

Cultivating Open Access

Through innovative services

fernando.melo@fccn.pt

Arquivo.pt – Open Access



<https://arquivo.pt>

Publicly available web archive

Research infrastructure

Reusable – APIs

Source code on Github

Free

- 1. Mobile version**
2. Arquivo.pt API
3. ImageSearch API

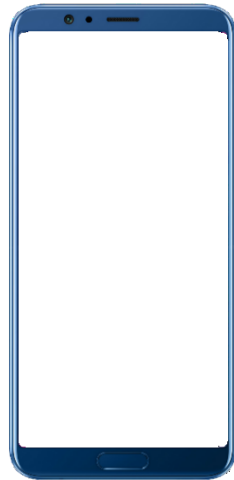
Arquivo.pt – available to anyone

Archived Web pages on mobile devices?

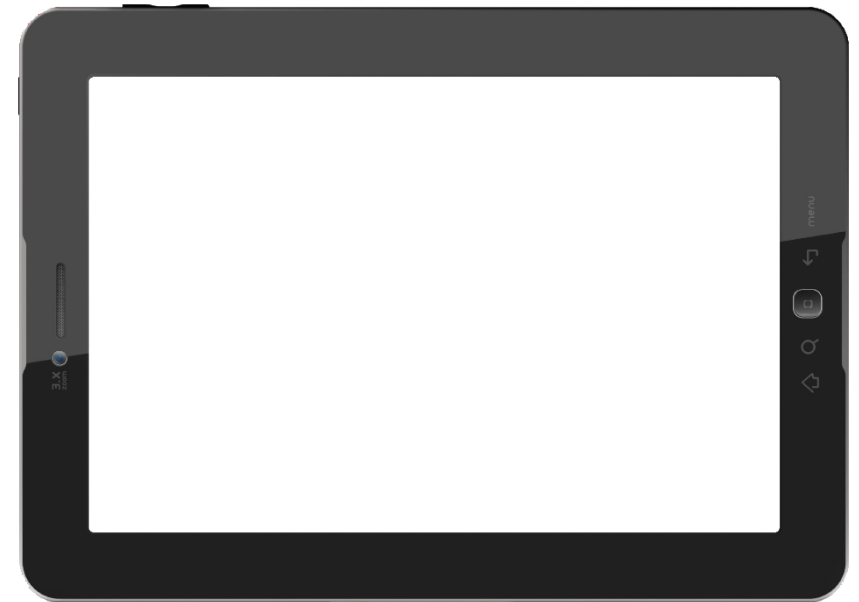
Arquivo.pt – available to anyone

Archived Web pages on mobile devices?

20%



5%



Source: Google Analytics of arquivo.pt

Arquivo.pt – mobile version

[arquivo.pt](#)



[m.arquivo.pt](#)

