GESIS Guides to Digital Behavioral Data
#12

Overview of Working with Data from YouTube

Annika Deubel[1], Johannes Breuer[1,2], Julian Kohne[2], M. Rohangis Mohseni[3]

1 Center for Advanced Internet Studies (CAIS), Bochum, Germany
2 GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
3 TU Ilmenau, Ilmenau, Germany

Publication date: April 24, 2024; Version 1.0

As the internet's largest video platform, YouTube and its users create vast amounts of data across a diverse range of content. Over the last years, these data have become of great interest to social scientists to study production and reception of content and social interactions. When starting out, however, accessing YouTube data can present significant challenges. This guide introduces researchers to YouTube as a source of data and provides an overview on how to get started with collecting and working with YouTube data. In addition to instructions on accessing the YouTube API, this guide contains an overview of tools that can be used for collecting YouTube data and recommendations for processing and handling it.

This guide is written for readers who want to get a first overview of what kind of data can be collected from YouTube, how to access them, and how to work with these data. Readers do not need any technical or coding skills; however, they should have a basic understanding of what an API is and what web scraping means.

Keywords: data collection, web API, web scraping, social media data, platform data, YouTube, YouTube APIs, YouTube tools

1 YouTube and social science research

YouTube is currently by far the largest video platform on the internet and the world's second-most visited website, with over 113 billion monthly visits (Statista, 2023), second only to Google Search in terms of traffic, and second to Facebook in terms of user base. It is often used for entertainment and music, but many also use it as a source of informa-

tion and news. Regarding its main functionalities, YouTube allows users to upload or live-stream and watch videos, like or dislike them, and comment on them. Given its popularity and relevance as a source of various types of content, YouTube has also become a platform of interest for social science research.

Broadly speaking, social science research on YouTube can be categorized into studies that focus on communicators or creators, content, or the platform audience (Breuer et al., 2023). These perspectives can also be combined. Similar to the content on YouTube, the topics investigated in social science studies on the platform are quite diverse. While there is some macro-perspective research on the structure of the platform as a whole and the quantity and development of its content, such as the recent study by McGrady et al. (2023), most studies focus on specific types of content, user groups, or phenomena.

Research on the content of videos on YouTube has, for example, focused on education (Kohler & Dietrich, 2021; Utz & Wolfers, 2022), health information (Bopp et al., 2019; Gaus et al., 2021), or politics (Bringula et al., 2023; Lai et al., 2024). Other research investigates the recommendation algorithm on YouTube, e.g., regarding news vs. entertainment (Huang & Yang, 2024), misinformation (Tang et al., 2021), or radicalization (Haroon et al., 2022). Several studies have also specifically looked at user comments, e.g., by investigating hate speech, political ideology, gender differences or sentiment for specific topics (Döring & Mohseni, 2019; McLellan et al., 2022; Rauchfleisch & Kaiser, 2020).

As in the topics, there also is quite some variation in the data used as well as data collection, processing, and analysis methods in social science research about YouTube. While many studies on YouTube, especially those on its use and effects, rely on self-reports (from surveys or interviews), others also use data from YouTube itself. Thus, the platform is not only of interest as a subject but also as a data source for social science research.

What kinds of data exist on YouTube?

Data from YouTube can come in various forms and formats. We will briefly describe some of the types of data available from YouTube, while focusing on textual and numeric data stored in the databases underlying the YouTube website. Many of those are accessible via the Application Programming Interfaces (APIs) that YouTube offers (see sections 2 and 3 in this guide). However, we will also mention other types of YouTube data that are not available via platform APIs. One important thing to consider is that content on YouTube and, hence, also the data associated with this content is organized in a nested and partly hierarchical structure. **Figure 1** depicts different types of content and data on YouTube and their relationships.

