gesis Leibniz Institute for the Social Sciences

Automatic Sampling and Analysis of YouTube Data

Tools for collecting YouTube data

Johannes Breuer, Annika Deubel, & M. Rohangis Mohseni

February 14th, 2023

Leibniz Association

How to Collect *YouTube* Data

There are many different ways in which data from *YouTube* and other social media can be collected (see Breuer et al., 2020):

- Manually (e.g., via copy & paste and manual content analysis)
- Using existing data, such as *YouNiverse: Large-Scale Channel and Video Metadata from English YouTube* (also see the accompanying preprint by Ribeiro & West, 2021)
- Automatically via the *YouTube* API or web scraping

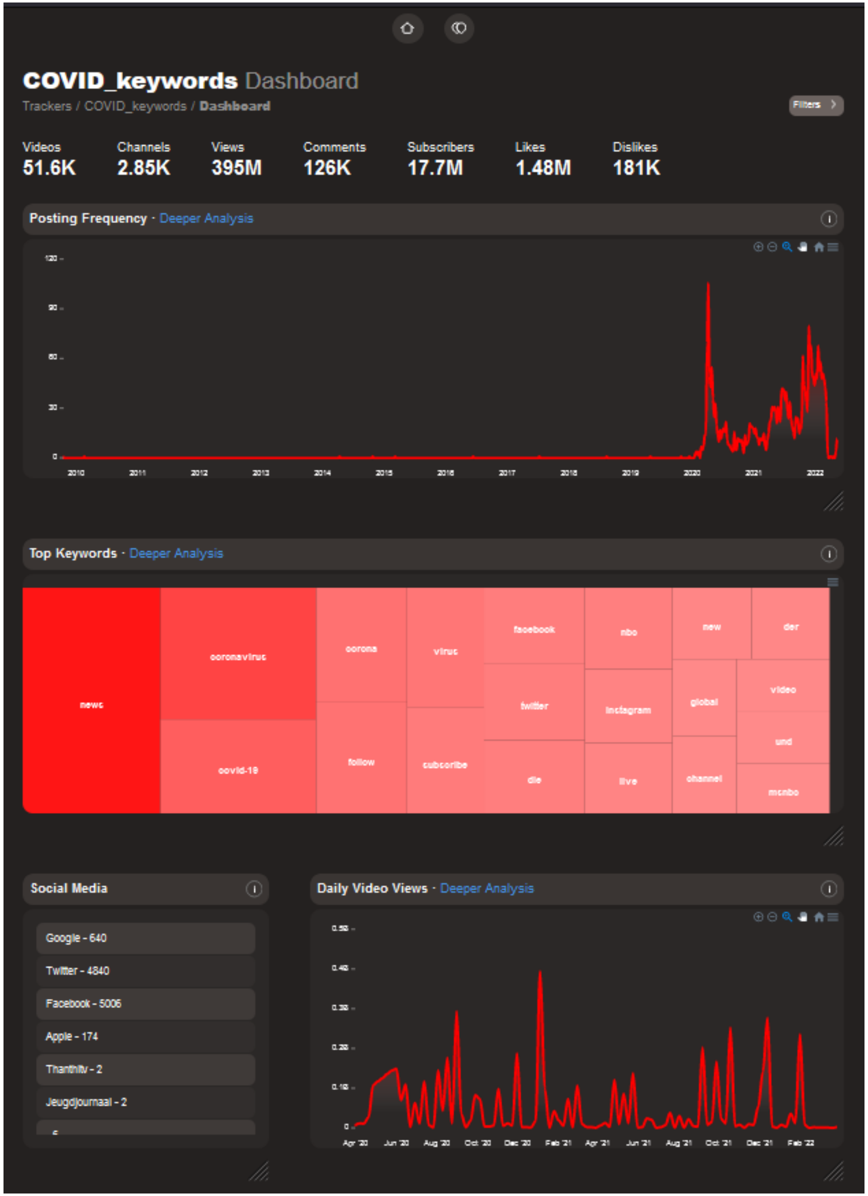
Identifying Relevant Channels or Videos

If new data is collected, it is necessary to identify relevant channels and videos for the sample.

