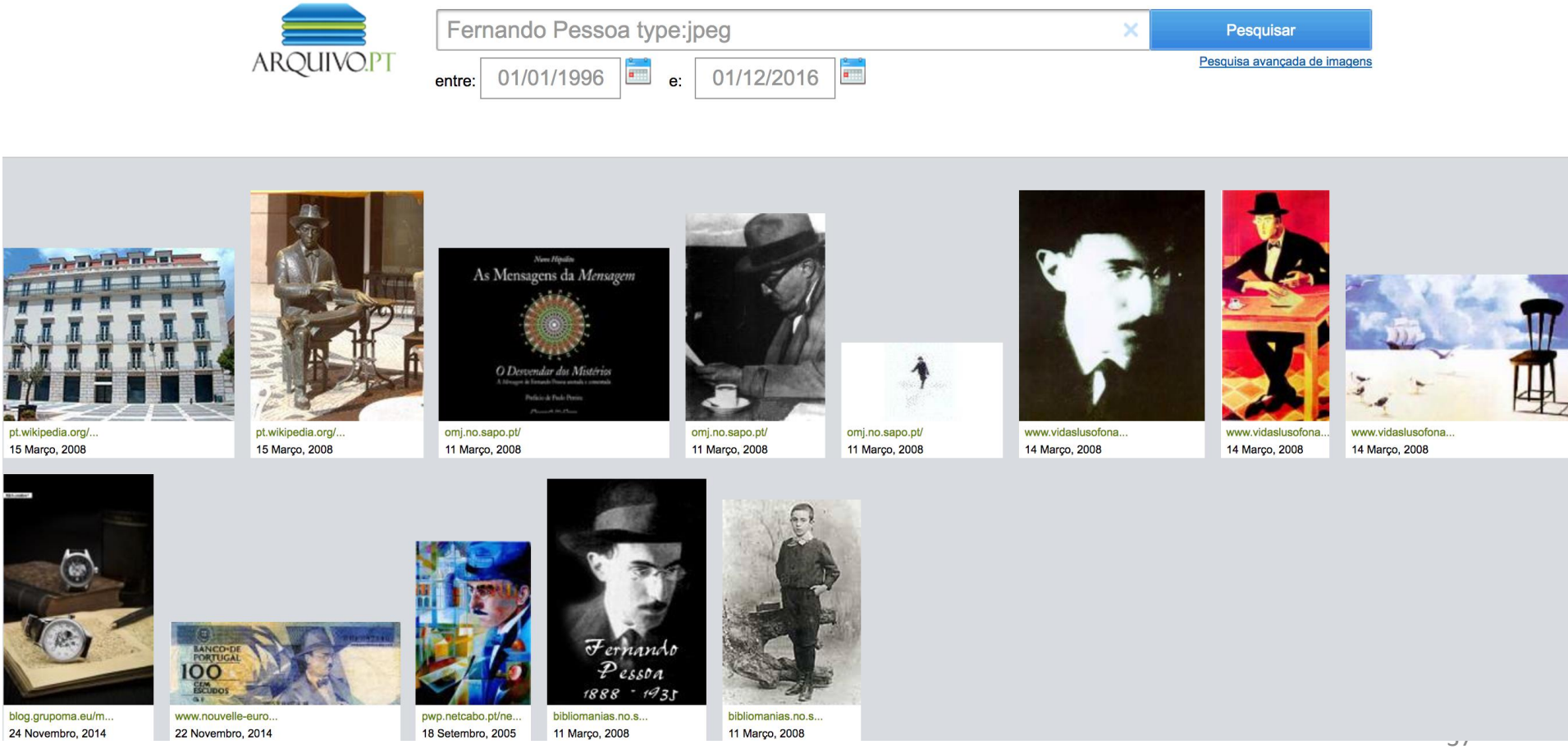


Sorted by **oldest**



Search Image API – Advanced Search Operators

type: returns only images of the mimetype



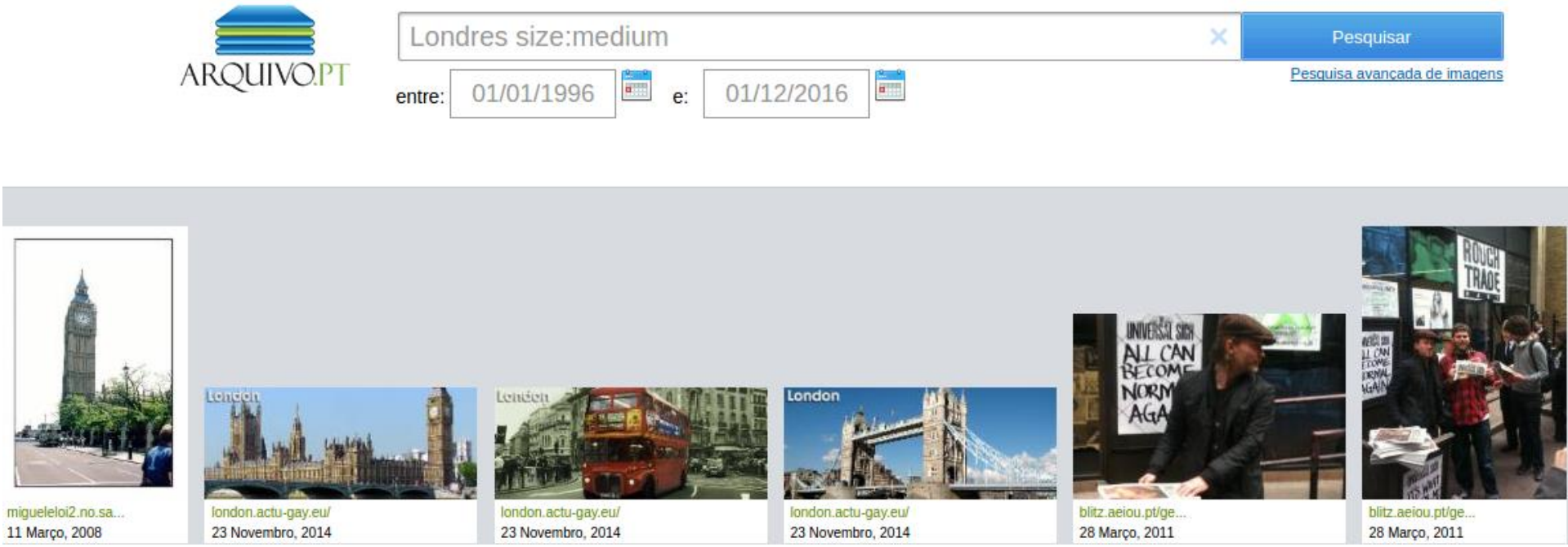