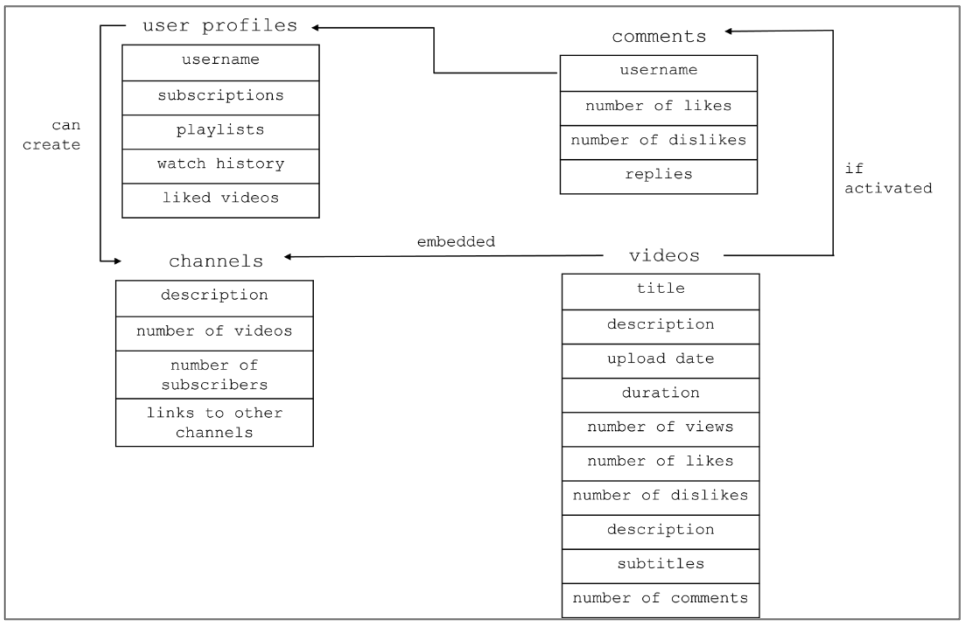

Figure 1: Different types of content and information on YouTube and their relations

User profiles: For each user, YouTube stores, e.g. information on subscriptions, liked videos, watch history and user preferences, which are used for personalization and recommendation features. These data are mostly private, i.e., only visible to the user and not accessible for external users through the API.

Channels: Users can, but do not have to create one or more channels. Channels also have their own attributes, such as a description, the number of subscribers, the number of videos, and other linked channels. Figure 2 displays this information for the GESIS YouTube channel.

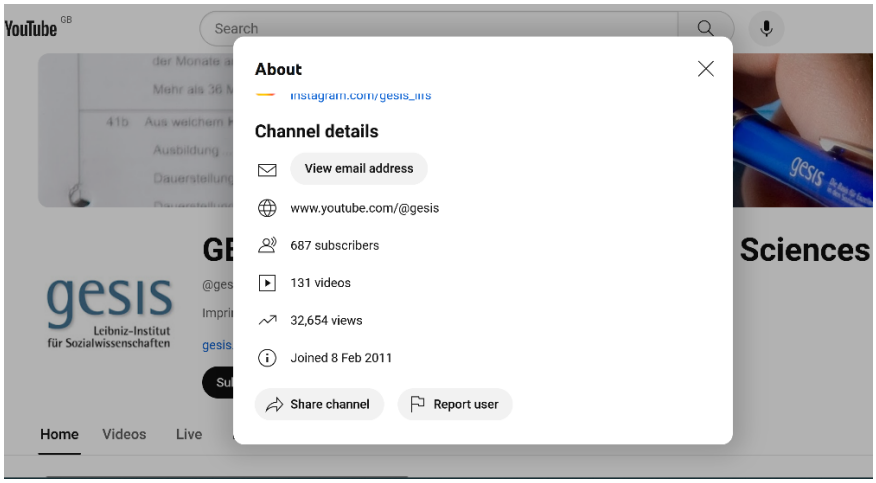

Figure 2: Exemplary depiction of public channel data

Video data: For every video published by a channel, YouTube maintains various kinds of metadata. This includes title, description, duration, upload date, view count, likes, and tags. If activated, a video is also associated with subtitles and captions (of which there

can be different types) and comments (as well as replies to comments). Also, the videos themselves can be considered and used as sources of data as they contain visual and auditory information.

Comments: If activated by the producer or channel, videos can contain comments and replies to comments. On YouTube, everyone who holds a user profile or a channel can comment on another video unless the uploader disables the feature. Comments can also be liked and disliked by other users.

Subtitles/captions: Videos on YouTube can have different types of subtitles and captions. Usually, all videos have subtitles created with automatic speech recognition (ASR). Those are always in English, even if the video language is not English. Videos can also have manually created subtitles and subtitles in multiple languages.

Recommendations: Based on different factors, such as user personalization, geolocation, relevance, and popularity, YouTube's recommendation system provides suggestions for other videos that users may be interested in. Recommendations are generated dynamically and, hence, not accessible via APIs.

2 Options for collecting YouTube data

Researchers have different options for collecting YouTube data (Breuer et al., 2020). Two important dimensions that these options differ on are the type of data they can collect and the resources they require.

Manual approaches

For many types of YouTube data, it is possible to collect them manually. This can entail the manual recording of subscriber, viewer, or like counts, copy-pasting video descriptions or comments from the YouTube website or the manual coding of visual or auditory content from videos. While such manual approaches are quite versatile regarding the kind of data they can generate, their major limitation is that they do not scale well and require a substantial number of resources in terms of human labor.

Existing data collections

While archiving and sharing YouTube data presents challenges (Breuer et al., 2023), some researchers have created extensive collections of YouTube data that are available for use. One notable collection is the YouNiverse collection, which comprises metadata for over 136,000 channels and 72.9 million videos published between May 2005 and October 2019. This collection also includes channel-level time-series data of weekly subscriber and view

counts. The main advantage of using existing data collections is, of course, that researchers do not have to collect data themselves. On the other hand, it may not be suitable for answering their specific research questions.

