# Report on data availability of social media platforms

**Version: 1.0.0**

**Addison Suhr**
Research Data Analyst

**Boyd Thai Hoang Nguyen**
Data Scientist/Developer

**Australia Digital Observatory**

digitalobservatory@qut.edu.au

**3 February 2025**

# 1 Executive summary

This report outlines the feasibility of collecting data from a number of social media platforms. The investigated platforms include both mainstream sites (e.g. Facebook, Reddit, Instagram) and decentralised platforms (e.g. Mastodon, BlueSky). Feasibility is assessed based on the availability of Application Programming Interface (API), ease of access, and other practical considerations for data harvesting. The report was completed as part of the Australian Internet Observatory (AIO) Work Package II.I.

The central finding from the report is that the high-volume social media data access for research has become increasingly difficult to obtain. Platforms are moving to gatekeep and restrict use of their data, while sites that support open data sharing are too small to provide comprehensive data for academic research. Researchers needing large data volume for their projects can consider multi-platform data collection or web archiving.

# 2 Introduction

Social media data have been used for myriad research purposes. Platforms such as Twitter, Facebook, Instagram, Reddit, and Wikipedia have facilitated academic studies spanning sociology, public health, economics, and political science. However, recently there has been a tendency among these platforms towards locking down their data. For example, following its acquisition by Elon Musk, X (formerly Twitter) closed the Academic API that had been a vital data source for social media research[1]. Likewise, Reddit has drastically restricted access to its previously liberal API.

Since then the research community has widened their search for data sources of similar qualities - high-volume, comprehensive, up-to-date, and easy to collect. This report provides a horizon scan of existing social media platforms, with a focus on the *feasibility* of collecting data from them. Aside from well-known, centrally managed sites such as Facebook and Reddit, the report also looks at decentralised and niche spaces, including those within the Fediverse

The *Fediverse* refers to a federated (inter-operationally connected) collection of decentralised social networks that can communicate with one another using a common protocol[2]. Mastodon is one of the most popular platforms of the Fediverse. Networks within the Fediverse send and receive data from each other via typically open-source protocols such as ActivityPub, which promotes openness, public accessibility, and enables public data to flow freely between servers. As a result, platforms that support the Fediverse increasingly gained research interests thanks to their potential for systematic, high-volume data sharing

[1] https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research

[2] https://www.theverge.com/24063290/fediverse-explained-activitypub-social-media-open-protocol

and collection. However, despite the exodus to decentralised platforms following Elon Musk's acquisition of Twitter, adoption remains low. Around 925,000 users post on Mastodon every month, a fraction of X's 611 million active tweeters[3].

## 2.1 Feasibility of data collection

The properties of Twitter's old Academic API form the basis on which we assess the feasibility of collecting data from other platforms. In specific, three major factors were considered: availability of official APIs, ease of access and costs, and data volume.

### 2.1.1 Availability of official APIs

An official public Application Programming Interface (API) from the platform enables users to retrieve its data in a controlled manner, while ensuring compliance with its terms and policies. In general, APIs are the most reliable and legitimate means of collecting data from platforms.

Twitter's old Academic API was exemplary in its comprehensiveness, transparency, and volume. Approved researchers, guided by clear and thorough API documentation, could query the entire public Twitter for relevant data and download them to their computers. The same standards are used to assess other platforms. A lack of public APIs, whether intentional or not, poses considerable challenges to large-scale data collection.

On the other hand, Fediverse platforms, by using standard and open communication protocols, in general allow user data to be shared and collected more freely, with the data being under the server owner's control instead of a central corporation. Access to data, as a result, does not depend on a single platform's policies and permissions. However, it should be noted that some servers might still impose specific restrictions or require authentication for data collection.

### 2.1.2 Ease of access and costs

Having an API does not mean easy access to the platform's data. Hurdles such as costs, lack of endpoints, and other restrictions might stand in the way of getting useful data. For example, the price of collecting 300,000 tweets from X API v2 now stands at 5,000 US dollars per month[4]. Tiktok's Research API is not available in Australia[5], and YouTube Data API's application for research access requires technical specifics that is largely inapplicable for most academic research projects[6].