Arquivo.pt – mobile homepage



Mobile – search by text



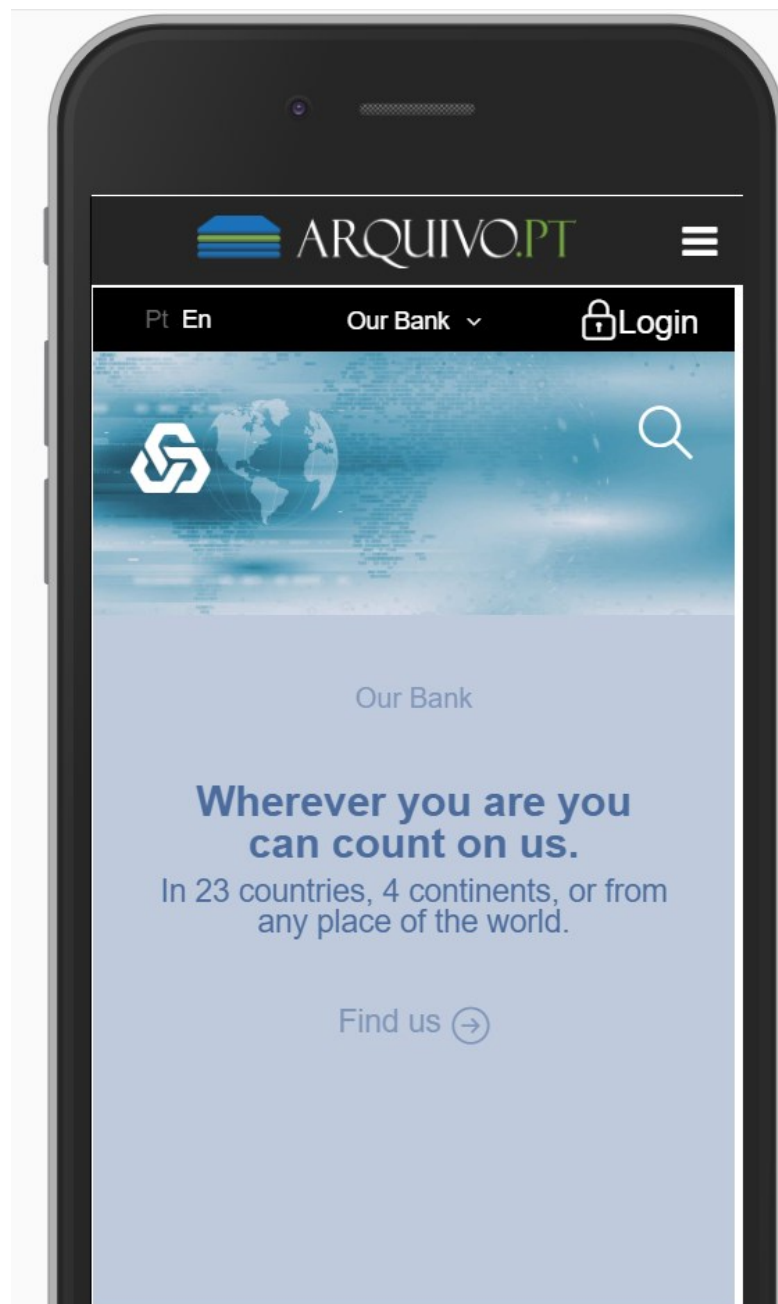
Mobile – advanced text search



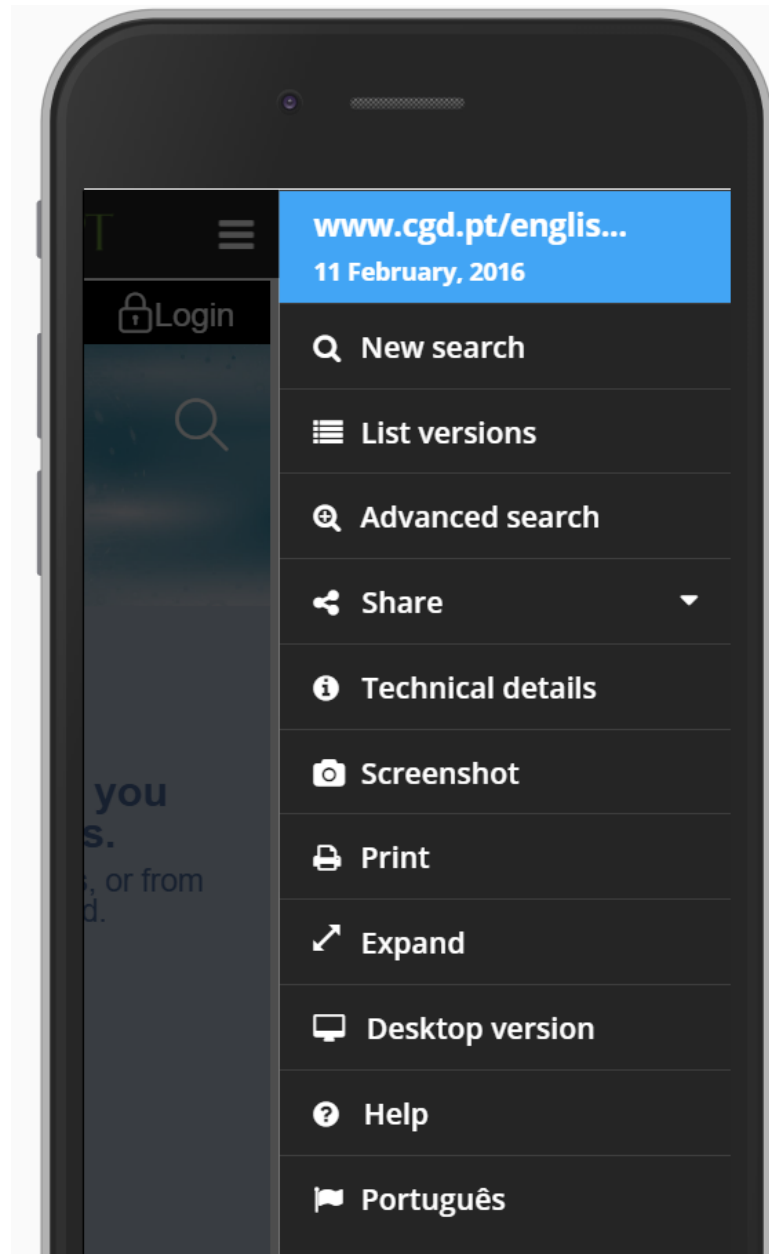
Mobile – search by URL



Mobile – replay archived pages



Mobile – replay menu



1. Mobile version
- 2. Arquivo.pt API**
3. ImageSearch API

Arquivo.pt API

arquivo.pt/api

Arquivo.pt API

arquivo.pt/textsearch

Response in **JSON** format.

URL search



URL search – request

[http://arquivo.pt/textsearch?
versionHistory=http://edition.cnn.com](http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com)

List preserved versions of URL:

[**http://edition.cnn.com**](http://edition.cnn.com)

offset=0 (first result)

maxItems=50 (number of results)

URL search – **response** item

"title": "CNN - Breaking News, U.S., World, Weather, Entertainment & Video News",
"originalURL": "<http://edition.cnn.com/>",
"linkToArchive": "<https://arquivo.pt/wayback/20161031162930/http://edition.cnn.com/>",
"tstamp": "20161031162930",
"contentLength": "33725",
"digest": "IRDFU7WXJY5SXIP5PC476BGW4Z4Q24SF",
"mimeType": "text/html",
"linkToScreenshot": "<https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2F20161031162930/http://edition.cnn.com/>",
"date": "1477931370",
"encoding": "UTF-8",
"linkToNoFrame": "<https://arquivo.pt/noFrame/replay/20161031162930/http://edition.cnn.com/>",
"status": "200",
"collection": "AWP22",
"linkToExtractedText": "<https://arquivo.pt/textextracted?m=http%3A%2F%2Fedition.cnn.com%2F%2F20161031162930/http://edition.cnn.com/>",
"linkToMetadata": "<https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fedition.cnn.com%2F%2F20161031162930/http://edition.cnn.com/>"

URL search - **offset**

[http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com
&**offset=100**](http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com&offset=100)

offset=100 (start in result 100)
maxItems=50 (number of results)

URL search- **maxitems**

[http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com
&**maxItems=1000**](http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com&maxItems=1000)

URL search – **from** and **to**

[http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com](http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com&maxItems=1000)
[&maxItems=1000](#)
[**&from=2010**](#)
[**&to=2016**](#)

URL search – **from** and **to**

[http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com
&from=20010224174130&to=20150424181416](http://arquivo.pt/textsearch?versionHistory=http://edition.cnn.com&from=20010224174130&to=20150424181416)

List versions of URL:

<http://edition.cnn.com>

date between:

24 de February 2001

and:

24 de April 2015

Text search



Text search

<http://arquivo.pt/textsearch?q=euro 2004>

Default parameters:

offset=0 (first result)

itemsPerPage=10 (number of results)

Text search- **mimetype PDF**

[http://arquivo.pt/textsearch?q=euro 2004
&type=pdf](http://arquivo.pt/textsearch?q=euro 2004&type=pdf)

Search with words:

euro 2004

In file format:

PDF

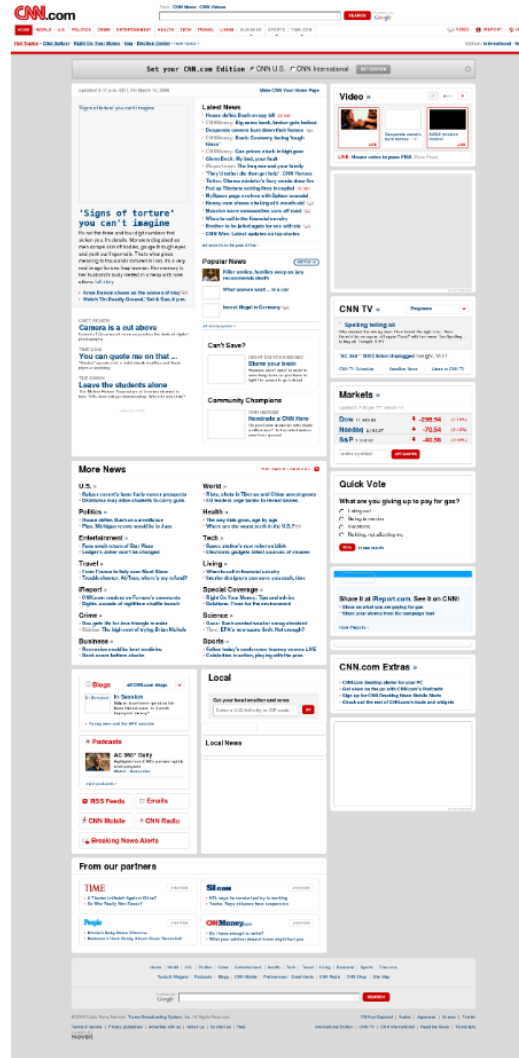
Text search response: **results summary**

```
{  
  "serviceName": "Arquivo.pt - the Portuguese web-archive",  
  "linkToService": "https://arquivo.pt",  
  "next_page": "https://arquivo.pt/textsearch?versionHistory=http://sapo.pt&&prettyPrint=true&offset=50",  
  "previous_page": "https://arquivo.pt/textsearch?versionHistory=http://sapo.pt&&prettyPrint=true&offset=0",  
  "total_items": "3617",  
  "request_parameters": {  
    "from": "19960101000000",  
    "to": "20181231235959",  
    "prettyPrint": "true"  
  }  
}
```

Result Item –text search

```
{
  "title": "DIPLOMA EURO 2004",
  "originalURL": "http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "linkToArchive": "https://arquivo.pt/wayback/20081023024830/http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "tstamp": "20081023024830",
  "contentLength": "64667",
  "digest": "f03e47a98174480c9b13bef89ed2426b",
  "mimeType": "application/pdf",
  "linkToScreenshot": "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20081023024830/http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "date": "1224730110",
  "linkToNoFrame": "https://arquivo.pt/noFrame/replay/20081023024830/http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "snippet": "DIPLOMA <em>EURO</em> <em>2004</em>\nContactos com esta&ccedil;&otilde;es de PORTUGAL  
(Continente e Regi&otilde;es Aut&oacute;nomas dos\nA&ccedil;ores e<span class=\\\"ellipsis\\\"> ... </span>  
UTILIZAR DURANTE O <em>EURO</em> <em>2004</em>\nTodos os Radioamadores portugueses, poder&atilde;o (e dev  
solicitar &agrave; ANACOM\num indicativo especial para utilizar durante o per&iacute;odo da realiza&ccedil;&atilde;o do .  
</em>\n<em>2004</em>,  
mediante<span class=\\\"ellipsis\\\"> ... </span>",
  "collection": "AWP3",
  "linkToExtractedText": "https://arquivo.pt/textextracted?m=http%3A%2F%2Fct1end.netpower.pt%2Fdiplomas%2Fdiploma\_euro2004\_portugues.pdf",
  "linkToMetadata": "https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fct1end.netpower.pt%2Fdiplomas%2Fdiploma\_euro2004\_portugues.pdf",
}
```

