# On the Design and Implementation of a Multiplatform Framework for Social Media Data Collection

Anderson Frasão[1], Tiago Heinrich[2], Vinicius Fulber-Garcia[1]

[1] Universidade Federal do Paraná, Curitiba, Paraná, Brasil,
{aacfrasao,vinicius}@inf.ufpr.br
[2] Max Planck Institute for Informatics, Saarbrücken, Saarland, Germany.
theinric@mpi-inf.mpg.de

**Abstract.** Social media has become a central source of data for both academia and industry. However, access to this data is restricted by official API limitations, opaque commercial solutions, and fragmented scraping tools. This article presents a framework design for collecting, homogenizing, and storing cross-platform social media data. The design combines automated collection with structured extraction of text, images, videos, audio, metadata, URLs, and comments, organizing these elements into a standardized database. This standardization reduces the engineering effort required to consolidate data, facilitating comparative analyses across diverse social networks with different processing methods. In addition, the design foresees time-based controls, logging, and management of multiple accounts and cookies, making the collection process more robust in the face of account blocking, access limits, and changes to social network platforms.

**Keywords:** Social Network, Social Media, Data Collection

## 1 Introduction

Over the past two decades, social networks have become central environments for human interaction and large-scale digital data production, enabling the observation of social phenomena at unprecedented scale [19]. The multimodal data (combining text, images, videos, and metadata) hosted by these social networks supports diverse studies on engagement, diffusion, sentiment, and collective behavior [9], as well as research on scams and fraud in online environments [8], being a rich source for both academia and industry.

To support research and applications, there is an increasing reliance on automated data collection tools, including APIs, extraction libraries, and automation routines, to handle the scale and heterogeneity of social media data [6]. However, access to such data has become progressively restricted in the so-called "post-API" context, marked by the closure of public APIs and academic programs such as Twitter/X and CrowdTangle [17]. Even when access is formally granted, it

encompasses an opaque accreditation process and the use of technically limited APIs, resulting in a condition of "permissioned independence" [1].

Beyond access constraints, APIs also introduce technical and epistemological limitations. Empirical audits reveal missing or selectively returned content, such as in TikTok's Research API, where a significant portion of public metadata could not be retrieved [7]. Similar issues across platforms, rate limiting, undocumented sampling, and filtering, affect data representativeness and must be considered in analysis [4].

As a result, the academia and industry often resort to scraping or ad hoc solutions, which are difficult to scale, reproduce, and document [10, 22, 26]. In addition, many social phenomena of interest are inherently multi-platform. Actors frequently adapt their strategies to the affordances, moderation policies, and audience structures of each platform, making analyses restricted to a single environment insufficient to capture broader patterns. Consequently, comparative and integrated analyses across platforms are necessary to identify recurring behaviors, platform-specific adaptations, and shared operational structures.

However, conducting such cross-platform analyses remains technically challenging due to heterogeneous data structures, access restrictions, and the lack of unified collection infrastructures. In response to this scenario, this paper presents a generic design for a multiplatform framework for social media data collection. The framework design seeks to enable the development of unified, transparent, and extensible frameworks for the integrated, multimodal, and cross-platform collection of social media data. It is aimed at supporting comparative and cross-platform analysis by abstracting heterogeneous access mechanisms and organizing content into a standardized data model.

As a proof of concept, we also present the Social Media Collect (`SoMeCollect`) framework, a solution for collecting data from four particular social networks, YouTube, Instagram, TikTok, and Twitter/X, based on the premises, requirements, and protocols of the proposed design. By automating the collection of users, posts, interactions, and associated media, SoMeCollect reduces technical overhead and allows researchers to focus on identifying patterns, behaviors, and dynamics that emerge in different social media environments. Results from empirical experiments and the analysis of data retrieved by SoMeCollect enabled us to validate the solution and identify a series of opportunities and challenges in holistically collecting data from social network platforms.

