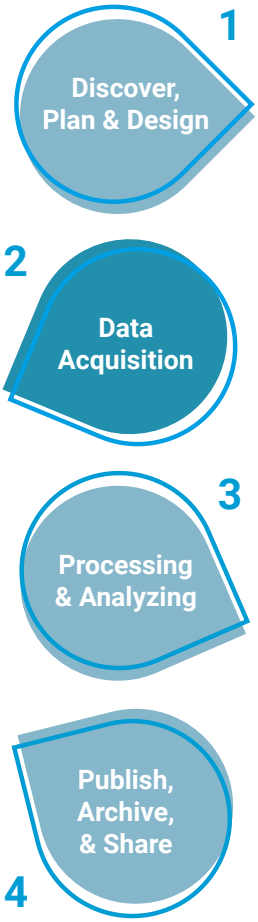
CAIS RESEARCH FOR THE DIGITAL AGE

Navigating Research Data and Methods

November, 2023
Volume 1.1



Collecting Social Media Data

Tools for Obtaining Data from Social Media Platforms

Annika Deubel, Johannes Breuer, Katrin Weller

Social media platforms provide an internet-based form of communication for users to have conversations, share information, and create web content. The vast amounts of data produced by the increasing integration of social media into people's daily lives is not only of interest to platform providers, but also to researchers in various scientific fields of digitalization research. The data from social media platforms can be used to gain insights into various behavioral, usage, and interaction patterns.

The analysis of this data has become increasingly popular in fields such as sociology, political science, linguistics, and computer science. For example, it allows for understanding collective sentiment, user behavior during global events, monitoring public opinion, identifying key topics in public discussions, and detecting popular routes followed by social media users.

Compared to other types of data that are commonly used in the social and behavioral sciences, such as survey or interview data, social media data are non-designed, i.e., they occur naturally and in large quantities. Researchers can, hence, harvest a massive amount and diverse range of user-generated content and other digital trace data without the need for data collection procedures that rely on self-reports. Once equipped with the resources for accessing and acquiring data, researchers can potentially scale data harvesting without expending a great deal of resources and often with less lead time compared, e.g., to survey research.

Introduction

In general, social media data can be obtained via three main options. Firstly, researchers can collect the data themselves, typically either through web scraping or the use of application programming interfaces (APIs). Secondly, privileged access can be gained by cooperating with companies that produce or hold these data. As a third option, the data can be purchased from market research companies or data resellers (Breuer et al., 2020).

Among all options, collecting data through APIs or web scraping have been the most widespread among researchers working with social media data (Acker & Kreisberg, 2020; Breuer et al., 2020). Web Scraping therein includes all methods and techniques for extracting data and information from the web (Dewi et al., 2019). In contrast, API is a back-end interface built for developers that directly communicates with the database of a service. They are thus often described as "glue" that sticks together different computer systems (Sloan & Quan-Haase, 2017). Many social media platforms make data available through APIs, with which they are governed by their technical and legal regulations. This allows researchers to retrieve, store, and process digital traces left by users of the respective platforms for empirical analysis (Perriam et al., 2020).

One of the main advantages of API-based data collection is the easy access. Although a certain technical and programming expertise may be required, a few lines of code are often sufficient for simple API requests. Besides, there is a variety of open-source tools which can be used without the necessity of prior programming experience. By contrast, the aforementioned access method of web scraping generally involves code-based frameworks that are much more difficult to learn (Lomborg & Bechmann, 2014). Not only is the workflow much more complex and time-consuming than API extraction, but it also differs for every website due the pages' unique characteristics (which also tend

to change over time). Although web scraping is generally more flexible than using APIs, this type of data collection is usually prohibited by the Terms of Services (ToS) of most social media platforms (Halavais, 2019). However, a recent U.S. court ruling indicates that scraping publicly accessible internet data is legal for academic research (Whittaker, 2022). It should still be treated carefully, as it may contain sensitive information and cause the platform to block the IP that is used to scrape data.

Despite its advantages, API-based data collection also brings about several downsides that researchers should be aware of when considering using them. While APIs make data publicly available, they are not open in the sense of giving full and unlimited access to the entire database. APIs are limited to public data[1] and usually also restricted in volume and variety. In most cases, only samples of the data can be obtained with the process behind the sampling usually being unknown or at least not fully transparent to researchers (Morstatter et al., 2013).