Result Item –linkToScreenshot



Result Item –text search

```
{
  "title": "DIPLOMA EURO 2004",
  "originalURL": "http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "linkToArchive": "https://arquivo.pt/wayback/20081023024830/http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "tstamp": "20081023024830",
  "contentLength": "64667",
  "digest": "f03e47a98174480c9b13bef89ed2426b",
  "mimeType": "application/pdf",
  "linkToScreenshot": "https://arquivo.pt/screenshot?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20081023024830/http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "date": "1224730110",
  "linkToNoFrame": "https://arquivo.pt/noFrame/replay/20081023024830/http://ct1end.netpower.pt/diplomas/diploma\_euro2004\_portugues.pdf",
  "snippet": "DIPLOMA <em>EURO</em> <em>2004</em>\nContactos com esta&ccedil;&otilde;es de PORTUGAL  
(Continente e Regi&otilde;es Aut&oacute;nomas dos\nA&ccedil;ores e<span class=\\\"ellipsis\\\"> ... </span>  
UTILIZAR DURANTE O <em>EURO</em> <em>2004</em>\nTodos os Radioamadores portugueses, poder&atilde;o (e dev  
solicitar &agrave; ANACOM\num indicativo especial para utilizar durante o per&iacute;odo da realiza&ccedil;&atilde;o do  
</em>\n<em>2004</em>,  
mediante<span class=\\\"ellipsis\\\"> ... </span>",
  "collection": "AWP3",
  "linkToExtractedText": "https://arquivo.pt/textextracted?m=http%3A%2F%2Fct1end.netpower.pt%2Fdiplomas%2Fdiploma\_euro2004\_portugues.pdf",
  "linkToMetadata": "https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fct1end.netpower.pt%2Fdiplomas%2Fdiploma\_euro2004\_portugues.pdf",
}
```

Result Item –snippet

1996

2017

CNN.com - Breaking News, U.S., World, Weather, Entertainment...
<http://www.cnn.com/>

CNN.com - Breaking News, U.S., World, Weather, Entertainment & Video News Web | CNN News | CNN ...
14 Março, 2008 - [Listar versões](#)

Breaking News, U.S., World, Weather, Entertainment & Video
<http://edition.cnn.com/>

Breaking News, U.S., World, Weather, Entertainment & Video News - CNN.com Breaking News Toggle ...
14 Agosto, 2015 - [Listar versões](#)

CNN Breaking News (cnnbrk) on Twitter
<http://twitter.com/cnnbrk>

CNN Breaking News (cnnbrk) on Twitter Skip past navigation On a mobile phone? Check out m.twitter ...
19 Dezembro, 2009 - [Listar versões](#)

Google News
<http://news.google.com/>

Santa Clara County San Jose Mercury News The Associated Press - Seattle Times - Reuters - CNN all 1,585 news articles » Sky News
Mother, son missing in forced chemotherapy case CNN - 9 hours ago ...
25 Junho, 2009 - [Listar versões](#)

```
{
  "title": "CNN.com - Breaking News, U.S., World, Weather, Entertainment & Video News",
  "originalURL": "http://www.cnn.com/",
  "linkToArchive": "https://arquivo.pt/wayback/20080314183530/http://www.cnn.com/",
  "tstamp": "20080314183530",
  "contentLength": "91841",
  "digest": "08b05035e0fc42adbb999f9fa39e458a",
  "mimeType": "text/html",
  "linkToScreenshot": "https://arquivo.pt/screenshot/?url=https%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20080314183530/http://www.cnn.com/",
  "date": "1205519730",
  "encoding": "windows-1252",
  "linkToNoFrame": "https://arquivo.pt/noFrame/replay/20080314183530/http://www.cnn.com/",
  "snippet": "<em>CNN</em>.com - Breaking <em>News</em>, U.S., World, Weather, Entertainment & Video <em>Morning</em><span class=\"ellipsis\"> ... </span>",
  "collection": "AWP2",
  "linkToExtractedText": "https://arquivo.pt/textextracted?m=http%3A%2F%2Fwww.cnn.com%2F%2F20080314183530",
  "linkToMetadata": "https://arquivo.pt/textsearch?metadata=http%3A%2F%2Fwww.cnn.com%2F%2F20080314183530"
}
```

Download text of URL

[http://arquivo.pt/textextracted?
m=edition.cnn.com/20080215055940](http://arquivo.pt/textextracted?m=edition.cnn.com/20080215055940)