## 2 Related Works

Existing solutions for social media data collection include commercial services, official APIs, and open-source scraping tools, but each has apparent limitations. Platforms such as `Make`, `Bright Data`, `Apify`, and `Octoparse` offer automation and ready-to-use scrapers, although they work as "black boxes", with little transparency and a reliance on paid models [15, 3, 23, 2, 18]. Several official API alternatives, such as TikTok and Twitter/X, impose restrictive accreditation requirements, request limits, and paid plans, which reduce accessibility

[24, 27]. Alternative tools, such as `Instaloader`, `snscrape`, and `yt-dlp`, expand the scope of social media data collection but lack standardization, as each produces its own outputs and lacks native integration across platforms [12, 21, 29]. All the described limitations are reinforced by multiple academic works, which highlight that social media data collection is usually fragmented, dependent on network-specific pipelines, and subject to instability, silent changes, and format heterogeneity, which increase the consolidation effort and hinder comparative analyses [13].

All the considered academic studies reach two particularly convergent conclusions: data collected across independent tools requires manual integration, and the absence of a unified pipeline limits broader analysis. Recent works explore different analytical approaches, including structuring Instagram data for public-interest investigations, evaluating marketing campaigns, studying digital behavior, and analyzing engagement during COVID-19 [16]. For example, on TikTok, research combines scraping, official APIs, and theoretical data models to examine social networks, interactions, and relational dynamics [11, 30]. For Twitter/X, the focus is on complete timeline scraping, the use of multiple IPs, and collection to feed sentiment classifiers [28]. On YouTube, analyses explore the computational performance of collecting metadata for music, culture, and video [14, 20, 25]. Taken together, the literature shows that analyses have ranged from social and cultural investigations to behavioral modeling and network studies, reinforcing the need for more integrated and extensible collection structures.

To address the challenges presented, the framework proposed in this work consists of an execution pipeline that seeks to mitigate these limitations by offering a unified, multiplatform collection architecture, all connected to a common storage layer. Notably, the implementation of the proposed pipeline, SoMeCollect, provides specific modules for Instagram, TikTok, Twitter/X, and YouTube. The collected data is organized in a standardized database, with compatible schemas across networks for users, media, and comments, facilitating comparative queries and joint analyses. In addition, the pipeline was designed to handle multimodality in a structured way, associating texts, media files, and metadata in the same analytical unit, and to be reproducible and expandable, allowing adaptation to changes in APIs and the inclusion of new platforms through specific and isolated changes in specific modules of the architecture/framework, without the need for complete reengineering processes.

## 3 Designing a Multiplatform Social Media Data Collection Framework

This section describes the architecture, operation, and core components of the proposed data collection framework for operating within multiple social network platforms. The objective is to detail how data is obtained from different platforms, how multimodal elements are extracted, and how this information is organized in a structured manner for subsequent analysis.

### 3.1 Base Pipeline

The framework is based on a multi-platform collection pipeline divided into four execution stages: (i) configuration, (ii) data access and collection, (iii) data cleaning and standardization, (iv) data storage and persistence. In the first stage (configuration), the user defines the set of profiles, channels, or target publications to explore on each social network, based on a predefined criterion (*e.g.*, institutional accounts, specific influencers, particular brands, or hashtags).

In the following stage (data access and collection), independent collection modules are activated for all the supported social networks. At this point, it is important to note that social networks have heterogeneous data and access interfaces, and that they offer varying levels of support for automated access. Thus, the routines for accessing a particular social network should be developed as an independent module, providing a common set of operations and a standard interface to be executed internally in the framework. In this way, these modules act as software drivers in the pipeline, activated according to user requirements defined in the first stage. The minimum set of operations that each module must support consists of:

- **Platform Access**: the module must be able to access a social network platform (typically through the Internet) and be able to interact with it;
- **Platform Authentication**: the module must properly authenticate within the social network platform, *e.g.*, user accounts or developers' keys;
- **Search and Deep Search**: the module must be capable of searching for given targets in the social network platform, as well as trigger deep search routines to enrich or enlarge data recovered in previous searches (for example, recovering information about the users that commented on a found publication);
- **Transform Data**: the module must implement routines to support the third stage of the pipeline, receiving raw data, processing it, and returning standardized data compliant to the model adopted by the framework implementation.