Data donation

Since individual-level YouTube user data are not publicly available, researchers interested in such data must collaborate with platform users (Halavais, 2019) and employ the method of data donation (Boeschoten et al., 2022), meaning that users export their personal data (or parts of it) from YouTube and share it with the researchers. While several technical frameworks for data donation have been developed and made available for the research community (see, e.g., Araujo et al., 2022; Boeschoten et al., 2023), this method requires substantial effort on the side of the researchers as well as the study participants. Another limitation of this method is that the self-selection of participants and the required effort on their part can introduce various biases in the sample and, thus, the data (Pfiffner & Friemel, 2023; Silber et al., 2024). A strength of data donation is that it can provide in-depth data about users, representing what Menchen-Trevino (2013) called "vertical trace data". From an ethical perspective, data donations also have the advantage of being transparent for study participants and requiring active opt-in consent.

Web scraping

Web scraping includes all methods and techniques for extracting data and information from the web (Dewi et al., 2019). This approach is sometimes also called "screen scraping". In short, this method entails that researchers store the HTML files used by the web browser to display YouTube content and extract the parts they are interested in with XML parsers. This data collection method offers the advantage of being able to extract virtually any visible information from a website.

In theory, there are no strict limitations on the amount or type of publicly available information that can be gathered. However, in practice, websites often implement measures to deter automated scraping, such as CAPTCHA challenges, mandatory user login, or unnecessarily complex document structures. With the recent success of Large Language Models (LLMs) that are largely trained on publicly available data, this has become even more pronounced. User-created content has become even more of a valued asset to social media platforms, showcased, e.g., by recent efforts of reddit to employ technical measures to limit the amount of data that can be gathered either through the API or web scraping (reddit, 2023).

Additionally, many websites explicitly prohibit web scraping in their Terms of Service (ToS). The legality of web scraping can vary depending on several factors, including the researcher's country of residence, the type of data being scraped, the purpose of the scraping, and how the scraped data is shared with third parties (Whittaker, 2022). A key

strength of web scraping is its versatility with the type of data it can generate. In contrast to API-based data collection on YouTube, it is, e.g., possible to collect video subtitle data via web scraping (e.g. Kramer, 2021). Besides the types of data listed in **Figure 1**, web scraping also allows for collecting data on video recommendations via approaches that have been described as "algorithmic auditing", "simulated users" or "sock-puppet accounts" (for an example not related to YouTube, see Hase et al., 2023).

A tool that has been specifically created by the digital human rights organization "AlgorithmWatch" to investigate the YouTube video recommendation algorithm is DataSkop. Notably, this specific tool relies on data donations instead of web scraping. Platforms like YouTube are often not in favor of scraping-based algorithmic auditing or simulated user approaches to studying their recommender systems. The main disadvantage of web scraping, besides legal questions (also see German Data Forum (RatSWD), 2020), is the programming and technical expertises that are required.

YouTube Application Programming Interfaces (APIs)

An Application Programming Interface (API) is "a way for two or more computer programs or components to communicate with each other. It is a type of software interface, offering a service to other pieces of software" (Reddy, 2011). Researchers can use APIs to access and automatically collect large amounts of data from various sources, such as social media platforms (Perriam et al., 2020). A helpful resource for this is the online guide "APIs for social scientists".

Notably, APIs also have limitations. In the case of YouTube, an example for this is the availability of video subtitles. Although the YouTube platform provides at least automatically created subtitles for most videos, the YouTube API does not allow accessing the subtitles of videos (apart from those that the person using the API has uploaded themselves). APIs usually also place restrictions on the amount of data that can be gathered within a specific timeframe. Further, short-term changes or shutdowns of API functionalities can lead to a fragile data access pipelines that may be too unstable for longitudinal studies (Perriam et al., 2020). Considering such API changes or closures, some researchers have diagnosed an "APIcalypse" (Bruns, 2019) or a "post-API age" (Freelon, 2018).

Despite these limitations, this method has several advantages. In addition to being allowed by the platform, working with an API is generally more straightforward than web scraping and results in more structured data. Hence, we will focus on YouTube data collection via API in more detail for the remainder of this guide.