Search Image API – Advanced Search Operators

***size**:* returns only images with specified size.

- icon
- small
- medium
- large

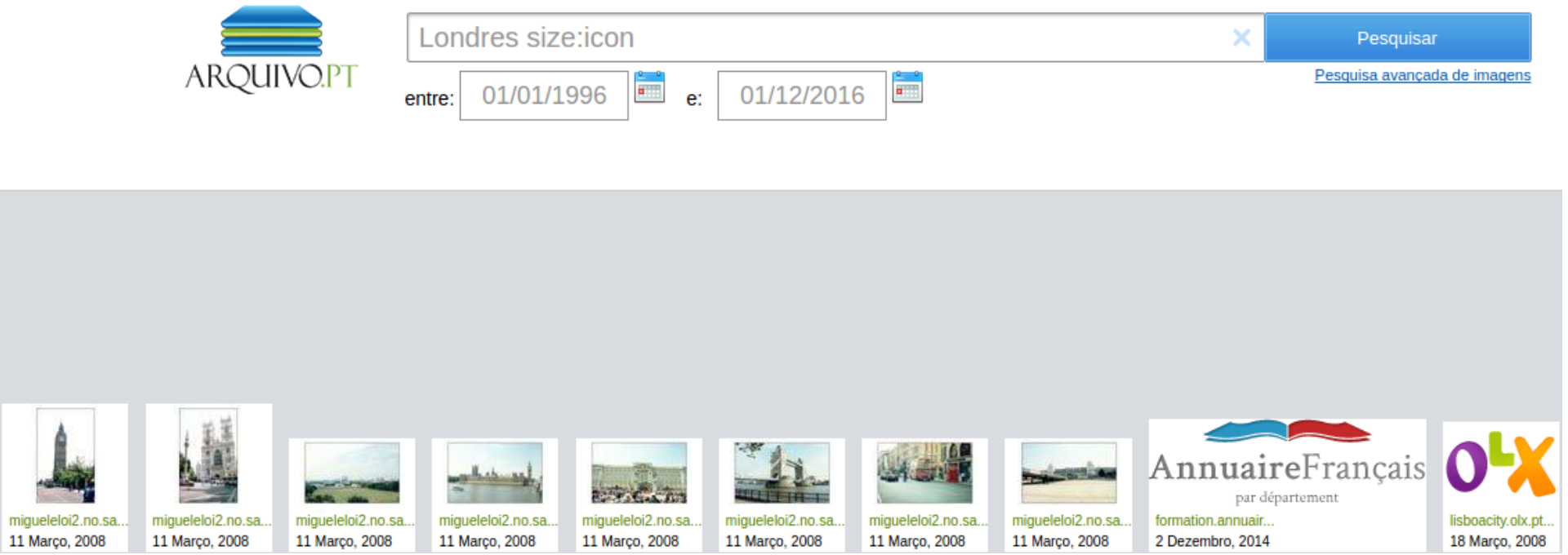
Search Image API – Advanced Search Operators

