ARQUIVO.PT

# Enhancing access to research the Geocities historical collection.

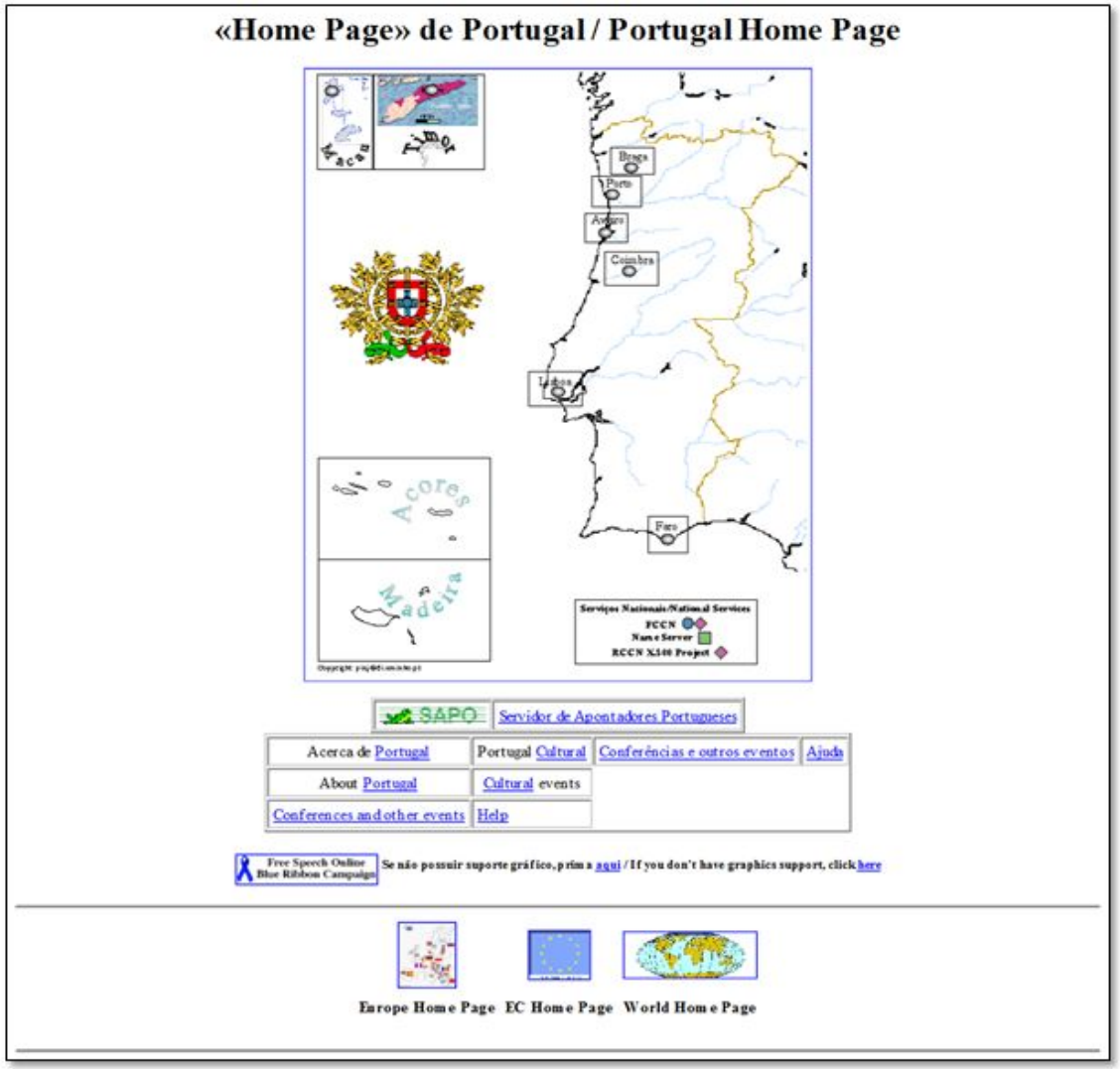
Pedro.Gomes@fccn.pt

Daniel.Gomes@fccn.pt

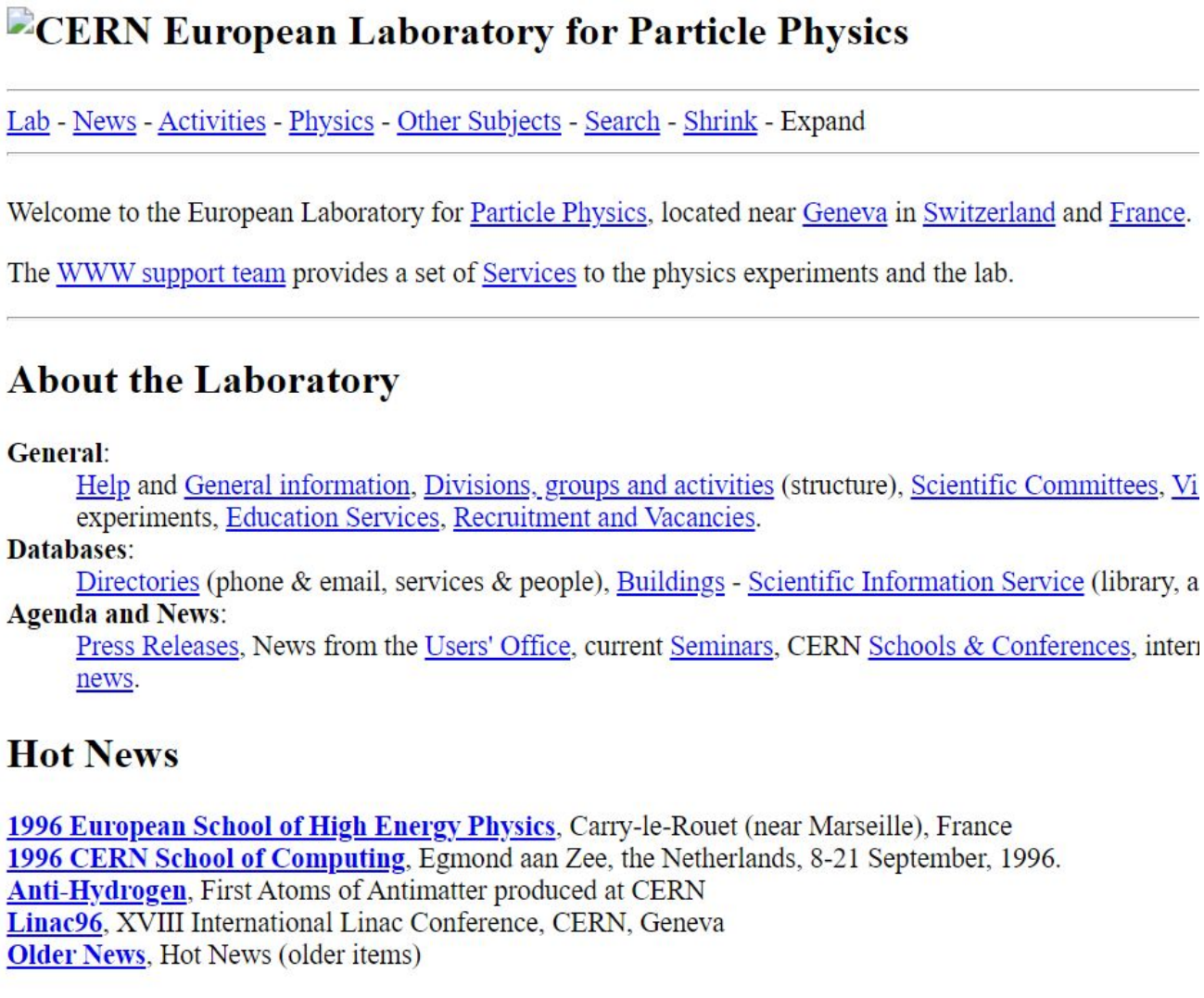
FCT Fundação para a Ciência e a Tecnologia
Computação Científica Nacional
FCCN

