Automatic Processing

fernando.melo@fccn.pt

Tutorial outline

Research the Past Web using Web archives

1. Search and access
   - The Past Web: examples and use cases
   - Public online services
2. Publish and preserve
   - Recommendations to publish preservable information
   - How to preserve information collected from the Web
3. **Automatic processing**
   - Interoperability protocols
   - Application Programming Interfaces (API)

arquivo.pt/training

Wayback Syntax

Fast access to Web archived content

Wayback Syntax – it works on other web archives

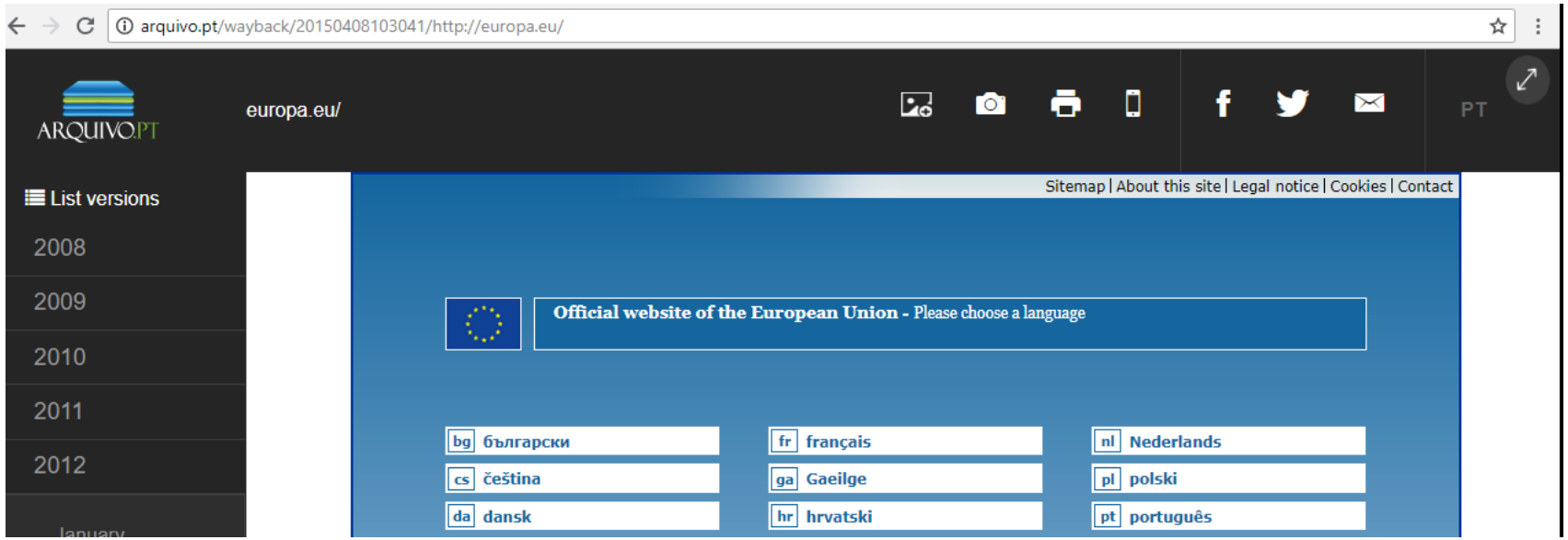
http://arquivo.pt/wayback/{**Timestamp**}/{**URL**}

http://web.archive.org/web/{**Timestamp**}/{**URL**}

http://webarchive.org.uk/wayback/archive/{**Timestamp**}/{**URL**}

Wayback Syntax

http://arquivo.pt/wayback/{**Timestamp**}/{**URL**}



{**Timestamp**} = 20150408103041
{**URL**} = http://europa.eu

Wayback Syntax

http://arquivo.pt/wayback/{**Timestamp**}/{**URL**}

14 Digits

{**Timestamp**} = 2015 04 08 10 30 41

Wayback Syntax

http://arquivo.pt/wayback/{**Timestamp**}/{**URL**}

14 Digits

{Timestamp} = 2015 04 08 10 30 41

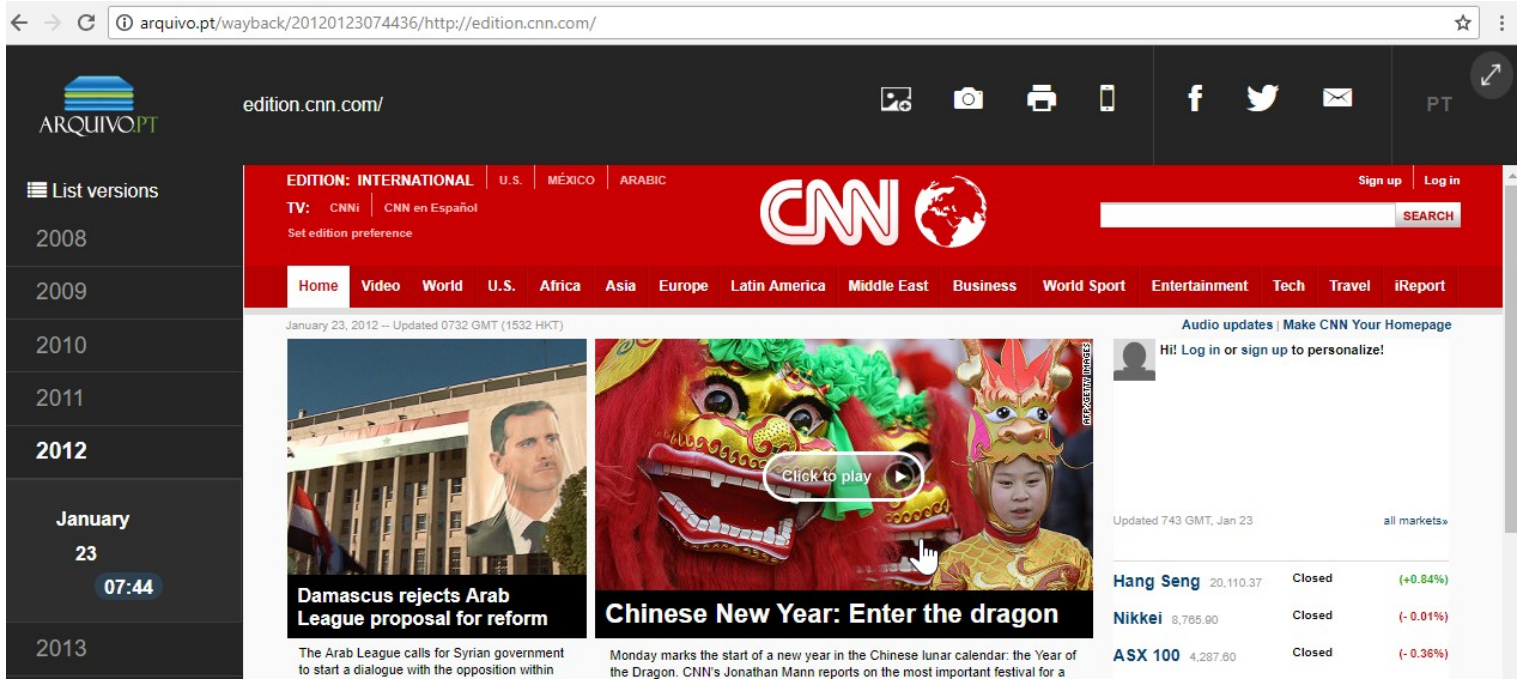
year
month
day
hours
minutes
seconds

Wayback Syntax - closest date

What if we don’t know the exact timestamp?