Extracted text

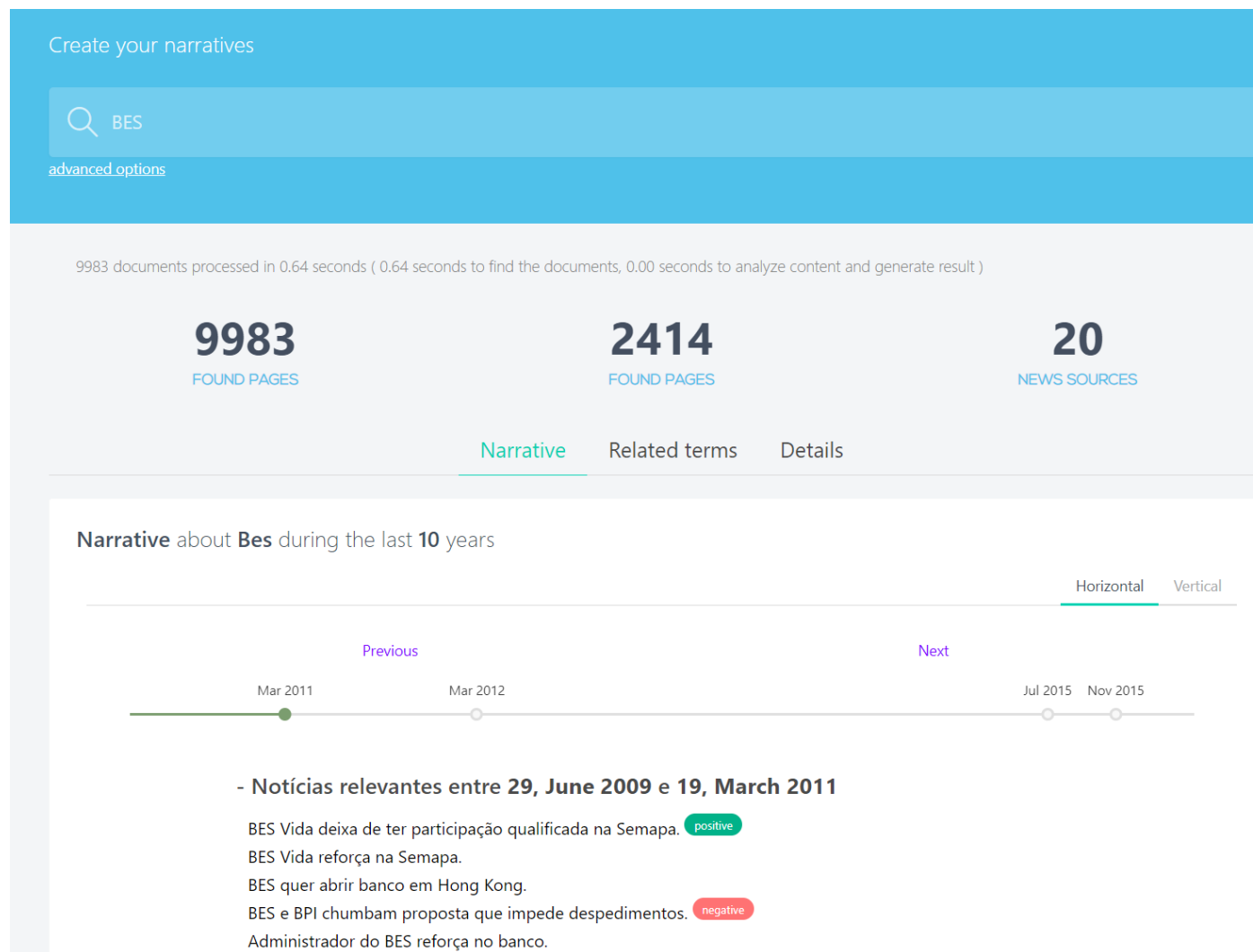
edition.cnn.com-20080215055940 - Notepad