In the data cleaning stage (third stage), the raw data collected is cleaned and modeled to fit into a set of common, predefined entities and attributes to eliminate data formatting differences across networks. The standardized data is finally stored in a database (the data storage and persistence stage), which serves as a consolidation layer. From the standardized base, it is finally possible to process and derive different analytical snapshots, export the data to be used in other tools (*e.g.*, notebooks, machine learning pipelines, or qualitative analysis software), and reproduce experiments in new contexts by simply replacing the list of profiles or the collection time interval.

### 3.2 Implementation Requirements

Implementing a multi-platform collection framework for social media requires continuous, structured, and reproducible access to data from the target plat-

form. In line with previous work that emphasizes the challenges of extracting information in heterogeneous environments with access limitations [17, 26, 10, 22], this work establishes the need for a framework capable of integrating, standardizing, and storing multimodal data from different social networks to overcome restrictions imposed by APIs, scraping instability, and divergent return formats.

In this sense, the framework should act as an intermediary between social network platforms and users, abstracting from technical particularities and providing a set of standardized entities regardless of the data source. In addition, it must handle authentication mechanisms, cookies, multiple accounts, access limits, and frequent changes to platform interfaces, ensuring that the data collection remains consistent over time. Based on these needs, the following functional requirements for the framework were defined:

- **FR1** - Collect multimodal data (text, image, video, audio, URLs, and metadata) from each supported platform;
- **FR2** - Normalize heterogeneous data into a standardized set of entities (users, media, and comments);
- **FR3** - Record, for each item collected, the essential metadata (engagement, dates, identifiers, and links between entities);
- **FR4** - Store all data in relational databases organized by platform, following a uniform logical schema;
- **FR5** - Tolerate access limitations imposed by APIs, cookies, temporary blocks, and anti-bot mechanisms.

In turn, non-functional requirements prioritize robustness, modularity, performance, and reproducibility. They are as follows:

- **NFR1** - Allow the addition, removal, and updating of collection modules without impacting other components;
- **NFR2** - Maintain a unified internal interface for database operations across all platforms;
- **NFR3** - Support extended collection runs, with configurable timing and pause mechanisms to prevent crashes;
- **NFR4** - Record all actions in the framework in standardized logs, facilitating auditing and debugging;
- **NFR5** - Ensure reliable data persistence, preventing loss due to connection failures, timeouts, or unexpected interruptions;
- **RNF6** - Be compatible with the Unix environment.

Based on the observation of functional and non-functional requirements, an implementation based on the described framework design is particularly well-suited to provide high-quality datasets for several applications, especially for academic research that relies on a high volume of data.

## 4 SoMeCollect: Social Media Collect

This section presents the implementation of a multi-platform data collection framework, following the design established in Section 3, called Social Media Col-

lect (SoMeCollect). The proposed implementation is structured into platform-specific blocks, supported by shared components for timing, logging, and database access [3]. Regarding social network support, SoMeCollect provides four collection modules for YouTube, Instagram, TikTok, and Twitter/X.

SoMeCollect enables users to **configure** target accounts and content for each supported platform. Each module iterates over predefined users and posts, triggering platform-specific collection routines. During the **collection** stage, the framework interacts with each social network using the most suitable mechanism: official APIs (YouTube), third-party libraries (Instagram and TikTok), or automated browsing and scraping (Twitter/X); details are provided in Section 4.1. Metadata for users, media, and comments is retrieved, and associated media files (images, videos, audio, thumbnails) are downloaded and stored in platform-specific directories. By default, SoMeCollect collects all related comments for a given publication, and each comment retrieval triggers a deep search to collect information about its author.

Data is **cleaned and standardized** at the last moment of collection into a unified internal data representation composed of three core entities: users, media, and comments. Despite differences across platforms, heterogeneous fields (*e.g.*, followers vs. subscribers, captions vs. descriptions) are mapped to a standardized schema[4], ensuring consistency in identifiers, textual content, engagement metrics, and file paths.

