Journal of the Statistical and Social Inquiry Society of Ireland
Vol. LIV

Symposium: Big Tech and Big Data - New Frontiers in Regulation

Transparent and Accountable Platforms with Open APIs[1]

Marco T. Bastos

University College Dublin
City St George's, University of London

(read before the Society, 30th April 2025)

Abstract: This article examines the data access policies implemented by Twitter that transformed the platform into an open resource for academic research. We revisit the many Application Programming Interfaces that Twitter offered to developers and researchers and their role in building trust with social media companies. This includes REST, Search, Streaming, Academic, and Compliance APIs in addition to databases curated by the company and shared with the research community before its contentious acquisition by Elon Musk in late 2022 and its ensuing rebranding as X in July 2023. We conclude by outlining the requirements for transparent and accountable social media platforms and discussing the opportunities afforded by the EU's Digital Services Act to advance this agenda.

Keywords: Open APIs, Twitter, DSA, Content Moderation, Systemic Risks
JELs: M14, M48, Z1, C8, C55

1. INTRODUCTION

Twitter's generous stance towards data access, particularly the public API released in 2006, was a pivotal development in the evolution of social media that allowed a relatively small company to emerge as the most studied and scrutinised social media platform. Twitter's provision of multiple public APIs, alongside premium and enterprise offerings, facilitated extensive data collection (Twitter, 2019) and led to its disproportionate representation in social media research (Blank, 2017). The social networking service emerged as a prime source of social science data due to its relatively large user base across the world, with a demographic that skewed towards young adults and the technorati while also appealing to journalists, artists, activists, and academics (Newman et al., 2021).

Public and open APIs such as those operated by Twitter remained an exception in the social media ecosystem, even during the regime of open data access that preceded the data lockdown of social media platforms (Walker et al., 2019, Bruns, 2019). In contrast to Twitter openness, Facebook's Public Feed API, later referred to as 'Feed' API, was restricted to a limited set of media publishers (Facebook, 2018). Twitter became a pervasive tool in election campaigns, with research relying on the platform data to understand how candidates, parties, and journalists reacted to, commented on, or interacted around politics (Jungherr, 2016). Scalable and reliable access to Twitter data provided the necessary transparency and reflexivity to the networked publics during a period of relative trust in social media companies.

The Cambridge Analytica data scandal, and the data lockdown of social media platforms that ensued, was an important milestone that changed the assumption that social network sites were natural challengers to the monopoly

[1] We are thankful for the comments provided by the editors on previous versions of this manuscript. This article draws on a chapter in Axel Bruns, Gunn Enli, Eli Skogerbo, Anders O. Larsson, and Tanja Bosch, eds., *The Routledge Companion to Social Media and Politics* (London: Routledge, 2025).

enjoyed by the mass media (Castells, 2012). The period was marked by an intense focus on social media manipulation and disinformation (Benkler et al., 2018), but also on protests and demonstrations amplified by the instantaneous reach of Twitter conversations. More importantly, it marked a transition from social media platforms as infrastructure that was integral to the network publics to advertisement-based services run by very large corporations.

As online social networks transitioned from spaces of activity and interaction into platforms marked by passive media consumption, they also altered the regime of data access in fundamental ways. This shift was marked by the rise of mobile computing over desktop applications, leading to the cloud-based centralisation of social platforms that offers a stark contrast to the fragmented ecosystem of social network applications from which Hootsuite and TweetDeck emerged in the 2010s, the latter of which would ultimately be acquired and rebranded 'X Pro' following the company takeover by X Corp. The ever-diminishing access to data offered to researchers was accompanied by a significant shift in the business model of online social networks toward media consumption and the monetisation of user data.

The growing restrictions imposed on data access for social media research, and the collapse of Twitter's public endpoints for historical data access, underscores a critical shift in the power dynamics between platforms and researchers. Twitter's early commitment to open APIs fostered a rich era of academic inquiry into online social dynamics, political communication, and information diffusion. The early days of relatively open and public platforms supported a measure of transparency into these systems through generous regimes of data access. This initial stance, however, proved to be an anomaly in the broader social media ecosystem, with Meta platforms adopting increasingly more draconian regimes of data access that turned online social networks from public infrastructures into services tailored for passive media consumption and designed for maximum advertisement revenue.

