The State of Digital Media Data Research, 2024

Megan A. Brown
Dr. Josephine Lukito
Jason Greenfield
Dr. Bin Chen
Sarah Graham
Dr. Sarah Shugars
Dr. Meredith L. Pruden

MDDC
Media & Democracy Data Cooperative

Executive Summary

The purpose of this report is to reflect on the state of digital media data research in 2024. This is the second in a series of reports on the state of digital media research, which we originally published in 2023. We reflect on changes to digital media research since our report in 2023.

Specifically, we highlight the following trends:

1. From 2023 to 2024, access to digital media data changed drastically. Researchers were largely priced out of the Twitter API, and Pushshift–a commonly used archive for Reddit data–went private to comply with Reddit's API policies. Meta also announced the imminent sunsetting of CrowdTangle, a transparency tool popular amongst researchers and journalists alike. At the same time, however, many platforms announced academic programs for data access, including the YouTube researcher program, TikTok's Research API, and the Meta Content Library.
2. Federated social media platforms became more popular. Following Elon Musk's purchase of Twitter, Twitter users flocked to Mastodon, Threads, BlueSky, and other federated (or soon to be federated) platforms. This presents unique challenges for researchers studying digital media data. As new platforms are created, researchers must build new tools to analyze them or wait for third parties or the platforms themselves to make data available.
3. Generative AI's explosion may change how we study digital media. First, researchers using computational methods to measure social media content have turned to OpenAI's ChatGPT and other Large Language Models (LLMs) to classify content. Second, researchers and civil society groups are increasingly concerned about the possibility for Generative AI to flood the information environment with fake content.
4. In February 2024, the EU Digital Services Act (DSA) went into effect, mandating that large platforms give researchers near real-time access to public data. We don't yet know how these policies will impact data access in the United States, and it remains unclear what this data access will look like in practice. In the United States, legislative efforts to mandate researcher access stalled.

While the last year brought many welcome and unwelcome changes to digital media data research, the findings in this report renew our encouragement that digital media data research should be guided by collaboration, transparency, preparation, and consistency.

Loss of broader data access

The past year has seen two major changes to the state of access to digital media data for research: the replacement of Twitter's free API with a paid version and the closure of Pushshift, a popular historic data API for accessing Reddit data. Due to the prevalence of their use in academic research and their relative ease of use, the decline of these two data access tools has been a great cause of concern for researchers, both within and outside of academia. Yet there is a tension between the loss of broader data access and more recent initiatives to grant researchers access to data, as the consequence of both platform and legislative initiatives. Researchers will inevitably replace these tools with other methods for accessing data, but that process is burdened by an increase in both cost and risk.

Twitter announced the end of free access to their API in February 2023. This announcement came only a few months after Elon Musk concluded his acquisition of Twitter (which he later renamed X) and marked the beginning of a new regime of access to Twitter data. This period included several changes to Twitter data access, including the deprecation of the V1 API, the discontinuation of the Twitter 2.0 API Academic Track access and previous authentication tokens used by academics with increased quotas, and the release of new pricing tiers for the V2 API. Because of these changes, the Twitter API no longer provides sufficient, affordable, access to data. At the same time, Elon Musk has demonstrated he is far more litigious against those who scrape Twitter than previous owners, suing at least four unnamed individuals and the nonprofit Center for Countering Digital Hate. Taken in tandem, these dynamics paint a rather unclear picture of what future research about Twitter will look like. This shift is especially acute given how heavily academics relied on Twitter for data access and the disproportionate amount of research that used this data as opposed to that from other platforms.

Academics used to benefit from widespread access to Reddit data via the Pushshift Data API. But in April 2023, Reddit announced an update to their relationship with Pushshift, ultimately resulting in the closure of researcher access (though the tool still remains available to moderators on the platform). Moreover, Reddit also announced changes in its API for all users, substantially increasing the cost of data collected. Together, these changes to both Reddit's relationship with Pushshift and the changes in API pricing mean that the scope and scale of data previously available to researchers has been dramatically reduced.