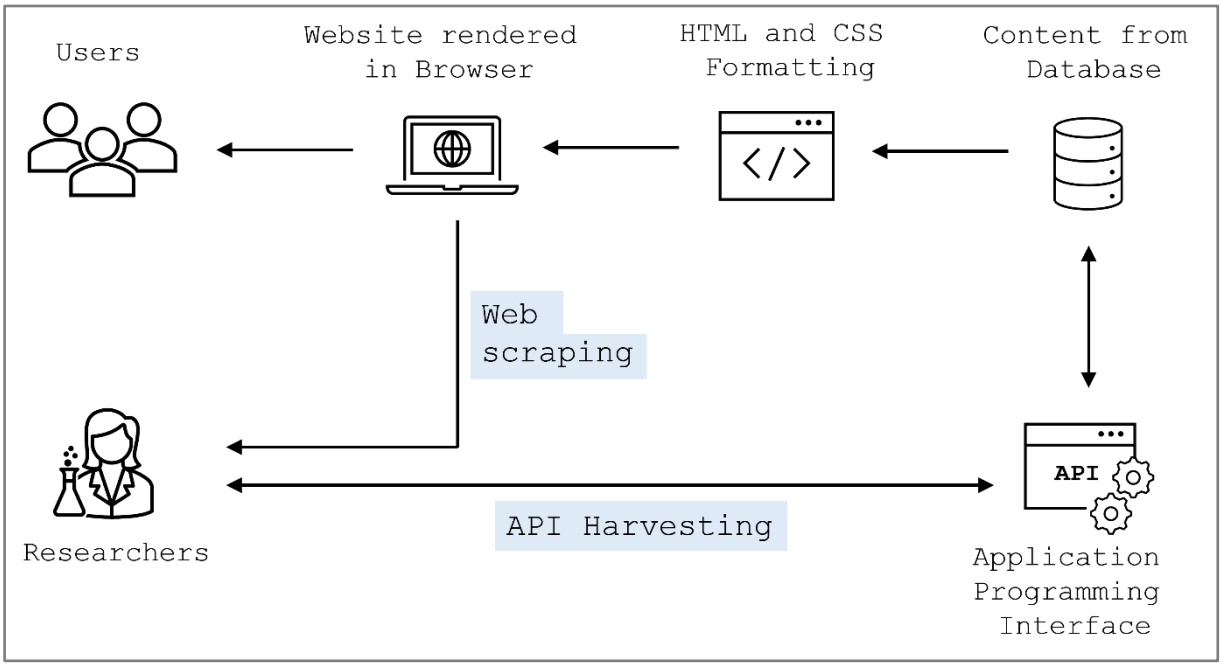

Figure 3: Schematic illustration of the differences between web scraping and API harvesting

YouTube offers several APIs that serve different purposes. The main APIs likely of interest to social science researchers are listed in **Table 1**. To collect YouTube data via API, researchers need to use the "YouTube Data API v3". It offers various data retrieval capabilities, such as searching for video IDs with specific keywords or obtaining metadata (e.g., comment count, view count, likes) and comments for one or multiple videos.

	Description and functionalities
YouTube Data API (v3)	It provides access to public YouTube data and, thereby, the retrieval of information about YouTube videos, channels, playlists, and other resources. The API provides functionalities, such as searching for videos, retrieving video metadata, uploading and managing videos, accessing user activity data, and managing playlists. Link
YouTube Analytics API	It provides access to analytics data and metrics associated with YouTube videos, channels, and playlists, such as viewing statistics or other popularity metrics. This API helps in analyzing and monitoring the performance of one's own YouTube channels and videos. Link
YouTube Live Streaming API	It enables to manage and control YouTube live streaming functionalities. The API allows creating, updating, and managing live broadcasts, monitoring live events, and retrieving information about live broadcasts, including chat messages, viewer statistics, and stream status. Link

Table 1: Overview of YouTube APIs and their functionalities.

3 Finding the right approach

When planning a data collection via the YouTube API, it is important for researchers to familiarize themselves with the API, the data it offers, and its limitations. The API documentation is a useful resource for that purpose. Another helpful resource for getting started is the respective chapter in the online guide "APIs for social scientists". In addition to that, researchers should be clear about what data they need and how they can gather it. For that, it helps to answer a few questions before starting with any kind of data collection:

- What kind of data do I want to collect?
- What do I want to do with the YouTube data?
- Does the API provide the data I need?
- What do I need to access specific data via the API?
- Can I or do I need to do further specifications in my API calls to narrow down the resulting data? If so, how can these be implemented?

To provide some more specific examples, in **Table 2**, we outline five common research scenarios for analyzing YouTube data. These scenarios are examples and not exhaustive.

I want to analyze …	What do I need?	What to consider?	What (type of) data do I get?	Research examples
topics	keywords	timeframe, number of videos, sorting	list of videos; metadata (e.g., video ID, title, number of likes, view count, channel name)	Bopp et al., 2019; Kohler & Dietrich, 2021
video channels	channel IDs	timeframe, number of videos, sorting	list of videos; metadata (e.g., video ID, title, number of likes, view count, channel name)	Miller, 2017
comments	video IDs	number of comments	comments and replies (text data)	Döring & Mohseni, 2019; Rauchfleisch & Kaiser, 2020; Thelwall, 2018
subtitles	video IDs or URLs	web scraping required, works best for English language videos	video captions (text data)	Soldner et al., 2019
recommendations	video IDs or 'seed video(s)'	depth, levels of recommended videos	list of videos and their metadata; (potential) network data	Huang & Yang, 2024: Ribeiro et al., 2020; Tang et al., 2021

Table 2: Typical scenarios for collecting YouTube data.