Wayback Syntax – closest date

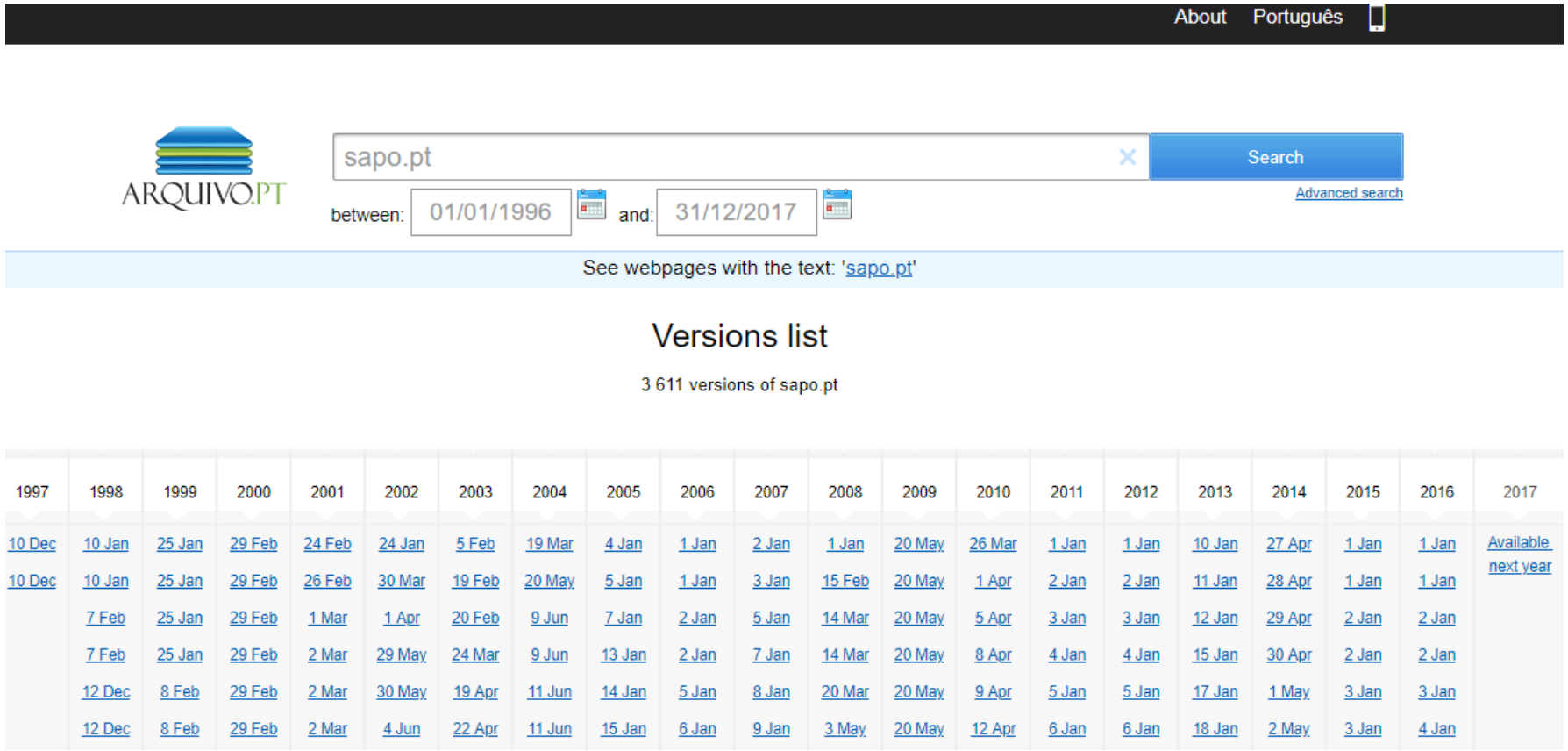
http://arquivo.pt/wayback/**20120000000000**/edition.cnn.com



http://arquivo.pt/wayback/**20120123074436**/http://edition.cnn.com/

Wayback Syntax – list all archived versions

http://arquivo.pt/wayback/*/sapo.pt



1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
10 Dec	10 Jan	25 Jan	29 Feb	24 Feb	24 Jan	5 Feb	19 Mar	4 Jan	1 Jan	2 Jan	1 Jan	20 May	26 Mar	1 Jan	1 Jan	10 Jan	27 Apr	1 Jan	1 Jan	Available next year
10 Dec	10 Jan	25 Jan	29 Feb	26 Feb	30 Mar	19 Feb	20 May	5 Jan	1 Jan	3 Jan	15 Feb	20 May	1 Apr	2 Jan	2 Jan	11 Jan	28 Apr	1 Jan	1 Jan	
	7 Feb	25 Jan	29 Feb	1 Mar	1 Apr	20 Feb	9 Jun	7 Jan	2 Jan	5 Jan	14 Mar	20 May	5 Apr	3 Jan	3 Jan	12 Jan	29 Apr	2 Jan	2 Jan	
	7 Feb	25 Jan	29 Feb	2 Mar	29 May	24 Mar	9 Jun	13 Jan	2 Jan	7 Jan	14 Mar	20 May	8 Apr	4 Jan	4 Jan	15 Jan	30 Apr	2 Jan	2 Jan	
	12 Dec	8 Feb	29 Feb	2 Mar	30 May	19 Apr	11 Jun	14 Jan	5 Jan	8 Jan	20 Mar	20 May	9 Apr	5 Jan	5 Jan	17 Jan	1 May	3 Jan	3 Jan	
	12 Dec	8 Feb	29 Feb	2 Mar	4 Jun	22 Apr	11 Jun	15 Jan	6 Jan	9 Jan	3 May	20 May	12 Apr	6 Jan	6 Jan	18 Jan	2 May	3 Jan	4 Jan	