- VTracker
- Socialblade
- YouTube Channel Crawler

VTracker

- Search for and tracking of videos
- Low-key analysis such as engagement, keyword trends, influence detection
- Creation of Dashboard for different metrics
- Data can't be collected for further analysis
- Still a bit buggy



COVID_keywords Dashboard

Trackers / COVID_keywords / Dashboard

Videos 51.6K

Channels 2.85K

Views 395M

Comments 126K

Subscribers 17.7M

Likes 1.48M

Dislikes 181K

Posting Frequency · Deeper Analysis

Top Keywords · Deeper Analysis

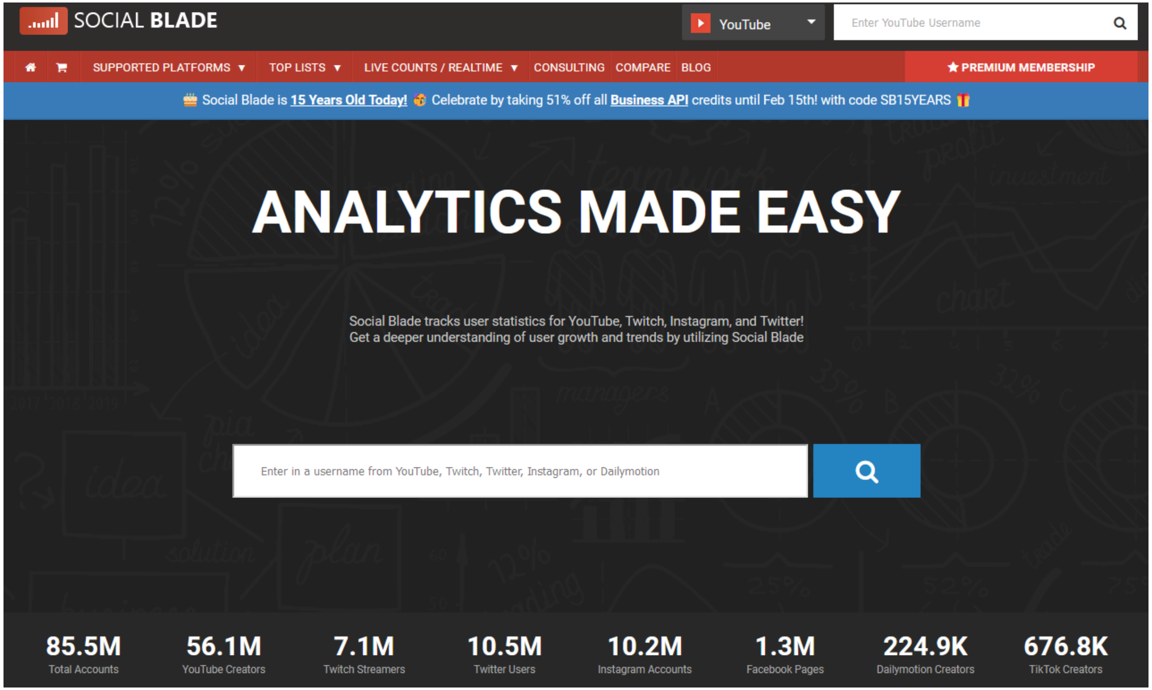
Social Media

- Google - 640
- Twitter - 4840
- Facebook - 5006
- Apple - 174
- Thanthitv - 2
- Jeugdjournaal - 2

Daily Video Views · Deeper Analysis

Socialblade

- Ranked lists of channels
- Useful if there are no content-related criteria for channel selection



Social Blade is 15 Years Old Today! Celebrate by taking 51% off all Business API credits until Feb 15th! with code SB15YEARS