Tools for collecting YouTube data

As with the collection methods, there are various tools for collecting YouTube data. This section presents different types of such tools: web-based tools, standalone tools, and packages for specific programming languages (see **Table 3**). The only web-based tool on our list is **YouTube Data Tools.** It offers a user-friendly interface and leverages a generous quota provided by its creator, enabling the download of a substantial number of comments. For more flexibility, standalone tools, such as **Webometric** and **Facepager,** can be used. However, these tools have a steeper learning curve and require the use of a YouTube account for API authentication. One limitation of these options is that they are primarily used for data collection and do not offer extensive processing and analysis capabilities.

This is where packages for programming languages prove more advantageous, as they can be combined with other packages to create a pipeline from data collection to cleaning, filtering, preprocessing, and statistical analysis. However, with these packages or libraries, the data collection process is more complex, and researchers need at least basic coding skills. As a benefit, such packages are generally more flexible with the types of data they can collect.

Notably, however, there are differences between packages in this respect. For instance, the R package "tuber" does not support sorting search results obtained via the API by anything other than relevance. A special case is the R package **YouTube Caption**. As the retrieval of subtitles is not possible through the YouTube API, this web scraping-based tool can be used to download subtitles in text form.

A list of tools can be found below in **Table 3**. However, please note that tools may become deprecated and, hence, cease to function over time. For introductions on how to use the tools, researchers should consult their documentation. For the R package **tuber** (and in parts also for *vosonSML*), we also provide explanations on how to collect YouTube data in our workshop materials on GitHub.

Questions to be answered for choosing a tool:

- Which tool provides the required data?
- Do I have coding skills, if yes, in which programming languages?

Tool	API-based	Authentification method	GUI	Programming required?, language	Form	Data types	Data formats
Facepager	✓	API Key	✓	⃠	standalone	video metadata, comments, channel and playlist information	csv file
Webometric 4.3	✓	API Key	✓	⃠	standalone	channel metadata, video metadata, comments	txt file
YouTube Data Tools	✓	⃠	✓	⃠	web-based	channel metadata, video metadata, comments	csv file
Python-YouTube	✓	API Key	⃠	✓ Python	package/ library	channel metadata, video metadata, comments	Python data formats; export as csv file
youte	✓	API Key	⃠	✓ Python	package/ library	channel metadata, video metadata, comments	Python data formats, export as csv file
python-youtube-api	✓	API Key	⃠	✓ Python	package/ library	channel metadata, video metadata, comments	Python data formats, export as csv file
youtube-easy-api	✓	API Key	⃠	✓ Python	package/ library	video metadata	Python data formats, export as csv file
tuber	✓	oAuth 2.0	⃠	✓ R	package/ library	channel metadata, video metadata	R data frame, export as csv file
vosonSML	✓	API Key	⃠	✓ R	package/ library	comments	R data frame, export as csv file
YouTube Caption	Web Scraping	⃠	⃠	✓ R	package/ library	video subtitles	R data frame, export as csv file
PHP8 YouTube API	✓	API Key	⃠	✓ Python	package/ library	channel metadata, video metadata, playlist info	JSON file

Table 3: Tools for collecting YouTube data.

Disclaimer: Please note that GESIS and the authors of this guide are not responsible for the functioning or ethical or legal issues that may potentially arise from using the tools listed above.

Preprocessing YouTube data

Depending on what types of data are collected and how, the data may require different degrees of preprocessing before it can be analyzed. Again, the use of packages and libraries for programming languages like R or Python has the advantage that this step can be combined with the data collection in the same software environment. For the specific case of user comments, researchers may, e.g., want to extract certain elements from those, such as user mentions, URLs, or emojis. For that purpose, we have developed a small R package called tubecleanR that is available via GitHub. The main function of this package takes an R dataframe containing comments collected with the packages **tuber** or **vosonSML** as input, and extracts and parses several pieces of information from the comments, including URLs and emojis. Of course, what types of preprocessing are required heavily depends on the research question and the analysis methods to be used.

Analyzing YouTube data

The ways YouTube in which data can or should be analyzed depend on their nature and the research questions they are supposed to answer. Given the heterogeneity in data types and potential research questions, it is not possible to provide an overview of all analysis methods that can be applied to YouTube data. Methods applied to YouTube comments in previous research include sentiment analysis, topic models, network analysis, stance or hate speech detection.