These data access challenges are likely to get worse rather than better, at least in the short-term. In March 2024, Meta announced the impending closure of CrowdTangle, yet another pivotal data source that many researchers relied on. While CrowdTangle support had been progressively depreciating, the closure of the tool, which is slated for August 2024, will likely end many current efforts to study Meta-owned platforms Instagram and Facebook. As many academic and civil society researchers have decried, the closure of CrowdTangle will also hinder researchers' ability to support election integrity efforts in the 2024 U.S. Presidential election and the 64 other elections happening world-wide this year.

These changes to broader data access highlight the delicate relationship between digital media researchers and the platforms that they study. This suggests that data access is at an inflection point: current approaches to data collection have been piecemeal and highly contingent on access to platforms' APIs, many of which were not necessarily intended for researcher data access. As noted in the 2023 report, this resulted in inequitable access, both in the sense that some

platforms were studied more than others and in the sense that some researchers had greater access than others (resulting in a rich-get-richer data model). Such losses to data access tools also highlight a growing need for more robust policies and methods for independent digital media research.

Table 1. Data Access Chart

For more details regarding data access, see the Institute for Data, Democracy & Politics' Platform Transparency Tracker.

Platform Data Source	Who has access?	Academics Only	Notes
Google records request	Researchers affiliated with EU-based organizations	✓	Google records request has options for Maps, Play, Search, Shopping, and YouTube
LinkedIn	Unclear	✓	
Meta Content Library	Global		Meta Content Library supplies access to both Facebook and Instagram
Meta CrowdTangle	Global		Meta announced that CrowdTangle will sunset in August, 2024
Pinterest Researchers Intake	Unclear	✓	
Reddit Data API	Global		
Pushshift Archive	Moderators on Reddit		Previously, Pushshift was publicly accessible, but this is no longer the case
Snap Researcher Data Access	Unclear	✓	
TikTok Research API	United States and Europe	✓	
X (Twitter) API	Global		
YouTube Researcher Program	Must be affiliated with an academic institution in specified countries	✓	

Rise of Academic-only APIs

To substitute the loss of data access, some social media platforms have announced researcher programs, which grant just academic researchers (often) free and modified access to public APIs. YouTube, TikTok, Reddit, and Meta currently offer researcher programs with varying levels of access, which we discuss below. For each of these platforms, researchers from accredited institutions must apply for access with a research process and use case explicated. Importantly, these APIs come at a time when researcher access to data is diminishing, as we highlight previously. Despite these academic researcher programs, challenges still remain for researcher data access. First, researcher API access comes with terms of service that often conflict with common scientific practices, which promote transparency, replication of findings, and independent publication of research. For example, many platform policies prohibit researchers from sharing full metadata and require them to only provide content or user identifiers for replication materials, often rendering research unreproducible, as platform content changes.

Despite increased transparency and researcher data access programs, researchers are still bound by the good will of platforms for access to data to conduct research in the public interest. Importantly, platforms are still the arbiters for data access applications, making decisions about who does and does not get access based on their research questions (with the exception of the Meta Content Library, where access is approved by the Social Media Archive at the Inter-university Consortium for Political and Social Research). This means that even though these specialized access programs exist, the actual provision of access is still at the will of platforms to decide who gets access and for what purpose. Moreover, these researcher programs typically mandate that researchers must be based at a college or university, leaving out many researchers.

Another challenge facing researchers using platform data APIs is the presence of data deletion requirements. These policies require that projects delete metadata for any content that has been removed from the platforms. What this means for researchers is that they must delete data as they are collecting it when that data disappears from the platform. This is likely to be a particularly acute problem for researchers studying harmful content such as hate speech or disinformation, as that data is more likely to be removed because of platform moderation policies. In cases where researchers release datasets of identifiers (to meet compliance requirements with platform data sharing policies), researchers aiming to reproduce or extend findings using the same data may face challenges, as rehydration of data using platform APIs excludes deleted content. YouTube is currently the only platform that allows researchers to retain deleted data, but this only applies once researchers have started analysis.