Consequently, the ability of independent researchers to scrutinise platform impact and systemic risks has been severely curtailed, with tangible consequences for free speech, democratic persuasion, and the moderation of large technical systems driven by algorithmic filtering. The data access lockdown also curbed the ability of social scientists to study online behaviours, algorithmic biases, and the democratic consequences of increasingly opaque digital spaces. As social media platforms have turned into centralised and opaque content providers, changing the regime of data access is likely to require regulatory mandates to ensure public accountability and informed discourse online.

Without external oversight, digital platforms may operate without sufficient accountability and enable practices that harm users and exacerbate societal issues. Independent researchers offer critical perspectives that help uncover systemic risks - including the spread of misinformation, political polarisation, algorithmic bias, and mental health impact - that internal studies may overlook or underreport due to conflicts of interest. This lack of transparency is particularly damaging as it prevents addressing not only salient and known issues with content moderation and platform governance; it also maximises the set of risks that remain unknown for policymakers and the citizenry due to the opaque nature of the systems in place.

As such, a comprehensive understanding of how platforms are influencing public discourse and behaviour is paramount for an informed citizenry and the open dialogue required for bridging ideological divides. The absence of data access for independent researchers also impedes evidence-based policymaking, leaving regulators without the intelligence needed to craft and implement effective digital governance systems. This creates a knowledge gap that delays or distorts responses to emerging harms. The dominance of proprietary research, with social platforms spousing large data science teams, also undermines public trust, as users and stakeholders have no assurance that platform-generated findings reflect reality, or that research findings that run counter to the commercial interests of these platforms are communicated to the public. The corrective drive of independent analysis also affords opportunities to improve platform design, content moderation, and user protection, without which social media systems can weaken democratic oversight and social resilience in the face of rapidly evolving digital environments.

2. OPEN APIs

Application Programming Interfaces (APIs) are critical endpoints designed to facilitate programmatic access to data and computational resources within cloud infrastructures, including those provided by social media platforms. Open APIs are often referred to as a 'public APIs' as these endpoints are available on the internet and free to access by any user. The initial release of the Twitter Developer API in 2006, shortly after the platform's public launch, offered

remarkably expansive access to Twitter content thorough a well-documented public API. While subject to latency constraints during data-intensive operations, such as retrieving social graphs or downloading large datasets, whitelisted users (i.e., pre-approved accounts) enjoyed virtually unrestricted access to Twitter data. This regime of relatively open data access proved critical in the success of the platform with developers, journalists, and the academic community.

Twitter later introduced limits to API access (Needleman, 2009) from a single IP address at approved or 'whitelisted' Twitter services to 20,000 requests per hour (each request could return several tweets). These revisions significantly impacted services relying on Twitter APIs for bulk data collection, particularly the requesting of follower-followee information essential for constructing the social graphs of user communities. Beyond the increasingly rare whitelisted accounts, general users faced strict limitations: the REST API permitted retrieval of only 3,200 recent posts in real-time and offered no historical access, while the Search API allowed 100 hits, providing approximately one week of historical data, albeit the historical data provided by the endpoint offered a sample notoriously for its unreliability.

The Streaming API presented a more robust alternative, providing access to an effectively unlimited stream of near real-time data, contingent on the query volume not exceeding 1% ('spritzer'), 10% ('gardenhose'), or 100% ('firehose') of the total Twitter public stream during the API call. Taken together, Twitter's suite of public APIs allowed researchers to access a vast trove of data and facilitated the monitoring of emerging social issues. This was particularly important in the context of politically contentious events, including elections and referenda, which frequently gained initial traction on Twitter before mainstream media coverage. Notable examples include the 2009 US Airways plane crash in the Hudson River, which established Twitter's role in breaking news dissemination, a trend that continued through subsequent events such as bombings and protests (Hermida, 2010). This accessibility fundamentally transformed the capacity for real-time analysis of public discourse and event narratives.

The introduction of Twitter API version 1.1 in 2012, and the ensuing deprecation of public API 1.0 in June 2013, introduced a range of new features that proved important for social media research. This update introduced crucial features such as querying native media, quotes, quoted tweets, and polls. It also introduced a more equitable data access regime that allowed registered accounts to query posts, lists of followers and followees, and other Twitter data in 15-minute intervals. These standardised limits proved instrumental for researchers, enabling systematic planning and scalable data collection at regular intervals and supported much of the research that emerged in the early 2010s probing large and very large graphs from Twitter, with representative sample sizes that would be inconceivable in the following decade (Huberman et al., 2009, Kwak et al., 2010).