Wayback Syntax - challenge

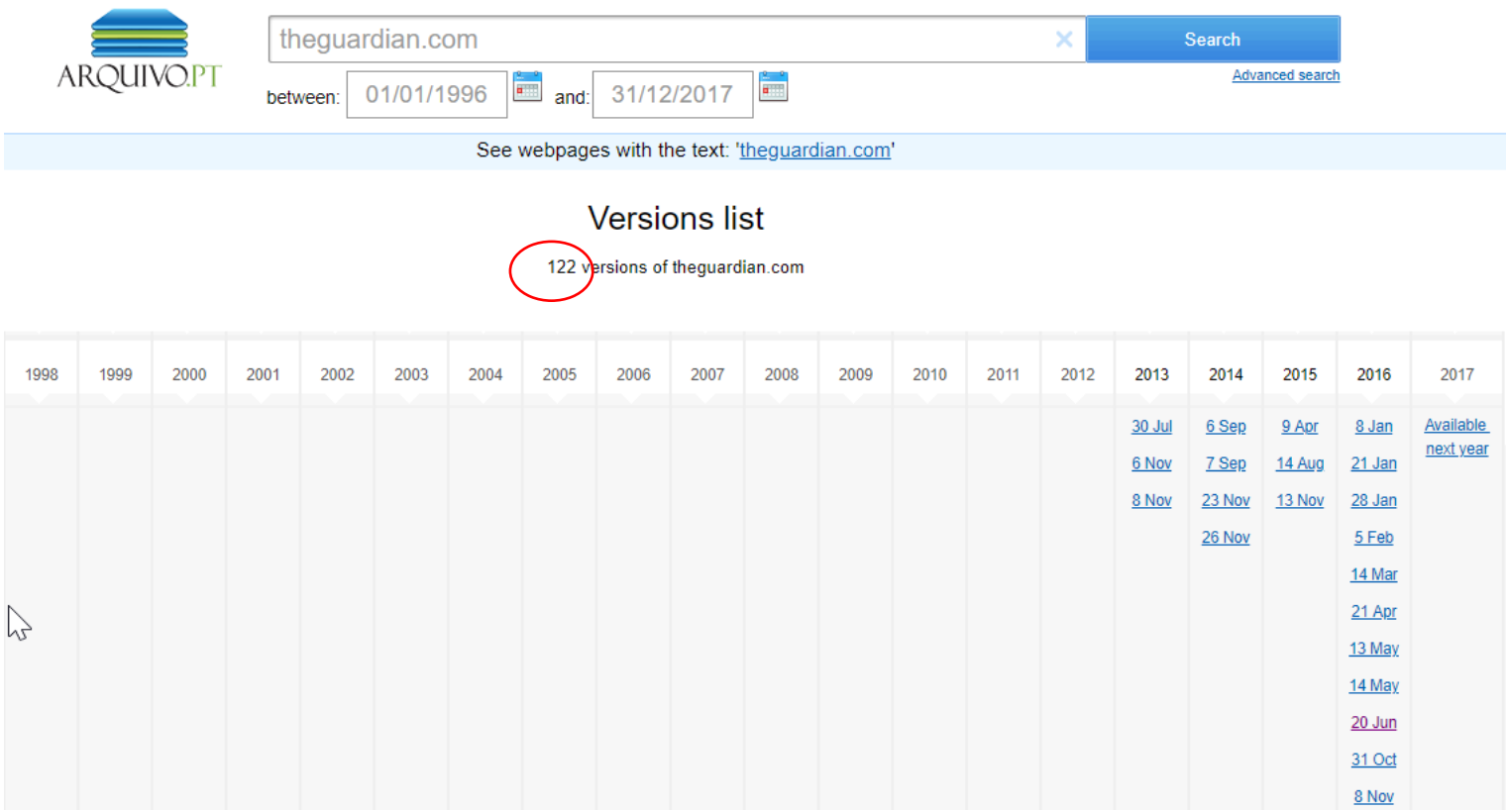
1. Closest version: October 2008 of the URL nytimes.com
2. Number of archived versions of the website theguardian.com

Wayback Syntax – challenge 1

http://arquivo.pt/wayback/**20081000000000**/nytimes.com

Wayback Syntax – challenge 2

http://arquivo.pt/wayback/*/theguardian.com



Wayback Syntax

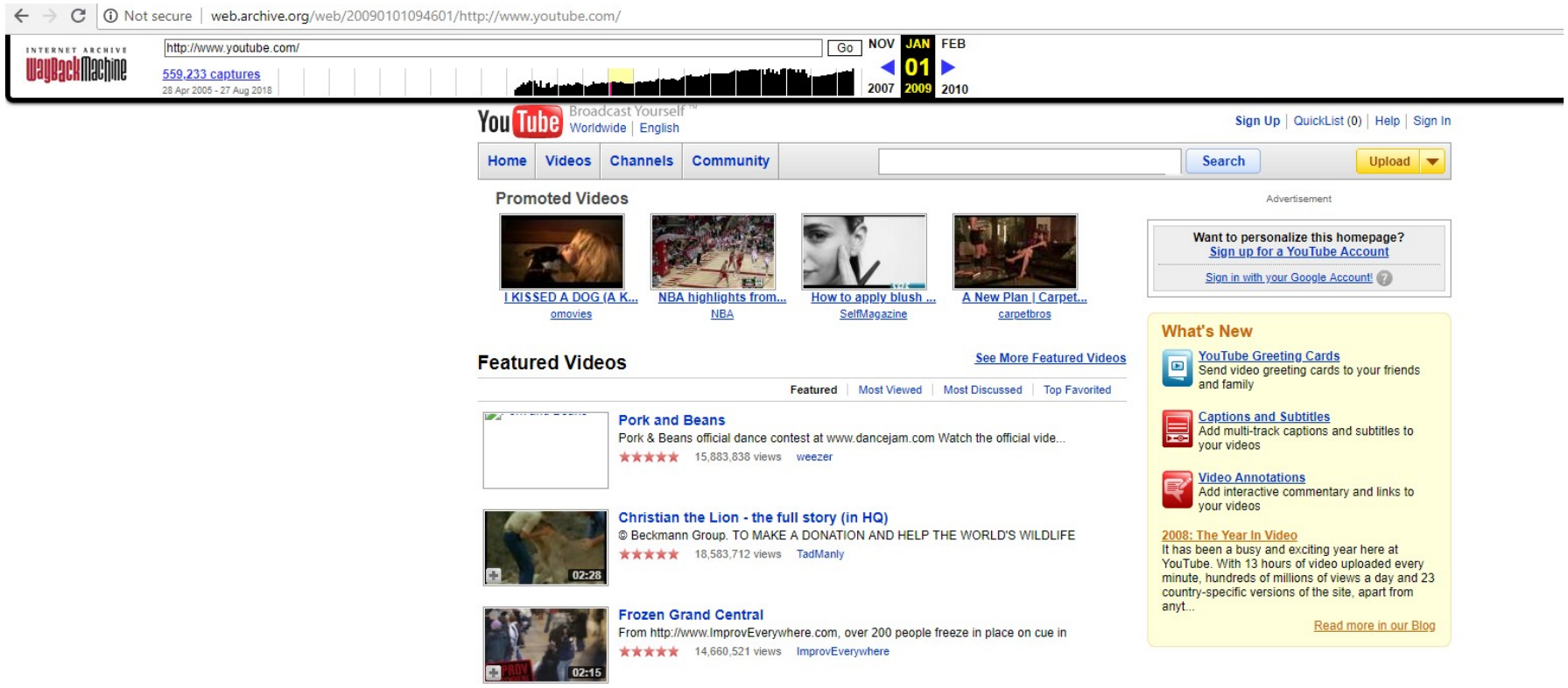
How about other Web archives?