For text data, such as comments or subtitles, there are many helpful introductions, tutorials and guidelines available online, focusing on different tools/programming languages and methods. For R, such resources are, e.g., "Text Mining with R: A Tidy Approach" by Julia Silge & David Robinson, "Automated Content Analysis" by Chung-hong Chan, or the documentation and tutorials for the *quanteda* package. As examples for specific applications, the vosonSML package provides different functionalities for network analyses and the package *peRspective* can be used for the automated detection of toxicity in YouTube comments (using the Perspective API). Recently, LLMs have received increased attention as tools for automated text analysis in the social sciences. The guideline by Törnberg (2024) can serve as a good starting point here. As YouTube comments are often in different languages, multilingual approaches might be required for this type of textual data. The materials from the GESIS Training workshop on multilingual text data by Hauke Licht and Fabienne Lind can be a useful resource for that purpose. Recent work by Rathje et al. (2024) suggests that LLMs can also be used for multilingual text analysis.

4 Limitations of YouTube data

While YouTube data has large potential for social science research, like other types of digital trace data, they have specific limitations that researchers need to consider. To

begin with, data collected via the API or web scraping only provides very limited information on individual users. For gathering detailed individual-level user information, data donation approaches are necessary. These also allow for a linking of these data with individual additional data, e.g., from surveys (Stier et al., 2020). In general, platform data can be affected by different types of biases on the sampling and the measurement level (Sen et al., 2021). Additionally, YouTube data can be noisy and incomplete, as they may contain irrelevant or duplicated content, and not all user activities or interactions are captured (Cesare et al., 2018).

5 Ethical and legal considerations

Besides the technical and methodological considerations discussed above, collecting and working with YouTube data also raises legal and ethical questions. With regard to legal aspects, two relevant things to consider are the platform's Terms of Service (ToS) and its developer policies. Notably, these were originally not designed with academic research in mind, making them challenging to interpret for researchers and their use of the API. Given that the Terms of Service (ToS) of a platform may undergo modifications over time, it is advisable for researchers to save a copy of the ToS as it existed during the data collection period. This precautionary measure ensures they have a reference point in case inquiries arise at a later stage.

Researchers may also encounter questions and issues related to copyright or intellectual property rights when working with video content. Since local legislation varies and can change, seeking legal counsel is generally advisable to ensure the legality of different methods of accessing and publishing data. While YouTube content is considered public once it is published, including metadata and user comments, it is important to note that also for public data ethical considerations regarding their use and distribution need to be made. Hence, it is generally recommendable to seek ethics review, e.g., via Institutional Review Boards (IRBs), also when collecting public data from YouTube.

Further legal and ethical questions arise when it comes to sharing/publishing data from YouTube (see Breuer et al., 2023). Again, how these can be addressed strongly depends on the type(s) of data and how they were collected. As is usually the case with legal and ethical questions, there can be no general guidance, but decisions need to be made on a case-by-case basis. Still, some existing resources on ethics in social media research and data sharing can also be informative for studies using YouTube data (see, e.g., Bishop & Gray, 2017; Franzke et al., 2020; Samuel & Buchanan, 2020; Townsend & Wallace, 2016; Williams et al., 2017).

Authors' Note: This guide is based on a workshop that the authors have previously taught within the GESIS Training program, and another guide on tools for collecting social media data (Deubel et al., 2023).

References

Araujo, T., Ausloos, J., Van Atteveldt, W., Loecherbach, F., Moeller, J., Ohme, J., Trilling, D., Van De Velde, B., De Vreese, C., & Welbers, K. (2022). OSD2F: An open-source data donation framework. *Computational Communication Research*, *4*(2), 372–387. https://doi.org/10.5117/CCR2022.2.001.ARAU

Bishop, L., & Gray, D. (2017). Chapter 7: Ethical Challenges of Publishing and Sharing Social Media Research Data. In K. Woodfield (Ed.), *Advances in research ethics and integrity* (Vol. 2, pp. 159–187). Emerald Publishing Limited. https://doi.org/10.1108/S2398-601820180000002007

Boeschoten, L., Ausloos, J., Möller, J. E., Araujo, T., & Oberski, D. L. (2022). A framework for privacy preserving digital trace data collection through data donation. Computational *Communication Research*, *4*(2), 388–423. https://doi.org/10.5117/CCR2022.2.002.BOES

Boeschoten, L., De Schipper, N. C., Mendrik, A. M., Van Der Veen, E., Struminskaya, B., Janssen, H., & Araujo, T. (2023). Port: A software tool for digital data donation. *Journal of Open Source Software*, *8*(90), 5596. https://doi.org/10.21105/joss.05596

Bopp, T., Vadeboncoeur, J. D., Stellefson, M., & Weinsz, M. (2019). Moving beyond the gym: A content analysis of YouTube as an information resource for physical literacy. *International Journal of Environmental Research and Public Health*, *16*(18), Article 18. https://doi.org/10.3390/ijerph16183335

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. (2023). Using YouTube data for social science research. In *Research Handbook on Digital Sociology* (pp. 258–279). Edward Elgar Publishing.