File Edit Format View Help

CNN.com International Home Asia Europe U.S. World World Business Technology Entertainment World Sport Travel
On TV Video I-Report CNN Mobile Hot Topics » Russia • South Korea • East Timor • America votes • Digital Biz •
more topics » Edition: U.S. | Arabic | Set Pref February 15, 2008 -- Updated 0536 GMT (1336 HKT) Make CNN Your
Home Page DEVELOPING STORY Gunman kills five at U.S. university A gunman has shot 22 people at a university in the
U.S. state of Illinois, killing five, before fatally shooting himself, police said.
The gunman used a shotgun and two handguns, according to police. developing story 'Thinking I was going to die' | Map 'Run!
Serbia defiant as Kosovo breakaway looms all most popular » Top Stories Chad declares state of emergency East Timor issues
h East Timor China promises \$1.25B in storm aid Europe » Serbia: Kosovo breakaway 'illegal' Pilot jailed over 9/11 can clai
le at satellite SciTechBlog: The sky is falling Funny News » Governor's son sells 'Don't Drop the Soap' Cops snag lotto tic
ia appeals to the U.N. 1:47 Former terror suspect 2:58 Political sport 2:06 Arab voices 5:08 Serbia's argument against bid
l turmoil - now, it's time to vote. How will Pakistan's vote affect the nation and the war on terror? TV Schedule | What's
rquee New You Buddy Blog Paging Dr. Gupta Political Ticker QuickVote QuickThoughts Salute to Troops Situation Room Warp CNN
CNN Radio | CNNAvantGo | Site Map CNN en Español | Arabic | Japanese | Korean | Turkish CNN U.S. | CNN U.S. T

Arquivo.pt API usecases

contamehistorias.pt/arquivopt/?lang_code=en



arquivodeopiniao.pt/en/author-search/

 Search by Author Keyword About pt

Authors



Direção

No. of Articles: 2980

Newspapers: Público | Diário de Notícias | Expresso | Sábado | Jornal i



Miguel Esteves Cardoso

No. of Articles: 2675

Newspapers: Público



Fernando Sobral

No. of Articles: 2358

Newspapers: Jornal de Negócios | Correio da Manhã



Camilo Lourenço

No. of Articles: 1724

Newspapers: Jornal de Negócios



Ferreira Fernandes

No. of Articles: 1571

Newspapers: Diário de Notícias



Helena Garrido

No. of Articles: 1432

Newspapers: Jornal de Negócios | Sábado | Correio da Manhã



Vasco Pulido Valente

No. of Articles: 1025

Newspapers: Público



Pedro Santos Guerreiro

No. of Articles: 1002

Newspapers: Jornal de Negócios | Expresso



Rui Tavares

No. of Articles: 829

Newspapers: Público



Armando Esteves Pereira

No. of Articles: 730

Newspapers: Correio da Manhã | Jornal de Negócios



[mais](#)

1. Mobile version
2. Arquivo.pt API
- 3. ImageSearch API**

arquivo.pt/imagesearch

Text to image search API

[http://arquivo.pt/imagesearch?
q=Lionel Messi](http://arquivo.pt/imagesearch?q=Lionel+Messi)

Search for images related with words
Lionel and **Messi**

```
{
  "imgSrcURLDigest": "05afdee1c01cd6793b88bee117d109408b127db0a9c05fbdccc003dea03be51a",
  "imgWidth": 140.0,
  "pageImages": 93,
  "imgDigest": "15dbcc9517556211ab8e639523968adc605dab4cd394a8afd3b7a43bafa5b4a9",
  "imgTstamp": 20121210173543,
  "pageTitle": "Latest news, sport and comment from the Guardian | The Guardian",
  "imgMimeType": "image/jpeg",
  "pageURLTokens": "http www guardian co uk ",
  "safe": 0.001,
  "pageTstamp": 20121210173250,
  "imgSrcBase64": "/9j/4AAQSkZJRgABAQAAQABAAD/2wBDAAgGBgcGBQgHBwcJCQgKDBQNDAsLDBkSEw8UHRofH...",
  "pageProtocol": "http",
  "collection": "FAWP11",
  "imgHeight": 130.0,
  "imgSrcTokens": "http static guim co uk sys images Football Pix pictures
  2011 4 6 1302108418156 Lionel Messi 002 jpg ",
  "imgAlt": " Lionel Messi",
  "pageHost": "www.guardian.co.uk",
  "pageURL": "http://www.guardian.co.uk/",
  "imgSrc": "http://static.guim.co.uk/sys-images/Football/Pix/pictures/2011/4/6/1302108418156/-Lionel-Messi-002.jpg",
},
```

[http://arquivo.pt/imagesearch?
q=Euro 2004&type=png](http://arquivo.pt/imagesearch?q=Euro 2004&type=png)

Search for images related with words **Euro 2004**,
in **PNG** image format.