In addition, the ToS of platforms and their API are subject to change and can, therefore, be(come) unreliable for continuous (or repeated) data access. Relying on APIs for acquiring data, hence, means that research is, to a large extent, dependent on decisions made by commercial companies (Bruns, 2019; Freelon, 2018). As a result of debates about data security and user privacy following the Cambridge Analytica scandal, many platforms, most notably Facebook and Instagram, have continuously restricted access to their data. This has led to a general instability of digital data sources (Gerlitz, 2016; Perriam et al., 2020). The risks and implications that come with what some have called a "Post-API age" (Freelon, 2018) or an "APIcalypse" (Bruns, 2019) are severe. API shutdowns or restrictions create a significant gap in data access among researchers and can also hamper the efforts of third-party developers who create tools for collecting and processing these data (Bruns, 2019; Freelon, 2018). Short-term changes or shutdowns of API functionalities also lead to a fragile data access structure that can be too unstable for longitudinal studies (Perriam et al., 2020).

Besides taking into account the risks associated with the use of APIs, before collecting any data, researchers should determine what data are needed for the respective research objective and choose an appropriate platform, sampling approach, and collection tool(s). Regarding the sampling procedure, the most common approaches are user-related, keyword-related (e.g., tracking a specific hashtag) or URL-related (Mayr & Weller, 2017; Stieglitz et al., 2018). Notably, due to technical, legal, or ethical limitations, however, it may be possible that not all available data can actually be acquired. There are several methods for programmatically collecting data from APIs or endpoints. These include commercial services and tools, as well as options that involve programming languages like Python or R, API tools, or free command line tools like cURL. Some of these methods may have limitations in terms of transparency or ease of modification.

To facilitate making informed choices for data access, we provide a commented list of tools that can extract social media in the following. As web scraping is harder to learn and usually prohibited by the platforms' ToS, the focus of the lists is on tools that make use of platform APIs. Since all data collection tools come with certain limitations, choosing the "right" tool strongly depends on the research objective and the researchers' technical capabilities. The aim of this document is to provide an overview and comparison of the tools. Please note that questions of research ethics and legal perspectives on working with data from social media platforms, such as data privacy, informed consent and data ownership, are not considered in this document.

Additional Resources

General resources

Mancosu, M., & Vegetti, F. (2020). What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data. *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120940703

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2021). A Total Error Framework for Digital Traces of Human Behavior on Online Platforms. *Public Opinion Quarterly*. 85(1), 399-422. https://doi.org/10.1093/poq/nfab018

Ulloa, R. [GESIS – Leibniz Institute for the Social Sciences]. (2021, November 11). *Meet the Experts: Dr. Roberto Ulloa - Introduction to Online Data Acquisition* [Video]. YouTube. Retrieved January 17, 2022, from https://www.youtube.com/watch?v=inUvEFLG5EA

API resources

Bauer, P. C., & Landesvatter, C. (2021). *APIs for social scientists: A collaborative*

1 The definition of „public data" varies between platforms.

review. https://bookdown.org/paul/apis_for_social_scientists/ (last access: 17/01/2022)

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665-668. https://doi.org/10.1080/10584609.2018.1477506

Lomborg, S., & Bechmann, A. (2014). Using APIs for data collection on social media. *The Information Society*, 30(4), 256-265. https://doi.org/10.1080/01972243.2014.915276

Web scraping resources

Amos, D. (2018). A Practical Introduction to Web Scraping in Python. *Real Python*. https://realpython.com/python-web-scraping-practical-introduction/

Breuss, M. (2021). Beautiful Soup: Build a Web Scraper With Python. *Real Python*. Retrieved March 22, 2022 from https://realpython.com/beautiful-soup-web-scraper-python/

McNulty, K. (2019). Tidy web scraping in R – Tutorial and resources. Towards Data Science. https://towardsdatascience.com/tidy-web-scraping-in-r-tutorial-and-resources-ac9f72b4fe47

Schweinberger, M. (2022). Web Crawling and Scraping using R. *Language Technology and Data Analysis Laboratory*. https://slcladal.github.io/webcrawling.html

Disclaimer

The team Research Data & Methods at CAIS is not responsible for the tools listed in this document and their operability. In addition, we cannot guarantee that the tools are in compliance with the ToS of the respective platforms. Due to frequent and unforeseeable changes in the APIs, some tools listed here may cease to function. Likewise, tools may not be maintained and actively developed anymore. The list will periodically be checked and updated. Feel free to contact rdm@cais-research.de for further information or if you have suggestions for changes.

