

# The 4CAT Capture & Analysis Toolkit



Stijn Peeters  
@stijnstijn  
@stijn@oulipo.social  
stijn.peeters@uva.nl

# 4CAT CAPTURE & ANALYSIS TOOLKIT

An **open-source** and **Web-based** research toolkit designed to **collect and analyse data** from various online sources.

Ongoing development!

<https://4cat.nl>

The screenshot shows the 4CAT web interface. At the top, there is a red header with the title "4CAT: Capture and Analysis Toolkit". Below the header is a dark blue navigation bar with links: "Create dataset", "Past results", "Data overview", "API Access", "Control Panel", "FAQ", and "About". A red notification bar below the navigation bar states: "The 4chan data source is available once more." The main content area is divided into two columns. The left column is titled "Create new dataset" and contains a form with the following elements: a text area with instructions about dataset creation and a note about automatic deletion; a "Data source:" dropdown menu set to "Reddit" with an "external" button and a link "How is this data collected?"; a text area with a notice about keyword searching; an "API version:" dropdown menu set to "Regular" with a help icon; and a "Subreddit(s):" text input field with a help icon. The right column is titled "Dataset status" and contains a "Queue" section showing "Currently processing 1 search query: facebook (1)" and a "Results" section which is currently empty.

# 4CAT CAPTURE & ANALYSIS TOOLKIT

## Create new dataset

Please be considerate of other users; 4CAT is a shared resource and large dataset queries may prevent others from using it if they take a very long time to complete. We recommend to start with smaller date ranges and specific queries and then cast a wider net if needed.

Note that datasets will be deleted automatically after 2 weeks. You can choose to keep the dataset for longer from the result page.

Data source:   [How is this data collected?](#)

Results are limited to 5 million items maximum. Be sure to read the [query syntax](#) for local data sources first - your query design will significantly impact the results. Note that large queries can take a long time to complete!

Board:

Post contains:

Subject contains:

Type to filter

- European countries
- Afghanistan
- Aland Islands

## Dataset status

Waiting for input...

## Queue

Currently processing 1 search query:

## Results

Capture

## Conversion

<input type="button" value="Run"/>	<b>Convert to Excel-compatible CSV</b> 1 Change a CSV file so it works with Microsoft Excel.
<input type="button" value="Run"/>	<b>Convert to TCAT JSON</b> 1 Convert a Twitter dataset to a TCAT-compatible format. This file can then be uploaded to TCAT.
<input type="button" value="Options"/>	<b>Merge texts</b> Merges the data from the body column into a single text file. The result can be used for word clouds, word trees, etc.

## Filtering

<input type="button" value="Run"/>	<b>Remove author information</b> Anonymises a dataset by removing content of any column starting with 'author'
<input type="button" value="Options"/>	<b>Filter by value</b> A generic filter that checks whether a value in a selected column matches a custom requirement. This will create a new dataset.
<input type="button" value="Options"/>	<b>Filter by date</b> Retains posts between given dates. This will create a new dataset.
<input type="button" value="Options"/>	<b>Filter by words or phrases</b> 1 Retains posts that contain selected words or phrases, including preset word lists. This creates a new dataset.
<input type="button" value="Options"/>	<b>Random sample</b> Retain a pseudorandom set of posts. This creates a new dataset.
	<b>Filter for unique posts</b>

Analysis

# CREATE DATASETS FROM VARIOUS DATA SOURCES



**Reddit**  
(via Pushshift)



**4chan**



**TikTok**  
(via Zeeschuimer)



**Twitter**  
(Search API)



**Tumblr**



**Instagram**  
(via Zeeschuimer)



**Telegram**



**BitChute**

**...and more**

## HOW DOES 4CAT WORK?

[demonstration]

## USE CASES / QUESTIONS TO ASK OF 4CAT

- How did the phrase 'OK Boomer' originate on Reddit?
- What kind of content are Taylor Swift fans posting on Tumblr under the #tswiftedits hashtag?
- What kind of videos do people link to in neo-nazi threads on 4chan?
- How is the 'Katyusha' song appropriated in TikTok posts about the war in Ukraine?
- Where do the memes people post on Reddit come from?

## WRAPPING UP

4CAT is on the web at [4cat.nl](http://4cat.nl)

And on Mastodon at [@4cat@mastodon.digitalmethods.net](https://mastodon.digitalmethods.net/@4cat)

And you can e-mail us at [4cat@oilab.nl](mailto:4cat@oilab.nl)

Questions very welcome!