REPÚBLICA PORTUGUESA
CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR

# Present Arquivo.pt

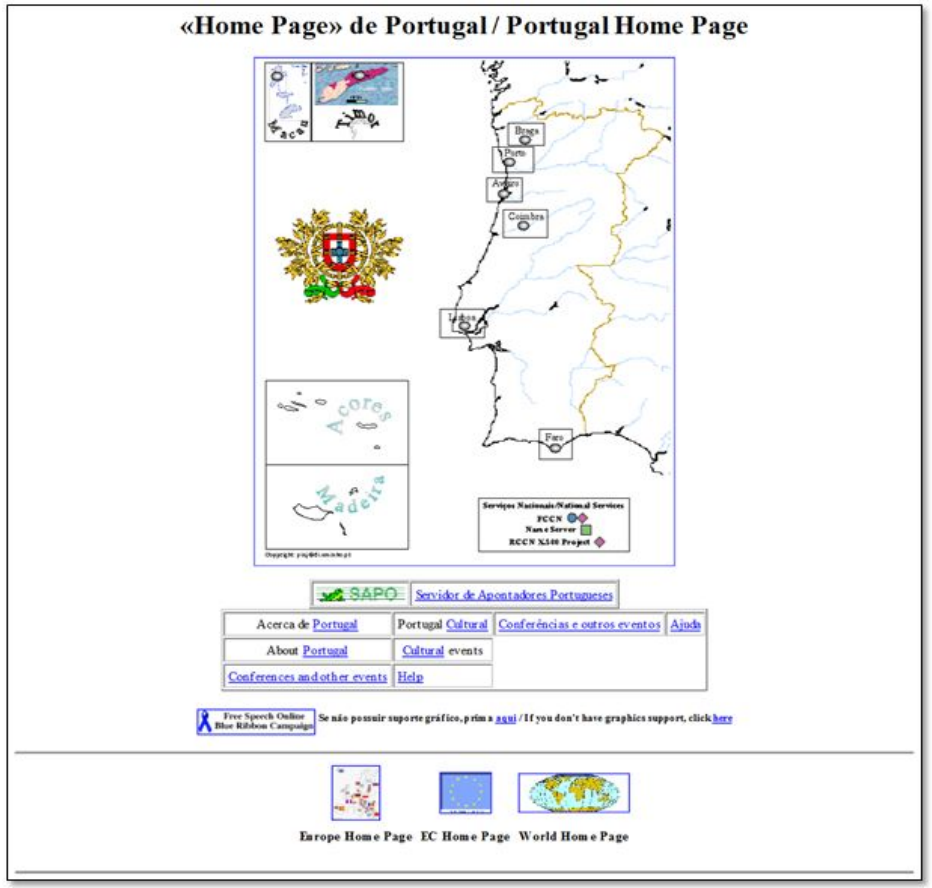



http://s700.uminho.pt/homepage-pt.html - 1996

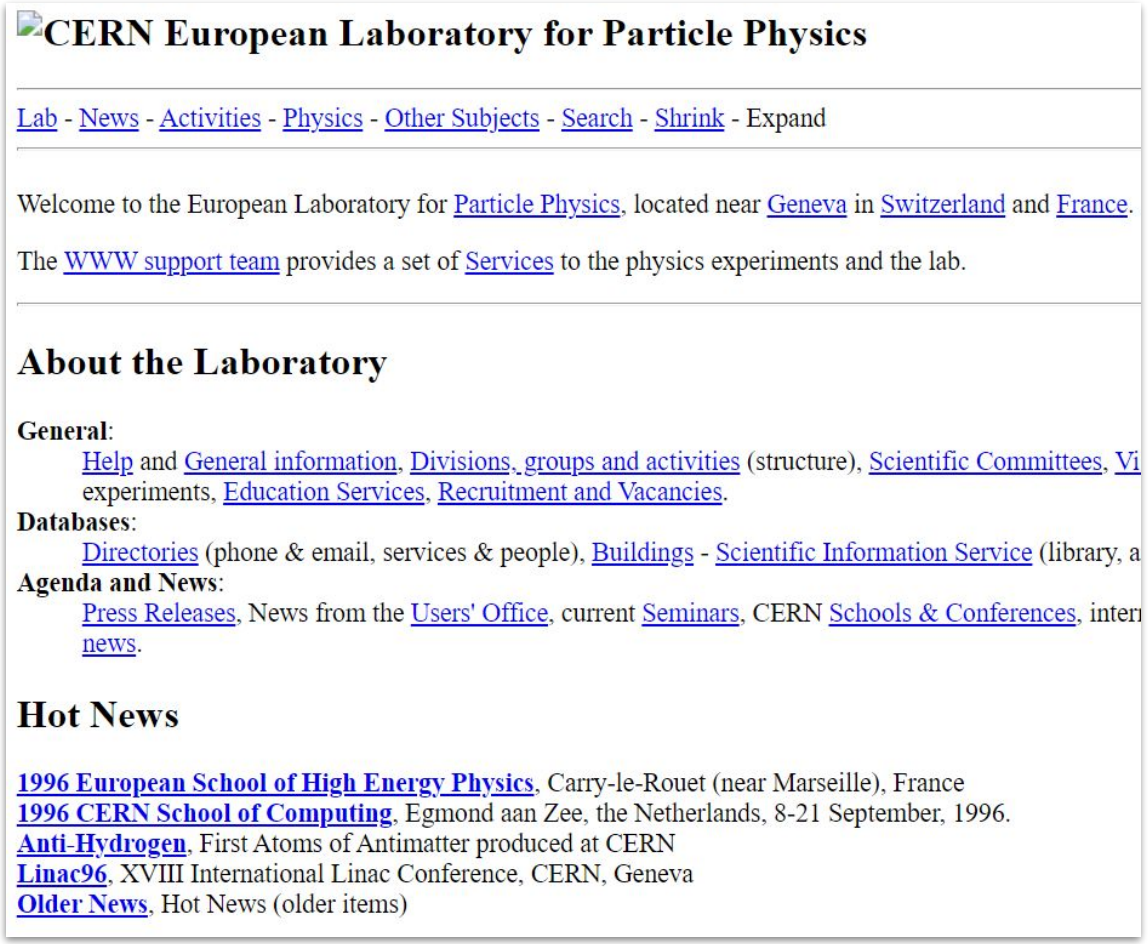



http://www.cern.ch - 1996

# Arquivo.pt preserves national and international historical web content



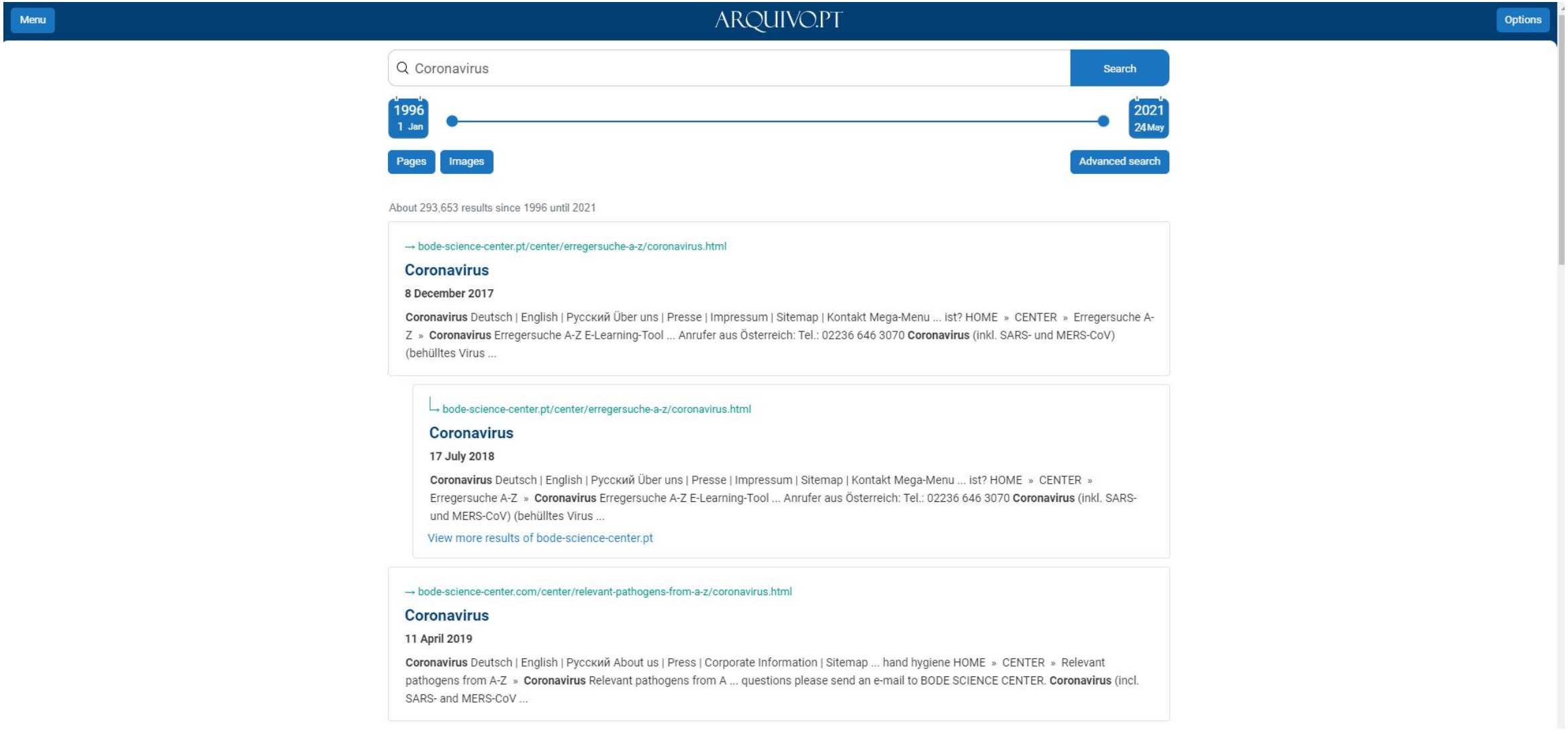

[s700.uminho.pt/homepage-pt.html](s700.uminho.pt/homepage-pt.html) - 1996