Acknowledgement

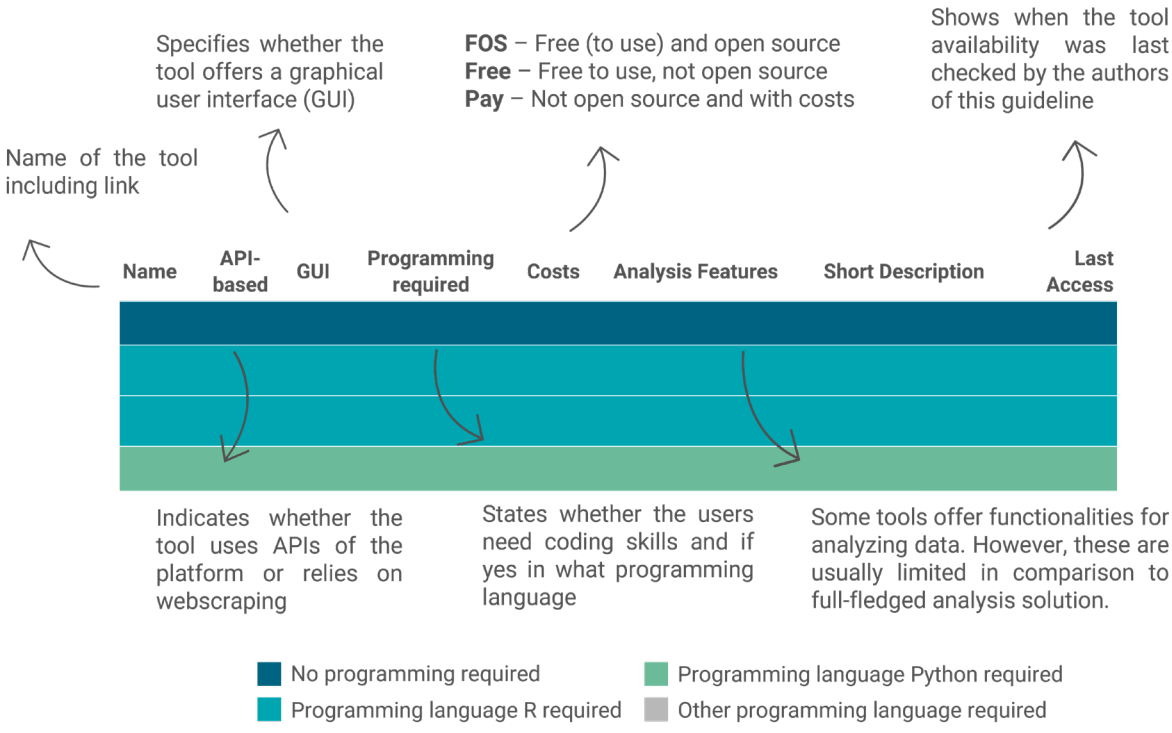
The majority of the tools listed in this document were identified via the following resources:

Davies, L. (2019, June 4). Social media data in research: a review of the current landscape. *SAGE ocean*. https://ocean.sagepub.com/blog/social-media-data-in-research-a-review-of-the-current-landscape

Repository of the Social Media Observatory of the Leibniz Institute for Media Research (n.d.). *Leibniz Hans Bredow Institute (HBI)*. https://smo-wiki.leibniz-hbi.de/

Social Media Research Toolkit (n.d.). *Social Media Lab*. https://socialmedialab.ca/apps/social-media-research-toolkit-2/

Explanation of the table columns



Multi-Platform Tools

Some tools can collect data from more than one platform. Hence, they may be particularly useful for conducting cross-platform research. Notably, the functionalities per platform may vary for these tools.

Name	Platforms	API-based	GUI	Programming required	Costs	Analysis Features	Short Description	Last Access
Crowdtangle	Facebook, Instagram, Reddit	Yes	Yes	No	Free	Yes (*Performance Analysis*)	A tool from Meta to help follow, analyze, and report on what's happening across social media	22/06/23
Facepager	Twitter/X, YouTube	Yes	Yes	Yes (*Python*)	FOS	No	Fetches publicly available data from YouTube, and Twitter based on APIs	22/06/23
NodeXL	Twitter/X, YouTube, Wikipedia, Flickr	Yes	Yes	No	Pay	Yes (*Content & Network Analysis*)	Access to network data streams and analysis with network metric (free version very limited)	22/06/23
Reaper	Facebook, Twitter/X, Reddit, YouTube, Tumblr, Pinterest	Yes	Yes	Optional (*Python*)	FOS	No	API-based tool for social media analysis with own GUI	22/06/23
vosonSML	Twitter/X, YouTube, Reddit	Yes	Optional	Yes (R)	FOS	Yes (*Text & Network Analysis*)	R package that provides tools for collecting data across popular platforms and generating networks	22/06/23
Webometric Analyst	YouTube, Twitter/X, Goodreads, Flickr, Mendeley	Yes	Yes	No	FOS	Yes (Network Analysis, Link Analysis, Statistical Processing)	Windows-based program that downloads data from the web through APIs; includes a wide range of processing options	22/06/23