Wayback Syntax - challenge

1. Closest version: 2009 of the URL youtube.com in Internet Archive
2. List all archived versions of the website gov.uk in UK Web Archive

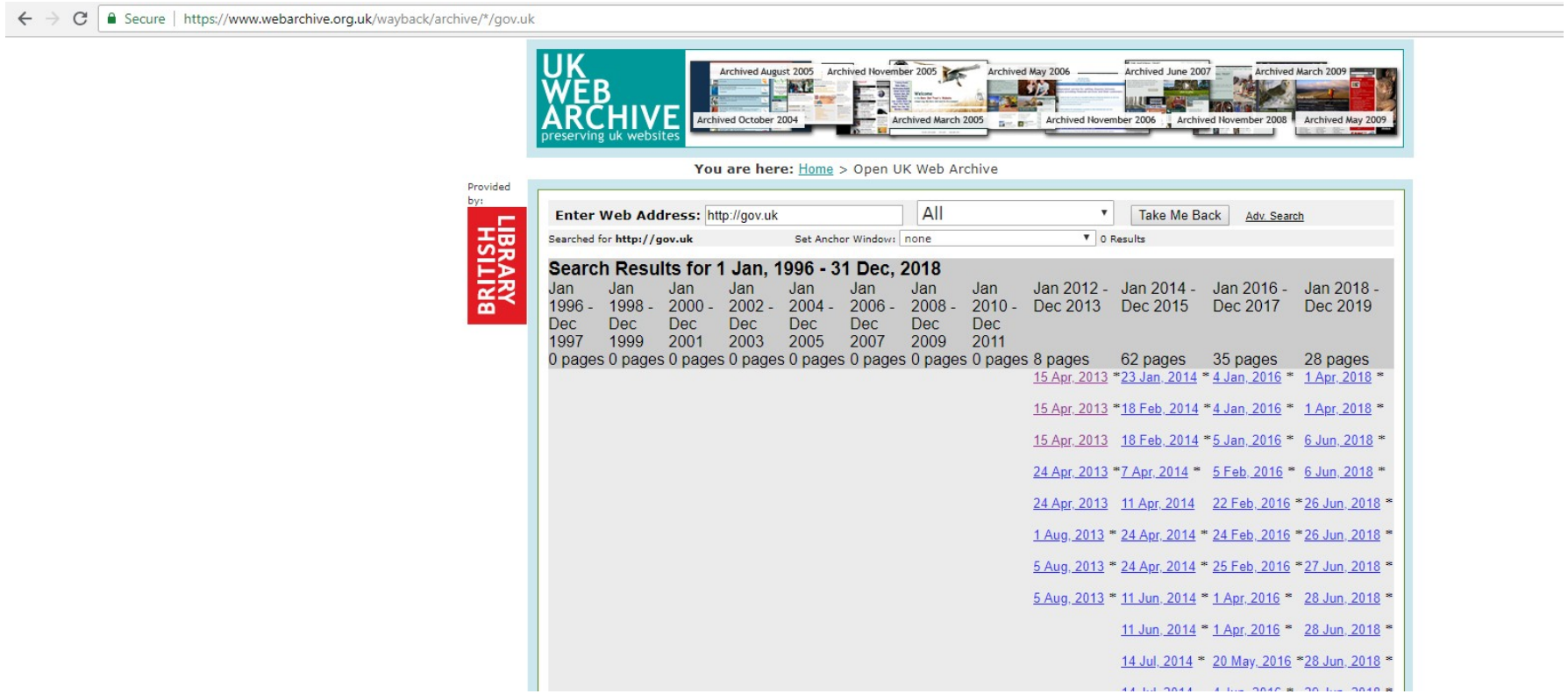
Wayback Syntax – challenge 1

http://web.archive.org/web/20090000000000/youtube.com



Wayback Syntax – challenge 2

http://webarchive.org.uk/wayback/archive/*/gov.uk



Application Programming Interfaces (APIs)

API – Application Programming Interface

Automatic access

Easy integration

Fast development of new applications

No need to understand core code

Web Archives APIs

Arquivo.pt API

+ Search by text or by URL
- Only works on Arquivo.pt

Memento TimeTravel API

+ Search in several Web archives
- Only search by URL

Arquivo.pt API

ARQUIVO.PT

Arquivo.pt API

arquivo.pt/textsearch

Automatic access

URL search

Text search

Metadata search

Arquivo.pt API

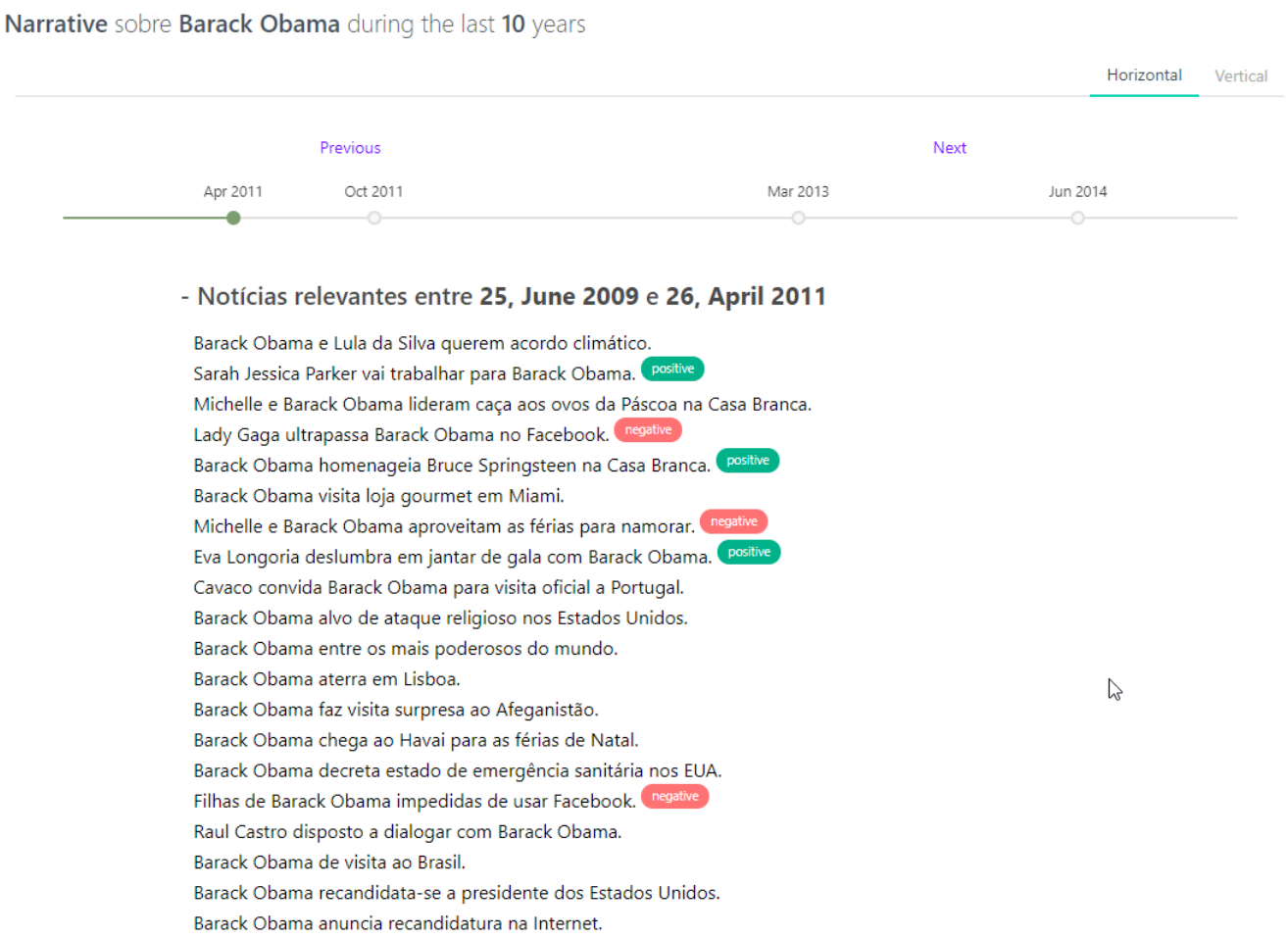
arquivo.pt/textsearch

API response in **JSON** format.

