



# 80 thousand pages on street art

exploring techniques to build thematic collections

[Ricardo.Basilio@fccn.pt](mailto:Ricardo.Basilio@fccn.pt), Arquivo.pt Web Curator

IIPC WAC 2024, Paris, 25-04-2024, SESSION #02





ARQUIVO.PT

[arquivo.pt](http://arquivo.pt)

# Arquivo.pt

- **Free online** service to research the Past Web
- Preserves **publicly accessible** information related with:
  - Portugal
  - **Research and Education** (international)
- Governmental service provided by Foundation for Science and Technology (Portugal)
- A digital research infrastructure
- Available at <https://arquivo.pt/>



# Goal

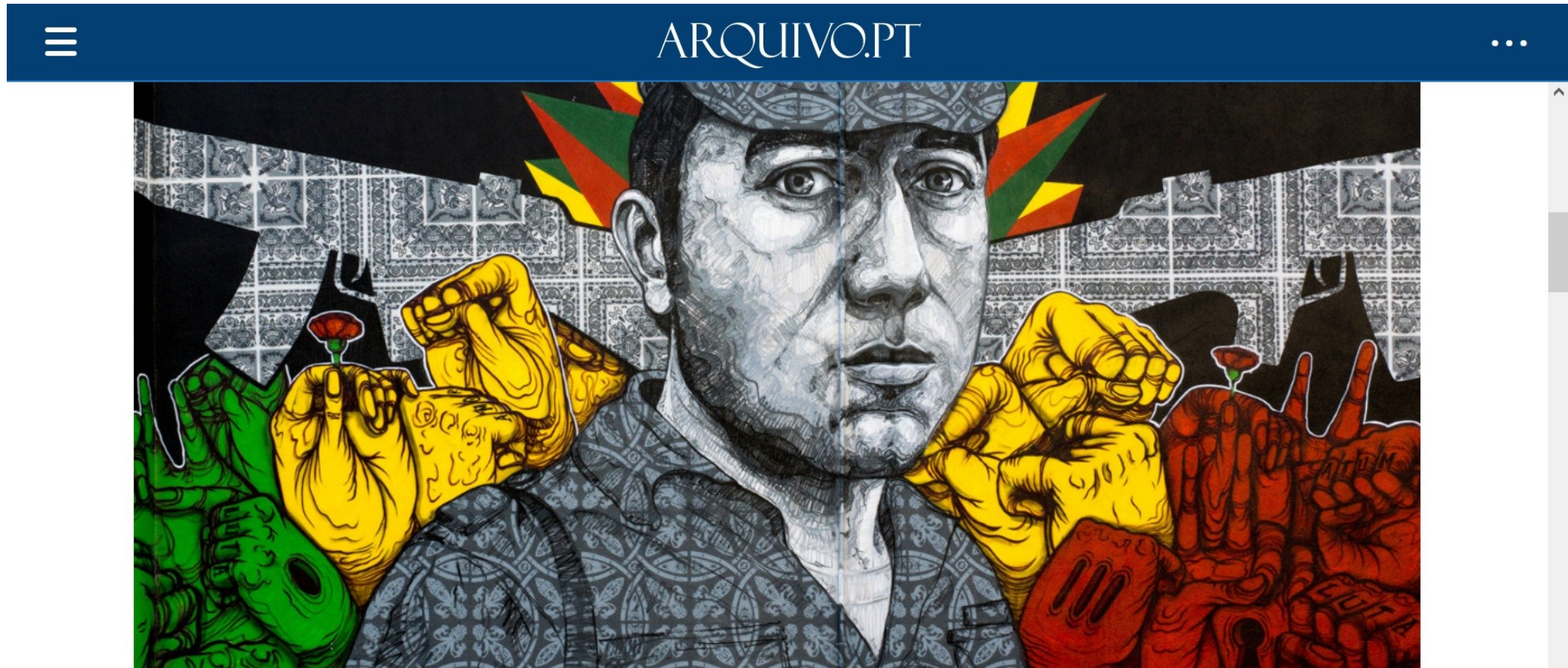
To describe **the techniques and tools** used to build thematic collections at Arquivo.pt, which are accessible to **non-specialists**.

# Agenda

1. Street Art in Lisbon
2. Problem: how to get thousands of pages related to street art
3. Recording and making the content available
4. Conclusions. Difficulties and lessons learned

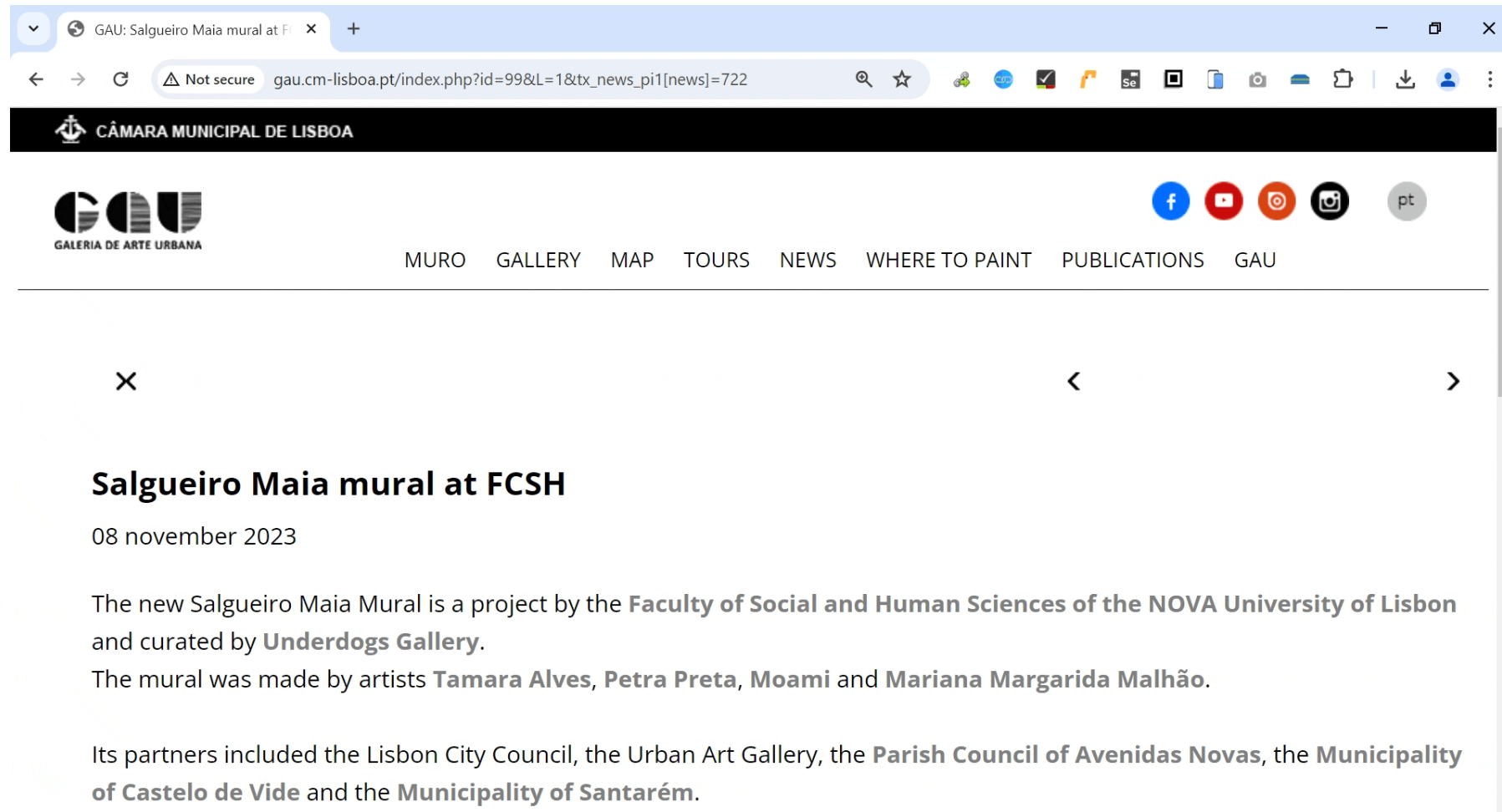
# 1- Street Art in Lisbon

# 1 - Street Art in Lisbon



[Mural on the NOVA University to commemorate the 25<sup>th</sup> April 1974, the Portuguese Revolution. From the website of the Contemporary History Institute, \[ihc.unl.pt\]\(http://ihc.unl.pt\), at \[arquivo.pt\]\(http://arquivo.pt\), 2017](#)