Twitter/X

Over the last years, Twitter has become the "model organism" (Tufekci, 2014, p. 2) of social media research, mainly due to the fact that the data are relatively easy to obtain through the platform's API. One thing to note here is that a developer account and an API Key (that can be obtained with a developer account) is required for using use the API. The signup to the Twitter Developer Portal can be found here. The API has numerous endpoints that can be used to access various types of data, such as posts, user profiles, followers, or mentions.

In the beginning of 2021, Twitter had launched an Academic Research product track for their new API version (V2) which was available for non-commercial academic research. After applying for and receiving authorization, researchers could gain access to the full history of public tweets and a higher monthly cap. In 2023, significant modifications have been implemented to the Academic Research API, rendering it highly restrictive. Twitter made an announcement in April 2023, stating that accessing the API would necessitate a paid subscription for being able to collect data. Access levels are divided into Free, Basic (100$/month), Pro (5000$/month) and Enterprise. In July 2023, Twitter's CEO Elon Musk renamed the company from Twitter to X.

API rate limits: The API v2 has a consumption cap which limits the number of tweets that can be received on a monthly basis. The rate limits also differ between the newly introduced access levels. As such, the Basic version has a post cap of 10,000, while the Pro version has a cap of 1,000,000. The free access level cannot pull posts.

Requests are made per 15-minute window, whereas every specific lookup has a different limitation. The requests overview can be found here. Some requests may have a monthly cap that do not fall under the tweet consumption cap described above. For example, the "get users" request has a monthly cap of 15,000 independent of any other rate limits. Please note that packages and tools are constantly being archived due to the fact that open-source solutions are not maintainable anymore due to high API costs by X.

Additional Resources

Barrie, C., & Ho, J. C., (2022). *Using the Twitter Academic API With R for Social Science Research* [How to Guide]. SAGE Research Methods: Doing Research Online. https://dx.doi.org/10.4135/9781529609233

Tornes, A., & Trujillo, L. (2021, January 26). Enabling the future of academic research with the Twitter API. *Twitter Developer Platform Blog*. https://developer.twitter.com/en/blog/product-news/2021/enabling-the-future-of-academic-research-with-the-twitter-api

Authorization process for the Twitter Academic Research Product Track (n.d.). *The Comprehensive R Archive Network*. https://cran.r-project.org/web/packages/academictwitteR/vignettes/academictwitteR-auth.html.