Bringula, R., Tabo, R. M. R., Alcazar, F. T. S. L., Delica, J. M. I., & Sayson, J. E. A. (2023). YouTube videos on the achievements of presidential candidates: Sentiment and content analysis. *Journal of Political Marketing*, 1–17. https://doi.org/10.1080/15377857.2023.2202617

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

Cesare, N., Lee, H., McCormick, T., Spiro, E., & Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography*, *55*(5), 1979–1999. https://doi.org/10.1007/s13524-018-0715-2

Deubel, A., Breuer, J., & Weller, K. (2023). *Collecting social media data: Tools for obtaining data from social media platforms* [Navigating Research Data and Methods]. Retrieved from www.cais-research.de/wp-content/uploads/Collecting-Social-Media-Data.pdf

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

Döring, N., & Mohseni, M. R. (2019). Fail videos and related video comments on YouTube: A case of sexualization of women and gendered hate speech? *Communication Research Reports*, *36*(3), 254–264. https://doi.org/10.1080/08824096.2019.1634533

Franzke, A. S., Bechmann, A., Zimmer, M., & Ess, C. M. (2020). Internet Research: Ethical Guidelines 3.0 Association of Internet Researchers. *Internet Research: Ethical Guidelines 3.0*. Retrieved from https://aoir.org/reports/ethics3.pdf

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

Gaus, Q., Jolliff, A., & Moreno, M. A. (2021). A content analysis of YouTube depression personal account videos and their comments. *Computers in Human Behavior Reports*, *3*, 100050. https://doi.org/10.1016/j.chbr.2020.100050

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

Haroon, M., Chhabra, A., Liu, X., Mohapatra, P., Shafiq, Z., & Wojcieszak, M. (2022). YouTube, The great radicalizer? Auditing and mitigating ideological biases in YouTube Recommendations. https://doi.org/10.48550/ARXIV.2203.10666

Hase, V., Boczek, K., & Scharkow, M. (2023). Adapting to affordances and audiences? A cross-platform, multi-modal analysis of the platformization of news on Facebook, Instagram, TikTok, and Twitter. *Digital Journalism*, *11*(8), 1499–1520. https://doi.org/10.1080/21670811.2022.2128389

Huang, S., & Yang, T. (2024). Auditing Entertainment Traps on YouTube: How Do Recommendation Algorithms Pull Users Away from News. *Political Communication.* Advance online publication. https://doi.org/10.1080/10584609.2024.2343769

Kohler, S., & Dietrich, T. C. (2021). Potentials and limitations of educational videos on YouTube for science communication. *Frontiers in Communication*, *6*. Retrieved from www.frontiersin.org/articles/10.3389/fcomm.2021.581302

Kramer, A. (2021). Dependency lengths in speech and writing: A cross-linguistic comparison via YouDePP, a pipeline for scraping and parsing YouTube captions. In A. Ettinger, E. Pavlick, & B. Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics 2021* (pp. 359–365). Association for Computational Linguistics. Retrieved from aclanthology.org/2021.scil-1.36

Lai, A., Brown, M. A., Bisbee, J., Tucker, J. A., Nagler, J., & Bonneau, R. (2024). Estimating the ideology of political YouTube videos. *Political Analysis*. Advance online publication. https://doi.org/10.1017/pan.2023.42

McGrady, R., Zheng, K., Curran, R., Baumgartner, J., & Zuckerman, E. (2023). Dialing for videos: A random sample of YouTube. *Journal of Quantitative Description: Digital Media*, *3*. https://doi.org/10.51685/jqd.2023.022

McLellan, A., Schmidt-Waselenchuk, K., Duerksen, K., & Woodin, E. (2022). Talking back to mental health stigma: An exploration of YouTube comments on anti-stigma videos. *Computers in Human Behavior*, *131*, 107214. https://doi.org/10.1016/j.chb.2022.107214

Menchen-Trevino, E. (2013). Collecting vertical trace data: Big possibilities and big challenges for multi-method research: Collecting Vertical Trace Data. *Policy & Internet*, *5*(3), 328–339. https://doi.org/10.1002/1944-2866.POI336

Miller, B. (2017). YouTube as educator: A content analysis of issues, themes, and the educational value of transgender-created online videos. *Social Media & Society*, *3*(2), 205630511771627. https://doi.org/10.1177/2056305117716271

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

Pfiffner, N., & Friemel, Thomas. N. (2023). Leveraging data donations for communication research: Exploring drivers behind the willingness to donate. *Communication Methods and Measures*, *17*(3), 227–249. https://doi.org/10.1080/19312458.2023.2176474

Rathje, S., Mirea, D.-M., Sucholutsky, I., Marjieh, R., Robertson, C., & Bavel, J. J. V. (2024). *GPT is an effective tool for multilingual psychological text analysis*. https://doi.org/10.31234/osf.io/sekf5

Rauchfleisch, A., & Kaiser, J. (2020). The German far-right on YouTube: An analysis of user overlap and user comments. *Journal of Broadcasting & Electronic Media*, *64*(3), 373–396. https://doi.org/10.1080/08838151.2020.1799690

Reddit. (2023). *Data API Terms—Reddit*. Retrieved from www.redditinc.com/policies/data-api-terms

Reddy, M. (2011). *API Design for C++*. Elsevier.