# 1 - Street Art in Lisbon

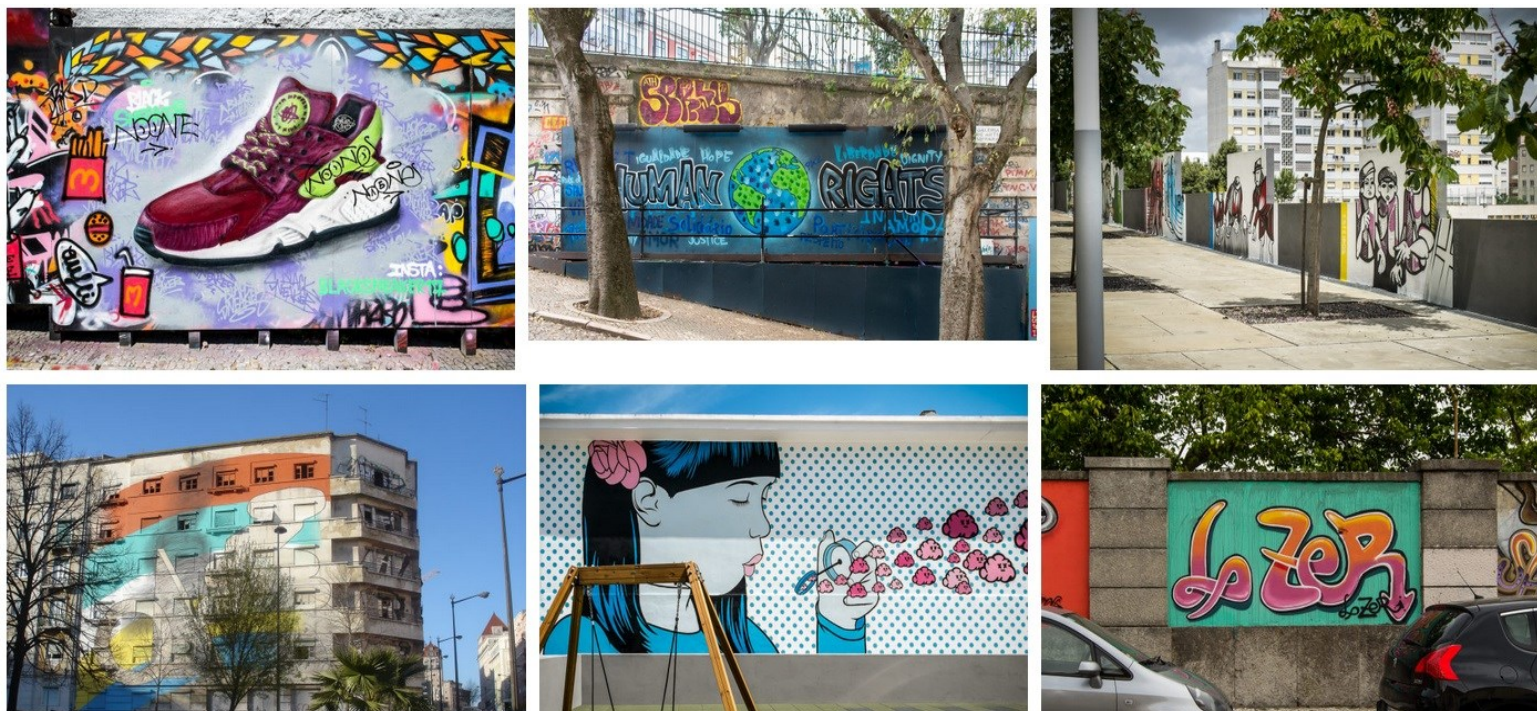


Mural on the NOVA University to commemorate the 25<sup>th</sup> April 1974, the Portuguese Revolution.



# 1 - Street Art in Lisbon

Projects  Artists  Year  Search



GAU stands for Urban Art Gallery, [gau.cm-lisboa.pt](http://gau.cm-lisboa.pt)

# 1 - Street Art in Lisbon



The 5th edition of MURO Festival de Arte Urbana LX\_2023 is coming up.

GAU stands for Urban Art Gallery. “MURO” means “Wall” [gau.cm-lisboa.pt](http://gau.cm-lisboa.pt)

# 1 - Street Art in Lisbon



Collaboration with IADE School, course of Creative Technologies.  
A student presenting their work about street art.

# 1 - Street Art in Lisbon

## Next...

- Feed [arquivo.pt](http://arquivo.pt) with more contents about street art
- Training
- Exhibitions and presentations

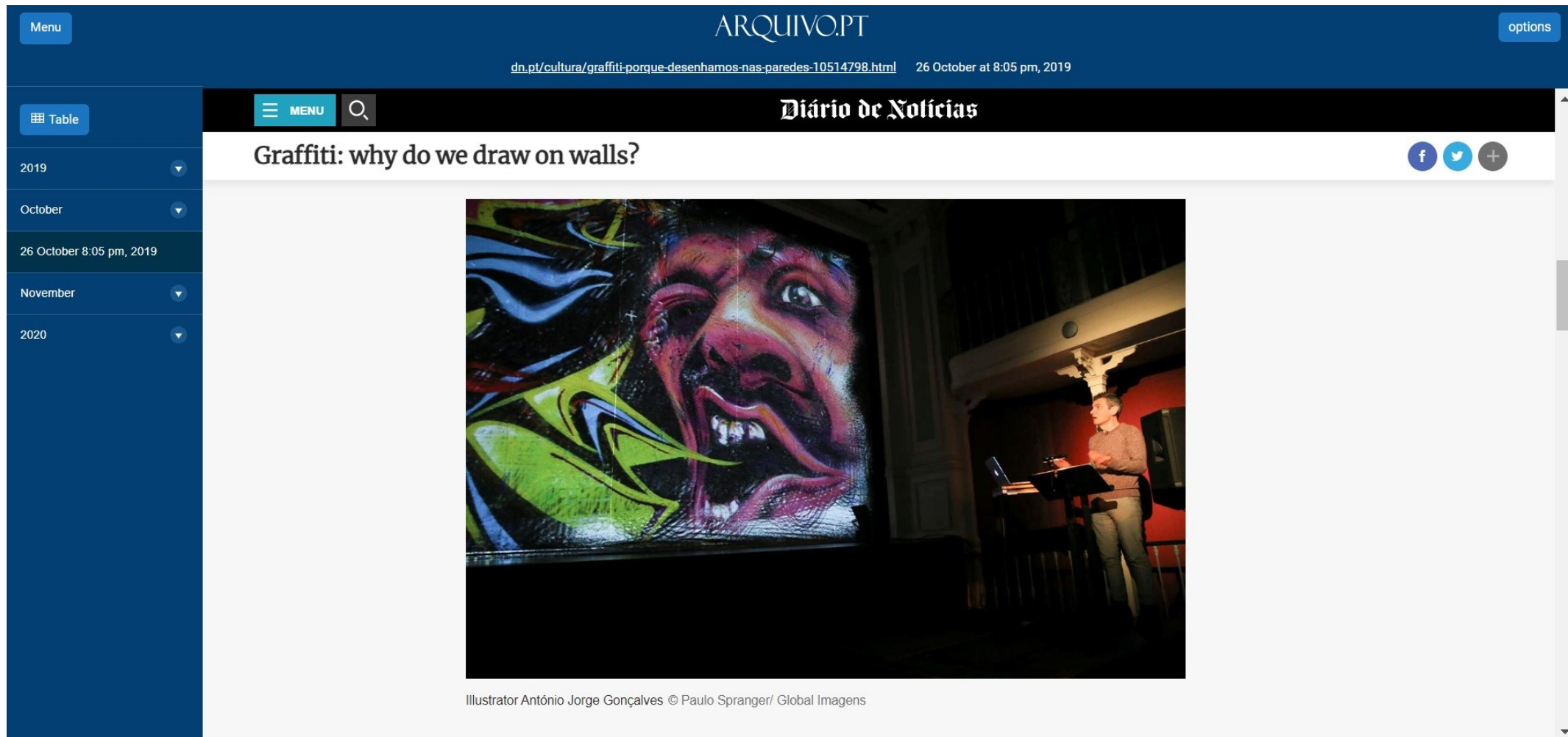
2- Problem: how to get thousands of pages related to street art

# 1 - Problem: how to get thousands of pages related to street art

```
street-art-portugal-world-urls-list.txt x
1 https://www.dn.pt/cultura/graffiti-porque-desenhamos-nas-paredes-10514798.html
2 http://100porcentobahia.blogspot.com/2016/05/bigod-entrevista.html
3 http://164.90.176.111/vdvv8/famous-sculpture-artists-today.html
4 http://17decembre.fr/cattwgory/couture/haut-couture/
5 http://1995-2015.undo.net/it/evento/48832
6 http://1995-2015.undo.net/it/mostra/62287
7 http://1995-2015.undo.net/it/mostra/71520
8 http://200.144.255.199/mostra/2013/12/ii_mostra_de_cinema_da_quebrada_no_cinusp/programacao
9 http://2014.xcoax.org/xcoax2014.pdf
10 http://2016.arcinemaargentino.com/ar-festival-de-cinema-argentino-apoio-e-agradecimentos/
11 http://28bienal.org.br/download-arquivo-jornal/28b3_baixa_1226152382.pdf
12 http://28mmphoto.over-blog.com/2015/10/ile-de-re-graffiti-m-chat-signe-thoma-vuille.html
13 http://5thingsilearnedtoday.com/blog/tag/street%2Bart
14 http://76140frontgauch.canalblog.com/archives/2019/12/03/37835602.html
15 http://81-90-53-65.addr.refertelecom.pt/centro-de-imprensa/arte-urbana-requalifica-espacos-ferroviarios
16 http://81-90-53-65.addr.refertelecom.pt/centro-de-imprensa/intervencao-de-arte-urbana-no-apeadeiro-de-marvila
17 http://823.hu/blog/fat-heat-interju-a-graffiti-vilagabol-i-1
18 http://8stmarket.com/travelogue/2017/08/31/8th-street-market-summer-camp-june-2017/
19 http://9blacklotus.com/
20 http://9eme.net/portfolio/colorama-festival-collectif-9emeconcept-artistes-urbain-streetart-digitalart-residence-francs
    colleurs-exposition-galerie/
21 http://a-casa-do-agricultor.visit-azores.info/en/
22 http://a-cooper1114-dc.blogspot.com/2012/05/45-designers-task-graffiti-street-art.html
23 http://a-sul.blogspot.com/2008/10/graffiti-o-folclore-do-regime-concuso.html
24 http://a-sul.blogspot.com/2008/10/seixal-graffiti-vandalismo.html
25 http://a-sul.blogspot.com/2012/11/
26 http://a-trompa.net/indice-tags
27 http://aawaara.blogspot.com/2012/06/comic-tour-of-tintin-land.html
28 http://abelgalois.blogspot.com/2007/01/graffiteros-espaoles.html
29 http://abnf.co/NJ-saddle_brook_bridge_and_graffiti_saddle_brook_nj.htm
```

List of more than 80 thousand URLs

# 1 - Problem: how to get thousands of pages related to street art



The screenshot shows a web browser displaying a news article from Diário de Notícias. The page is archived on Arquivo.pt, as indicated by the URL in the address bar: [dn.pt/cultura/graffiti-porque-desenhamos-nas-paredes-10514798.html](https://arquivo.pt/wayback/20191026190500/http://dn.pt/cultura/graffiti-porque-desenhamos-nas-paredes-10514798.html). The article title is "Graffiti: why do we draw on walls?". The main image is a vibrant, abstract graffiti piece featuring a face with a wide, open mouth, rendered in shades of purple, pink, and green. A person is visible in the background, standing at a podium and speaking into a microphone. The page includes a navigation menu on the left, a search bar, and social media sharing icons.