Arquivo.pt API use case

http://contamehistorias.inesctec.pt/arquivopt/?lang_code=en

Arquivo.pt API use case



URL search

(Arquivo.pt API)

The New York Times

nytimes.com

URL search **request**

arquivo.pt/textsearch?**versionHistory=nytimes.com**
&prettyPrint=true

List archived versions of the URL:

nytimes.com

offset=0 (first result)

maxItems=50 (number of results)

URL search **response**


```
{
  "title": "The New York Times - Breaking News, World News & Multimedia",
  "originalURL": "http://www.nytimes.com/",
  "linkToArchive": "http://arquivo.pt/wayback/20141223183313/http://www.nytimes.com/",
  "tstamp": "20141223183313",
  "contentLength": "38701",
  "digest": "HCUQ77BX4OWLNAU6VLMF7CV6RHIT3ZJ5",
  "mimeType": "text/html",
  "linkToScreenshot": "http://arquivo.pt/screenshot/?url=http%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20141223183
  "date": "1419359593",
  "encoding": "UTF-8",
  "linkToNoFrame": "http://arquivo.pt/noFrame/replay/20141223183313/http://www.nytimes.com/",
  "status": "200",
  "collection": "FAWP1920141223",
  "linkToExtractedText": "http://arquivo.pt/textextracted?m=http%3A%2F%2Fwww.nytimes.com%2F%2F20141223183313",
  "linkToMetadata": "http://arquivo.pt/textsearch?metadata=http%3A%2F%2Fwww.nytimes.com%2F%2F20141223183313"
}
```

URL search request parameter: **offset**

arquivo.pt/textsearch?**versionHistory=nytimes.com&offset=50**

offset=50 (firstresult)

maxItems=50 (number of results)

URL search request: **maxItems**

arquivo.pt/textsearch?versionHistory=nytimes.com
&maxItems=1000
&from=2012
&to=2014

maxItems=1000 (number of results)

URL search request: **from**

arquivo.pt/textsearch?versionHistory=nytimes.com
&**from=20100224174130**

List archived versions of URL

nytimes.com

With date equal or after:

2010 February 24, at 17h41m30s

URL search request: **from to**

arquivo.pt/textsearch?versionHistory=nytimes.com
&**from=20100224174130&to=20150223183001**

List archived versions of URL

nytimes.com

With date between:

2010 February 24, at 17h41m30s

And

2015 February 23, at 18h30m01s

URL search request: **fields**

arquivo.pt/textsearch?versionHistory=nytimes.com
&**fields=originalURL,tstamp**

Filter response to show only

URL of the page

Timestamp (**tstamp**) when it was preserved

URL search response: **fields**


```
{
  "originalURL": "http://nytimes.com/",
  "tstamp": "20161231190553"
},
{
  "originalURL": "https://www.nytimes.com/",
  "tstamp": "20161231180537"
},
{
  "originalURL": "http://www.nytimes.com/",
  "tstamp": "20161231180218"
}
```

Text search

(Arquivo.pt API)

© 2002 UEFA TM
UEFA
Euro 2004
PORTUGAL

Text search **request**

arquivo.pt/textsearch?**q=euro 2004**
&prettyPrint=true

Search words **euro** and **2004.**

offset=0 (first result)

maxItems=50 (number of results)

Text search request: **expression**

arquivo.pt/textsearch?q=**“euro 2004”**

Search results with expression
“euro2004”

RICARDO
1
ING DIRECT
EURO2004.com
T-Mobile
i'm lovin' it

1
EURO

Text search request: **exclude word**

arquivo.pt/textsearch?q=**euro 2004 -currency**

Search results with words

euro 2004

without word

currency

Text search request: **type**

arquivo.pt/textsearch?q=euro 2004&**type=pdf**

Search results with words

euro 2004

In files of type:

PDF

Text search response: **summary of results**


```
{
  "serviceName": "Arquivo.pt - the Portuguese web-archive",
  "linkToService": "http://arquivo.pt",
  "next_page": "http://arquivo.pt/textsearch?q=euro%202004&prettyPrint=true&offset=50",
  "previous_page": "http://arquivo.pt/textsearch?q=euro%202004&prettyPrint=true&offset=0",
  "total_items": "21664141",
  "request_parameters": {
    "q": "euro 2004",
    "prettyPrint": "true"
  },
```

Text search response: **result item**


```
{
  "title": "Euro 2004 - Estádios do Euro 2004",
  "originalURL": "http://helder2004.no.sapo.pt/",
  "linkToArchive": "http://arquivo.pt/wayback/20091217183109/http://helder2004.no.sapo.pt/",
  "tstamp": "20091217183109",
  "contentLength": "1718",
  "digest": "7bb196359aef86a45767a34b21a78a97",
  "mimeType": "text/html",
  "linkToScreenshot": "http://arquivo.pt/screenshot/?url=http%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20091217183109
  "date": "1261074669",
  "encoding": "windows-1252",
  "linkToNoFrame": "http://arquivo.pt/noFrame/replay/20091217183109/http://helder2004.no.sapo.pt/",
  "snippet": "<em>Euro</em> <em>2004</em> - Est&aacute;dios do <em>Euro</em>
   <em>2004</em>   Aqui Pode Encontrar Todos Os Palcos Do <em>Euro</em> <em>2004</em>
       ENTRAR<span class=\"ellipsis\"> ... </span>",
  "collection": "AWP6",
  "linkToExtractedText": "http://arquivo.pt/textextracted?m=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109",
  "linkToMetadata": "http://arquivo.pt/textsearch?metadata=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109"
}
```

Text search response field: **linkToMetadata**


```
{
 "title": "Euro 2004 - Estádios do Euro 2004",
 "originalURL": "http://helder2004.no.sapo.pt/",
 "linkToArchive": "http://arquivo.pt/wayback/20091217183109/http://helder2004.no.sapo.pt/",
 "tstamp": "20091217183109",
 "contentLength": "1718",
 "digest": "7bb196359aef86a45767a34b21a78a97",
 "mimeType": "text/html",
 "linkToScreenshot": "http://arquivo.pt/screenshot/?url=http%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20091217183109
 "date": "1261074669",
 "encoding": "windows-1252",
 "linkToNoFrame": "http://arquivo.pt/noFrame/replay/20091217183109/http://helder2004.no.sapo.pt/",
 "snippet": "<em>Euro</em> <em>2004</em> - Est&aacute;dios do <em>Euro</em>
  <em>2004</em>   Aqui Pode Encontrar Todos Os Palcos Do <em>Euro</em> <em>2004</em>
      ENTRAR<span class=\"ellipsis\"> ... </span>",
 "collection": "AWP6",
 "linkToExtractedText": "http://arquivo.pt/textextracted?m=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109",
 "linkToMetadata": "http://arquivo.pt/textsearch?metadata=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109"
}
```