[www.CERN.ch](www.CERN.ch) - 1996

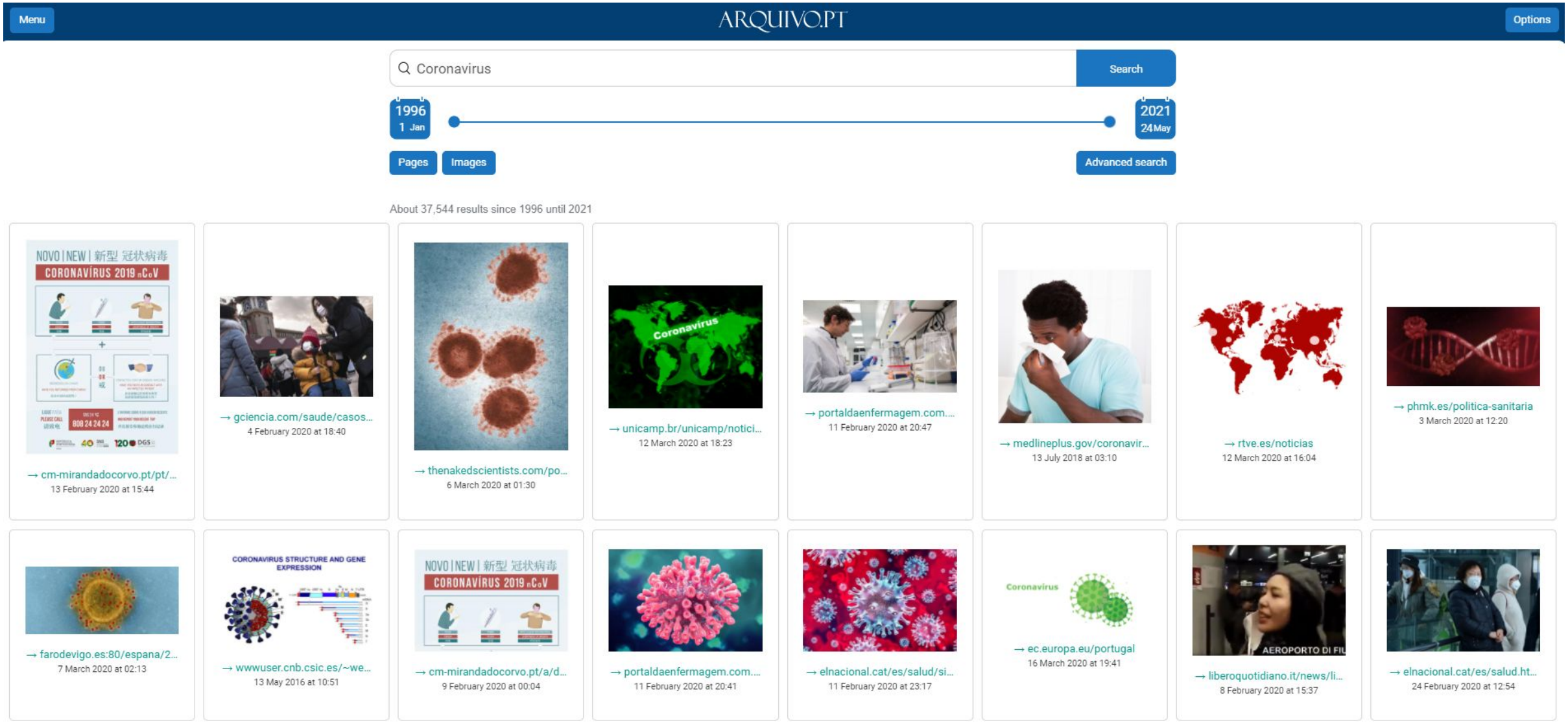
# Keep improving

## Full-text search engine & User interfaces



# To launch new services

## Like the Images Searching & High Quality Archiving Services



# And at the same time ...

Increase our collection and users community

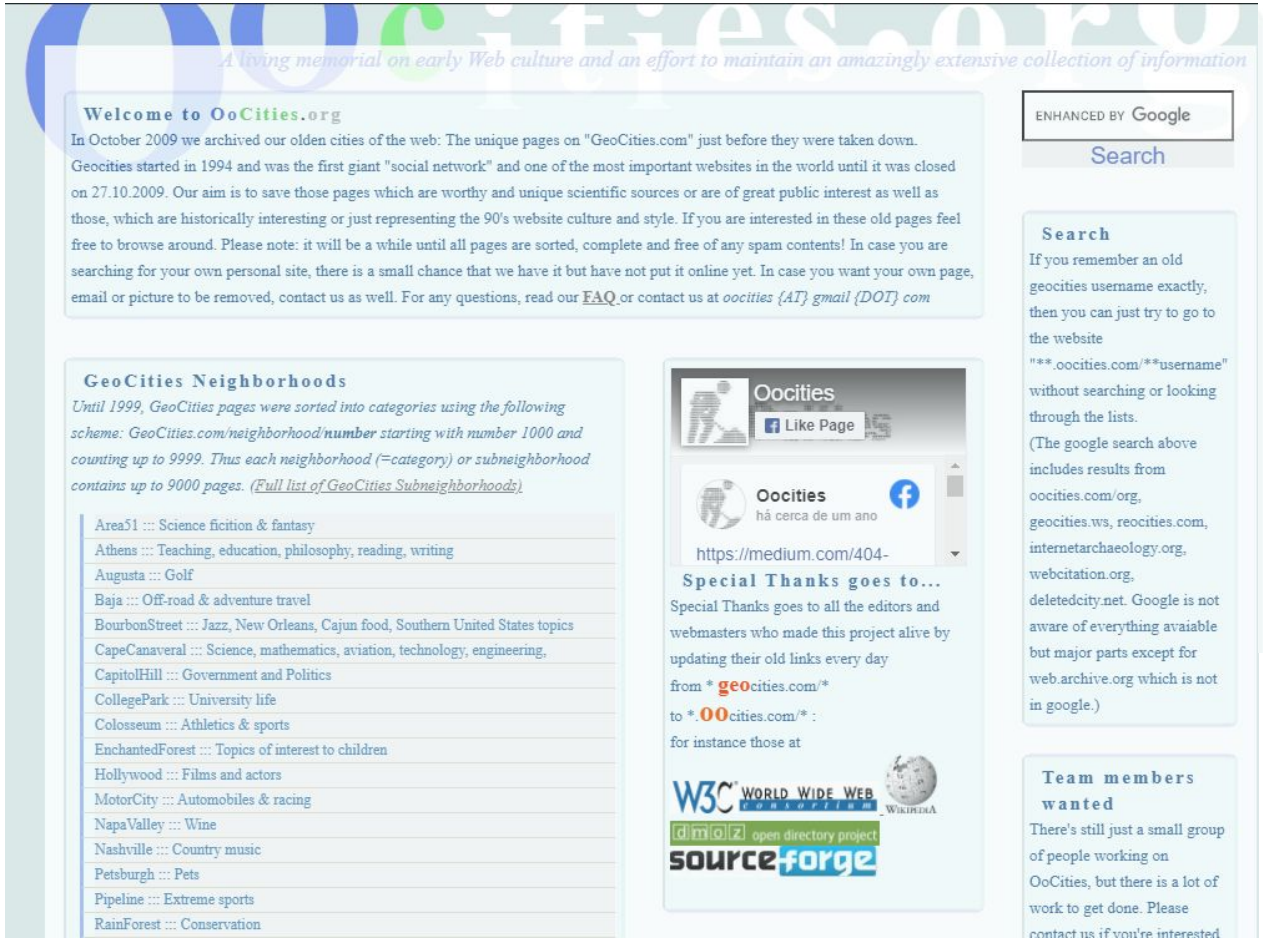
Photo courtesy of John Morgan, Flickr, CC BY2.0