**Persistence** is handled by a centralized database module that manages independent `SQLite` databases per platform. A common interface abstracts insert, update, and query operations, enforcing a shared logical schema across all platforms while preserving separate database files.

To ensure robustness and reduce the risk of deadlocks, the framework employs a unified logging and timing system. Dedicated loggers ensure consistent reporting, while random waits and periodic long pauses simulate human-like access patterns during intensive collection tasks.

Finally, the resulting databases provide a unified access point for exploration and downstream use. A standardized relational schema facilitates data export and integration with machine learning and visualization pipelines, without requiring changes to the collection logic.

### 4.1 Social Network Modules

Each social network is handled by a dedicated collection module, which implements the appropriate access strategy based on each platform's technical restrictions. The technical details of the implementation of the SoMeCollect social network modules are presented next:

**Instagram**. The Instagram module relies on authenticated access via the `instagrapi` library to collect public or semi-public profiles and their associated

[3] https://github.com/Carmofrasao/SoMeCollect
[4] https://github.com/Carmofrasao/SoMeCollect/blob/main/db/createDB.py

posts. The module retrieves user-level metadata, post-level textual and engagement information, and comments, which are subsequently normalized and linked via user and media identifiers.

**Twitter/X**. For Twitter/X, data collection is performed through web scraping using `Selenium`, complemented by `yt-dlp` for media retrieval. The module captures posts and replies visible on profile pages or query results, extracting textual content, engagement counters, temporal information, and basic author metadata, while preserving the relationships between tweets, users, and replies.

**YouTube**. The YouTube module is based on the official YouTube Data API, accessed via the Google API Client. The collection process explores channels, videos, thumbnails, and comments, storing structured metadata alongside the downloaded media to support multimodal analysis.

**TikTok**. Due to access restrictions and dynamically loaded content, the TikTok module combines `TikTokApi`, `Selenium`, and `yt-dlp`. It retrieves author metadata, post descriptions, engagement indicators, comments, and associated media, maintaining the same logical structure adopted by the other modules.

## 5 Tests, Results and Evaluation

Using SoMeCollect, it is possible to build a unified dataset of users, media, and comments from Instagram, TikTok, Twitter, and YouTube. Although data from each platform is stored in a separate SQLite database, all follow the same logical schema, enabling the execution of standard queries and cross-platform analyses. Thus, users can easily select profiles, collect posts within a time interval, and directly use the standardized tables to support statistical analyses and exploratory studies, drastically reducing or even eliminating the need for additional transformation or cleaning steps. Taking all these characteristics into account, this section presents tests, results, and discussions of the proposed multiplatform social media data collection framework design, particularly as implemented in the SoMeCollect solution.

### 5.1 Data Collection Efficiency

To validate SoMeCollect, tests were conducted by collecting 30 posts from each supported platform (Instagram, TikTok, Twitter/X, and YouTube), considering variations in engagement and comment volume. Efficiency was evaluated using two metrics: Data Acquisition Rate (kB/s) and Comment Acquisition Rate (c/s), with particular emphasis on comment collection, which accounts for most of the execution time (due to the deep searches triggered to collect the metadata of comment authors).

The results, presented in Table 1, show that YouTube achieved the highest performance, with comment and data collection rates up to 20 and 50 times higher, respectively, than the other platforms. The Instagram module performed

the worst, while the TikTok module achieved a relatively high comment acquisition rate but an overall data rate similar to the Instagram and Twitter/X modules.

| Metrics/Social Network | **Instagram** | **TikTok** | **Twitter/X** | **YouTube** |
|---|---|---|---|---|
| **Data Collection Rate (Posts)** | 0.01 kB/s | 0.02 kB/s | 0.02 kB/s | 1.02 kB/s |
| **Comment Collection Rate** | 0.02 c/s | 0.19 c/s | 0.07 c/s | 3.78 c/s |

**Table 1.** SoMeCollect Efficiency.

The data access mechanisms of each platform mainly explain these differences. The YouTube module benefits from a stable, optimized official API, whereas Instagram and Twitter/X modules rely on automated access methods that are subject to strict detection and blocking mechanisms. The TikTok module demonstrated intermediate performance despite its reliance on automation tools, thanks to effective rate-limiting strategies.