Twitter's diverse suite of Application Programming Interfaces (APIs) facilitated both programmatic and interactive collection of public data, each with distinct limitations regarding temporal coverage, latency, and limits on the number of requests. For instance, the REST API permitted access to a user's most recent 3,200 tweets. While this constraint posed minimal issues for studies focused on daily user activity or newly established accounts, it significantly hampered research designs requiring comprehensive historical data from prolific or long-standing users who would predictably exceed this limit. To effectively study Twitter, researchers were expected to familiarise themselves with the differential access regime afforded by each API endpoint, thereby shaping the scope and feasibility of various research projects.

Despite this limitation, Twitter research often leveraged the 3,200 tweets per user limit to go back in time and retrieve messages from a group of users over a relatively long period. As a RESTful web service, data was collected by requesting specific sets of data mostly centred on posts (i.e., tweets) or user events (user bio, following graph, etc.) Similarly, the Search API allowed data collection based on keywords or hashtags. The indexing was however incomplete, and searches going back in time were restricted to a sampling of tweets posted in the past 7 days, in addition to a stringent request limit (18,000) and a plethora of variables affecting the ever-changing index of tweets available via the Search API. Researchers could nonetheless repeatedly query the Search API to maximise the coverage of the collected data (Patrick, 2017). Common to the Search API and the larger endpoints available to the REST API was the request/response structure, with each HTTP request returning a chunk of data, so the process was somewhat simpler and decidedly different compared to querying the Streaming API.

Another important feature of the REST API was that it allowed researchers to cross-reference their datasets irrespective of which API was used to collect the data, which in addition could also be purchased from data resellers

or by scraping social media websites (Burgess and Bruns, 2015). While the Terms of Service prevented researchers from sharing complete datasets, Twitter allowed the sharing of each post's unique identification number, known as a 'snowflake.' The tweet ID could then be queried through the REST API to programmatically retrieve (or 'rehydrate') the original social media post, if still accessible (Bastos, 2021). This feature made it possible to inform other researchers which posts were included in the study and provided a method for comparing posts in different datasets. Deleted posts and accounts, however, could not be retrieved from the REST API, with the removal of user accounts further generating orphaned data. Similarly, modified posts and post metadata were not flagged by the APIs or web interfaces, so researchers could not determine if a post or its metadata had changed since it was originally posted. The other main shortcoming of rehydration was the time-consuming nature of the process, with the REST API rate-limiting queries to 150 requests per hour and returning a maximum of 100 tweets per request (Bastos, 2025a).

The Streaming API, on the other hand, was based on HTTP streaming instead of HTTP requests. In this implementation, the API pushes updates to web clients by keeping a persistent connection open that seamlessly pushes data to one's computer. Researchers could initiate a single request to secure a continuous data stream, archiving content matching their search criteria until the connection was manually terminated. This functionality made the Streaming API particularly valuable for monitoring ongoing events or tracking updates from specific user lists. A notable advantage of the Streaming API was its capacity to deliver substantial volumes of data, provided the query's response remained below 1% of the total Twitter public stream (Driscoll and Walker, 2014, Morstatter et al., 2013). This provided researchers with access to real-time, large-scale data that remained unattainable through other API endpoints.

In other words, the Streaming API theoretically provided access to all tweets matching the query criteria (keywords, username, hashtag, etc.), provided it did not include more than 1% of all messages tweeted in a given timeframe. The 1% limit applied to queries could also be used to retrieve a 1% 'random' sample of tweets, thereby maximising the data quota available to researchers (Morstatter et al., 2013). The free tier of the Twitter Streaming API, known as 'spritzer,' returned a 5% sample of the entire Twitter public stream from 2006 to 2010. After 2010, the spritzer returned a 1% sample of the complete (i.e., 100%) public Twitter stream, with the complete public stream being referred to as the 'firehose' output. An intermediate access level allowed users to retrieve a 10% sample of the public stream. This access level was referred to as 'gardenhose,' but Twitter would eventually transition to a business model in which non-paying users were pushed to the spritzer while commercial and enterprise users could purchase data via GNIP, Twitter's data-reseller, which rebranded the 'gardenhose' as 'decahose.' This transition marked an important shift towards the monetisation of social media data access that would accelerate with the commercialisation of Generative AI models requiring a steady data stream to train increasingly larger models, with OpenAI seeking to develop its own Twitter competitor for access to real-time data (Robison and Heath, 2025).