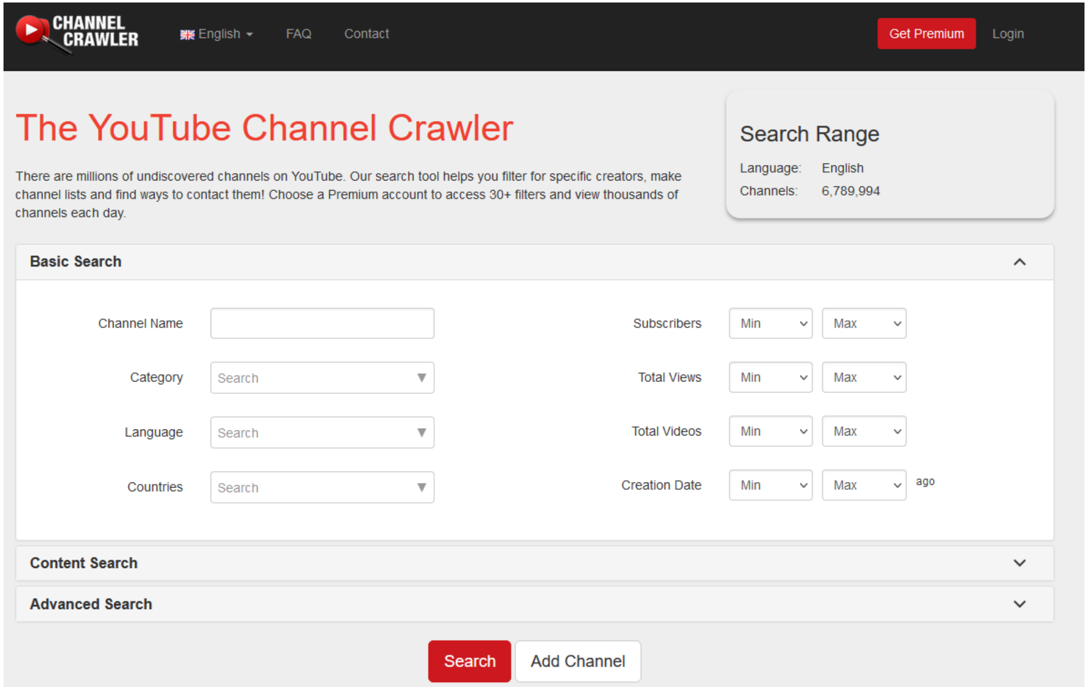
ANALYTICS MADE EASY

Social Blade tracks user statistics for YouTube, Twitch, Instagram, and Twitter!
Get a deeper understanding of user growth and trends by utilizing Social Blade

85.5M	56.1M	7.1M	10.5M	10.2M	1.3M	224.9K	676.8K
Total Accounts	YouTube Creators	Twitch Streamers	Twitter Users	Instagram Accounts	Facebook Pages	Dailymotion Creators	TikTok Creators

YouTube Channel Crawler

- Search for channels with the help of filters (e.g. language, likes)
- Useful if there are no content-related criteria for channel selection



The YouTube Channel Crawler

There are millions of undiscovered channels on YouTube. Our search tool helps you filter for specific creators, make channel lists and find ways to contact them! Choose a Premium account to access 30+ filters and view thousands of channels each day.

Search Range

Language: English
Channels: 6,789,994

Basic Search

Channel Name
Category: Search
Language: Search
Countries: Search
Subscribers: Min Max
Total Views: Min Max
Total Videos: Min Max
Creation Date: Min Max ago

Excluding Problematic Channels

- YouTube Wiki
  - Social background information on channels (only in German)
  - Useful to identify reasons for exclusion (e.g., fight between channels)

If the relevant channels are identified and potentially problematic channels are excluded, the next step would be to sample the comments.