Overall, each platform imposes specific constraints: the YouTube module is limited by API request quotas, the Instagram module requires multiple accounts to mitigate the risk of being blocked, and the TikTok and Twitter/X modules are sensitive to internal library limitations and frequent interface changes. Despite these challenges, SoMeCollect has proven capable of performing consistent and scalable data collection across all evaluated platforms.

### 5.2 Collected Data Analysis

The collected data can support a wide range of social media research tasks, including content analysis, user behavior modeling, and interaction dynamics. Due to the unified scheme adopted by SoMeCollect, various types of data can be analyzed together, enabling both specific investigations on the platform and comparative studies across platforms.

Hashtags provide an explicit signal of thematic framing, visibility strategies, and community alignment. By analyzing hashtag frequency, co-occurrence networks, and temporal evolution, researchers can identify emerging topics, coordinated campaigns, and shifts in discourse over time. Frequency analysis and co-occurrence networks can reveal how users organize content around shared themes and how visibility strategies differ across platforms.

The textual content of posts and comments supports natural language processing techniques such as sentiment analysis, topic modeling, discourse analysis, and stylistic characterization. Because SoMeCollect preserves the relational structure among users, media, and comments, researchers can study conversation patterns, response structures, and the evolution of discussions over time, thereby contributing to studies of online interactions and digital communication.

URLs extracted from posts, comments, and profiles enable the analysis of information flows beyond the platform's boundaries. Researchers can examine domain diversity, link repetition, and temporal patterns of external references,

supporting studies on information diffusion, media ecosystems, and the role of external sources in shaping online discussions.

Image-related data enables the investigation of visual communication strategies and content reuse. Using computer vision or perceptual hash techniques, researchers can study image similarity, "meme" propagation, and visual framing. These analyses contribute to understanding how visual elements complement textual messages and influence user engagement.

Video data supports studies on audiovisual narratives, engagement dynamics, and content dissemination. Metadata such as duration, format, and publication date can be correlated with interaction metrics to analyze audience response and inform platform-specific optimization strategies. It is particularly relevant for short-form video platforms, where time and format strongly influence visibility.

Engagement indicators (likes, shares, comments), timestamps, and user-level attributes provide essential contextual information for behavioral and temporal analysis. These features enable the modeling of activity patterns, content lifecycles, and interaction rhythms, supporting research on attention dynamics and user participation across platforms.

### 5.3 Discussion

The proposed framework design and the SoMeCollect implementation work, in practice, as a research infrastructure for empirical studies on social networks, reducing the barrier between interest in a phenomenon and the availability of adequate data for analysis. By abstracting the specific details of each platform into a standard relational schema, the system allows researchers to focus on the study's substantive questions without having to reimplement the same set of collection, cleaning, and integration routines for each project. This layer of abstraction is particularly relevant in research that requires comparison across platforms, as it ensures that queries are expressed in terms of conceptually equivalent entities.

The standardized database format also facilitates the application of multimodal analysis methods and NLP techniques. Since captions, descriptions, and comments are stored in well-defined fields, it is easy to extract sets for topic modeling, sentiment analysis, or the automatic identification of linguistic patterns. Similarly, the explicit link between text, media (images, videos, and audio), and metadata enables studies that combine visual characteristics, textual content, and engagement signals into a single analytical unit, opening up hybrid approaches with multimodal neural networks or hierarchical analysis models.

In addition, the use of relational databases by platform, with consistent naming conventions, facilitates longitudinal analyses and sociotechnical studies. Once the set of profiles of interest has been established, it is possible to repeat the collection at different times, consolidating historical series of metrics and content that can be compared, for example, before and after changes in moderation policies, major social events, or coordinated campaigns. The same infrastructure can be reused across multiple projects, helping build long-lasting data repositories and enabling other research groups to replicate results.