# Geocities

- Geocities.com was founded in 1994, and in 1999, Geocities was acquired by Yahoo! and it became the third most popular platform;
- In April 2009, Yahoo!, announced that it would shut down Geocities, deleting more than 38 million pages created by users all around the world;
- The Archive team announced a project to archive Geocities pages, resulting in a torrent file;

# Initiatives

- http://www.oocities.org/

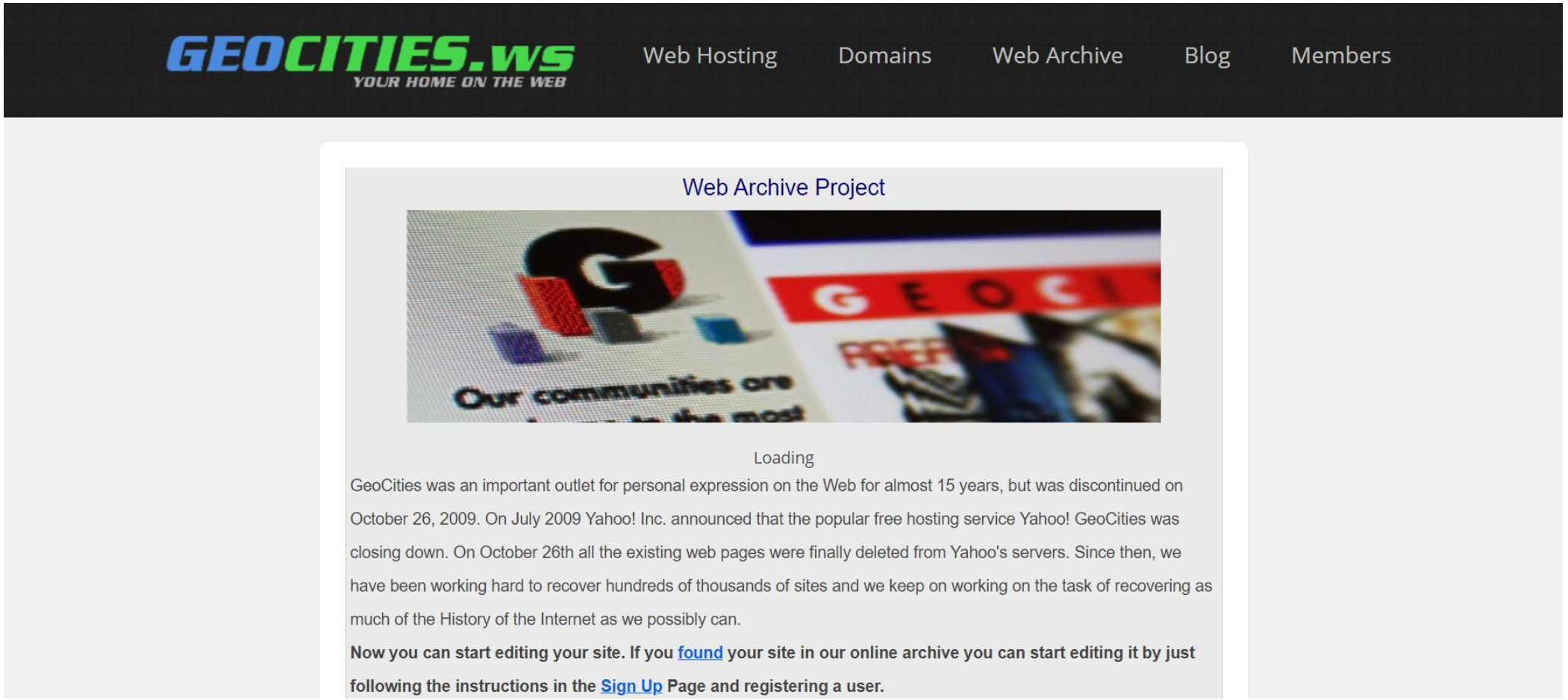



## Custom Search from Google

# Initiatives

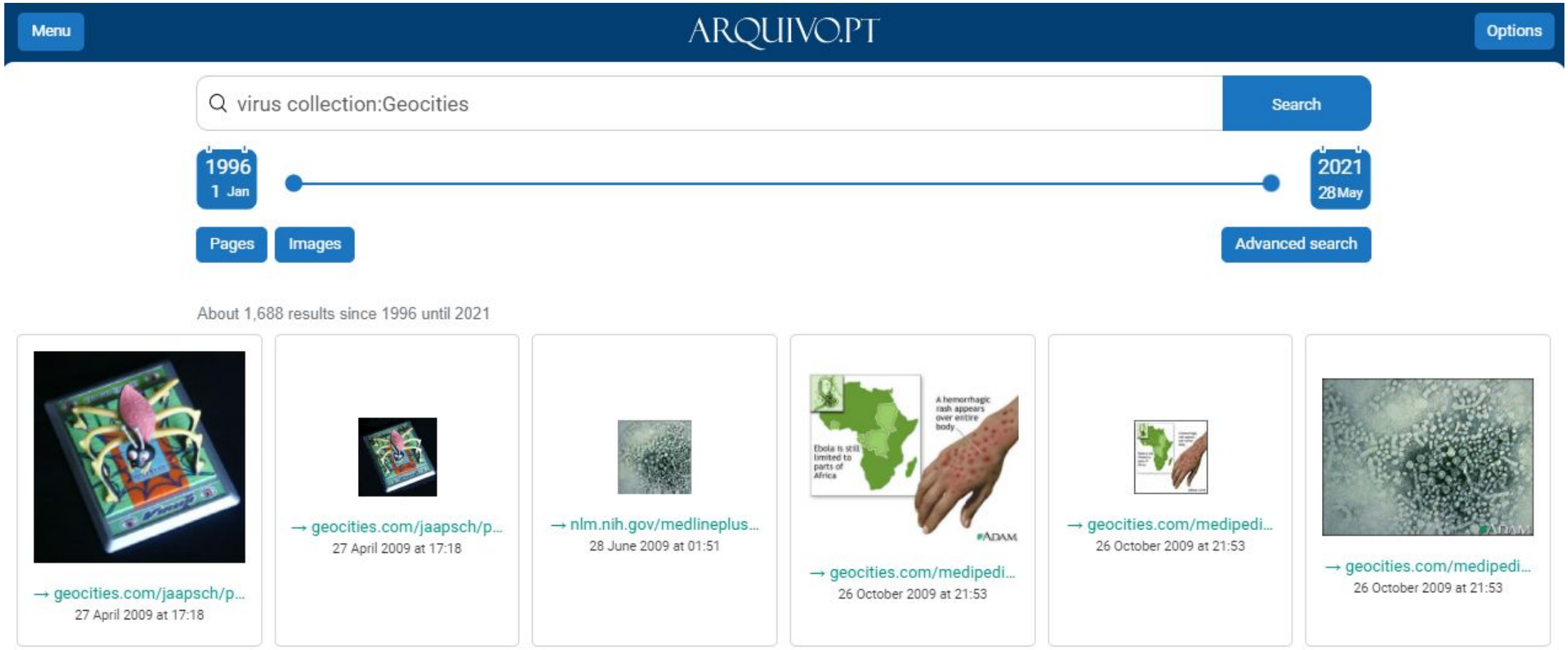
# Initiatives

- http://www.geocities.ws/archive/



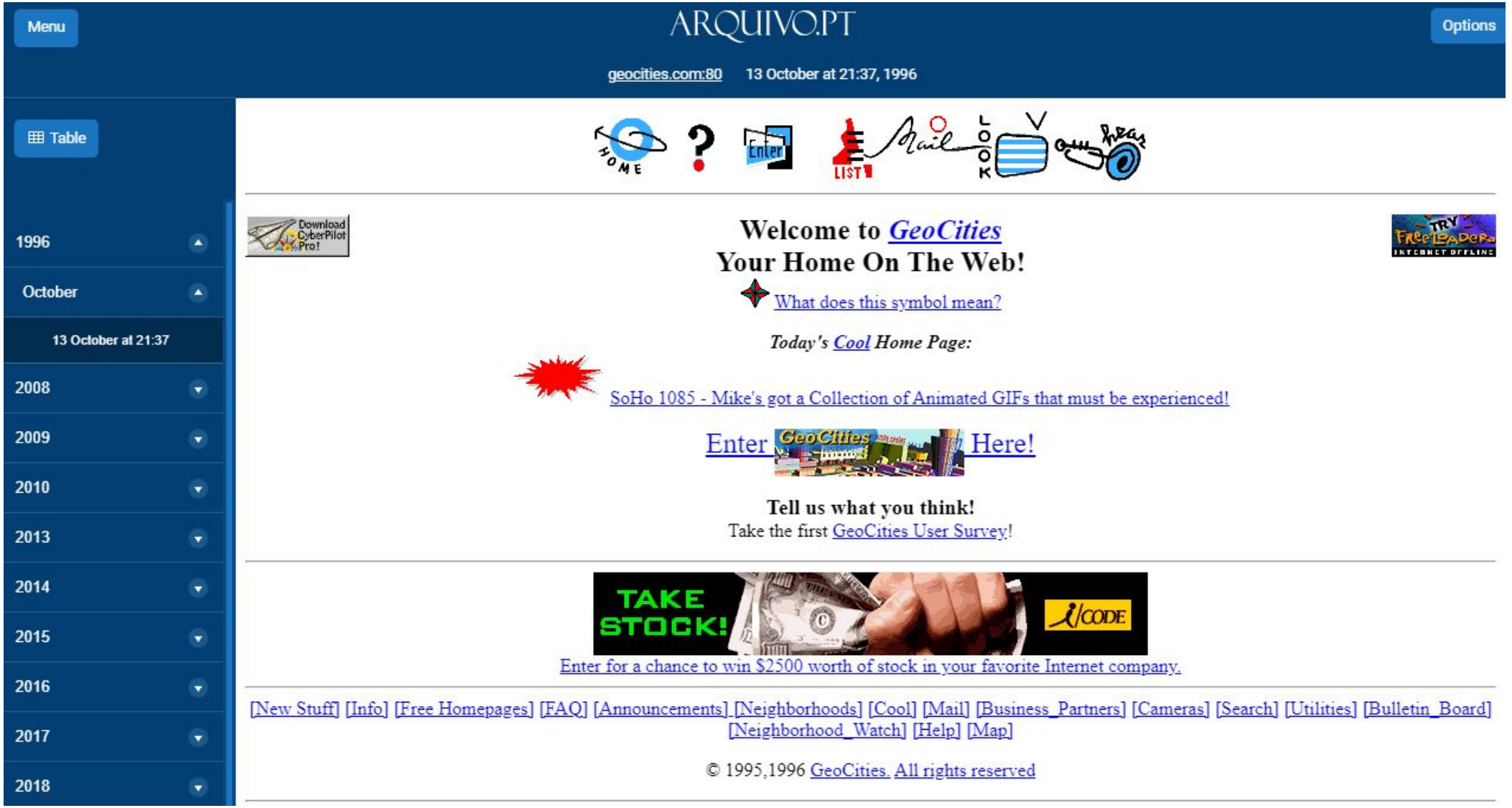
# Geocities Arquivo.pt

# Geocities Arquivo.pt

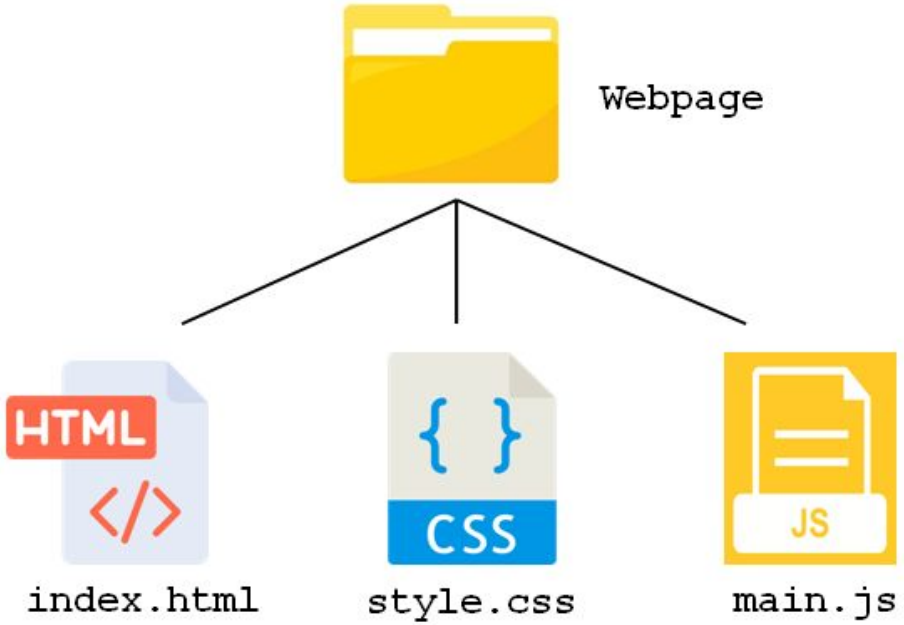


# Geocities Arquivo.pt

# Motivation


Menu
ARQUIVO.PT
Options
geocities.com:80
13 October at 21:37, 1996
Table
1996
October
13 October at 21:37
2008
2009
2010
2013
2014
2015
2016
2017
2018
Download CyberPilot Pro!
Welcome to GeoCities
Your Home On The Web!
What does this symbol mean?
Today's Cool Home Page:
SoHo 1085 - Mike's got a Collection of Animated GIFs that must be experienced!
Enter GeoCities Here!
Tell us what you think!
Take the first GeoCities User Survey!
TAKE STOCK!
Enter for a chance to win $2500 worth of stock in your favorite Internet company.
[New Stuff] [Info] [Free Homepages] [FAQ] [Announcements] [Neighborhoods] [Cool] [Mail] [Business_Partners] [Cameras] [Search] [Utilities] [Bulletin_Board] [Neighborhood_Watch] [Help] [Map]
© 1995,1996 GeoCities. All rights reserved