Metadata search response


```
{
  "title": "Euro 2004 - Estádios do Euro 2004",
  "originalURL": "http://helder2004.no.sapo.pt/",
  "linkToArchive": "http://arquivo.pt/wayback/20091217183109/http://helder2004.no.sapo.pt/",
  "tstamp": "20091217183109",
  "contentLength": "1203",
  "digest": "EYUZNPAREQBWSZLOMRM7FKJSHJXNQWTZ",
  "mimeType": "text/html",
  "linkToScreenshot": "http://arquivo.pt/screenshot/?url=http%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20091217183109
  "date": "1261074669",
  "encoding": "windows-1252",
  "linkToNoFrame": "http://arquivo.pt/noFrame/replay/20091217183109/http://helder2004.no.sapo.pt/",
  "status": "200",
  "collection": "AWP6",
  "linkToExtractedText": "http://arquivo.pt/textextracted?m=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109",
  "linkToMetadata": "http://arquivo.pt/textsearch?metadata=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109",
  "filename": "IAH-20091217182045-00726-awp01.fccn.pt.arc.gz",
  "offset": "99075262"
}
```

Text search response: **linkToExtractedText**


```
{
  "title": "Euro 2004 - Estádios do Euro 2004",
  "originalURL": "http://helder2004.no.sapo.pt/",
  "linkToArchive": "http://arquivo.pt/wayback/20091217183109/http://helder2004.no.sapo.pt/",
  "tstamp": "20091217183109",
  "contentLength": "1718",
  "digest": "7bb196359aef86a45767a34b21a78a97",
  "mimeType": "text/html",
  "linkToScreenshot": "http://arquivo.pt/screenshot/?url=http%3A%2F%2Farquivo.pt%2FnoFrame%2Freplay%2F20091217183109
  "date": "1261074669",
  "encoding": "windows-1252",
  "linkToNoFrame": "http://arquivo.pt/noFrame/replay/20091217183109/http://helder2004.no.sapo.pt/",
  "snippet": "<em>Euro</em> <em>2004</em> - Est&aacute;dios do <em>Euro</em>
   <em>2004</em>   Aqui Pode Encontrar Todos Os Palcos Do <em>Euro</em> <em>2004</em>
       ENTRAR<span class=\"ellipsis\"> ... </span>",
  "collection": "AWP6",
  "linkToExtractedText": "http://arquivo.pt/textextracted?m=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109",
  "linkToMetadata": "http://arquivo.pt/textsearch?metadata=http%3A%2F%2Fhelder2004.no.sapo.pt%2F%2F20091217183109"
}
```

linkToExtractedText: downloads as txt file

linkToExtractedText: txt file

http_--www.nytimes.com--20160620215512 - Notepad

File Edit Format View Help

```
The New York Times - Breaking News, World News & Multimedia NYTimes.com no longer supports Internet
Learned From Joe McCarthy’s Top Aide By JONATHAN MAHLER and MATT FLEGENHEIMER Roy Cohn, a fearsome l
Durayhib Military Base What the Islamic State Has Won and Lost By SARAH ALMUKHTAR, TIM WALLACE and D
Allows Use of Evidence Found After Illegal Stops By ADAM LIPTAK 1:52 PM ET The 5-3 ruling says such
gation 1:32 PM ET Officials Say Kabul Bombing Kills Security Contractors 12:18 PM ET Dustin Johnson
low: The G.O.P.’s Cynical Gay Ploy Cohen: Jo Cox and Britain’s Place in Europe Krugman: A Tale of Tw
uest to Get Americans to Care About Rugby Opinion Op-Ed: The ‘American Tragedy’ of Vietnam U.S. Deca
ess Day » Mediator: Mike Allen, Politico’s Newsletter Pioneer, Is Handing Over the Reins Justices Si
hows Hillary Clinton Holding Edge on Donald Trump Donald Trump Tweaks Language on Guns, Alcohol and
l Spawning, Hope for Endangered Reefs An Unwelcome Tourist Arrives in New Jersey: Clinging Jellyfish
Ghosts and Legacy Nonfiction: Susan Faludi’s ‘In the Darkroom’ Education » Yearbook Project Collects
r Away Feature: Can Netflix Survive in the New World It Created? Notebook: Why ‘Transcending Race’ I
t Viewed Trending Recommended for you Loading... Go to Home Page » Site Index The New York Times Sit
on Rate Mobile Applications Replica Edition International New York Times Site Information Navigation
```

Arquivo.pt: application programming example

github.com/arquivo/example-api

Click on example.html

```html
<html>
<head>
 <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/4.0.0/css/bootstrap.min.css">
 <meta name="viewport" content="width=device-width, initial-scale=1.0">
</head>
<body>
<div id="content"></div>
<script>
 function handler(response) {
  for (var i = 0; i < response.response_items.length; i++) {
    var item = response.response_items[i];
   document.getElementById("content").innerHTML +="<br/>"
   +"<div class=\"col offset-sm-2 col-sm-8 alert alert-primary\" style=\"word-wrap:break-word\">"
    + "<a href="+item.linkToArchive+"><h4>"+ item.title+"</h4></a>"
    + "<strong><h5 class=\"text-dark\">"+ item.tstamp.substr(0,4) +"</strong></h5>"
    + "<h5 class=\"text-muted\"><a>"+ item.originalURL+"</a></h5>";
   +"</div><br/>"  ;
  }
 }
</script>
<script src="http://arquivo.pt/textsearch?q=Barack%20Obama&maxItems=5&itemsPerSite=1&callback=handler"></script>
</body>
</html>
```

Head of example.html

```
<head>
  <link rel="stylesheet" href="https://maxcdn.bootstrapcdn.com/bootstrap/4.0.0/css/bootstrap.min.css">
  <meta name="viewport" content="width=device-width, initial-scale=1.0">
</head>
```