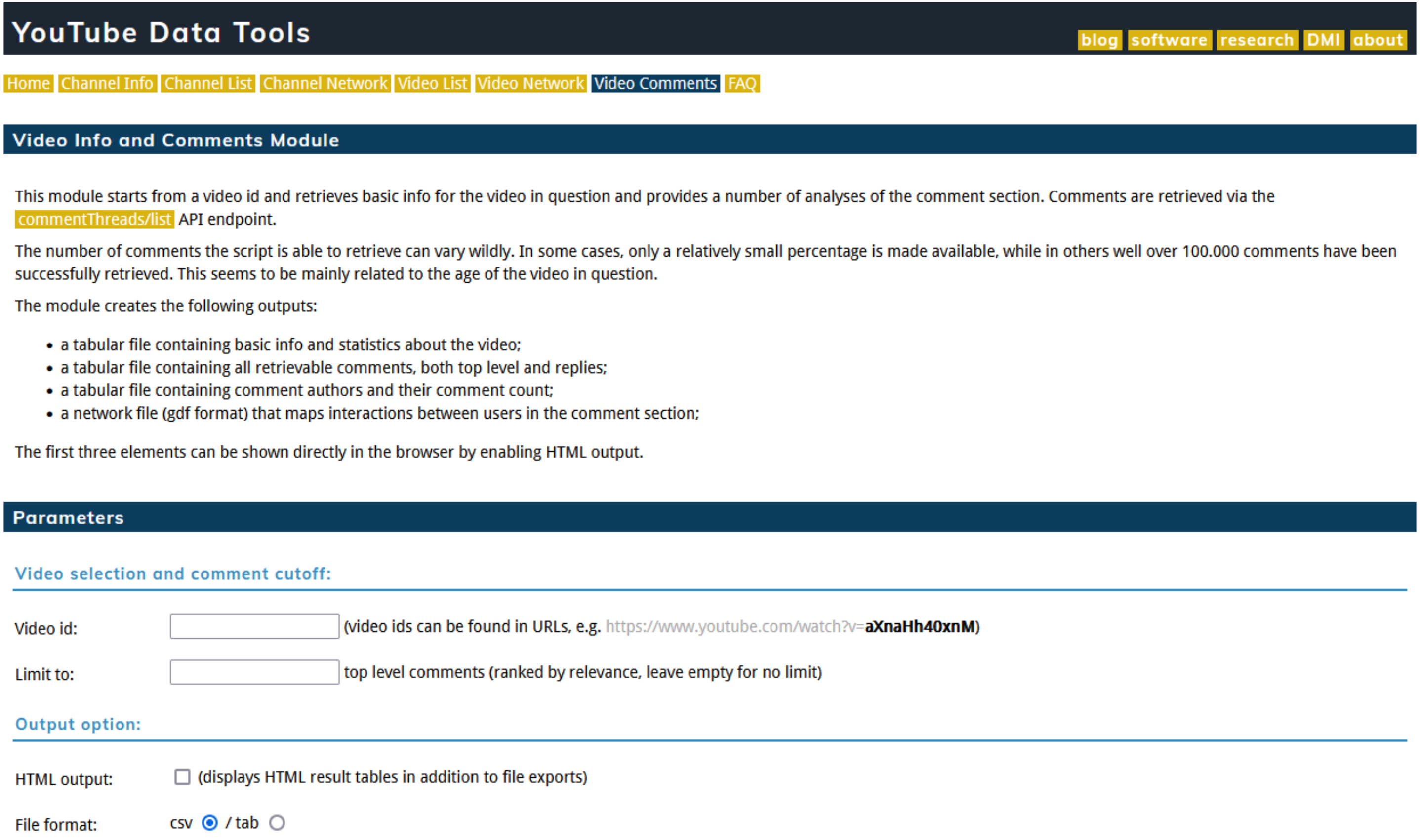
Some of the comment sampling tools also offer search functions that can be used in addition to or instead of the tools mentioned above.

Comparisons of Approaches for Collecting *YouTube* Data

Software	Type	Can collect	Comment Scope	Needs API Key
YouTube Data Tools 1.22	Website	Channel Info, Video Info, Comments	x top-level or all	No
Webometric 4.3	Standalone app	Channel Info, Video Info, Comments, Video Search	100 most recent or all	Yes
Tuber 0.9.9	R package	Channel Info, Video Info, Comments, Subtitles, All searches	20-100 most recent or all	Yes
vosonSML 0.29.13	R package	Video IDs, Comments	1-x top-level	Yes
youtubecaption 1.0.0	R package	Subtitles	n/a	No

YouTube Data Tools

YouTube Data Tools



Video Info and Comments Module

This module starts from a video id and retrieves basic info for the video in question and provides a number of analyses of the comment section. Comments are retrieved via the commentThreads/list API endpoint.