[Portuguese newspaper, dn.pt, preserved on Arquivo.pt, 2019. 1st URL on the list](https://arquivo.pt/wayback/20191026190500/http://dn.pt/cultura/graffiti-porque-desenhamos-nas-paredes-10514798.html)

1 - Problem: how to get thousands of pages related to street art

# Approaches

- I'm an expert and I choose what is important
- I'm not an expert but I'll ask for help
- **I don't need to be an expert, I just need to gather content  
and make it available**

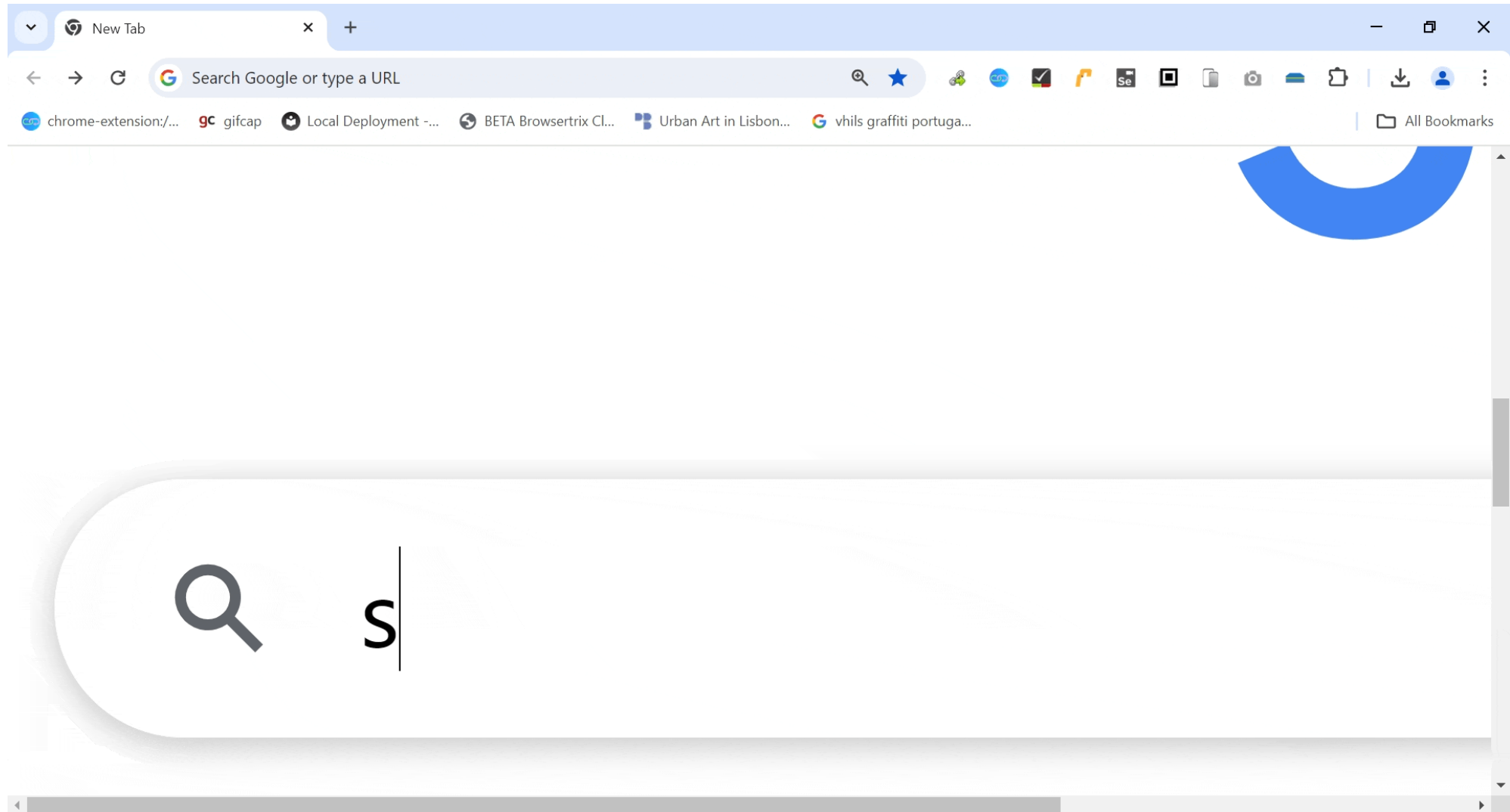


1 - Problem: how to get thousands of pages related to street art

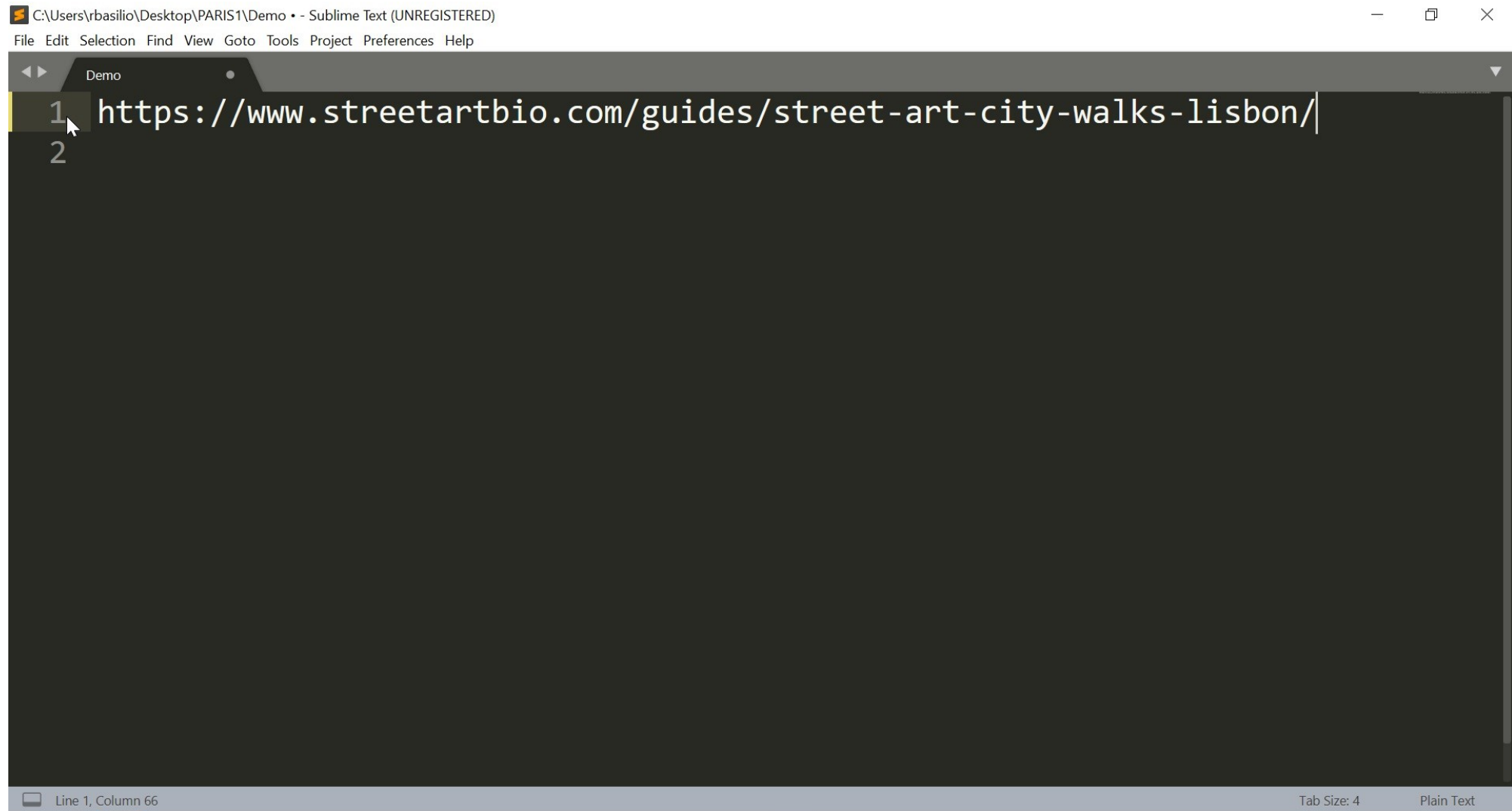
# Requirements to get started

- a computer with web access and a browser
- the ability to install extensions or applications
- a topic and a target community
- have a web archive in which to integrate the content