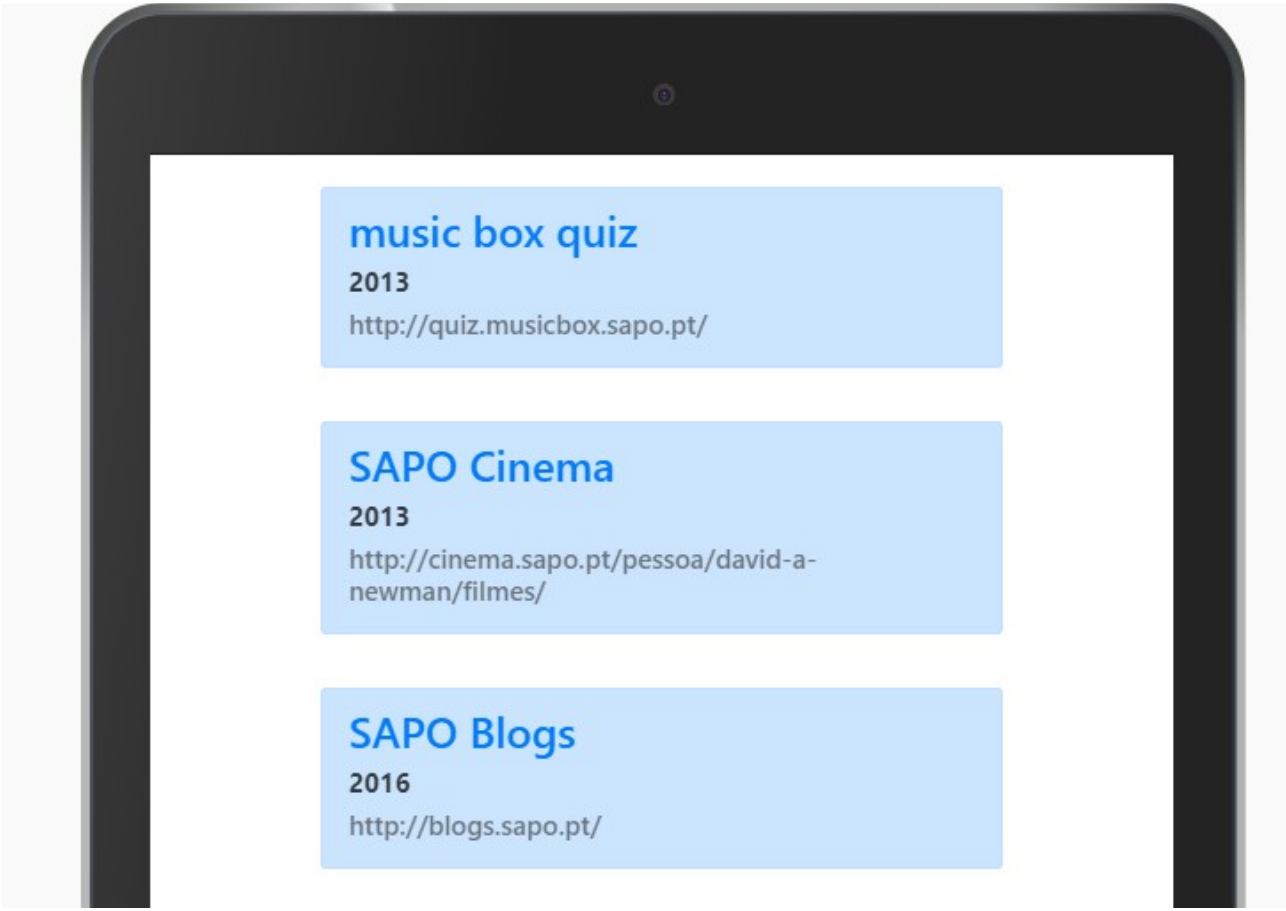
Handler function

```
<div id="content"></div>
<script>
 function handler(response) {
  for (var i = 0; i < response.response_items.length; i++) {
    var item = response.response_items[i];
   document.getElementById("content").innerHTML +="<br/>"
   +"<div class=\"col offset-sm-2 col-sm-8 alert alert-primary\" style=\"word-wrap:break-word\">"
    + "<a href="+item.linkToArchive+"><h4>"+ item.title+"</h4></a>"
    + "<strong><h5 class=\"text-dark\">"+ item.tstamp.substr(0,4) +"</strong></h5>"
    + "<h5 class=\"text-muted\"><a>"+ item.originalURL+"</a></h5>";
   +"</div><br/>"  ;
  }
 }
</script>
```

Call to Arquivo.pt API

```
<script src="http://arquivo.pt/textsearch?q=Barack%20Obama&maxItems=5&itemsPerSite=1
&callback=handler">
</script>
```

Arquivo.pt application example



Arquivo.pt API –**challenge**

Update the previous example

1. Show 15 results instead of only 5
2. List **500 versions** of the URL nytimes.com between the years of **2010** and **2011**

1. Show 15 results instead of only 5

<script src = "arquivo.pt/textsearch?**q**=Barack Obama &**maxItems**=15&**callback**=handler">

</script>

2. List **500 versions** of the URL nytimes.com between the years of **2010** and **2011**

<script src="http://arquivo.pt/textsearch?**versionHistory**=nytimes.com

&**from**=2010&**to**=2011&**maxItems**=500&**callback**=handler"></script>

Memento TimeTravel API

Memento TimeTravel 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.

These prior versions of web pages are named **Mementos**. **Mementos** can be found in **web archives** or in systems that support versioning

such as **wikis** and **revision control systems**.”

Memento TimeTravel API

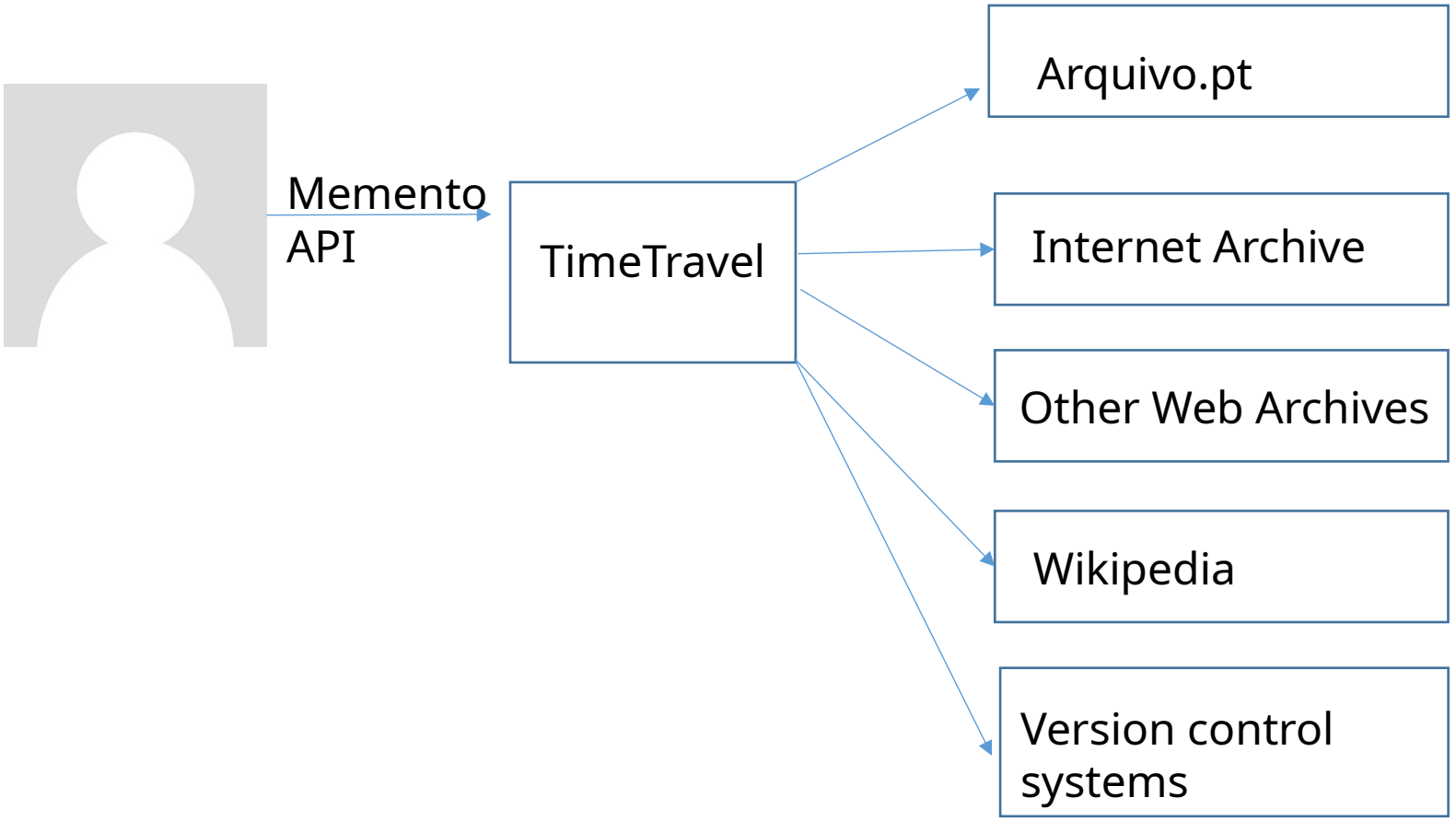
http://timetravel.mementoweb.org/guide/api/