Medium size



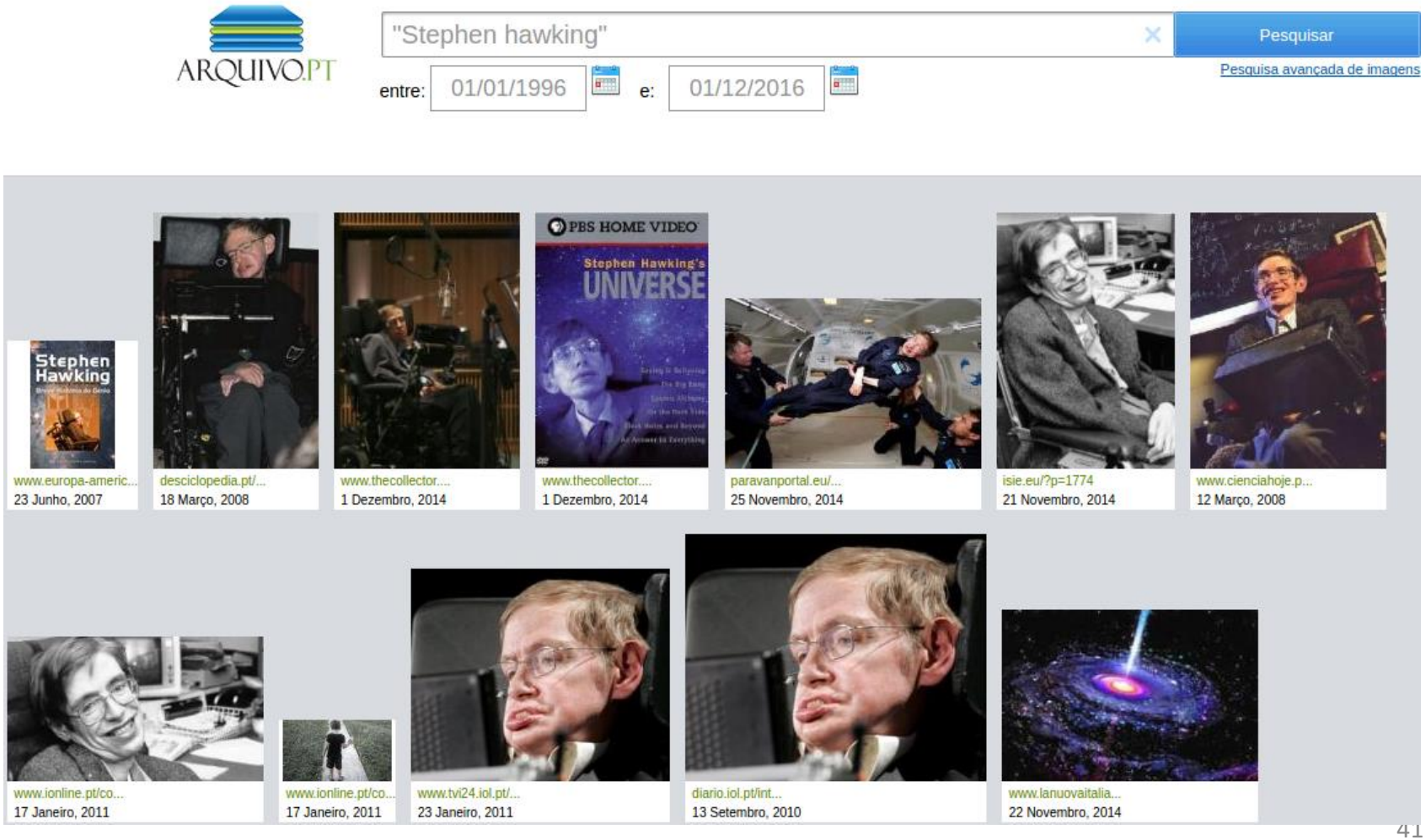
Search Image API – Advanced Search Operators

Icon size



Search Image API – Advanced Search Operators

“” : search images with a expression



Search Image API – Request Parameters

Client makes call to Search Image API with the following request parameters:

- **Query:** keyword search + advanced search
- **Stamp:** timestamp is the time gap where search will coincide.
- **Start:** number that indicates the first index of search.
- **safeImage:** parameter which indicates that we are restricting search only to safe Images
  - yes (safe for work)
  - no(not safe for work)
  - all (all)

Example of **request**:

http://p18.arquivo.pt/getimagesWS/?query=books&stamp=19960101000000-20151022163016&start=1

Search Image API – Response elements

Example **response**, in json

- "**totalResults**": 1,
- "**url**":"http://arquivo.pt/noFrame/replay/20141126190641im_/http://pics.ebaystatic.com/aw/pics/half/newsite/imgHalfLogo_126x54.gif",
- "**width**":"126.0",
- "**height**":"54.0",
- "**alt**":"Half.com by eBay: Buy and Sell new and used books, music, movies, games and more...",
- "**score**":4.5,
- "**timestamp**":"20141126190641",
- "**mime**":"image/gif",
- "**thumbnail**":"R0lGODlhfgA2APcAAP/////+/P7+/....",
- "**nsfw**": 0.1, (1 is safe / 0 is not safe)
- "**pageTitle**":"New Books - Buy, Sell, Search Books Online at Half.com"

Thank you

http://arquivo.pt/apis

João Nobre < joao.nobre@fccn.pt >
Fernando Melo < fernando.melo@fccn.pt >