[http://arquivo.pt/imagesearch?
q=Euro 2004&size=sm](http://arquivo.pt/imagesearch?q=Euro+2004&size=sm)

Search for images, with size **small**
related with words
Euro 2004,

size=md (medium image size)
size=lg (large image)

Lionel Messi



Search

between:

01/01/1996



and:

31/12/2017



[Advanced image search](#)

Web

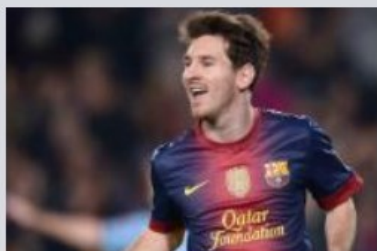
Images

Tools

experimental



www.sabado.pt/Mul...
25 November, 2012



www.abola.pt/nnh/...
24 November, 2012



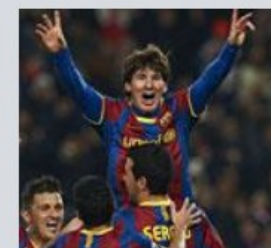
www.apostasdespor...
21 May, 2011



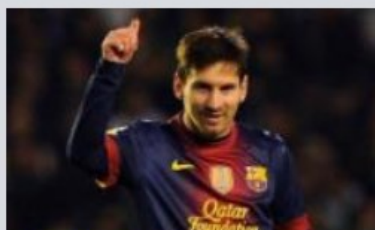
www.record.xl.pt/...
25 October, 2012



www.acorianoorien...
25 October, 2012



www.guardian.co.uk/
10 December, 2012



www.bbc.co.uk/
10 December, 2012



www.sfondi-deskto...
22 November, 2014



www.wallstock.eu/...
23 November, 2014



schweiz.myspace.c...
5 July, 2011



caras.sapo.pt/fam...
1 January, 2013



www.futebol365.pt...
12 November, 2016



www.futebol365.pt...
15 October, 2014



schweiz.myspace.c...
5 July, 2011



caras.sapo.pt/fam...
1 January, 2013



www.futebol365.pt...
12 November, 2016



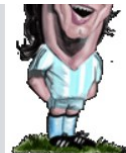
www.futebol365.pt...
15 October, 2014



www.futebol365.pt...
22 October, 2014



www.tvi24.iol.pt/...
5 October, 2012



karikamania.no.sa...
24 June, 2007



m.abola.pt/notici...
10 November, 2016



jogos-de-vestir.1...
15 November, 2013



www.garagetv.be/v...
6 August, 2010



sport.id.msn.com/...
28 December, 2009



flagra.pt/date/20...
22 October, 2014



jogosonlinegratis...
6 September, 2014



www.futebol365.pt...
12 November, 2016



www.futebol365.pt...
12 November, 2016



expresso.sapo.pt/...
24 October, 2014



www.futebol365.pt...
22 October, 2014



www.maisdesporto...
15 October, 2009



www.cmjornal.xl.p...
23 October, 2012



www.publico.pt/mu...
31 December, 2012



Page

Ronaldo | Mais Desporto - Todas as mo...
<http://www.maisdesporto.net/?tag=ronaldo>
10 October, 2009

Image

Messi: Votará no Cristiano Ronaldo pa...
jpeg 500 x 400
15 October, 2009

Visit
page

Show
image

Show
details

Share

Conclusions

100,000

Conclusions

Open access for everyone!

Allow cooperation

Increasing number of users

Mobile users also want access to preserved Web pages

Encourage people to use Arquivo.pt for research - through APIs and contests

Thank you



<http://arquivo.pt/api>

Fernando Melo <fernando.melo@fccn.pt>