GNIP, Twitter's data reseller, also offered access to a RESTful API called 'Historical PowerTrack,' an API that was functionally equivalent to the last endpoint provided by the company: the Twitter Academic API. This API provided global historical and real-time streaming data access through the Twitter API v2. This Academic Research API was an invaluable data source for researchers and an all-time highlight in Twitter's willingness to share data with academia. In addition to offering access to real-time and historical public data, it also included additional features and functionality supporting the collection of complete and more reliable datasets. But the main feature of the Academic API was the possibility of accessing the full archive of historical tweets, a process that had proven challenging even for seasoned data scientists. Retrospective data collection, however, continued to suffer from issues of orphaned and deleted data, as the Academic API could only retrieve content that still existed in the platform at the time of the request, therefore excluding users and tweets that had been removed from the platform. Despite these caveats, the Twitter Academic API could create nearly complete samples of Twitter data based on a wide variety of search terms (Pfeffer et al., 2023).

This was a significant change because the major restriction of the Streaming API was that it only worked proactively, that is to say, the HTTP connection would only return tweets posted from the moment the connection was established, so it was not possible to collect past tweets using that endpoint. Another limitation researchers had to contend with was that only low-volume queries could be reliably accommodated, with queries of trending or popular keywords and hashtags potentially exceeding 1% of the public stream and therefore being subjected to sampling. Streaming API users could also only query for a maximum of 400 keywords and, unlike the Search API, Boolean operators were not supported by this endpoint. Researchers in academic and industry seeking to monitor

users instead of keywords and hashtags were also faced with a limitation, as the Streaming API only allowed up to 5,000 Twitter accounts to be monitored per client (Morstatter et al., 2013).

Twitter also maintained a Compliance API, which, regrettably, was not publicly accessible and was only briefly available to researchers during the COVID-19 pandemic. The Compliance API was unique in that it provided a rolling record of every piece of content that was removed, blocked, suspended, or otherwise deleted from the platform. It included the Compliance Firehose featuring two event streams (user and tweet) divided into six real-time streams that required a minimum of six permanent connections to Twitter's API. Events related to posts or users were subjected to six permanent states: three applied to status objects (tweet deleted, tweet withheld, and tweet edited), two applied to user objects (geodata scrubbing and user withheld), and one to favourite objects (like/favourite deleted). It additionally included eight persistent states: two applied to status objects (status dropped and undropped) and six applied to user objects (user deleted and undeleted, user protected and unprotected, and user suspend and unsuspend). The metadata included the timestamp of every alteration made to content (e.g., from account suspension to account deletion).

The Compliance API is likely to represent the most comprehensive dataset available for research into content moderation, misinformation, disinformation, and broader questions surrounding online regulation and speech. It provided a robust indicator for the ongoing tweet deletion rate (decay) that when available to researchers exceeded the mark of 15% of the entire Twitter public stream (Bastos, 2021). It also provided a structural view of the inherently dynamic and ephemeral nature of social media data, which makes it difficult for any two researchers to collect the same dataset in real time. Instead of being designed for research purposes, the Compliance API was created to provide developers with tools to maintain Twitter data in compliance with the Twitter Developer Agreement and Policy. Interaction with the API differed in that the Compliance API would expect compliance jobs with text files of user or tweet identifiers, and it would then return the identifier, action, timestamp of the action, reason, and redaction time.

3. BEYOND APIs

The other irreplaceable source of data provided by Twitter was the Twitter Moderation Research Consortium (TMRC), a database offered to the academic community designed to support research on political communication, particularly on disinformation, state-sponsored influence operations, social media propaganda, and content moderation. This expansive repository included tweets and embedded rich media (images and videos), but also user account information and the profile images of fake user accounts that had been subjected to various forms of content moderation measures by the company. Of particular note, the TMRC database not only included textual content but also incorporated information on profile images subjected to content moderation measures (George et al., 2024).

Within this repository, researchers had access to aggregated and granular data relating to user accounts flagged, removed, or subjected to enforcement measures, categorised according to criteria such as inappropriate content, graphic images, or violations of Twitter's policies and guidelines. The database included key metrics such as the number of accounts taken down, the number of tweets, languages used by the group of false accounts, key hashtags, account activity temporal range, user-reported locations, and technical indicators of location (George et al., 2024). As such, this database offered critical insights into the efficacy of content moderation strategies while also foregrounding the challenges that social media platforms must contend with in ensuring user safety and adherence to community standards.