# Challenges




**WARC**

Web ARChive file format

ISO 28500:2009

webrecorder/**warcit**

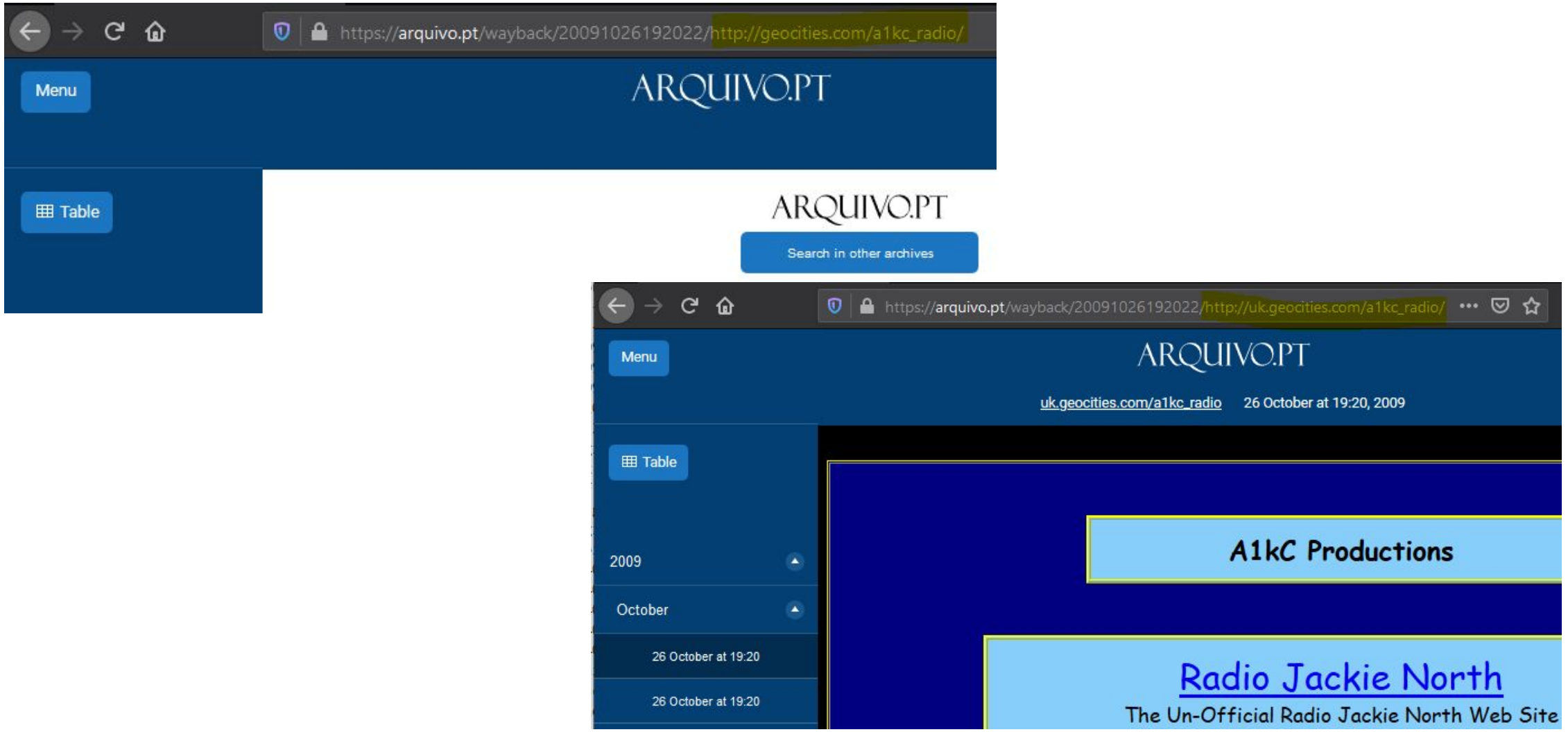
Convert Directories, Files and ZIP Files to Web Archives (WARC)

# Challenges

Choose the URL default for each folder:

- http://www.geocities.com vs http://geocities.com
- http://uk.geocities.com vs http://www.geocities.com

# Challenges


https://arquivo.pt/wayback/20091026192022/http://geocities.com/a1kc_radio/
Menu
ARQUIVO.PT
Table
ARQUIVO.PT
Search in other archives
https://arquivo.pt/wayback/20091026192022/http://uk.geocities.com/a1kc_radio/
Menu
ARQUIVO.PT
uk.geocities.com/a1kc_radio
26 October at 19:20, 2009
Table
2009
October
26 October at 19:20
26 October at 19:20
A1kC Productions
Radio Jackie North
The Un-Official Radio Jackie North Web Site

# Challenges

- Identify the correct timestamp for each webpage

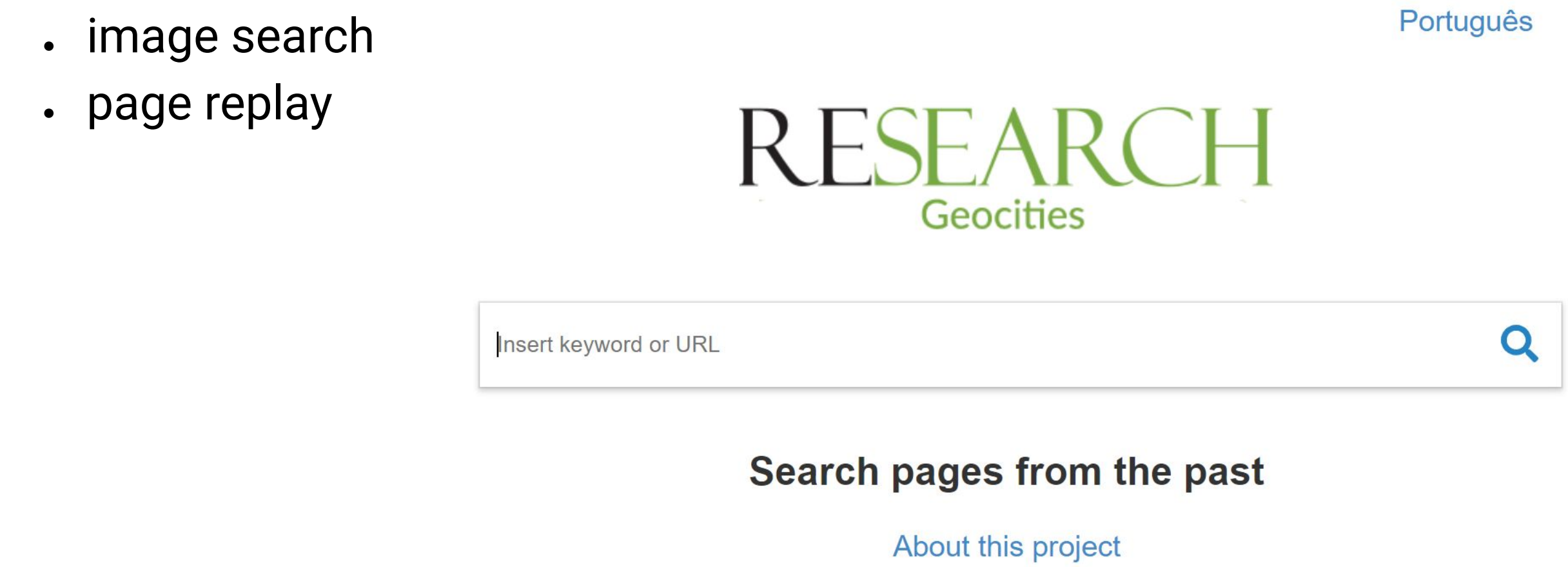
# Steps Algorithm

1. Unzip all files. Examples:
   a. Lowercase → YAHOOIDS → a → a → aa_chan01
   b. Subsites → YAHOOIDS → geocities → uk.geocities.com → a_arif_ali
2. For each subfile (e.g., aa_chan01):
   a. Get the new timestamp from the last update of the files inside in each folder (e.g., 2009 - 11 - 05).
   b. warcit --fixed-dt timestamp http://www.geocities.com/ path_folder.
      i. If there is a subsite, “http://www.geocities.com/” will be replace by “http://uk.geocities.com/”.
3. Merge each individual WARC file and removing the “revisit” records since will have the current timestamp (i.e., 2021).