Memento interoperability protocol.

https://mementoweb.org/guide/rfc/

https://tools.ietf.org/html/rfc7089

Memento TimeTravel API



Memento TimeTravel JSON API

http://timetravel.mementoweb.org/api/json/20140429175654/http://nytimes.com

Search for **mementos** of the URL

http://nytimes.com

Near the date:

2014 April 29 at 17:56:54

```
"mementos":{
  "next":{
    "datetime":"2014-04-29T22:33:16Z",
    "uri":[
      "https://swap.stanford.edu/20140429223316/http://www.nytimes.com/"
    ]
  },
  "last":{
    "datetime":"2018-05-30T14:00:33Z",
    "uri":[
      "http://webarchive.proni.gov.uk/20180530140033/https://nytimes.com/"
    ]
  },
  "prev":{
    "datetime":"2014-04-25T09:59:17Z",
    "uri":[
      "https://swap.stanford.edu/20140425095917/https://www.nytimes.com/"
    ]
  },
  "first":{
    "datetime":"1996-11-12T18:15:13Z",
    "uri":[
      "https://web.archive.org/web/19961112181513/http://www.nytimes.com:80/",
      "http://wayback.archive-it.org/all/19961112181513/http://www.nytimes.com/",
      "http://web.archive.bibalex.org:80/web/19961112181513/http://www.nytimes.com/"
    ]
  },
  "closest":{
    "datetime":"2014-04-29T17:56:54Z",
    "uri":[
      "http://arquivo.pt/wayback/20140429175654/http://www.nytimes.com/"
    ]
  }
```

Memento TimeTravel: **closest** memento


```
"closest":{
  "datetime":"2014-04-29T17:56:54Z",
  "uri":[
    "http://arquivo.pt/wayback/20140429175654/http://www.nytimes.com/"
  ]
}
```

Memento TimeTravel JSON API

http://timetravel.mementoweb.org/timemap/json/
http://nytimes.com

Search for **timemaps** for the URL

http://nytimes.com

Memento TimeTravel JSON API
timemap response list

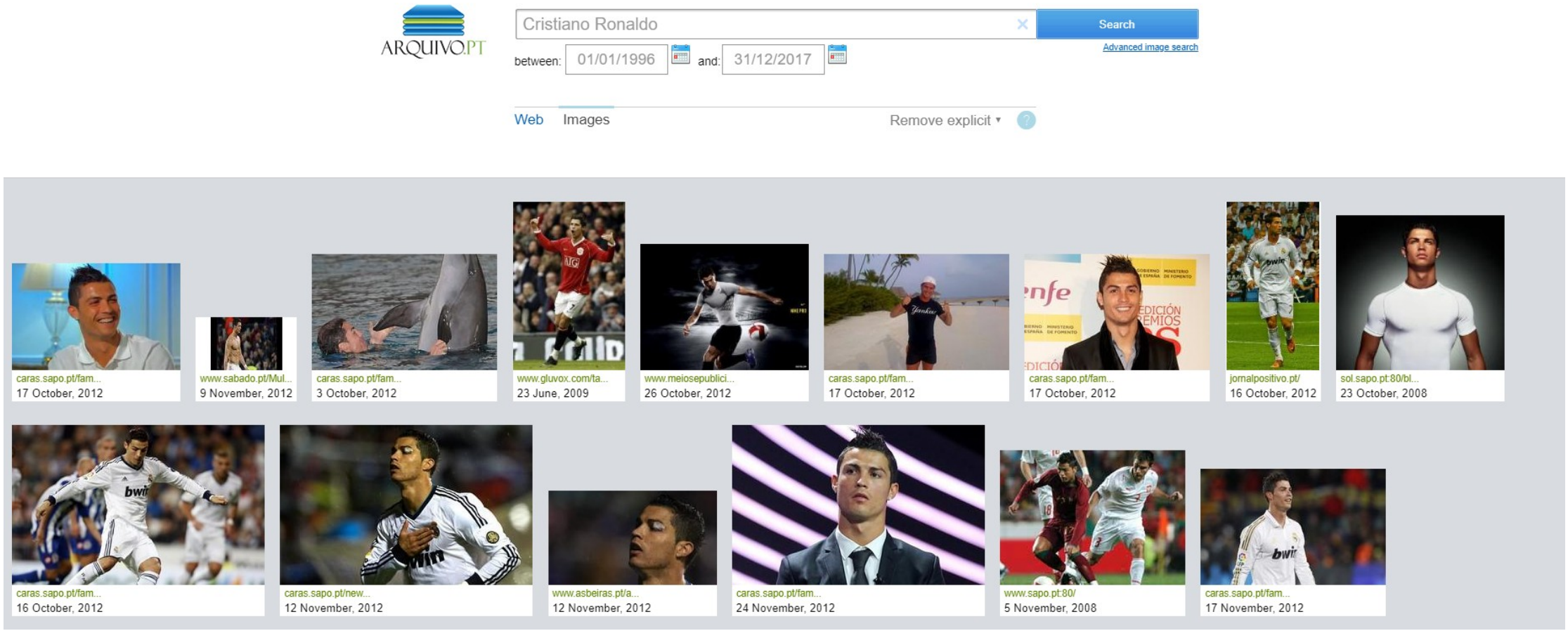

```
{
  "original_uri":"http://europa.eu",
  "timegate_uri":"http://timetravel.mementoweb.org/timegate/http://europa.eu",
  "timemap_index":[
    {
      "uri":"https://web.archive.org/web/timemap/link/http://europa.eu",
      "memento_compliant":"yes",
      "archive_id":"ia"
    },
    {
      "uri":"http://webarchive.proni.gov.uk/timemap/http://europa.eu",
      "memento_compliant":"yes",
      "archive_id":"proni"
    },
    {
      "uri":"http://web.archive.bibalex.org/web/timemap/link/http://europa.eu",
      "memento_compliant":"yes",
      "archive_id":"ba"
    },
    {
      "uri":"http://timetravel.mementoweb.org/webcite/timemap/link/http://europa.eu",
      "memento_compliant":"no",
      "archive_id":"webcite"
    },
```

Memento TimeTravel JSON API
timemap response

https://web.archive.org/web/19971008182708/http://www.sapo.pt:80/;
rel="first memento";
datetime="Wed, 08 Oct 1997 18:27:08 GMT",

<https://web.archive.org/web/19971210144509/http://www.sapo.pt:80/>;
rel="memento";
datetime="Wed, 10 Dec 1997 14:45:09 GMT",

Image Search API



Fernando Melo <fernando.melo@fccn.pt>