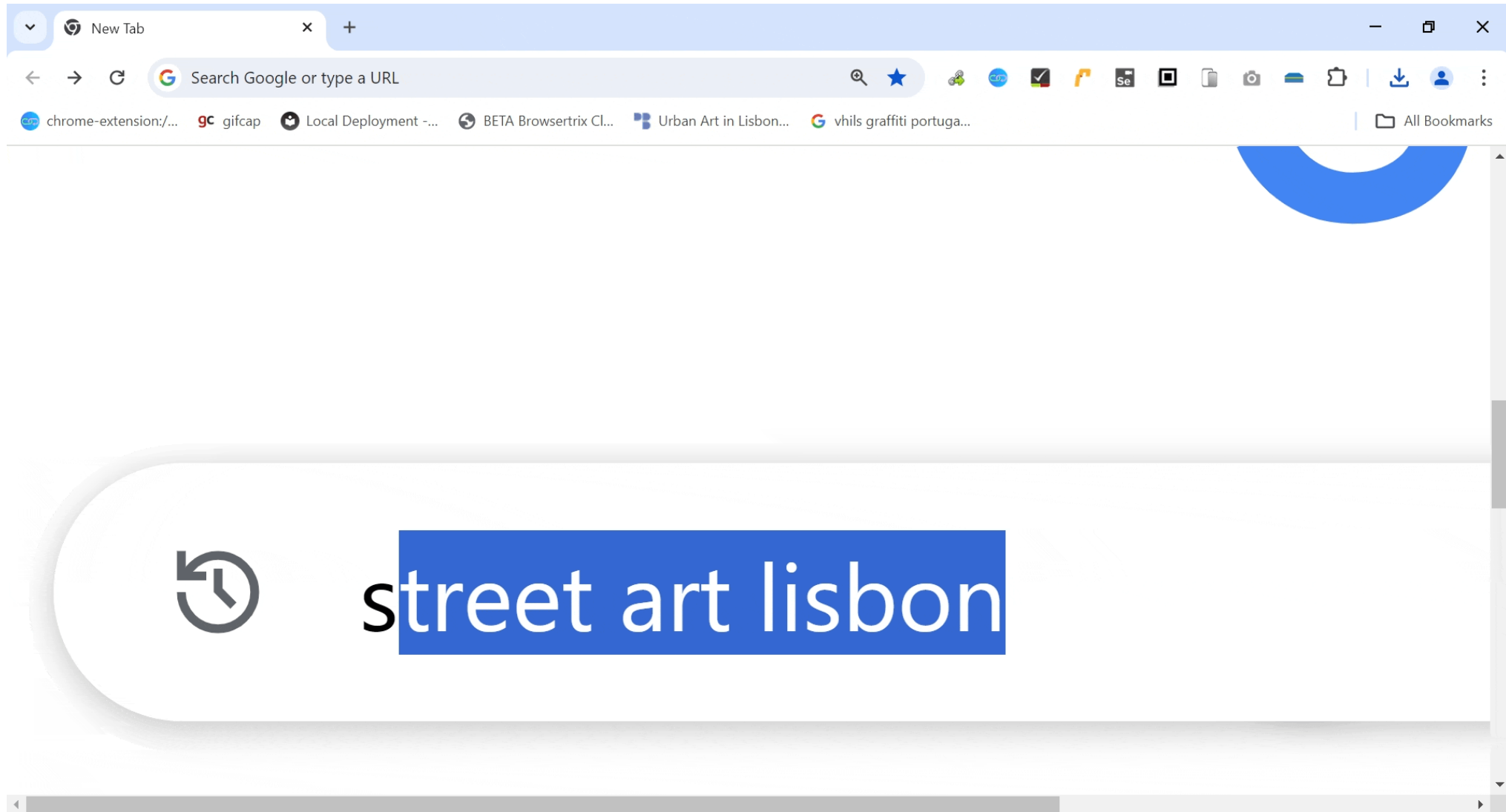
# 1 - Problem: how to get thousands of pages related to street art



# 1 - Problem: how to get thousands of pages related to street art

A screenshot of a Sublime Text editor window. The title bar reads "C:\Users\rbasilio\Desktop\PARIS1\Demo - - Sublime Text (UNREGISTERED)". The menu bar includes "File", "Edit", "Selection", "Find", "View", "Goto", "Tools", "Project", "Preferences", and "Help". The editor has a dark theme and a single tab titled "Demo". The main text area contains the URL "https://www.streetartbio.com/guides/street-art-city-walks-lisbon/" on line 1, column 66. A mouse cursor is positioned at the end of the URL. Line numbers 1 and 2 are visible on the left side. The status bar at the bottom shows "Line 1, Column 66", "Tab Size: 4", and "Plain Text".

# 1 - Problem: how to get thousands of pages related to street art



[Link Grabber, a extension for Chrome that extracts links from a list of results](#)

# 1 - Problem: how to get thousands of pages related to street art



C:\Users\rbasilio\Desktop\PARIS1\Demo - Sublime Text (UNREGISTERED)

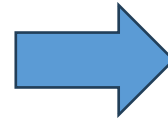
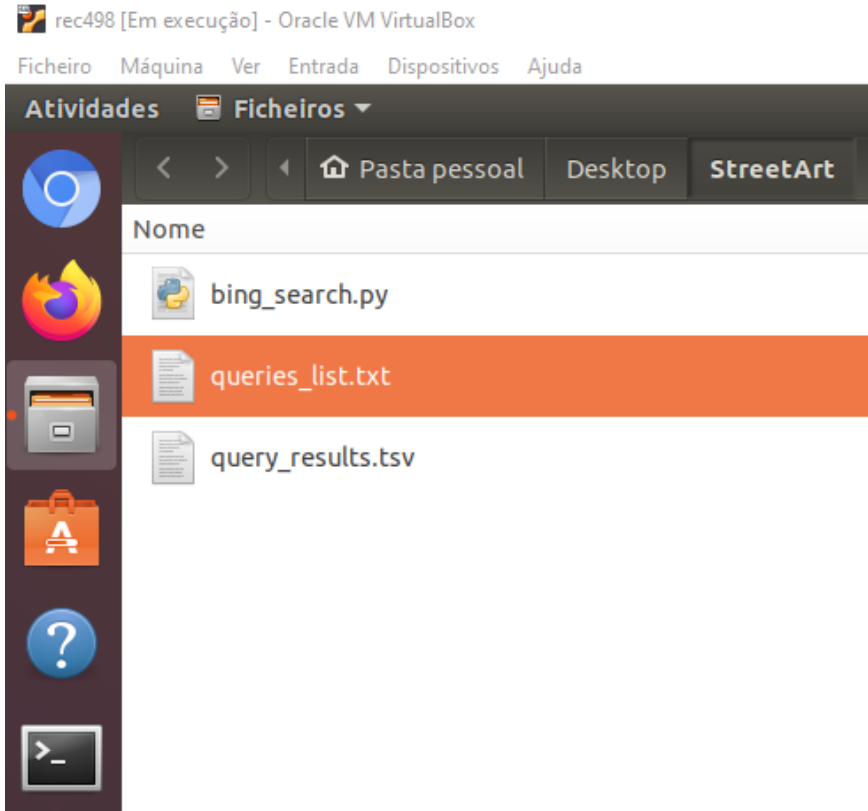
File Edit Selection Find View Goto Tools Project Preferences Help

```
Demo
65 https://www.hotelroma.pt/en/blog/street-art-tour-lisbon/
66 https://www.kuantokusta.pt/p/9277110/street-art-lisbon-vol-2-9789898729408
67 https://www.rbe.mec.pt/np4/2149.html
68 https://www.pinterest.pt/franciscapd10/street-art-lisboa/
69 https://www.tripadvisor.pt/Attraction_Review-g189158-d3906692-Reviews-Lisbon_Street_Art_Tours-Lisbon
_Lisbon_District_Central_Portugal.html
70 https://www.tripadvisor.pt/AttractionProductReview-g189158-d23744846-LISBON_Street_Art_Tour-Lisbon_L
isbon_District_Central_Portugal.html
71 https://www.belasartes.ulisboa.pt/PT/lisbon-street-art-urban-creativity-2/
72 https://run.unl.pt/bitstream/10362/16140/1/Lisbon%20Street%20Art%20Livro.pdf
73 https://urbanrevolution.pt/en/home/
74 https://www.wook.pt/livro/street-art-lisbon-vol-1/15949988
75 https://www.zestbooks.pt/product-page/street-art-lisbon-vol-2
76 https://www.zestbooks.pt/street-art-lisboa
```

Line 76, Column 43      Tab Size: 4      Plain Text

# 1 - Problem: how to get thousands of pages related to street art

## Using Bing API (on Linux)



## Queries

```
street art
street art 2upla
street art 2Carryon
street art Acar
street art Vihls
street art Add Fuel
street art Adres
street art aFase|
```

[github.com/arquivo/bing-search](https://github.com/arquivo/bing-search)

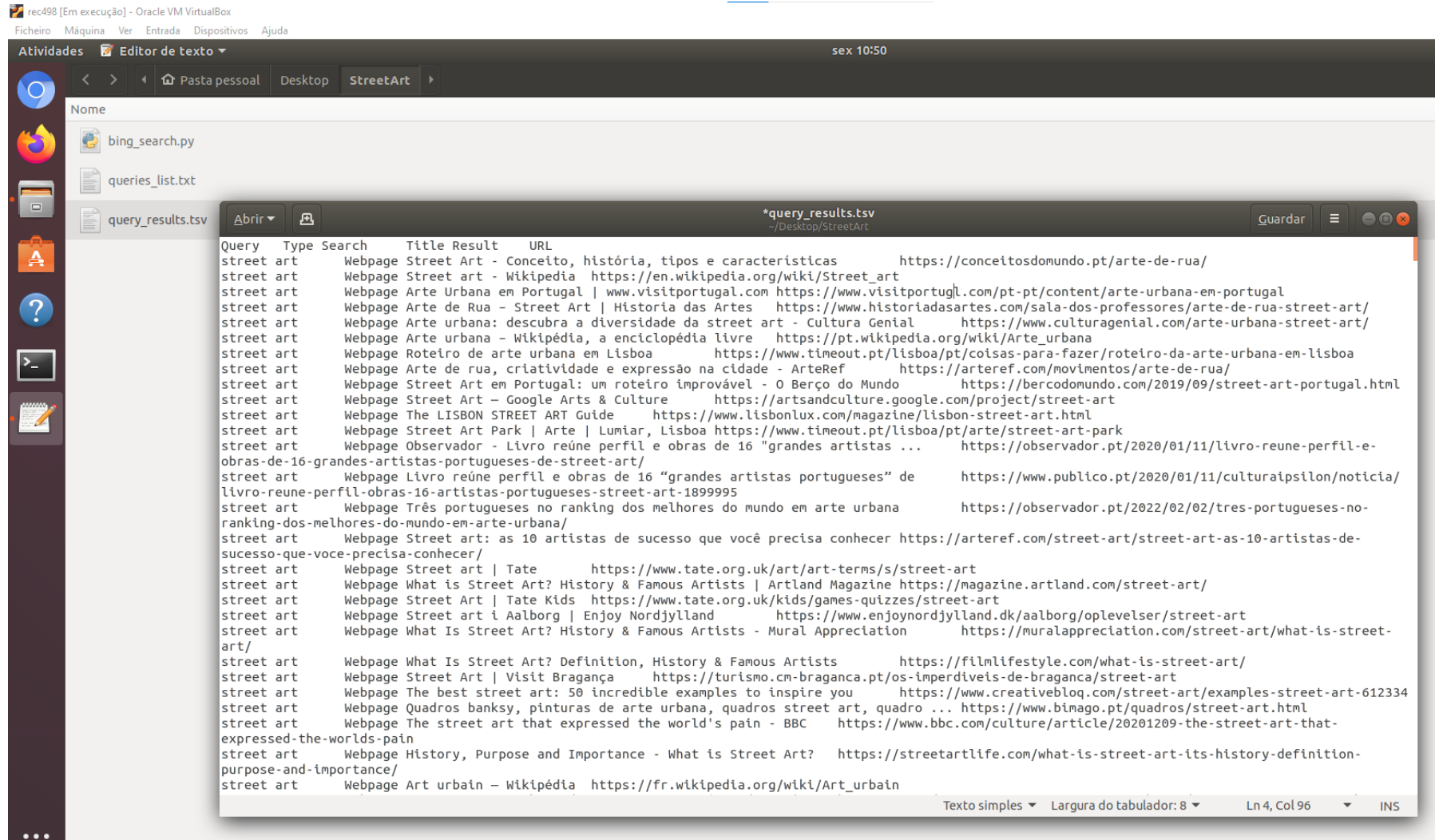
# 1 - Problem: how to get thousands of pages related to street art