# Geocities collection

[arquivo.pt/searchGeocities](arquivo.pt/searchGeocities)

- full text page search
- image search
- page replay



# Geocities Numbers

- From: 2009-04-24
- Until: 2010-10-13
- Disk space: 280 GB (warc.gz)
- 35 millions collected files

**Examples of Geocities preserved websites**

Geocities homepage in 1996

Geocities homepage in 2008

Geocities homepage in 2009

Country & WesternBeauties

Queen – Manu's Musica

Artbook Online Gallery – Collage + digital art

Computers

Ilustrious Portuguese Navigators

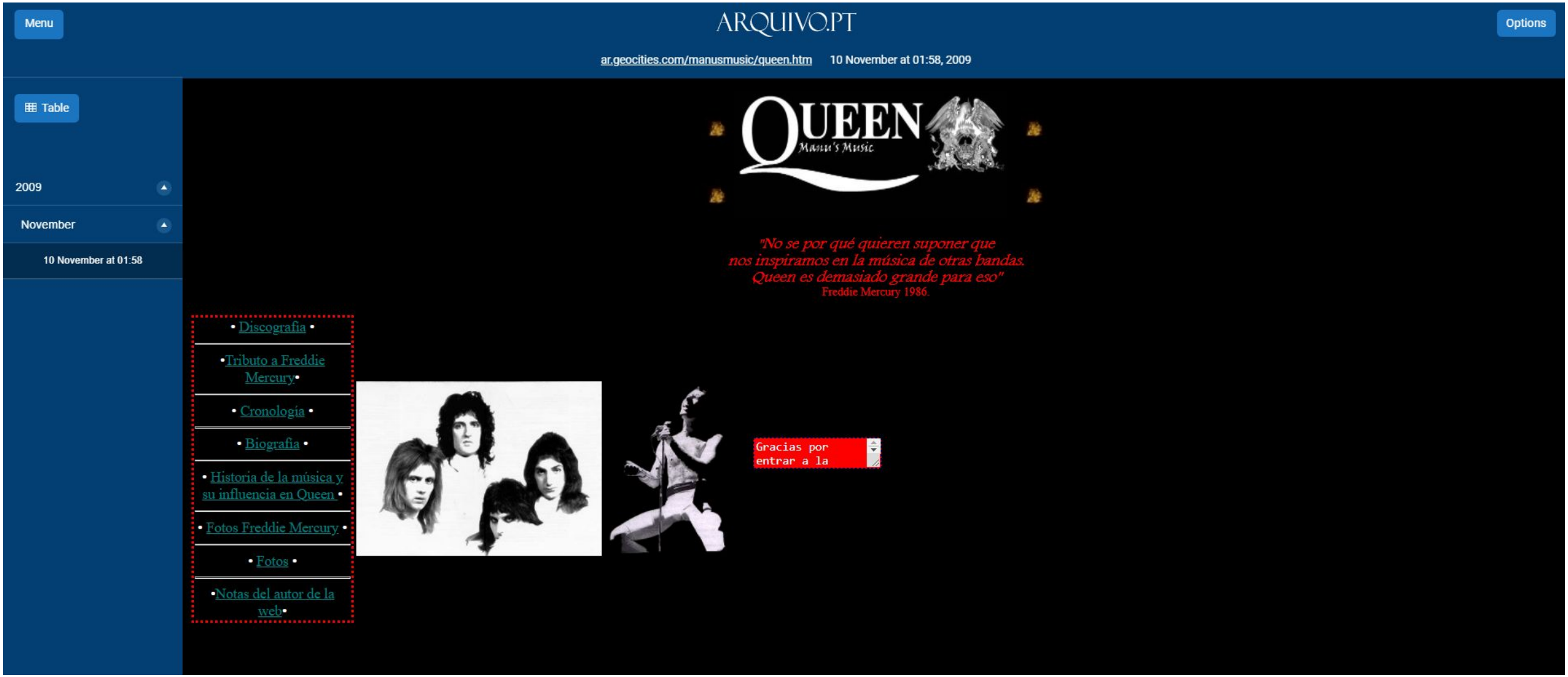
Unofficial F3000

Midis Internacionais (music files in midi)

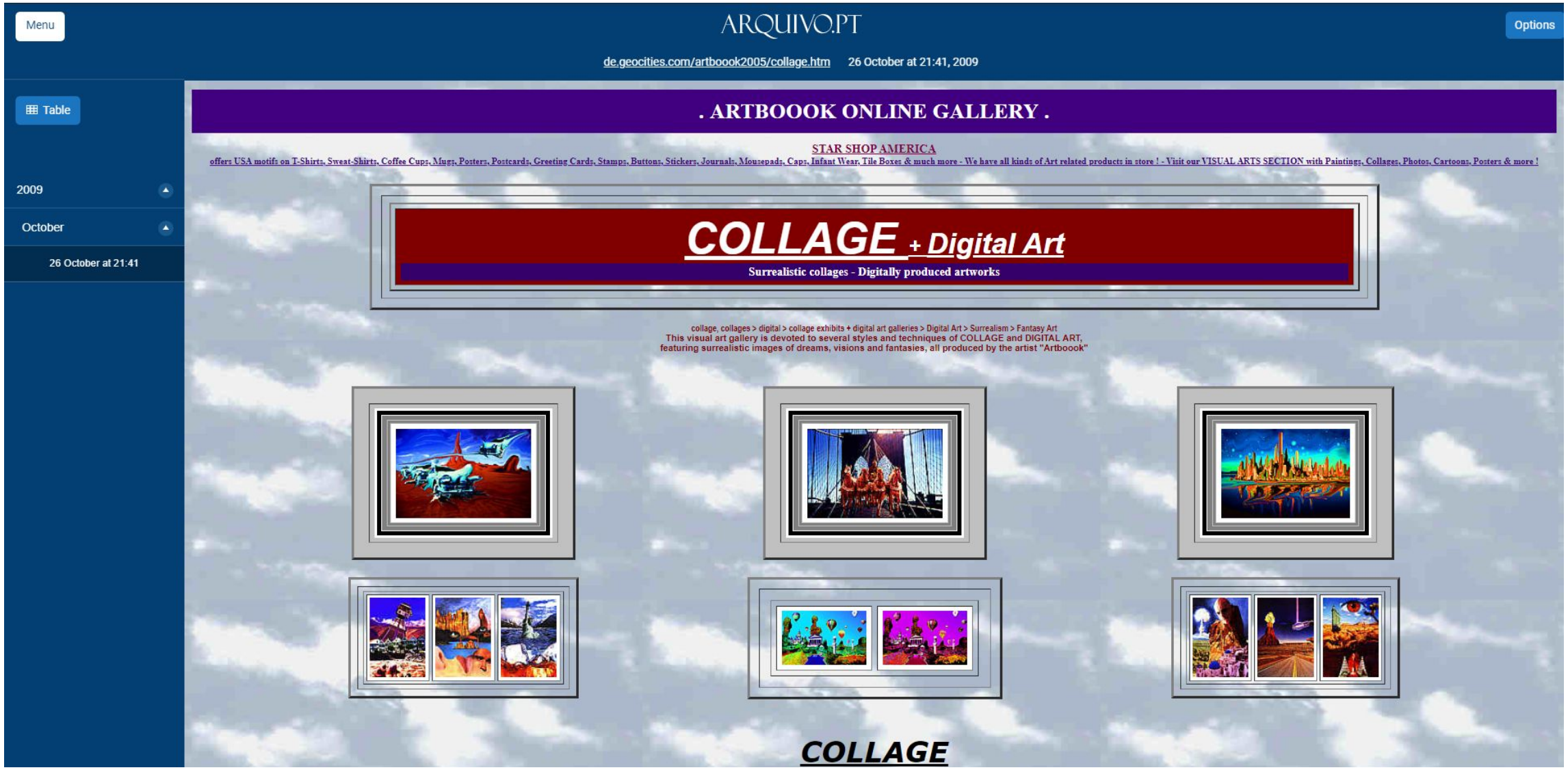
Our Hommer. All about dogs

My personal diary

# Examples


Menu
ARQUIVO.PT
Options
ar.geocities.com/manusmusic/queen.htm
10 November at 01:58, 2009
Table
2009
November
10 November at 01:58
QUEEN
Manu's Music
"No se por qué quieren suponer que
nos inspiramos en la música de otras bandas.
Queen es demasiado grande para eso"
Freddie Mercury 1986.
• Discografia •
•Tributo a Freddie Mercury•
• Cronología •
• Biografia •
• Historia de la música y su influencia en Queen •
• Fotos Freddie Mercury •
• Fotos •
•Notas del autor de la web•
Gracias por
entrar a la