### 2.1.3 Data volume

How much data are available is a major concern of researchers studying topics of a national or global scale. What made Twitter's old Academic API so valuable was its access to a "firehose" of high-volume data. All public tweets, if they had not been removed by the

[3] https://adamconnell.me/social-media-platforms/; https://explodingtopics.com/blog/x-user-stats
[4] https://developer.x.com/en/products/x-api
[5] https://developers.tiktok.com/products/research-api/
[6] https://support.google.com/youtube/contact/yt_researcher_certification

platform or their authors, were retrievable via the API. Such liberality is now rarely seen among social media platforms. Reddit, a platform previously known for its open data approach, has placed rate limits on their API and requested the closure of the historical Reddit archive Pushshift[7].

# 3 Findings

A total of 21 platforms were investigated, spanning the whole spectrum of size, user demographics, and network architecture. A table summary of findings can be found in the Appendix.

Many platforms (e.g. Bluesky, Threads, Truth Social, Mastodon) resemble Twitter in looks and interaction style, and most have been beneficiaries of the migration away from Twitter/X. **Bluesky**[8], a partially decentralised[9] microblogging platform that provides a public API for open data access, is one of the more promising platforms for systematic data collection. ATProto[10] - the protocol used by BlueSky - is designed for decentralised social networking, thus enabling easy data streaming and collection. Moreover, the architectural structure with which Bluesky and AT-Proto have been developed also enables far greater growth without the processing or communication overheads incurred by many older Fediverse protocols and platforms. With 26 million users[11], Bluesky stands as a potential alternative to X.

**Threads**[12] - Meta's Twitter alternative - has demonstrated willingness to engage with the Fediverse, providing users in some countries and regions the ability to *federate*, or connect to the Fediverse, their profiles. These federated accounts and their data, thus, are publicly accessible via ActivityPub. However, federation remains an opt-in choice, and non-federated accounts are still gated from public access due to a lack of official public APIs, while Meta have actively sought to shut down community reverse engineering attempts. Together, these mean that data collection from Threads remains a challenging endeavour.

**Truth Social**[13], one of the major alt-right social media platforms, does not support federation nor provide an official public API. Fortunately, data collection from the platform was made possible thanks to reverse engineering efforts by the community. These open source

[7] https://www.reddit.com/r/modnews/comments/134tjpe/reddit-data-api-update-changes-to-pushshift-access/

[8] https://bsky.app/

[9] Currently, most users are still on Bluesky's official server. A fully decentralised service will allow individuals to host their own servers.

[10] https://docs.bsky.app/docs/advanced-guides/atproto

[11] https://backlinko.com/bluesky-statistics#key-bluesky-stats

[12] https://www.threads.net/

[13] https://truthsocial.com/

tools nonetheless are highly susceptible to changes on the main platform and require community maintenance, which may not be guaranteed.

**Mastodon**[14] is a dominant player in the Fediverse. Mastodon's userbase expanded after the influx from Twitter/X, although it still caters to a relatively small community. Supporting the ActivityPub protocol common among the Fediverse, Mastodon allows its data to be accessed by other services and external users. However, data activities remain subject to the specific server's policies, which may enforce authentication or restrictions on mass data collection.

**Reddit**[15] is one of the most important platforms for social interaction data, particularly given its unique structure of topic-focused communities (subreddits) and hierarchical comment threads. Previously, Reddit data were readily accessible via the site's liberal API as well as Pushshift, an archive of all Reddit posts. API pricing changes were implemented in 2023, whereby Reddit moved away from its open approach to data access and imposed strict rate limits and API access fees. While some community-maintained scraping tools exist, Reddit's enhanced anti-scraping measures and legal stance against unauthorised data collection pose substantial challenges for researchers. Fortunately, academic access is still possible. An unofficial compressed archive of all Reddit data is also available for download, although the format and sheer size of these files requires significant computational resources and skills to process.

**Micro.blog**[16] (blogging), **Pixelfed**[17] (image sharing), and **PeerTube**[18] (video sharing) are smaller Fediverse platforms that support open data transfer. These are niche communities that might not contain sufficient content for research into broad issues. As a consequence, the data can be useful in supplementing other sources.

# 4 Conclusion and future directions

Collecting data from social media platforms is increasingly difficult, fraught with myriad legal and technical challenges. Closed commercial platforms boast a broad user base, although they usually employ measures that prohibit data collection, either by not providing an official public API, restricting access, or actively preventing and stopping attempts at scraping their data. Decentralised networks, due to their use of open protocols, allow easier transfer of data, but the user communities are small in scale and often niche in scope. Specific networks may also impose terms and policies that hinder high-volume data collection.