## Using Bing API (on Linux)

```
ricardo@rec498: ~/Desktop/StreetArt
Ficheiro Editar Ver Procurar Terminal Ajuda
ricardo@rec498:~/Desktop/StreetArt$ python3 bing_search.py -q queries_list.txt -k [REDACTED] -n 100
```

[github.com/arquivo/bing-search](https://github.com/arquivo/bing-search)

# 1 - Problem: how to get thousands of pages related to street art



Using Bing API to find content: [github.com/arquivo/bing-search](https://github.com/arquivo/bing-search)

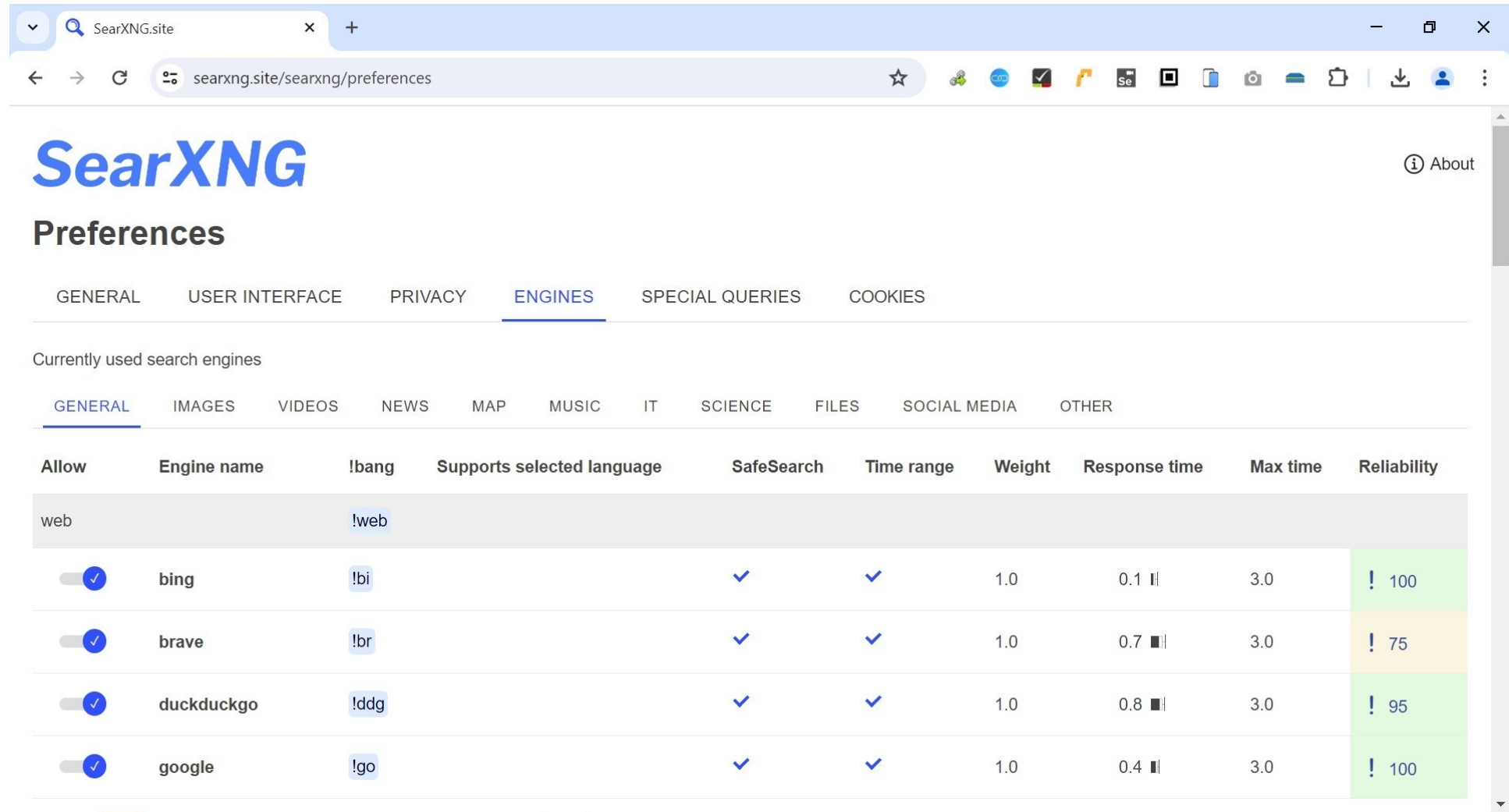


# 1 - Problem: how to get thousands of pages related to street art

	A	B	C	
1	Query	Type Search	Title Result	URL
2	street art	Webpage	Street Art - Conceito, história, tipos e características	<a href="https://conceitos">https://conceitos</a>
3	street art	Webpage	Street art - Wikipedia	<a href="https://en.wikipe">https://en.wikipe</a>
4	street art	Webpage	Arte Urbana em Portugal   www.visitportugal.com	<a href="https://www.visit">https://www.visit</a>
5	street art	Webpage	Arte de Rua – Street Art   Historia das Artes	<a href="https://www.hist">https://www.hist</a>
6	street art	Webpage	Arte urbana: descubra a diversidade da street art - Cultura Genial	<a href="https://www.cult">https://www.cult</a>
7	street art	Webpage	Arte urbana – Wikipédia, a enciclopédia livre	<a href="https://pt.wikipe">https://pt.wikipe</a>
8	street art	Webpage	Roteiro de arte urbana em Lisboa	<a href="https://www.time">https://www.time</a>
9	street art	Webpage	Arte de rua, criatividade e expressão na cidade - ArteRef	<a href="https://arteref.co">https://arteref.co</a>
10	street art	Webpage	Street Art em Portugal: um roteiro improvável - O Berço do Mundo	<a href="https://bercodom">https://bercodom</a>
11	street art	Webpage	Street Art — Google Arts & Culture	<a href="https://artsandcu">https://artsandcu</a>
12	street art	Webpage	The LISBON STREET ART Guide	<a href="https://www.lisbo">https://www.lisbo</a>
13	street art	Webpage	Street Art Park   Arte   Lumiar, Lisboa	<a href="https://www.time">https://www.time</a>
14	street art	Webpage	Observador - Livro reúne perfil e obras de 16 "grandes artistas ...	<a href="https://observado">https://observado</a>
15	street art	Webpage	Livro reúne perfil e obras de 16 “grandes artistas portugueses” de	<a href="https://www.pub">https://www.pub</a>
16	street art	Webpage	Três portugueses no ranking dos melhores do mundo em arte urbana	<a href="https://observado">https://observado</a>
17	street art	Webpage	Street art: os 10 artistas de sucesso que você precisa conhecer	<a href="https://arteref.co">https://arteref.co</a>

Using Bing API to find content: list of results (demo)