On the other hand, the framework inherits structural limitations inherent to the social network domain. First, access to data remains subject to each platform's terms of service, which may restrict the type of content obtained, the frequency of requests, or the form of data storage and redistribution. Second, changes to APIs, web interfaces, or automated protection mechanisms can cause sudden interruptions in collection, requiring continuous maintenance of scraping modules and authentication routines.

Finally, occasional collection failures (such as connection interruptions, incomplete responses, or undetected layout changes) may result in partially incomplete samples, requiring caution in interpreting the results and documenting the conditions under which each data set was obtained. Even with these restrictions, the proposed standardization and clear separation between collection and analysis contribute to making these limitations more visible, controllable, and communicable in the context of academic research.

## 6 Ethical Considerations

The collection and processing of data from social network platforms requires special attention to ethical, legal, and responsibility principles in the use of information. Although this work proposes a framework design and implementation aimed at academic research, its operation necessarily involves access to content published by users, which requires caution regarding privacy, the context of use, and how this data may be subsequently analyzed and shared.

Firstly, the framework was designed to update exclusively on public content, accessible without violating technical barriers, advanced authentication, or explicit permissions. No security bypass mechanisms, reverse engineering of private APIs, or collection of sensitive information is used. The data captured reflects only what is already publicly available on the platforms, respecting the reasonable expectation of visibility established by users at the time of publication.

Furthermore, the pipeline does not perform automated inference on sensitive attributes (such as political orientation, health, ethnic identity, or other protected characteristics). The system only collects and structures the multimodal elements made available by social networks, without promoting classifications that could generate additional risks to the users analyzed.

From a storage perspective, the framework keeps all data locally, in independent databases per platform, without sending it to external services or third-party infrastructures. Researchers are advised to adopt secure storage and disposal practices, avoiding any redistribution of content that violates the platforms' terms of service or compromises the privacy of individuals involved. Internal logs, in turn, are limited to technical information essential to the framework's operation and do not record user content.

It should also be emphasized that the main objective of the framework is to reduce asymmetries in data access amid increasing restrictions and opacity on platforms, while providing methodological support for studies of public interest, including investigations into security, scams, and collective behavior. Ethical and

responsible use remains a fundamental condition to ensure that its contribution to research does not result in harm to users or violation of their right to privacy, especially in compliance with current legislation, such as the General Data Protection Law (LGPD, Brazil) and the General Data Protection Regulation (GDPR, European Union).

Finally, although some social networking platforms discourage web scraping, legislation and regulations addressing open data and data access for research purposes in countries where these platforms operate may permit the use of such tools in the absence of an official open API for the same purpose [5]. All these aspects should be carefully considered before using and disseminating data collected with the proposed framework.

## 7 Conclusion

This paper presented a framework for multiplatform social media data collection, implemented as the solution SoMeCollect, for retrieving and storing cross-platform data from social networks. In particular, SoMeCollect interacts with four social networks: Instagram, TikTok, Twitter/X, and YouTube. Starting from a scenario marked by restricted access to data, limitations of official APIs, and the fragility of solutions based on isolated scraping, the proposal offers a unified infrastructure capable of collecting, normalizing, and persisting information about users, media, and comments in a structured way. By adopting a standardized relational schema and specific collection modules for each platform, the structure reduces the engineering effort required for empirical social media studies and promotes the reproducibility of experiments.

From a technical and methodological standpoint, the main contribution lies in combining multimodal data collection (text, images, videos, audio, metadata, and URLs) with a consistent data model across platforms. This standardization allows queries and analyses to be formulated in terms of conceptually equivalent entities, opening the door to comparative studies across networks, longitudinal analyses, and applications of NLP, data mining, and multimodal analysis methods. In addition, integrating time control mechanisms, log recording, and multiple account and cookie management makes the collection process more robust in the face of blocks, access limits, and incremental platform changes.