Name	API-based	GUI	Programming required	Costs	Analysis Features	Short Description	Last Access
Audiense	Yes (V1.1/V2)	Yes	No	Pay	Yes (*Consumer Insights*)	Utilize social data to understand audiences (marketing focused)	12/09/23
Chorus Analytics	Yes (V1.1/V2)	Yes	No	FOS	Yes (Visual Analytics)	Free, evolving, data harvesting and visual analytics suite to facilitate and enable social science research	12/09/23
DMI-TCAT	Yes (V1.1/V2)	No	Yes (*PHP*)	FOS	Yes (*Keyword Analysis*)	Robust and reproducible data capture and analysis, and interlinks with existing analytical software	12/09/23
Facepager	Yes (V1.1/V2)	Yes	No	FOS	No	Fetches publicly available data from YouTube, and Twitter based on APIs	12/09/23
Python-twitter	Yes (V1.1/V2)	No	Yes (*Python*)	FOS	No	A Python wrapper around the Twitter API	12/09/23
RTwitterV2	Yes (V2)	No	Yes (*R*)	FOS	No	Code to loop through timelines and the academic search API	12/09/23
SMO-TMAS	Yes (V1.1/V2)	Yes	No	FOS	Yes (*Text & Network Analysis*)	Shiny app that makes it easy to collect and analyze small Twitter data sets (<18.000 tweets)	12/09/23
Scweet	No (*Scraper*)	No	Yes (*Python*)	FOS	No	Simple and unlimited Twitter scraper using Python	12/09/23
Social Bearing	Yes (V1.1/V2)	Yes	No	Pay	Yes (*Consumer Insights*)	Insights & analytics for tweets & timelines	12/09/23
TAGS	Yes (V1.1)	Yes	No	FOS	No	Google Sheet template which lets you setup and run automated collection of search results	12/09/23
twacapic	Yes (V2)	No	Yes (*Python*)	FOS	No	API client wrapper that automates common tasks (e.g., get all tweets by a list of users and poll for new tweets regularly)	12/09/23
twarc	Yes (V1.1)	No	Yes (*Python/Bash*)	FOS	No	Command line tool and Python library for archiving Twitter JSON data	12/09/23
twarc2	Yes (V1.1/V2)	No	Yes (*Python/Bash*)	FOS	No	Collects data at the command line from the Twitter API	12/09/23
tweepy	Yes (V1.1)	No	Yes (*Python*)	FOS	No	Easy-to-use Python library for accessing the Twitter API	12/09/23
twitteRacademic	Yes (V2)	No	Yes (*R*)	FOS	No	Use R to retrieve data from Twitter API 2.0 academic research track	12/09/23
twitter-explorer	Yes (V1.1)	No	Yes (*Python*)	FOS	Yes (*Network Analysis*)	The twitter explorer combines collection, transformation, and visualization of Twitter data in an easily accessible interface	12/09/23
Twitterscraper	No (*Scraper*)	No	Yes (*Python*)	FOS	No	A simple script to scrape Tweets using Python to retrieve Twitter content without API restrictions	12/09/23
twittercrawler	Yes (V1.1)	No	Yes (*R*)	FOS	No	Provides a way to collect network data through Twitter's standard v1.1 APIs	12/09/23
Twurl	Yes (V1.1/V2)	No	Yes (*Ruby*)	FOS	No	OAuth-enabled curl for the Twitter API	12/09/23
Twython	Yes (V1.1/V2)	No	Yes (*Python*)	FOS	No	Premier Python library providing an easy (and up to date) access to data	12/09/23
vosonSML	Yes (V1.1)	No	Yes (*R*)	FOS	Yes (*Text & Network Analysis*)	Provides a suite of tools for collecting and constructing networks from social media data	12/09/23
Voson.tcn	Yes (V2)	No	Yes (*R*)	FOS	No	Collect tweets and generate networks for threaded conversations	12/09/23

YouTube

Although YouTube is the largest online video platform, the data it generates have been used much less frequently in academic research (Thelwall, 2018). The YouTube Data API provides access to data, such as video information, comments, playlists, or channel statistics. In order to use the API, a Google account and a project that has to be created via the developer console are required (Breuer et al., in press; Kaplan & Klein, 2022). As of July 2022, the platform provider has introduced the YouTube Researcher Program, which is an initiative that provides eligible researchers scaled access and insights to YouTube's public data as well as support and technical guidance from YouTube.

API rate limits: The YouTube API has a quota of 10,000 units per day, and each API method is associated with a different quota cost. The quota cost for each request can be assessed using the YouTube Quota Calculator. Through the YouTube Researcher Program, eligible researchers can obtain as much quota as they require for their research. Data access may be restricted depending on your research project by YouTube.The extended quota and program access is only operative for the duration of you research and will be revoked after completion (YouTube, 2022). This also comes with support and technical guidance. The application process, however, is a long process for which enough time must be considered before the data collection.

Additional Resources

Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social Research Methodology, 21*(3), 303-316. https://doi.org/10.1080/13645579.2017.1381821

YouTube (n.d.). YouTube Researcher Program. https://research.youtube/

YouTube Data API (n.d.). *Google Developers*. https://developers.google.com/YouTube/v3?hl=de

Name	API-based	GUI	Programming required	Costs	Analysis Features	Short Description	Last Access
Facepager	Yes	Yes	No	FOS	No	Collects publicly available data from JSON-based APIs	14/06/23
PHP YouTube API	Yes	No	Yes (*PHP*)	FOS	No	Designed to let devs easily fetch public data (video, channel, playlists info) from YouTube	14/06/23
Python-YouTube	Yes	No	Yes (*Python*)	FOS	No	A Python wrapper around for YouTube Data API V3	15/06/23
python-youtube-api	Yes	No	Yes (*Python*)	FOS	No	Python wrapper to fetch videos and comments based on keywords or channel Id.	15/06/23
Tuber	Yes	No	Yes (*R*)	FOS	No	R package for access to the YouTube API. Get comments, like counts, or search for videos with content	14/06/23
vosonSML	Yes	No	Yes (*R*)	FOS	Yes (*Text & Network Analysis*)	R package that provides a suite of tools for collecting and analyzing public Twitter, YouTube, and Reddit data	14/06/23
VTracker	Yes	Yes	No	Free	Yes (*e.g. Trend Exploration, Influencer Detection*)	Monitors, tracks, and analyzes YouTube videos based on keywords	14/06/23
yt_scraper	No	No	Yes (*Python*)	Free	No	Web scraper for collecting YouTube search results by netzpolitik.org	15/06/23
youte	Yes	No	Yes (*Python*)	FOS	No	Command-line utility to help collect video metadata from YouTube API	13/06/23
YouTube Caption	No	No	Yes (*R*)	FOS	No	R package to download and save YouTube subtitle data	15/06/23
YouTube Data Tools	Yes	Yes	No	FOS	No	Collection of modules that can extract data from YouTube using YouTube apiv3	14/06/23
youtube-easy-api	Yes	No	Yes (Python)	FOS	No	Python wrapper for the YouTube API 3.0 that lets users collect videos and metadata.	15/06/23