# 1 - Problem: how to get thousands of pages related to street art

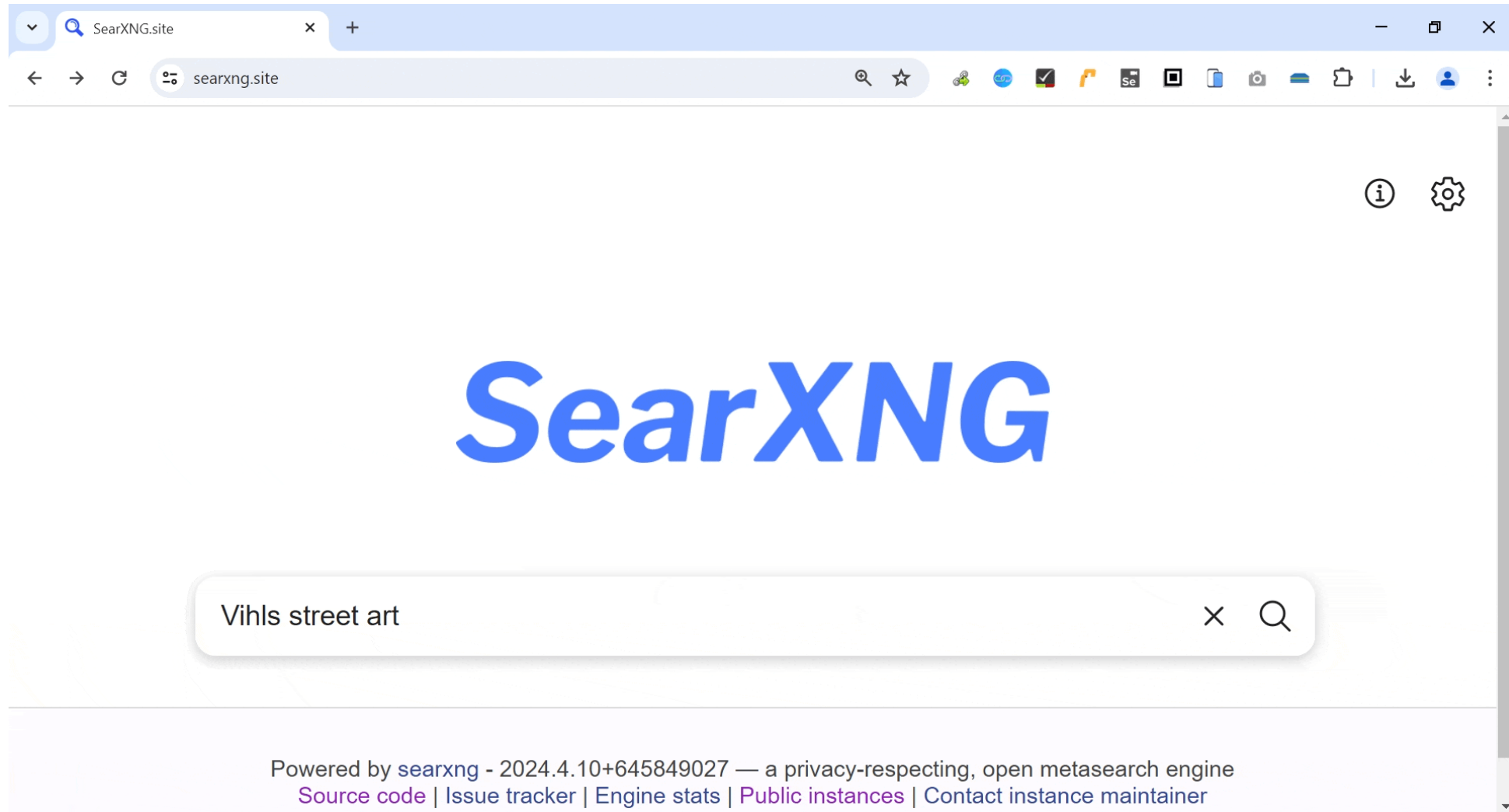


The screenshot shows the SearXNG Preferences page, specifically the 'ENGINES' tab. The page title is 'SearXNG Preferences' and it includes an 'About' link. The 'ENGINES' tab is selected, showing a table of currently used search engines. The table has columns for 'Allow', 'Engine name', '!bang', 'Supports selected language', 'SafeSearch', 'Time range', 'Weight', 'Response time', 'Max time', and 'Reliability'. The 'Reliability' column shows a percentage of engines that are working, indicated by a green background for 100% and a yellow background for 75%.

Allow	Engine name	!bang	Supports selected language	SafeSearch	Time range	Weight	Response time	Max time	Reliability
	web	!web							
<input checked="" type="checkbox"/>	bing	!bi		✓	✓	1.0	0.1	3.0	! 100
<input checked="" type="checkbox"/>	brave	!br		✓	✓	1.0	0.7 ■	3.0	! 75
<input checked="" type="checkbox"/>	duckduckgo	!ddg		✓	✓	1.0	0.8 ■	3.0	! 95
<input checked="" type="checkbox"/>	google	!go		✓	✓	1.0	0.4	3.0	! 100

[SearXNG, tool to expand a search on many search engines. Also with API](#)

# 1 - Problem: how to get thousands of pages related to street art



[SearXNG, tool to expand a search on many search engines engines. Also with API](#)

# 1 - Problem: how to get thousands of pages related to street art

```
C:\Users\rbasilio\Desktop\PARIS1\Demo • - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help

Demo
65 https://www.hotelroma.pt/en/blog/street-art-tour-lisbon/
66 https://www.kuantokusta.pt/p/9277110/street-art-lisbon-vol-2-9789898729408
67 https://www.rbe.mec.pt/np4/2149.html
68 https://www.pinterest.pt/franciscapd10/street-art-lisboa/
69 https://www.tripadvisor.pt/Attraction_Review-g189158-d3906692-Reviews-Lisbon_Street_Art_Tours-Lisbon
_Lisbon_District_Central_Portugal.html
70 https://www.tripadvisor.pt/AttractionProductReview-g189158-d23744846-LISBON_Street_Art_Tour-Lisbon_L
isbon_District_Central_Portugal.html
71 https://www.belasartes.ulisboa.pt/PT/lisbon-street-art-urban-creativity-2/
72 https://run.unl.pt/bitstream/10362/16140/1/Lisbon%20Street%20Art%20Livro.pdf
73 https://urbanrevolution.pt/en/home/
74 https://www.wook.pt/livro/street-art-lisbon-vol-1/15949988
75 https://www.zestbooks.pt/product-page/street-art-lisbon-vol-2
76 https://www.zestbooks.pt/street-art-lisboa

Line 76, Column 43 Tab Size: 4 Plain Text
```

[SearXNG, tool to expand a search on many search engines engines. Also with API](#)

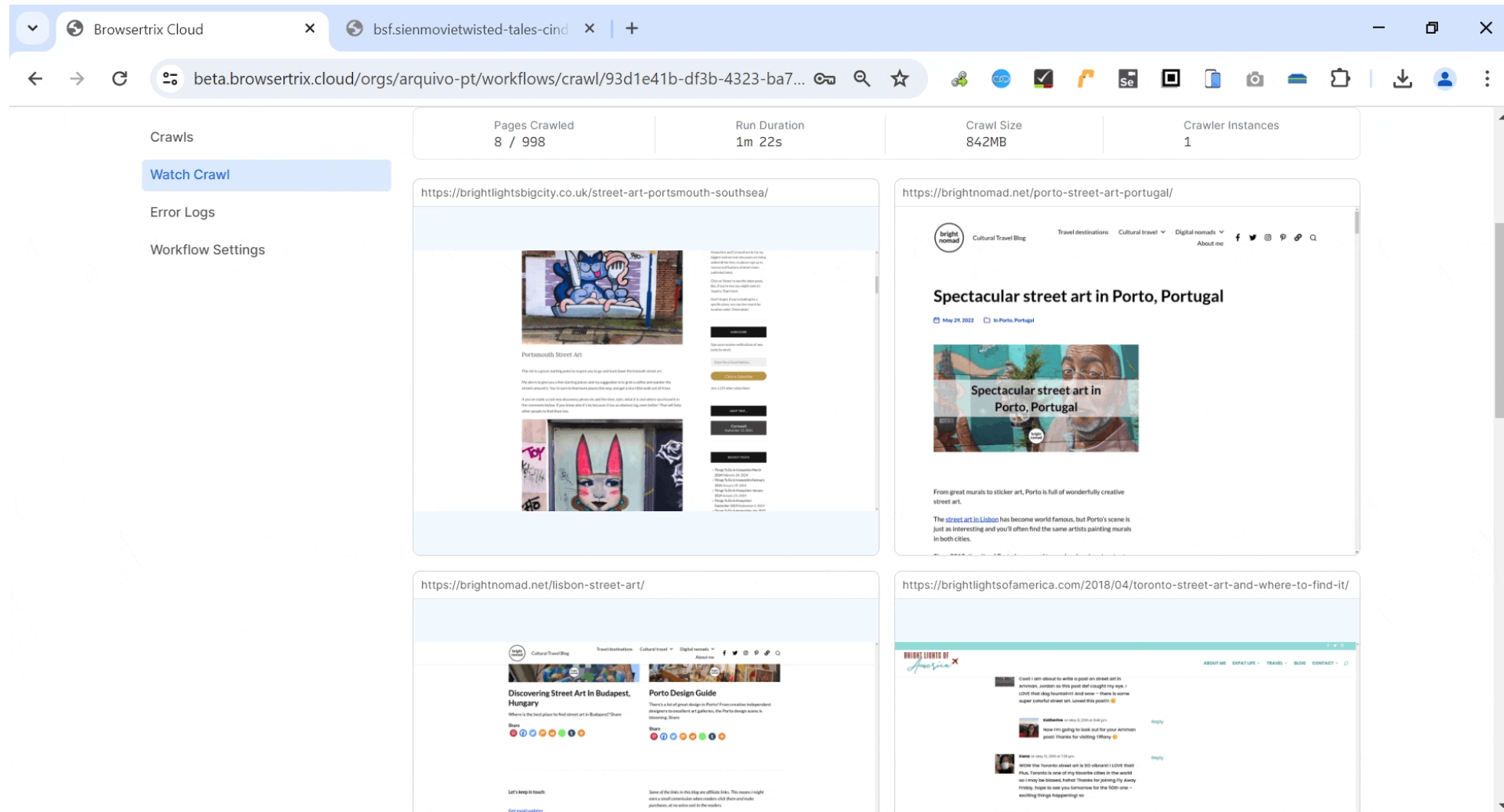
## 2 - Problem: how to get thousands of pages related to street art

# Next...

- Selection of content about street art by experts
- Collaboration with the office for street art in Lisbon (GAU team)
- Contributes to curated collections (the IIPC collaborative collection "Street Art" on Archive-It)
- Online exhibitions at Arquivo.pt

# 3 - Recording and making the content available

# 3 - Recording and making the content available



The screenshot displays the Browsertrix Cloud interface for a crawl workflow. At the top, the browser address bar shows the URL: `beta.browsertrix.cloud/orgs/arquivo-pt/workflows/crawl/93d1e41b-df3b-4323-ba7...`. Below the address bar, a summary bar provides the following statistics:

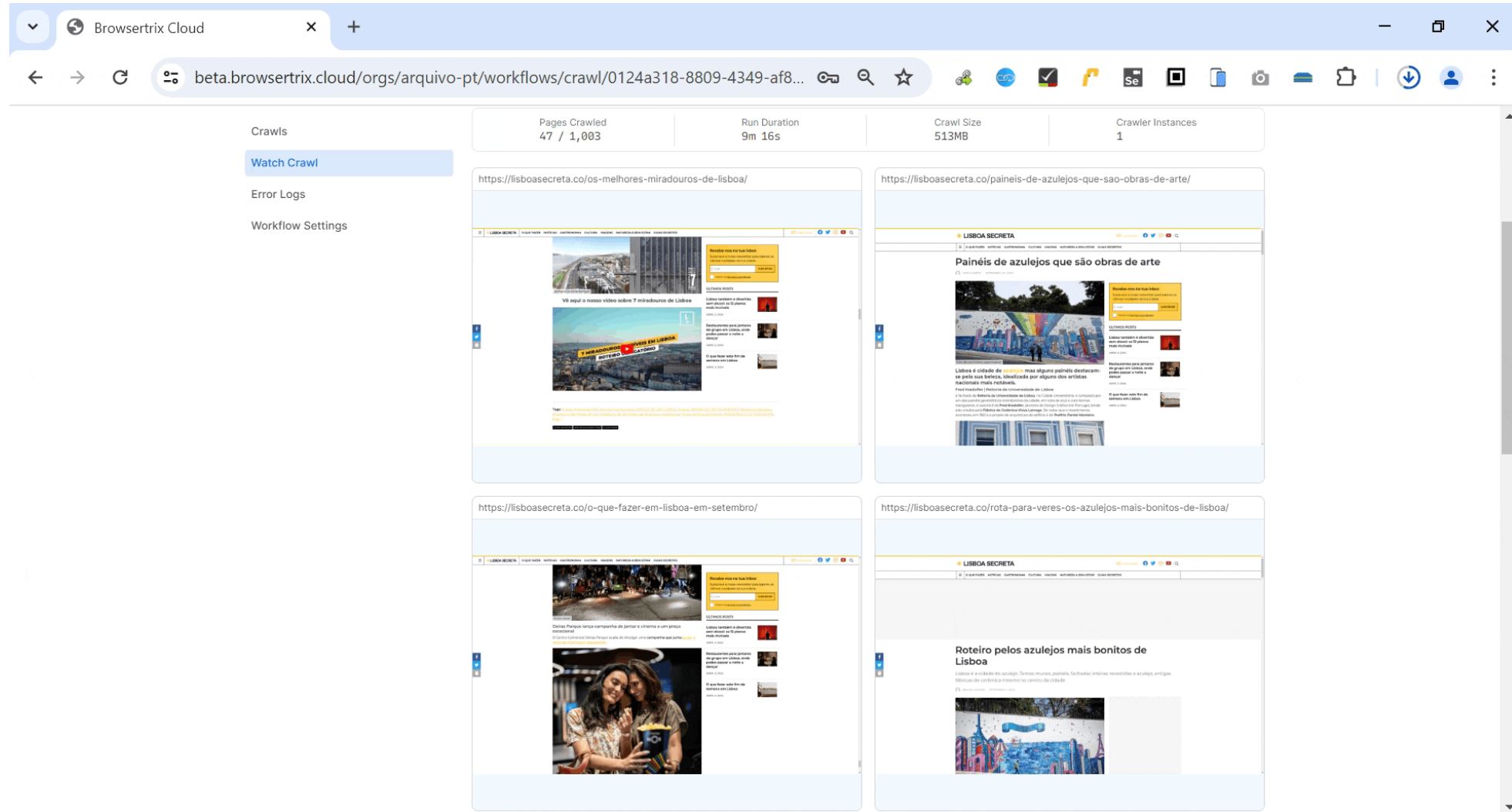
- Pages Crawled: 8 / 998
- Run Duration: 1m 22s
- Crawl Size: 842MB
- Crawler Instances: 1

On the left side, a navigation menu includes: Crawls, Watch Crawl (highlighted), Error Logs, and Workflow Settings. The main content area displays a grid of four recorded web pages:

- `https://brightlightsbigcity.co.uk/street-art-portsmouth-southsea/`: A page about Portsmouth Street Art featuring a cartoon illustration of a blue cat.
- `https://brightnomad.net/porto-street-art-portugal/`: A page titled "Spectacular street art in Porto, Portugal" with a header image of a mural.
- `https://brightnomad.net/lisbon-street-art/`: A page titled "Discovering Street Art in Budapest, Hungary" and "Porto Design Guide".
- `https://brightlightsofamerica.com/2018/04/toronto-street-art-and-where-to-find-it/`: A page titled "BRIGHT LIGHTS OF America" featuring a comment section with user feedback.

Demo A: recording using Browsertrix-cloud (beta).

# 3 - Recording and making the content available



The screenshot displays the Browsertrix Cloud interface for a crawl workflow. At the top, the browser address bar shows the URL: `beta.browsertrix.cloud/orgs/arquivo-pt/workflows/crawl/0124a318-8809-4349-af8...`. Below the address bar, a summary bar provides the following statistics:

- Pages Crawled: 47 / 1,003
- Run Duration: 9m 16s
- Crawl Size: 513MB
- Crawler Instances: 1

On the left side, there is a navigation menu with the following options:

- Crawls
- Watch Crawl (highlighted)
- Error Logs
- Workflow Settings

The main content area shows a grid of four article thumbnails from the website 'LISBOA SECRETA'. The thumbnails are:

- Top-left: `https://lisboasecreta.co/os-melhores-miradouros-de-lisboa/` - Article about the best viewpoints in Lisbon.
- Top-right: `https://lisboasecreta.co/paineis-de-azulejos-que-sao-obras-de-arte/` - Article about azulejo panels that are works of art.
- Bottom-left: `https://lisboasecreta.co/o-que-fazer-em-lisboa-em-setembro/` - Article about things to do in Lisbon in September.
- Bottom-right: `https://lisboasecreta.co/rota-para-veres-os-azulejos-mais-bonitos-de-lisboa/` - Article about the best routes to see the most beautiful azulejos in Lisbon.

Demo B: recording using Browsertrix-cloud (beta).



## 3 - Recording and making the content available

# Options and settings

- **URL List (single page)**
- Page Behavior Timeout: 5 minutes
- Auto-Scroll Behavior
- **Delay Before Next Page: 10 seconds**
- Crawler Instances: 1x

# 3 - Recording and making the content available

## Thanks to Webrecorder.net

Browsertrix Cloud

Overview **Crawling** Archived Items Collections Browser Profiles Org Settings

### Crawl Workflows

[New Workflow](#)

Search all Workflows by name or Crawl Start URL Sort by: Latest Crawl

[All](#) Scheduled No schedule Show Only Mine

Name & Schedule	Latest Crawl	Total Size	Last Modified
StreetArt-27 Manual run by ricardobasiliofcsh	Complete 08/04/24, 10:35 in 1h	3.68GB 1 crawl	ricardobasiliofcsh 08/04/24, 07:57
StreetArt-26 Manual run by ricardobasiliofcsh	Complete 08/04/24, 04:38 in 4h	7.78GB 1 crawl	ricardobasiliofcsh 07/04/24, 23:14
StreetArt-25 Manual run by ricardobasiliofcsh	Complete 07/04/24, 23:57 in 2h	9.12GB 1 crawl	ricardobasiliofcsh 07/04/24, 20:22

Recordings of 1000 URLs using Browsertrix-cloud (beta).

# 3 - Recording and making the content available

## Thanks to Webrecorder.net

REPLAY WEBPAGE / Street Art 80000 / Pages

search://view=pages

Pages Resources

Search by Page URL, Title, or Text

STREET ART 80000	Date	Page Title
<i>80\,000 pages about street art in Portugal and around the world put in recording on Browsertrix\ - cloud\ . This is also a proof of concept about how to collect a large amount of content within a given topic\ . This collection will be integrated in Arquivo\,pt\ .</i>  <i>100 of 14768 Pages Shown</i>	04/04/2024 14:34:49	<b>Berlin: festivais de música, graffiti, lugares. Feriados alemães - Aprenda alemão online - Comece a Grande Feira Alemã no Reno</b> <a href="https://collectiontravels.ru/pt/velikobritaniya/berlin-muzykalnye-festivali-graffiti-mesta-nemeckie/">https://collectiontravels.ru/pt/velikobritaniya/berlin-muzykalnye-festivali-graffiti-mesta-nemeckie/</a>
	03/04/2024 18:12:14	<b>"Déplacé.e.s" by JR @ Mugombwa, Rwanda – Barbara Picci</b> <a href="https://barbarapicci.com/2023/01/30/streetart-deplace-e-s-by-jr-mugombwa-rwanda/">https://barbarapicci.com/2023/01/30/streetart-deplace-e-s-by-jr-mugombwa-rwanda/</a>
	03/04/2024 12:33:50	<b>Michiko &amp; Hatchin – 15/16 [Graffiti in Vain/Etude of Crimson Inconstancy] – Throwback Thursday - All Things Anime</b> <a href="https://anime.atsit.in/archives/52243">https://anime.atsit.in/archives/52243</a>

Replay on with ReplayWeb.page: Browsertrix-cloud (beta): [tinyurl.com/street-art-browsertrix](https://tinyurl.com/street-art-browsertrix)