The number of comments the script is able to retrieve can vary wildly. In some cases, only a relatively small percentage is made available, while in others well over 100.000 comments have been successfully retrieved. This seems to be mainly related to the age of the video in question.

The module creates the following outputs:

- a tabular file containing basic info and statistics about the video;
- a tabular file containing all retrievable comments, both top level and replies;
- a tabular file containing comment authors and their comment count;
- a network file (gdf format) that maps interactions between users in the comment section;

The first three elements can be shown directly in the browser by enabling HTML output.

Parameters

Video selection and comment cutoff:

Video id: (video ids can be found in URLs, e.g. https://www.youtube.com/watch?v=**aXnaHh40xnM**)

Limit to: top level comments (ranked by relevance, leave empty for no limit)

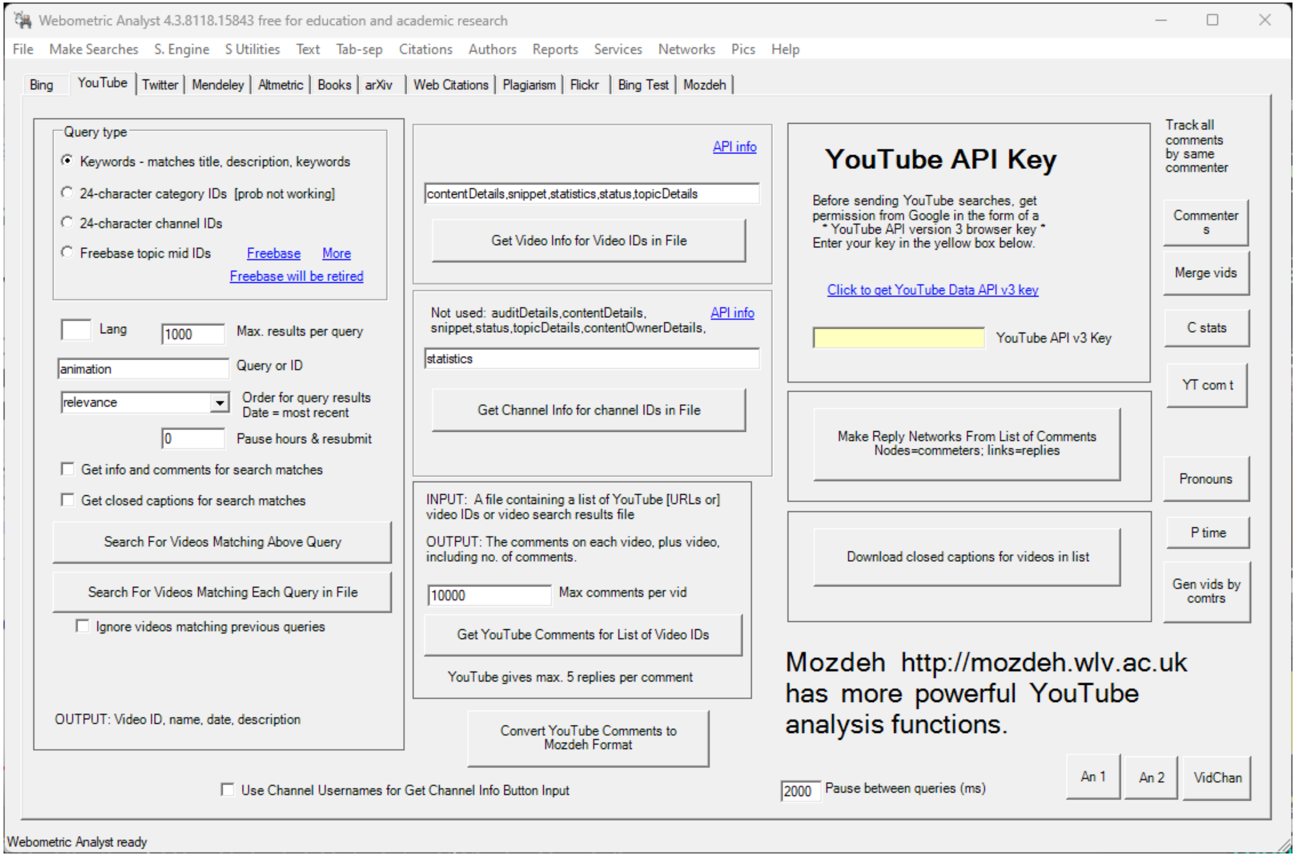
Output option:

HTML output: □ (displays HTML result tables in addition to file exports)

File format: csv ● / tab ○

Webometric

Webometric 4.3


Webometric Analyst 4.3.8118.15843 free for education and academic research
File
Make Searches
S. Engine
S Utilities
Text
Tab-sep
Citations
Authors
Reports
Services
Networks
Pics
Help
Bing
YouTube
Twitter
Mendeley
Altmetric
Books
arXiv
Web Citations
Plagiarism
Flickr
Bing Test
Mozdeh
API info
API info
Search For Videos Matching Above Query
Search For Videos Matching Each Query in File
Get Video Info for Video IDs in File
Get Channel Info for channel IDs in File
Get YouTube Comments for List of Video IDs
Convert YouTube Comments to Mozdeh Format
Make Reply Networks From List of Comments Nodes=commeters; links=replies
Download closed captions for videos in list
Commenters
Merge vids
C stats
YT com t
Pronouns
P time
Gen vids by comtrs
An 1
An 2
VidChan

Query type

● Keywords - matches title, description, keywords
○ 24-character category IDs [prob not working]
○ 24-character channel IDs
○ Freebase topic mid IDs Freebase More
Freebase will be retired

Lang: □ 1000 Max. results per query

animation Query or ID

relevance Order for query results Date = most recent

0 Pause hours & resubmit

□ Get info and comments for search matches

□ Get closed captions for search matches

□ Ignore videos matching previous queries

OUTPUT: Video ID, name, date, description

contentDetails,snippet,statistics,status,topicDetails

Not used: auditDetails,contentDetails, snippet,status,topicDetails,contentOwnerDetails,

statistics

INPUT: A file containing a list of YouTube [URLs or] video IDs or video search results file

OUTPUT: The comments on each video, plus video, including no. of comments.

10000 Max comments per vid

YouTube gives max. 5 replies per comment

□ Use Channel Usernames for Get Channel Info Button Input

YouTube API Key

Before sending YouTube searches, get permission from Google in the form of a * YouTube API version 3 browser key * Enter your key in the yellow box below.

Click to get YouTube Data API v3 key

YouTube API v3 Key

Track all comments by same commenter

Mozdeh http://mozdeh.wlv.ac.uk has more powerful YouTube analysis functions.

2000 Pause between queries (ms)

Webometric Analyst ready

Exemplary Comparison of the Different Tools

Software	Ease of Use	Disadvantages	No. of Comments
YouTube Data Tools 1.30	High	Lacking flexibility, less information	54,850
Webometric 4.1	Low	Only first 5 follow-up comments, no error feedback, undetectable time-outs	51,095
Tuber 0.9.9	Low	Only first 5 follow-up comments	51,084
vosonSML 0.29.13	Low	Lacking flexibility, only comments	52,679

Example data source: Dayum Video

A Note on Using FOSS

The tools listed are free and open source software (FOSS). Using FOSS has many advantages (availability, adaptability, etc.). However, one risk associated with using FOSS is that tools are not maintained anymore and cease to function. After all, people create and maintain these tools in their spare time or as side projects and this work is often not recognized enough (esp. within academia). For this reason it is important to acknowledge the work that goes into these tools by properly citing them.

```r
citation("tuber")
```

```
##
## To cite package 'tuber' in publications use:
##
##   Gaurav Sood (2020). tuber: Access YouTube from R. R package version 0.9.9.
##
## Ein BibTeX-Eintrag für LaTeX-Benutzer ist
##
##   @Manual{,
##     title = {tuber: Access YouTube from R},
##     author = {Gaurav SOod},
##     year = {2020},
##     note = {R package version 0.9.9},
##   }
```

Any questions so far?