The database resulted from initiatives implemented by Twitter's then-Head of Trust and Safety, Del Harvey, who established and oversaw the company's efforts to safeguard elections and deal with other problematic content on the platform that could jeopardise healthy conversations on Twitter (Harvey and Roth, 2018). The initiative evolved into Twitter's Civic Integrity policy and included a range of actions to identify activity that potentially interfered, caused confusion, or undermined public confidence in an election or civic process (Twitter, 2021). This initiative would eventually mature into the Twitter Moderation Research Consortium (TMRC). Starting in 2017-2018 as a reaction to the influence operations carried out by the Kremlin-linked Internet Research Agency 'troll factory' (IRA), it shared data with the academic community, initially under the umbrella of Twitter's Elections Integrity initiative, which identified and ultimately removed false accounts, Twitterbots, and sock-puppets (Elections Integrity, 2018, Roth, 2019).

The first data release included 2,752 accounts the company attributed to the IRA (Bastos and Farkas, 2019). This list was expanded in early 2018 to include 3,814 IRA-linked accounts. The TMRC continued to be updated over the next years, and the final dataset released by the TMRC included 115,474 unique Twitter accounts, millions of individual tweets, and more than one terabyte of media removed from the platform due to breaches of the Terms of Service. The subset of the data with detailed user profile information included accounts that posted in excess of 100 million tweets (25 million in TMRC14 and TMRC15 and 34 million in the 2018-2019 releases) linked to 57 influence operations carried out in several countries following the seminal campaign deployed by the IRA (Bastos, 2025b). Much like Twitter's many public APIs, access to the TMRC database was terminated following the transition of ownership to Elon Musk.

The database included only networks for which there was significant evidence indicating that state-affiliated entities tried to manipulate and distort public conversations. The influence operations taken down by the TMRC include small networks in Bangladesh that engaged in coordinated platform manipulation with a focus on regional political themes, and networks in the United Arab Emirates and Egypt that primarily targeted Qatar and Iran while amplifying messaging supportive of the Saudi government. This is in addition to accounts linked to Saudi Arabia's state-run media apparatus that engaged in coordinated efforts to amplify messaging beneficial to the Saudi government. Other small campaigns included in the database were the operations of Partido Popular in Spain and a separate small network associated with the Catalan independence movement, specifically Esquerra Republicana de Catalunya, but also networks in Ecuador tied to the PAIS Alliance political party, which primarily engaged in spreading content about President Moreno's administration.

The TMRC also included large information operations in Russia, Iran, and Venezuela targeting other countries and/or domestic audiences by leveraging 'spammy' content focused on divisive political themes, with behaviour that mimics the influence operation orchestrated by the IRA. The Iranian cohort posted nearly two million tweets with an angle that benefited the diplomatic and geostrategic views of the Iranian state. This playbook was also identified in a group of 4,248 accounts operating uniquely from the United Arab Emirates directed at Qatar and Yemen that employed false personae and tweeted about regional issues such as the Yemeni Civil War and the Houthi Movement. It also included very large influence operations counting over 200,000 accounts manned by the People's Republic of China (PRC) and dedicated to sowing political discord in Hong Kong and undermining the legitimacy and political positions of local protest movements. These accounts were suspended for a range of violations of Twitter's platform manipulation policies, including platform manipulation and spam, coordinated activity, fake accounts, attributed activity, distribution of hacked materials, ban evasion, and what Twitter would refer to as 'violative content.'

The contentious acquisition of Twitter by X Corp marked the pinnacle of corporate ownership of digital media data and social media ownership, with Meta setting the initial trend by drastically restricting access to Facebook and Instagram data in 2018. Five years later, in February 2023, Twitter also ceased to offer any free access to APIs v2 and v1.1 (X Developers, 2023). Paid tiers were introduced and immediately rendered many previous use cases financially prohibitive. This was undoubtedly an unfortunate development as public and open APIs made it possible for researchers to retrieve large datasets and curate databases associated with politically and sociologically meaningful events (Bruns, 2019). With limited to no API access, researchers have to resort to scraping web interfaces to retrieve data (Freelon, 2018), a process that is labour-intensive and drastically limits the amount of information that can be collected and processed. The removal of researcher access to APIs also constrains social sciences research to human-intensive means of data collection that cannot produce large or representative samples of real-world events, including social movements and elections, but also state- and non-state-sponsored disinformation campaigns (Bruns, 2019).