## References

1. Acker, A., et al.: Social media data archives in an api-driven world. Archival Science **20**(2), 105–123 (2020)
2. apify: Get real-time web data for your ai (nov 2025), https://apify.com/
3. Bright Data: The web's data, unlocked (nov 2025), https://brightdata.com/
4. Brooker, P., et al.: Have we even solved the first 'big data challenge?'practical issues concerning data collection and visual representation for social media analytics. In: Digital methods for social science: An interdisciplinary guide to research innovation, pp. 34–50. Springer (2016)

5. Brown, M.A., et al.: Web scraping for research: Legal, ethical, institutional, and scientific considerations. Big Data & Society **12**(4), 20539517251381686 (2025)
6. Debreceny, R.S., et al.: Research in social media: Data sources and methodologies. Journal of Information Systems **33**(1), 1–28 (2019)
7. Entrena-Serrano, C., et al.: Tiktok's research api: Problems without explanations. arXiv preprint arXiv:2506.09746 (2025)
8. Frasão, A., et al.: O cenário atual de golpes em redes sociais: Uma revisão da literatura. Anais do Computer on the Beach **16**, 406–413 (2025)
9. Ghani, N.A., et al.: Social media big data analytics: A survey. Computers in Human behavior **101**, 417–428 (2019)
10. Goswami, A., et al.: Challenges in the analysis of online social networks: A data collection tool perspective. Wireless Personal Communications pp. 4015–4061 (2017)
11. Gruber, J.B.: traktok—making tiktok data accesible for research (2025)
12. Instaloader: Instaloader (nov 2025), https://instaloader.github.io/
13. Kaur, P.: Sentiment analysis using web scraping for live news data with machine learning algorithms. Materials Ttoday: proceedings **65**, 3333–3341 (2022)
14. Kready, J., et al.: Youtube data collection using parallel processing. In: 2020 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW). pp. 1119–1122. IEEE (2020)
15. Make: Ai automation you can visually build and orchestrate in real time (nov 2025), https://www.make.com/en
16. Martín-Gómez, L., et al.: Business benefits of instagram scraping: Questionable uses of data. In: International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence. pp. 219–227. Springer (2021)
17. Mimizuka, K., et al.: Post-post-api age: Studying digital platforms in scant data access times. arXiv preprint arXiv:2505.09877 (2025)
18. Octoparse: Easy web scraping for anyone (nov 2025), https://www.octoparse.com/
19. Schroeder, R.: Big data and the brave new world of social media research. Big Data & Society **1**(2), 2053951714563194 (2014)
20. Scott, R.E.: Data scraping youtube for the study of lieder reception. Nineteenth-Century Music Review **19**(3), 655–667 (2022)
21. snscrape: snscrape (jun 2023), https://github.com/JustAnotherArchivist/snscrape
22. Stieglitz, S., et al.: Social media analytics - -challenges in topic discovery, data collection, and data preparation. International journal of information management **39**, 156–168 (2018)
23. Thunderbit: The next gen ai web scraper (nov 2025), https://thunderbit.com/
24. TikTok for developers: Develop for >communities (nov 2025), https://developers.tiktok.com/
25. Vlachos, S., et al.: Exploring instagram and youtube data. In: Data Management Technologies and Applications. p. 90. Springer Nature (2023)
26. Weller, K., et al.: Uncovering the challenges in collection, sharing and documentation: the hidden data of social media research? In: International AAAI Conference on Web and Social Media. vol. 9, pp. 28–37 (2015)
27. X: X api (nov 2025), https://developer.x.com/en/docs/x-api
28. You, J., et al.: A complete and fast scraping method for collecting tweets. In: 2021 IEEE international conference on big data and smart computing (BigComp). pp. 24–27. IEEE (2021)
29. yt-dlp: Yt-dlp a feature-rich command-line audio/video downloader (nov 2025), https://github.com/yt-dlp/yt-dlp
30. Zelle, Y., et al.: Scitok-a web scraping tool for social science research. In: International Conference on Human-Computer Interaction. pp. 103–109. Springer (2023)