Facebook

As a response to the Cambridge Analytica scandal, Facebook essentially closed access to their Pages and Graph API which previously allowed researchers to extract posts, comments and associated metadata from public Facebook pages (Bruns, 2019). Currently, the only possibility for collecting individual-level user data from Facebook is through creating an own application that each user must give permission to access their data. Some types of Facebook data can be accessed through the CrowdTangle API, a public insights tool which was acquired by Facebook in 2016. Apart from Facebook, CrowdTangle also allows users to follow and analyze public content on the social media platforms Twitter, Instagram, and

Reddit. As of 2019, the CrowdTangle API and user interface are also available for academics and researchers after authorization. However, access to CrowdTangle is controlled and Facebook does not allow data from CrowdTangle to be published. While researchers may publish aggregate results from their data analysis, the original data cannot be shared publicly.

API rate limits: By default, the CrowdTangle API has a limitation of 6 calls per minute for all data except for links, for which the rate limit is 2 calls per minute. The rate limits of the Facebook APIs amount to 200 calls per hour.

Additional Resources

Archibong, I. (2018, April 4). API and Other Platform Product Changes. Meta for Developers. https://developers.facebook.com/blog/post/2018/04/04/facebook-api-platform-product-changes/

Bruns, A. (2018, April 25). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*. https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

Facebook Ad Library (n.d.). *Facebook Developers*. https://www.facebook.com/ads/library/?active_status=all&ad_type=political_and_issue_ads&country=DE&media_type=all

Facebook Graph API (n.d.). *Facebook Developers*. https://developers.facebook.com/docs/graph-api

Fan, Christina (n.d.). CrowdTangle for Academics and Researchers. *CrowdTangle*. https://help.crowdtangle.com/en/articles/4302208-crowdtangle-for-academics-and-researchers

Fan, Christina (n.d.). API Cheat Sheet: Shortcuts to understanding and using the CrowdTangle API. *CrowdTangle*. https://help.crowdtangle.com/en/articles/3443476-api-cheat-sheet

Schroepfer, M. (2018, April 4). An Update on Our Plans to Restrict Data Access on Facebook. *Meta*. https://about.fb.com/news/2018/04/restricting-data-access/

Name	API-based	GUI	Programming required	Costs	Analysis Features	Short Description	Last Access
CrowdTangle	Yes	Yes	Yes (*Python*)	Free	Yes (*Benchmarking, Performance Analysis*)	A tool owned by Meta to help follow, analyze, and report on what's happening across social media	23/06/23
Facebook Scraper	No (*Scraper*)	No	Yes (*Python*)	FOS	No	Scrape Facebook public pages without an API key	23/06/23
facepager	Yes	Yes	No	FOS	No	Fetches publicly available data from Facebook, YouTube, Twitter on the basis of APIs. Can also connect to the CrowdTangle Facebook API.	23/06/23
FBCrawler	No (Scraper)	No	Yes (*Python*)	FOS	No	Facebook crawler with python3.x	23/06/23

Instagram

Just as the Facebook API, Instagram's public API has been essentially shut down for external access by academic researchers in the wake of the Cambridge Analytica scandal (Bruns, 2019; Meier-Barthold, 2022). The Instagram Graph API, hence, has similar restrictions as the Facebook Graph API. There are two main APIs for Instagram: the Instagram Basic Display API and the Instagram Graph API. Both offer functionalities for collecting public data that may be of interest for researchers.

API rate limits: The Instagram Graph API and Instagram Basic Display API have a rate limiting of 200 calls per hour. Prior to the Cambridge Analytica scandal, up to 5,000 calls per hour were possible.