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 131–141. https://doi.org/10.1145/3351095.3372879

Ribeiro, M. H., & West, R. (2021). *YouNiverse: Large-scale channel and video metadata from English-speaking YouTube* (arXiv:2012.10378). arXiv. https://doi.org/10.48550/arXiv.2012.10378

Samuel, G., & Buchanan, E. (2020). Guest Editorial: Ethical issues in social media research. *Journal of Empirical Research on Human Research Ethics*, *15*(1–2), 3–11. https://doi.org/10.1177/1556264619901215

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*(S1), 399–422. https://doi.org/10.1093/poq/nfab018

Silber, H., Breuer, J., Felderer, B., Gerdon, F., Stammann, P., Daikeler, J., Keusch, F., & Weiß, B. (2024). *Asking for traces: A vignette study on acceptability norms and personal willingness to donate digital trace data* [Preprint]. SocArXiv. https://doi.org/10.31235/osf.io/2aum8

Soldner, F., Ho, J. T., Makhortykh, M., Van der Vegt, I., Mozes, M., & Kleinberg, B. (2019, June). Uphill from here: Sentiment patterns in videos from left-and right-wingYouTube news channels. In *Workshop on Natural Language Processing and Computational Social Science (Vol. 3, pp. 84-93). ACL.* https://aclanthology.org/W19-2110/

Statista (2023). *Most popular websites worldwide as of November 2023, by total visits.* Retrieved from https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an merging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Tang, L., Fujimoto, K., Amith, M. (Tuan), Cunningham, R., Costantini, R. A., York, F., Xiong, G., Boom, J. A., & Tao, C. (2021). "Down the Rabbit Hole" of vaccine misinformation on YouTube: Network exposure study. *Journal of Medical Internet Research*, *23*(1), e23262. https://doi.org/10.2196/23262

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

Törnberg, P. (2024). *Best Practices for Text Annotation with Large Language Models* (arXiv:2402.05129). arXiv. https://doi.org/10.48550/arXiv.2402.05129

Townsend, L., & Wallace, C. (2016). *Social media research: A guide to ethics*. University of Aberdeen.

Utz, S., & Wolfers, L. N. (2022). How-to videos on YouTube: The role of the instructor. *Information, Communication & Society*, *25*(7), 959–974. https://doi.org/10.1080/1369118X.2020.1804984

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

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149–1168. https://doi.org/10.1177/0038038517708140

All links in the text and the reference list were retrieved on April 22, 2024.

About the authors

Annika Deubel (M.Sc.) is a doctoral researcher and a member of the team Research Data & Methods at the Center for Advanced Internet Studies (CAIS) in Bochum, Germany. Her research interests include health communication and information on social media platforms, digital trace data and computational methods.

Johannes Breuer (Ph.D.) is a senior researcher and leader of the team Digital Society Observatory in the Department Computational Social Science at GESIS − Leibniz Institute for the Social Sciences in Cologne, Germany and the team Research Data & Methods at the Center for Advanced Internet Studies (CAIS) in Bochum, Germany. His research interests include the use and effects of digital media, digital trace data and computational methods, and open science and meta-science. More information: www.johannesbreuer.com

Julian Kohne (M.Sc.) is a doctoral researcher in the team Designed Digital Data in the Department of Computational Social Science at GESIS – Leibniz Institute for the Social Sciences in Cologne, Germany, and the Department of Molecular Psychology at Ulm University, Germany. His work at GESIS contributes to developing an app for collecting survey data and digital behavioral data using smartphones. In his dissertation at Ulm University, he is using donated WhatsApp chat log data to investigate communication in close interpersonal relationships. More information: www.juliankohne.com/

M. Rohangis Mohseni (Ph.D.) is a postdoctoral researcher in the Media Psychology and Media Design Group at Technische Universität Ilmenau in Germany. He is currently working on his habilitation on the topic of sexist online hate speech. His research interests include electronic media effects and moral behavior.
More information: www.rmohseni.de and orcid.org/0000-0001-7686-8322

Suggested citation

Deubel, A., Breuer, J., Kohne, J., & Mohseni, M. R. (2024): *Overview of Working with YouTube Data* (= GESIS Guides to Digital Behavioral Data, 12). Cologne: GESIS – Leibniz Institute for the Social Sciences

Series editors

Danica Radovanović, Maria Zens, Katrin Weller, Claudia Wagner

Publisher

GESIS Leibniz Institute for the Social Sciences