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

[Search Bar]

- Charlie Hebdo
- between: 01/01/2015 and 31/12/2015

[Images of Charlie Hebdo cartoons and images related to the magazine]
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

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
ARQUIVO.PT 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
Arquivo.pt OpenSearch API:

Full-text search
URL search
OpenSearch – Full-text search
Full-text search request


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


start=10 (next 10 results)
Full-text search request - 2000 results

&start=0&hitsPerPage=2000
Full-text search request – search an expression


Search results with the expression: “euro 2004”
Full-text search request results – excluding a word


Search results containing words euro, 2004 but **without** the word **coin**
Full-text search request - search by mimetype pdf

  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
<?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

```xml
<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>
  <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**


**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

WebApp

Memento API

OpenSearch API

Image Search API

Data + Indexes
What is search image API for?
Search Image API - How to use?

Keyword on based image search

Search for "Fernando Pessoa" with dates from 01/01/1996 to 31/12/2016.
Search Image API – Advanced Search Operators

**site:** site-only search

```
sports site:www.nytimes.com
```

Between: 01/01/1996 and 01/12/2016
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

Londres size:medium

entre: 01/01/1996 e: 01/12/2016
Search Image API – Advanced Search Operators

**Icon size**

Londres size:icon
text: 01/01/1996 e: 01/12/2016
Search Image API – Advanced Search Operators

“”: search images with an expression

Example:

بحث الصور مع التعبير "Stephen hawking".
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,
- "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 >