Further Resources

Instagram Graph API (n.d.). *Facebook Developers*. https://developers.facebook.com/docs/instagram-api

Instagram Basic Display API (n.d.). *Facebook Developers*. https://developers.facebook.com/docs/instagram-basic-display-api

Name	API-based	GUI	Programming required	Analysis Features	Costs	Short Description	Last Access
Instagram Java Scraper	No	No	Yes (*Java*)	FOS	No	Scrapes account information, photos and videos	23/06/23
Instagram Private API	Yes	No	Yes (*Python*)	FOS	No	Python wrapper for the private API. No 3rd party dependencies	23/06/23
Instaloader	No	No	Yes (*Python*)	FOS	No	Download pictures/videos with captions and other metadata	23/06/23
Instalooter	No	No	Yes (*Python*)	FOS	No	Downloads any picture or video associated from profiles	23/06/23
Instaphyte	No	No	Yes (*Python*)	FOS	No	Fast and simple hashtag scraper for exploratory analysis	23/06/23
Reaper	Yes	Optional	Optional (*Python*)	FOS	No	API-based tool for social media analysis with own GUI	23/06/23
Rinstapkg	Yes	No	Yes (*R*)	FOS	No	R package that connects to the API using tidy principles	23/06/23

Reddit

Reddit has become one of the most prominent social platforms on the web and a popular environment for discussing prominent and controversial events (Proferes et al., 2021). Its subreddit structure makes finding relevant data relatively easy, presenting an additional advantage compared, e.g., to Twitter (Amaya et al., 2021; Proferes et al., 2021). In the past, Reddits' open and free API contributed towards being a popular and valuable data source for researchers, similar to Twitter.

Starting July 1, 2023, Reddit has started to restrict free API access by introducing reduced rate limits. There now exists an Enterprise Level Tier that costs $0.24 per 1K API calls. Upon request, researchers may be provided with higher rate limits.

API rate limits: The free Data Reddit API has a rate limit of 100 queries per minute, and 100 items can be retrieved per request.

Additional Resources

API access rules (n.d.). *GitHub*. https://github.com/reddit-archive/reddit/wiki/API

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media+ Society, 7*(2). https://doi.org/10.1177/20563051211019004

Reddit API Documentation (n.d.). *Reddit*. https://www.reddit.com/dev/api/

Sivak, E. (n.d.). Reddit as a source of data. https://sicss.io/2021/materials/hse/reddit_.html

Name	API-based	GUI	Programming required	Costs	Analysis Features	Short Description	Last Access
Reaper	No (*Scraper*)	Yes	No	FOS	No	An open source and collaborative framework for extracting the data from websites	25/06/23
PRAW	Yes	No	Yes (*Python*)	FOS	No	Library for API access to Reddit	25/06/23
Pushshift	Yes	No	Yes (*Python*)	FOS	No	Enhanced functionality and search capabilities for searching Reddit comments and submissions	25/06/23
Rreddit	Yes	No	Yes (*R*)	FOS	No	An R package to extract Reddit data	25/06/23
RedditExtractoR	Yes	No	Yes (*R*)	FOS	No	R wrapper for the Reddit API	25/06/23

Other platforms

In the previous sections, we focused on the social media platforms with the largest global user base: Twitter, YouTube, Facebook, and Instagram. There are, however, also tools available for other platforms, such as Telegram, Wikipedia or TikTok. While we cannot cover all of those, we list some examples in the following.

Additional Resources

Chat Export Tool, Better Notifications and More (2018, August 27). *Telegram*. https://telegram.org/blog/export-and-more

Khaund, T., Hussain, M. N., Shaik, M., & Agarwal, N. (2020). Telegram: Data Collection, Opportunities and Challenges. *Communications in Computer and Information Science*. Springer, Cham. https://doi.org/10.1007/978-3-030-76228-5_37

Kindig, C. (2021). Wikipedia as a Valuable Data Science Tool. *Towards Data Science*. https://towardsdatascience.com/wikipedia-as-a-valuable-data-science-tool-6769991b43b7

Prabhu, T.N. (2020). Wikipedia API for Python. *Towards Data Science*. https://towardsdatascience.com/wikipedia-api-for-python-241cfae09f1c

TikTok for developers – Quickstart Objective (n.d.). *TikTok*. https://developers.tiktok.com/doc