4. BACK TO OPEN APIs

Elon Musk's takeover of Twitter in 2022 led to changes to the platform and its user base, but it also led to consequential changes to researchers' data access. The billionaire publicly voiced his concerns regarding Twitter's opaque algorithmic ranking system and its perceived restrictions on free speech (Albergotti, 2022). Following the acquisition, Musk implemented changes to the systems that filtered spambots and reversed bans on Twitter accounts previously removed from the platform for spreading divisive or inflammatory content. Other changes included payment for verified accounts, the dismissal of multiple executives and staff members, and the transition of the company into private ownership (Vállez et al., 2024). Research in the area identified that the acquisition of Twitter by Elon Musk, and his pledge to promote free speech on the platform by overhauling verification and moderation

policies, was associated with a significant increase in engagement with contentious posts and growth in the influence of actors on the political right (Barrie, 2022). For researchers, it marked the end of Twitter's generous and transparent policy towards data access, a development that negates Musk's initial pledge for openness and free speech.

Researchers have suggested that restrictions on data access may lead to the consideration of alternative methods (Venturini and Rogers, 2019) and the assumption that a 'post API Age' has dawned where data would be collected against platforms' Terms of Service through techniques like web scraping (Freelon, 2018). These suggestions offer a roadmap to the resources researchers may leverage to collect data and implement their projects, but they cannot replace the central role of APIs in providing scalable and reproducible access to digital trace data (Bastos, 2024). They are certainly not drop-in replacements, as the volume, type (text, image, videos, interface, etc.), fidelity, timeliness, platform filtering, and extent of available metadata vary considerably across these methods. Similarly, high-volume data retrieved from APIs cannot directly substitute low-volume web scraping data. Even if the volume of data collected using web scraping and APIs were identical, the metadata available via API requests is considerably different from metadata that is visible on the user-facing portions of a social media platform's website that are used for web scraping.

The assumption that research has entered a 'Post API Age' (Freelon, 2018) also seems somewhat misplaced. While social media platforms have ostensibly curbed access to their APIs, social media services, streaming platforms, and the plethora of services that constitute modern web applications cannot operate without APIs, which remain central to cloud and web-based applications. As such, it would be more accurate to refer to a 'post-research-API access age,' as Application Programming Interfaces (APIs) remain essential to mobile and cloud-based technologies underpinning the social web. Indeed, it is difficult to see how cloud-based business development, and web applications in general, could perform operations requiring personalisation and scalability without resorting to APIs. Equally important, there are notable exceptions to this trend, including the many social networking sites that emerged as X/Twitter alternatives, including Bluesky and Mastodon. Mastodon uses the ActivityPub protocol for federation, which allows users to run and manage their own instance of the social media server. Due to its federated nature, data collection is restricted to single instances, with no indexing or a globally unique identifier for users and posts (although this can be inferred by combining instance name and snowflake ID). Access to the network social graph is therefore restricted by design, and the data immediately available to researchers are the instance name and the followers of the authenticated user, as well as the instances through which users may have reposted content from the authenticated user and their followers on that same instance.

Bluesky has some decentralisation features, with the AT Protocol that underpins Bluesky allowing for Personal Data Servers (PDS) through which users can host their own Bluesky content. But Bluesky's protocol also includes a centralised index and a radically open API that channels the spirit of the early days of Twitter. Indeed, the service is built under the assumption that all content (except for out-of-protocol content like Direct Messages) is intrinsically open. Bluesky's relays or core indexers fetch repository updates and forward these updates into network-wide data streams as a firehose of content that can be monitored, archived, and studied. Unlike Twitter's Streaming API, however, Bluesky's firehose cannot be currently filtered by keyword, hashtag, geographical bounding box, or username. Researchers must therefore self-host the raw DAG-CBOR data stream, which amounts to several terabytes of daily content, or alternatively monitor the smaller JSON output from Jetstream.

There are also several initiatives designed to secure researchers' access to social media APIs, chief of which is Article 40 of the EU's Digital Services Act (DSA), which mandates that vetted researchers must be able to request data from Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) to conduct research on systemic risks in the EU member states. The DSA mandates that platforms not only offer researchers access to public and non-public databases like the TMRC, but also publicly available data 'without undue delay,' a disposition that seems to establish a right to API-like access to public data, likely building on previous work based on Twitter's many APIs (Windwehr and Selinger, 2024). While API endpoints for data collection were primarily designed for programmers building application software that adds to the services offered by social platforms, as opposed to being resources designed from the ground up to meet the needs of researchers, the implementation of the DSA may result in APIs being purposefully designed for reproducible scientific research.