# Examples


Menu
ARQUIVO.PT
Options
de.geocities.com/artboook2005/collage.htm
26 October at 21:41, 2009
Table
2009
October
26 October at 21:41
. ARTBOOOK ONLINE GALLERY .
STAR SHOP AMERICA
offers USA motifs on T-Shirts, Sweat-Shirts, Coffee Cups, Mugs, Posters, Postcards, Greeting Cards, Stamps, Buttons, Stickers, Journals, Mousepads, Caps, Infant Wear, Tile Boxes & much more - We have all kinds of Art related products in store ! - Visit our VISUAL ARTS SECTION with Paintings, Collages, Photos, Cartoons, Posters & more !
COLLAGE + Digital Art
Surrealistic collages - Digitally produced artworks
collage, collages > digital > collage exhibits + digital art galleries > Digital Art > Surrealism > Fantasy Art
This visual art gallery is devoted to several styles and techniques of COLLAGE and DIGITAL ART,
featuring surrealistic images of dreams, visions and fantasies, all produced by the artist "Artboook"
COLLAGE

# Examples

Menu

ARQUIVO.PT

Options

uk.geocities.com/peterochocki/computers/1980comp/appmac.htm 20 October at 10:35, 2009

Table

2009

October

20 October at 10:35

Back to Home Page

Help

Notes

## The Apple Macintosh

| | |
|---|---|
| Manufacturer | Apple Computers (US) |
| Model | Macintosh<br>(From a variety of apple.) |
| Date Launched | January 1984 |
| Price | £1600 |
| Microprocessor type | Motorola 68000 @ 8 MHz |
| ROM size | 64 kilobytes |
| Standard RAM | 128 kilobytes |
| Maximum RAM | 128 kilobytes |
| Keyboard type | Typewriter style |

# Examples

Menu
ARQUIVO.PT
Options
geocities.com/SoHo/3808/navega_e.html
13 October at 05:48, 2010
Table
2010
October
13 October at 05:35
13 October at 05:48
Illustrious Portuguese Navigators
Gil Eanes
João Gonçalves Zarco
Bartolomeu Dias
Vasco da Gama
Pedro Álvares Cabral
Fernão de Magalhães
Gonçalo Velho Cabral
Diogo Cão
Sail Home

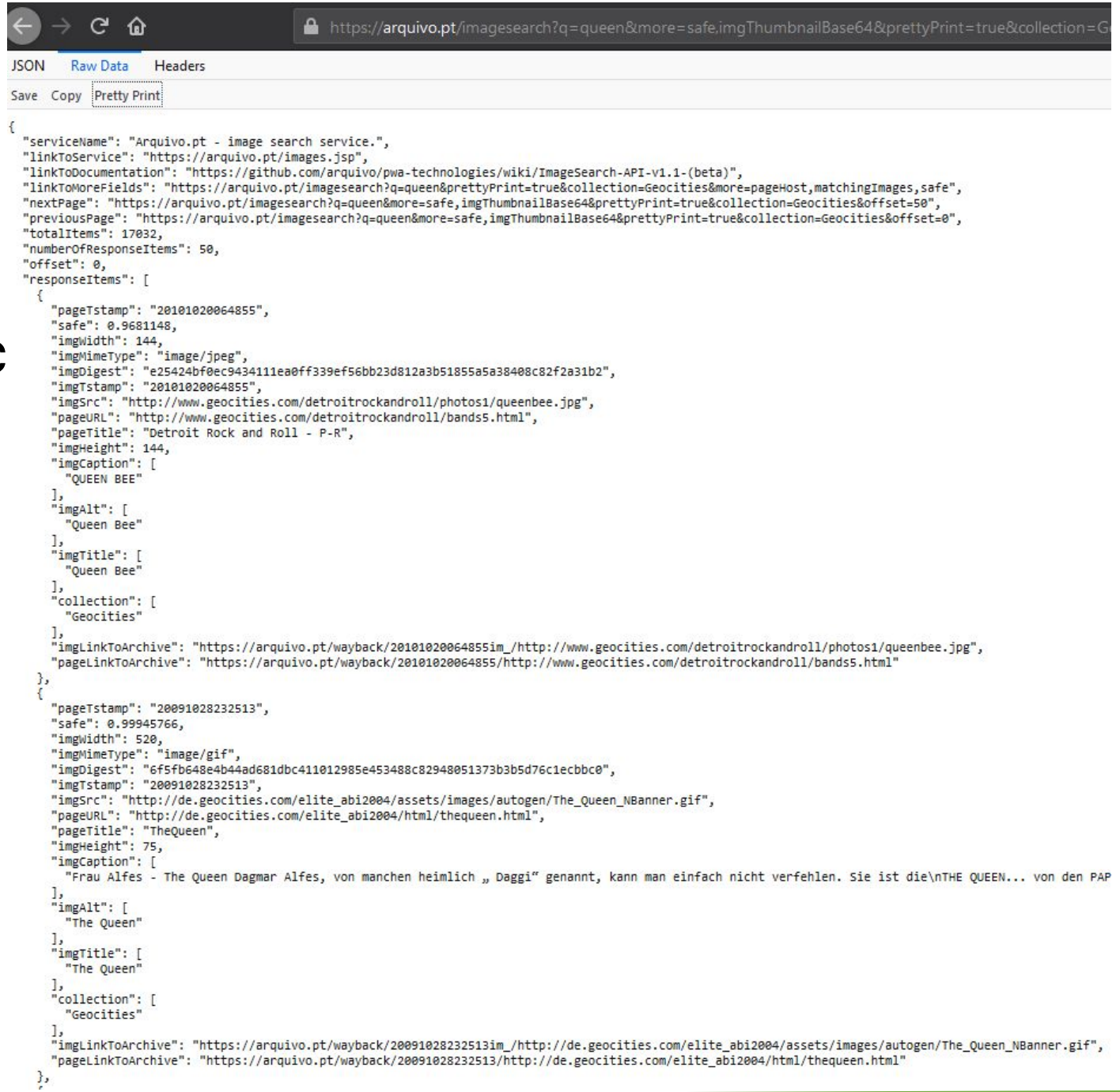
# API

[arquivo.pt/api](arquivo.pt/api)

- Available for automatic processing

APIs

- CDX-server API
- Memento API
- Page Search API
- Image Search API



# Conclusion

- Originate innovative scientific contributions.
- National and international researchers could benefit from the access tools of Arquivo.pt.
- This process can be applied easily to any kind of collection.

# New

- **Special Collection** → Coronavirus COVID-19 outbreak
  - IIPC Collaborative Collection
  - Automatically queried a web search engine
    - e.g., “coronavirus site:news-site.pt“
  - Human review
- **New Service Save Page Now**
- **New Version of Image Search**
  - Search among 1.8 billion images archived since 1994
  - Image captions extracted from text close to images in HTML

# Arquivo.pt Award 2021

## Values of the Awards

- 1st place: 10 000 €
- 2nd place: 3 000 €
- 3rd place: 2 000 €

# Thank you !

More information

[arquivo.pt/searchGeocities](arquivo.pt/searchGeocities) > [About this project](About this project)