Name	Platform	API-based	GUI	Programming required	Costs	Analysis Features	Short Description	Last Access
Reaper	Reddit, Tumblr, Pinterest	No (*Scraper*)	Yes	No	FOS	Yes	An open source and collaborative framework for extracting the data you need from websites	25/06/23
PykTok	TikTok	No	No	Yes (*Python*)	FOS	No	A simple module to collect video, text, and metadata from TikTok	25/06/23
scholarnetwork	Google Scholar	No	No	Yes (R)	FOS	Yes (Network Visualisation)	Extract and Visualize Google Scholar Collaboration Networks	25/06/23
Telegram Analytics	Telegram	Yes	Yes	Yes (*Python*)	FOS	Yes	Index of Telegram channels (not all countries/languages included)	25/06/23
TikTok-Api	TikTok	Yes	No	Yes (*Python*)	FOS	Yes	Python Wrapper for the TikTok API	25/06/23
Wikipedia	Wikipedia	Yes	No	Yes (*Python*)	FOS	Yes	Python library that facilitates accessing and parsing data from Wikipedia	25/06/23
Wikipedia-API	Wikipedia	Yes	No	Yes (*Python*)	FOS	Yes	Python Wrapper that supports extracting texts, sections, links, categories, translations, etc.	25/06/23

References

Acker, A., & Kreisberg, A. (2020). Social media data archives in an API-driven world. *Archival Science, 20*(2), 105–123. https://doi.org/10.1007/s10502-019-09325-9

Amaya, A., Bach, R., Keusch, F., & Kreuter, F. (2021). New Data Sources in Social Science Research: Things to Know Before Working With Reddit Data. *Social Science Computer Review, 39*(5), 943–960. https://doi.org/10.1177/0894439319893305

Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (2020). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society, 22*(11), 2058–2080. https://doi.org/10.1177/1461444820924622

Breuer, J., Kohne, J., & Mohseni, M. R. (in press). Using YouTube Data for Social Science Research. In *Research Handbook of Digital Sociology*. Edward Elgar Publishing.

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society, 22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Dewi, L. C., Meiliana, & Chandra, A. (2019). Social Media Web Scraping using Social Media Developers API and Regex. *Procedia Computer Science*, 157, 444–449. https://doi.org/10.1016/j.procs.2019.08.237

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Gerlitz, C. (2016). What Counts? Reflections on the Multivalence of Social Media Data. *Digital Culture & Society*, *2*(2), 19–38. https://doi.org/10.14361/dcs-2016-0203

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society, 22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Kaplan, M., & Klein, J. (2022). Youtube API. In *APIS for Social Scientists—A collaborative review*. https://bookdown.org/paul/apis_for_social_scientists/youtube-api.html

Lomborg, S., & Bechmann, A. (2014). Using APIs for Data Collection on Social Media. *The Information Society*, *30*(4), 256–265. https://doi.org/10.1080/01972243.2014.915276

Loveluck, L. (2023, Juni 22). How new Twitter rules could hinder war crimes research and rescue efforts. *Washington Post*. https://www.washingtonpost.com/technology/2023/06/20/twitter-policy-elon-musk-api/

Mayr, P., & Weller, K. (2017). Think before you collect: Setting up a data collection approach for social media studies. In *The SAGE Handbook of Social Media Research Methods* (S. 107–124). SAGE Publications Ltd.

Meier-Barthold, M. (2022). Instagram Basic Display API. In *APIS for Social Scientists—A collaborative review*. https://bookdown.org/paul/apis_for_social_scientists/youtube-api.html

Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose. *arXiv:1306.5204 [physics]*. http://arxiv.org/abs/1306.5204

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, *23*(3), 277–290. https://doi.org/10.1080/13645579.2019.1682840

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., & Zimmer, M. (2021). Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society*, *7*(2), 205630512110190. https://doi.org/10.1177/20563051211019004

Sloan, L., & Quan-Haase, A. (2017). *The SAGE Handbook of Social Media Research Methods*. SAGE.

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, *39*, 156–168. https://doi.org/10.1016/j.ijinfomgt.2017.12.002

Thelwall, M. (2018). Social media analytics for YouTube comments: Potential and limitations. *International Journal of Social*

Research Methodology, *21*(3), 303–316. https://doi.org/10.1080/13645579.2017.1381821

Tufekci, Z. (2014). *Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls*. 10.

Whittaker, Z. (2022). Web scraping is legal, US appeals court reaffirms. *TechCrunch*. https://social.techcrunch.com/2022/04/18/web-scraping-legal-court/

YouTube. (2022). *YouTube Research—Program Terms & Conditions*. YouTube. https://research.youtube/policies/terms/

How to cite

Deubel, A., Breuer, J., Weller, K. (2023). Collecting Social Media Data: Tools for Obtaining Data from Social Media Platforms. *Navigating Research Data and Methods, Vol. 1*. CAIS Center for Advanced Internet Studies. https://www.cais-research.de/wp-content/uploads/Collecting-Social-Media-Data.pdf