The mandatory obligations imposed on VLOPs by the DSA offer a window of opportunity to demand Open APIs and data access for independent researchers. This is clearly defined in the regulation, which explicitly delineates the

function of independent researchers and the requirement that they "conduct research that contributes to the detection, identification, and understanding of systemic risks and the assessment of risk mitigation measures." In addition to that, Recital 85 recognises the evolving nature of risks in these systems and mandates that VLOPs preserve all relevant documentation, including raw data and algorithmic testing data, to "show the evolution of the risks identified" (Regulation 2065, 2022). This legal disposition provides a strong basis for requesting access to open and well-documented APIs and for conducting sustained, long-term audits of social media recommender systems (McNally and Bastos, 2025). This regulatory provision is particularly significant given the developments that led to a near complete data access lockdown for independent researchers.

Twitter, of course, no longer exists, having been succeeded by X following the contentious acquisition of the social media platform by X Corp, a parent company established by Elon Musk in 2023 as the successor to Twitter, Inc. Musk's purported reasons for purchasing the company were grounded on concerns about malicious use of social media platforms to spread misinformation and disinformation, but also the much needed transparency and oversight of these services. In reality, however, the acquisition of Twitter by Elon Musk hindered platform transparency and fostered the spread of problematic content, including misinformation and disinformation. The shift in business model and the comprehensive staff layoffs, particularly in trust and safety teams, followed a policy of data access restrictions for independent researchers. This was in sharp contrast to Twitter's open API policies that facilitated extensive academic study, a development that made it all but impossible to monitor and analyse trends in misinformation and content moderation effectively.

These problems were compounded by the mass firings of content moderation staff, the dismantling of dedicated teams (e.g., Election Integrity team) that monitored harmful content, the monetisation of the verification system to provide legitimacy to misinformation peddlers, and the algorithmic amplification of Musk's own divisive content. As a result of these changes, the platform, now known as X, is facing significant scrutiny from regulatory bodies, particularly in the European Union under the Digital Services Act (DSA) for alleged failures to curb illegal content and disinformation. Indeed, and despite the initial rhetoric about transparency and free speech, the acquisition of Twitter by Elon Musk offers a sobering case study on the reduction in data accessibility for external scrutiny leading to the proliferation of misinformation and disinformation in public discourse. We can only hope that the DSA and further regulatory developments may bring back the much needed external oversight of social media platforms, which can only be properly achieved with transparent, scalable, and free access to data that are inherently public (McNally and Bastos, 2025).

References

Albergotti, R. 2022. Elon Musk wants Twitter's algorithm to be public. It's not that simple. *The Washington Post.* Washington, DC: William Lewis.

Barrie, C. 2022. Did the Musk takeover boost contentious actors on Twitter? *arXiv preprint arXiv:2212.10646*.

Bastos, M. T. 2021. This Account Doesn't Exist: Tweet Decay and the Politics of Deletion in the Brexit Debate. *American Behavioral Scientist,* 65**,** 757-773.

Bastos, M. T. 2024. *Brexit, Tweeted: Polarization and Social Media Manipulation,* Bristol, Bristol University Press.

Bastos, M. T. 2025a. So Long Twitter, and Thanks for All the Tweets. *In:* Bruns, A., Enli, G., Larsson, A. O., Robinson, J. Y. & Bosch, T. (eds.) *The Routledge Companion to Social Media and Politics.* London: Routledge.

Bastos, M. T. 2025b. Visual Identities in Troll Farms: The Twitter Moderation Research Consortium. *Social Media + Society,* 11.

Bastos, M. T. & Farkas, J. 2019. "Donald Trump Is My President!": The Internet Research Agency Propaganda Machine. *Social Media + Society,* 5.

Benkler, Y., Faris, R. & Roberts, H. 2018. *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*, Oxford University Press.

Blank, G. 2017. The digital divide among Twitter users and its implications for social research. *Social Science Computer Review,* 35**,** 679-697.

Bruns, A. 2019. After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society,* 22**,** 1544-1566.