# 3 - Recording and making the content available

- 20240401211823703-c52e6d45-037-0.wacz
- 20240402172803832-ff1c09f5-389-0.wacz
- 20240402200121003-d5d8116f-255-0.wacz
- 20240403100308193-a8836d80-248-0.wacz
- 20240403141048115-5d9c536a-3db-0.wacz
- 20240403205534109-2ee82b98-054-0.wacz
- 20240404102347993-93d1e41b-df3-0.wacz
- 20240404235538944-af855aff-645-0.wacz
- 20240405171407643-7765b099-538-0.wacz
- 20240405214950234-aafc0982-f50-0.wacz
- 20240406121833779-338c598a-b6e-0.wacz
- 20240406201632567-06132efa-d4a-0.wacz
- 20240407113922274-0124a318-880-0.wacz
- 20240407173630453-6181a6fe-c3b-0.wacz
- 20240407193517495-f984314d-bd8-0.wacz
- 20240408093437661-c9a5feba-7c3-0.wacz
- 20240409172507006-aceb713c-ab1-0.wacz
- manual-20240402104152-d2f3fde4-0dc.log
- manual-20240403082003-72f8f1df-da6.log
- manual-20240403110902-ea632ee3-d04.log
- manual-20240403183912-2ee82b98-054.loc
- 20240401223905031-4148099a-c56-0.wacz
- 20240402180402224-ff1c09f5-389-0.wacz
- 20240402224312598-c0cbbfa6-9b4-0.wacz
- 20240403103035315-cf772b98-a84-0.wacz
- 20240403141816708-5d9c536a-3db-0.wacz
- 20240404021447742-88d1600e-e57-0.wacz
- 20240404141844388-25e3c38f-620-0.wacz
- 20240405103254104-c0585076-8bb-0.wacz
- 20240405172545987-7765b099-538-0.wacz
- 20240406011556141-8b7b73f8-f75-0.wacz
- 20240406123419644-338c598a-b6e-0.wacz
- 20240407015251435-13557c0f-c41-0.wacz
- 20240407124401706-0124a318-880-0.wacz
- 20240407175153992-6181a6fe-c3b-0.wacz
- 20240407225517301-856cc9c2-b6d-0.wacz
- 20240408211847317-f3a78aec-ca1-0.wacz
- 20240409193841282-381958a2-7da-0.wacz
- manual-20240402183018-d5d8116f-255.log
- manual-20240403094008-a8836d80-248.log
- manual-20240403112449-5d9c536a-3db.log
- manual-20240403212946-88d1600e-e57.loc
- 20240402121337040-d2f3fde4-0dc-0.wacz
- 20240402194404333-d5d8116f-255-0.wacz
- 20240403090438276-72f8f1df-da6-0.wacz
- 20240403112200191-ea632ee3-d04-0.wacz
- 20240403172838633-88c23ea5-611-0.wacz
- 20240404100723808-93d1e41b-df3-0.wacz
- 20240404182614972-bf1cd127-c65-0.wacz
- 20240405140829766-fa6b9251-525-0.wacz
- 20240405211848346-aafc0982-f50-0.wacz
- 20240406013931160-8b7b73f8-f75-0.wacz
- 20240406182459230-06132efa-d4a-0.wacz
- 20240407111838532-0124a318-880-0.wacz
- 20240407131248008-0124a318-880-0.wacz
- 20240407190313785-f984314d-bd8-0.wacz
- 20240408033541662-767f084a-7de-0.wacz
- 20240409003341062-0acc878e-c5b-0.wacz
- manual-20240401213540-4148099a-c56.log
- manual-20240402213345-c0cbbfa6-9b4.log
- manual-20240403100726-cf772b98-a84.log
- manual-20240403145532-88c23ea5-611.log
- manual-20240404081311-93d1e41b-df3.loc

Content **exported** and stored on a local drive: WACZ files and error logs (pages not recorded)

# 3 - Recording and making the content available



Content exported and stored on a local drive

# 3 - Recording and making the content available



[WARCs files were integrated and made available on Arquivo.pt.](#)

# 3 - Recording and making the content available

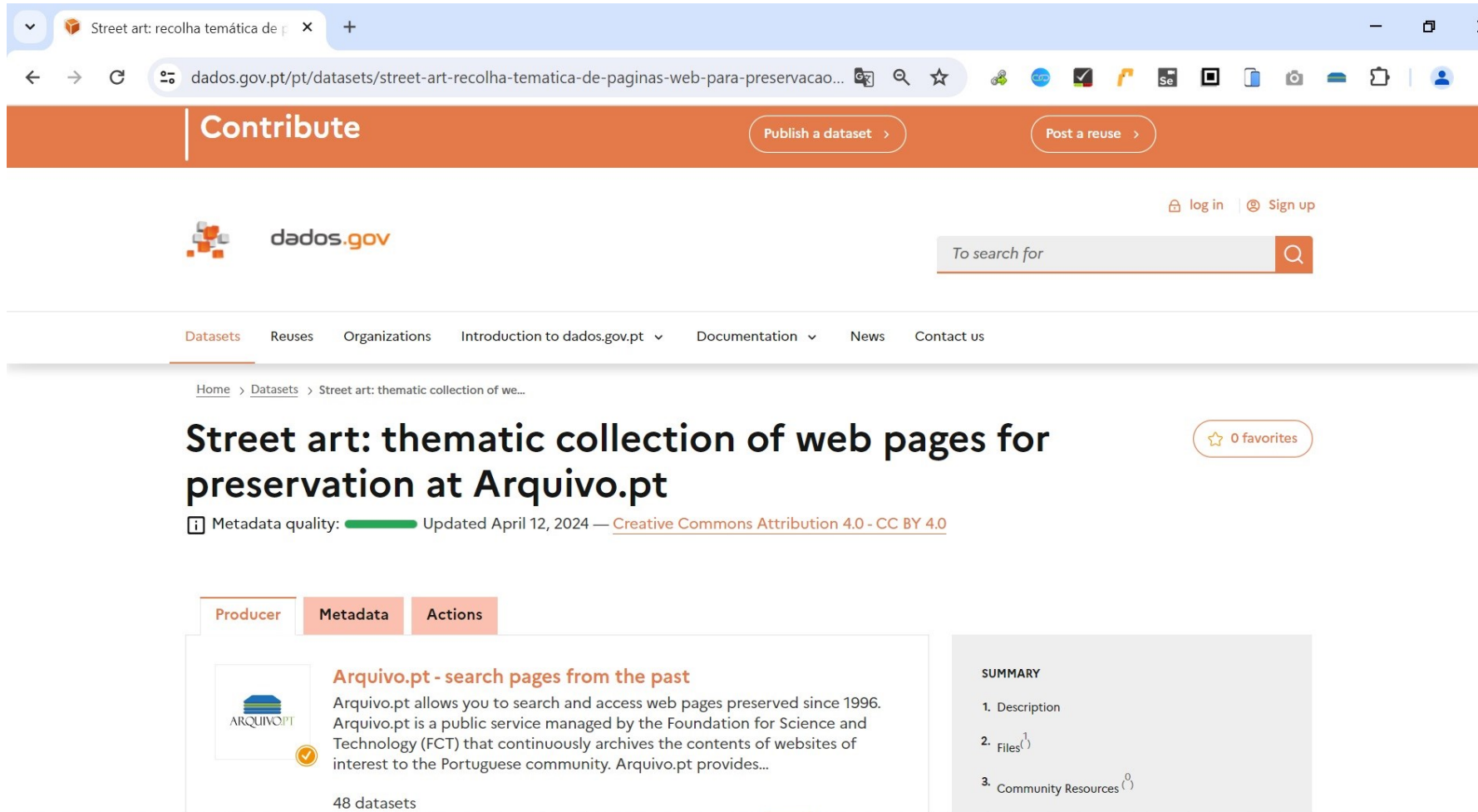


ARQUIVO.PT



[Page from the blog Ai Que Medo, aquimequedo.com.br, 2023](https://aquimequedo.com.br)

# 3 - Recording and making the content available



The screenshot shows a web browser window displaying the 'dados.gov.pt' website. The page title is 'Street art: thematic collection of web pages for preservation at Arquivo.pt'. The page features a navigation bar with 'Contribute', 'Publish a dataset', and 'Post a reuse' buttons. Below the navigation bar, there is a search bar and a menu with options like 'Datasets', 'Reuses', 'Organizations', 'Introduction to dados.gov.pt', 'Documentation', 'News', and 'Contact us'. The main content area includes a breadcrumb trail 'Home > Datasets > Street art: thematic collection of we...', a title 'Street art: thematic collection of web pages for preservation at Arquivo.pt', and a '0 favorites' button. Below the title, there is a metadata quality indicator (a green bar) and the text 'Updated April 12, 2024 — Creative Commons Attribution 4.0 - CC BY 4.0'. The page is divided into three tabs: 'Producer', 'Metadata', and 'Actions'. The 'Producer' tab is active, showing the Arquivo.pt logo and a description: 'Arquivo.pt - search pages from the past. Arquivo.pt allows you to search and access web pages preserved since 1996. Arquivo.pt is a public service managed by the Foundation for Science and Technology (FCT) that continuously archives the contents of websites of interest to the Portuguese community. Arquivo.pt provides...'. Below the description, it says '48 datasets'. To the right, there is a 'SUMMARY' section with a list of items: '1. Description', '2. Files (1)', and '3. Community Resources (1)'.

[URLs made available at the Public Administration Open Data portal](#)



# 4 – Conclusions.

## Difficulties and lessons learned

# 4 - Difficulties and lessons learned

## Difficulties

- Keep or adapt the scope (Portugal or international?)
- Clean the list of URL's (spam, nsfw, broken URLs, no SSL, social media)
- Limit to 1000 URLs for a recording workflow x80

## 4 - Difficulties and lessons learned

# Lessons

- **Search engines** prepare the selection by human experts
- **AI technologies** to build collections are the future
- **Single page** recording save resources and space
- **Community engagement** is the beginning and the goal of creating archive collections on the Web

# You can count on Arquivo.pt

SavePageNow: [arquivo.pt/savepagenow](https://arquivo.pt/savepagenow)

Automatic access via API: [arquivo.pt/api](https://arquivo.pt/api)

# Arquivo.pt Award 2024 (annual): open applications

- **10 000€** for the winner
- Works that make use of Arquivo.pt
- Applications until **6 May 2024**
- High Patronage of **President of Portugal**
- **Share!**
- Know more: [arquivo.pt/award](https://arquivo.pt/award)



Contact us

[contacto@arquivo.pt](mailto:contacto@arquivo.pt)