[14] https://joinmastodon.org/
[15] https://www.reddit.com/
[16] https://micro.blog/
[17] https://pixelfed.org/
[18] https://joinpeertube.org/

These obstacles mean that scraping data via APIs may no longer be a feasible option for researchers, especially those who rely on high-volume data on a broad time scale. One alternative paradigm is *data donation*, that is, inviting users to share their digital data. Users of social media platforms generally have access to their own data as Data Download Packages, a legal requirement as part of the European Union's General Data Protection Regulations[19]. Researchers can partner with these individuals to facilitate the voluntary sharing of data. Data donation has enabled several research initiatives to obtain sensitive or private social media data[20]. However, lack of scalability remains an obvious limitation. Another potential solution is *web archiving*[21], whereby web content is preserved in an archival format, ready for future use and access. While automated web archiving for large-scale data collection remains to be investigated, there is a big community dedicated to this method and ample resources are readily available. Researchers are encouraged to consider the applicability of web archiving to their studies and how this paradigm can be used to overcome the obstacles associated with API data collection.

# 5 Appendix

[19] https://gdpr-info.eu/art-15-gdpr/; https://gdpr-info.eu/art-20-gdpr/

[20] https://dl.acm.org/doi/10.1145/3491101.3503569; https://www.library.ucsf.edu/about/archives/social-media-collecting-policy/

[21] https://netpreserve.org/web-archiving/

**Table 1:** *Summary of social media platforms*

| Platform | Content Type | Architecture | Official API Availability | Practical Considerations | Protocol |
|---|---|---|---|---|---|
| Bluesky | Microblogging | Distributed | Yes | Data transfer via open protocols | AT-Proto |
| Facebook | Personal networking | Centralised | No | Limited Academic access | Proprietary |
| Flipboard | News media | Centralised | Yes | Closed API which supports open data transfer protocols | Proprietary, with ActivityPub support |
| Hive Social | Microblogging | Centralised | No | | Proprietary |
| Instagram | Image sharing, Short-form video | Centralised | Yes | Limited API that requires registration and approval from Meta. Limited Academic access | Proprietary |
| Kik | Live video, Group chat | Centralised | No | | Proprietary |
| Kick | Live video | Centralised | No | | Proprietary |
| Lemmy | Forum | Distributed | Yes | Data transfer via open protocols | ActivityPub |
| Mastodon | Microblogging | Distributed | Yes | Data transfer via open protocols | ActivityPub |
| Micro.blog | Microblogging | Distributed | Yes | Data transfer via open protocols | ActivityPub |
| Misskey | Microblogging | Distributed | Yes | Open API, and ActivityPub | Proprietary, with ActivityPub support |

| Platform | Content Type | Architecture | Official API Availability | Practical Considerations | Protocol |
|---|---|---|---|---|---|
| PeerTube | Video | Distributed | Yes | Data transfer via open protocols | ActivityPub |
| Pixelfed | Image sharing | Distributed | Yes | Data transfer via open protocols | ActivityPub |
| Pleroma | Microblogging | Distributed | Yes | Data transfer via open protocols | ActivityPub |
| Reddit | Forum | Centralised | Yes | Priced, rate limited | Proprietary |
| Rumble | Video | Centralised | No | | Proprietary |
| Threads (Meta) | Microblogging | Centralised | Partial | Public access is limited to accounts who have elected to federate their accounts | Proprietary, with opt-in ActivityPub support |
| TikTok | Short-form video | Centralised | Yes | Academic access available (currently not in Australia) | Proprietary |
| Truth Social | Microblogging | Centralised | No | | Proprietary |
| Twitter | Microblogging | Centralised | Yes | Can be costly depending on data volume | Proprietary |
| WordPress.org blogging | Blog | Distributed | Partial | Public access limited to federated users | Proprietary, with opt-in ActivityPub support |

| Platform | Content Type | Architecture | Official API Availability | Practical Considerations | Protocol |
|---|---|---|---|---|---|
| YouTube | Video sharing | Centralised | Yes | General access is rate limited; Academic access available upon application | Proprietary |

# 6 Citation

To cite this report:

- Suhr, A., Nguyen, T.H., and Australian Digital Observatory (2025). *Report on data availability of social media platforms*. https://www.digitalobservatory.net.au/resources/report-on-data-availability-of-social-media-platforms/