Burgess, J. & Bruns, A. 2015. Easy Data, Hard Data: The Politics and Pragmatics of Twitter Research after the Computational Turn. *In:* ELMER, G., LANGLOIS, J. & REDDEN, J. (eds.) *Compromised Data From Social Media to Big Data.* London: Bloomsbury Academic.

Castells, M. 2012. *Networks of Outrage and Hope: Social Movements in the Internet Age,* Cambridge, Polity Press.

Driscoll, K. & Walker, S. 2014. Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication,* 8**,** 20.
Elections Integrity 2018. Data archive. October 17, 2018 ed.: Twitter, Inc.
Facebook 2018. Public Feed API. Facebook, Inc.
Freelon, D. 2018. Computational research in the post-API age. *Political Communication,* 35**,** 665-668.
George, N., Sham, A., Ajith, T. & Bastos, M. T. 2024. Forty Thousand Fake Twitter Profiles: A Computational Framework for the Visual Analysis of Social Media Propaganda. *Social Science Computer Review***,** 08944393241269394.
Harvey, D. & Roth, Y. 2018. An update on our elections integrity work.
Hermida, A. 2010. Twittering The News: The emergence of ambient journalism. *Journalism Practice,* 4**,** 297-308.
Huberman, B., Romero, D. & Wu, F. 2009. Social networks that matter: Twitter under the microscope. *First Monday,* 14.
Jungherr, A. 2016. Twitter use in election campaigns: A systematic literature review. *Journal of information technology & politics,* 13**,** 72-91.
Kwak, H., Lee, C., Park, H. & Moon, S. 2010. What is Twitter, a social network or a news media? *In:* 19th International Conference on World Wide Web, 2010 New York, NY, USA. ACM, 591-600.
McNally, N. & Bastos, M. 2025. The News Feed is not a Black Box: A Longitudinal Study of Facebook's Algorithmic Treatment of News. *Digital Journalism*.
Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. M. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. *In:* 7th International Conference on Weblogs and Social Media (ICWSM13), July 8-11 2013 Cambridge, Massachusetts, USA.
Needleman, R. 2009. Twitter puts new limits on API calls: Who's affected. *CNET.*
Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T. & Nielsen, R. K. 2021. Reuters Institute Digital News Report 2020. *Reuters Institute for the study of Journalism*.
Patrick, R. 2017. Nonprobability Sampling and Twitter: Strategies for Semibounded and Bounded Populations. *Social Science Computer Review,* 36**,** 195-211.
Pfeffer, J., Mooseder, A., Lasser, J., Hammer, L., Stritzel, O. & Garcia, D. 2023. This Sample seems to be good enough! Assessing Coverage and Temporal Reliability of Twitter's Academic API. *In:* Proceedings of the International AAAI Conference on Web and Social Media, 2023. 720-729.
Regulation 2065 2022. Regulation 2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act). *In:* Parliament, E. (ed.) *2000/31/EC.* Brussels: EU.
Robison, K. & Heath, A. 2025. OpenAI is building a social network. *The Verge.* New York, NY: Vox Media.
Roth, Y. 2019. Empowering further research of potential information operations.
Twitter 2019. More about restricted uses of the Twitter APIs.
Twitter 2021. Civic integrity. Twitter, Inc.
Vállez, M., Boté-Vericad, J.-J., Guallar, J. & Bastos, M. T. 2024. Indifferent About Online Traffic: The Posting Strategies of Five News Outlets During Musk's Acquisition of Twitter. *Journalism Studies,* 25**,** 1249-1271.
Venturini, T. & Rogers, R. 2019. "API-based research" or how can digital sociology and journalism studies learn from the Facebook and Cambridge Analytica data breach. *Digital Journalism,* 7**,** 532-540.
Walker, S., Mercea, D. & Bastos, M. T. 2019. The Disinformation Landscape and the Lockdown of Social Platforms. *Information, Communication and Society,* 22**,** 1531–1543.
Windwehr, S. & Selinger, J. 2024. Can we fix access to platform data? Europe's Digital Services Act and the long quest for platform accountability and transparency. *Internet Policy Review, March,* 27**,** 2024.
X Developers. 2023. *Starting February 9, we will no longer support free access to the Twitter API, both v2 and v1.1. A paid basic tier will be available instead* [Online]. X/Twitter. Available: http://web.archive.org/web/20230907185848/https://twitter.com/XDevelopers/status/1621026986784337922 [Accessed February 2 2023].