Other challenges remain. Some platforms (e.g. YouTube, TikTok) currently require courtesy copies of publications written from platform data. While these are not systems requiring outright publication permissions from platforms, they do increase the likelihood that platforms can craft PR materials discrediting research before it is ever published. Moreover, none of these programs have systems to allow research projects–vetted for ethics, methodology, and potential impact–to perform research on more private platform data, as was allowed by platform-selected researchers in the U.S. 2020 Elections project by Meta. While researcher access programs welcome changes to researcher data access more broadly, there is still work to be done to ensure that platform transparency programs are supporting thorough public-interest research rather than transparency theater.

Policies on Data Access

This year, researchers studying and using digital media data continued to contend with a constantly changing data access landscape. The past 12 months have seen substantial changes to data access, in large part led by platform decisions to shut down enterprise and researcher data access programs. Research projects have stalled and placed researchers at risk of legal liability as researcher access to platform data has grown more limited, unreliable, or wholly cost-prohibitive.

Yet, researchers across academia and civil society have shown that data access is key to producing high-quality empirical evidence to understand the information environment. Legislative efforts have shown some potential to overcome these data access barriers by requiring social and digital media platforms to share data through various provisions and enforcement mechanisms. These legislative proposals and debates take place amid a global movement for online safety and transparency measures. Here, we highlight three key legislative proposals in the European Union (EU) and the United States, which have led the policy landscape for researcher access to data in the past year.

First, the European Union's Digital Services Act (DSA) went into effect in February 2024. Under Article 40 of the DSA, vetted researchers can request data from very large online platforms (VLOPs) or search engines (VLOSEs) to research systemic risks within the EU. The DSA enables researchers to access previously undisclosed data, including all public data in a near-real-time searchable format (akin to Facebook's soon-to-be-sunset CrowdTangle) and additional private data to study systemic risks to elections, democracy, health, and youth well-being. Initial evaluations of the DSA's success are mixed and the extent to which researchers can rely on meaningful access to data is yet to be shown.

Second, this year saw the reintroduction of the bipartisan Platform Accountability and Transparency Act (PATA) to the U.S. Congress. PATA requires social media companies to provide vetted, independent researchers and the public with access to certain platform data. Under this access protocol, researchers would submit proposals to the National Science Foundation, where, if approved, social media companies would be required to provide necessary data. In addition to access, PATA also seeks to legally protect researchers who are responsible for collecting and analyzing public data in the course of their work.

Third, the Kids Online Safety Act (KOSA) focuses on governing how minors engage with technology platforms. The Act itself is intended to safeguard youth audiences; however, the proposed mechanisms for enforcement and privacy reduction have fueled debates regarding KOSA's efficacy. Notably, a previous draft of the Act proposed data access for independent researchers to conduct public interest research through an application process with the National Institute of Standards and Technology (NIST). However, in July 2023, this provision for independent researcher access was removed after facing industry and political pushback in favor of internal auditing conducted by the National Academies of Science, Engineering, and Medicine. States are also proposing state-level legislation to limit youth's exposure to and use of social media, and it remains unclear how these regulations may impact researcher data access.

This year, more than 50 global elections will occur, coinciding with dramatic changes in the digital information environment, while a range of legislation, including the DSA, PATA, KOSA, and more, are up for debate. These policies focus on increasing transparency and broadening access to digital media data for researchers across academia.

Federated Social Media

The rise of federated social media, such as Mastodon and Bluesky, represents a significant shift in the landscape of digital platforms. Federated social media operate on a decentralized model, diverging fundamentally from the centralized nature of traditional social networks like Twitter; these social networks are often collectively called the "Fediverse." Unlike Twitter, where all users interact within a single, unified service controlled by one entity, federated social networks consist of multiple independent servers, or "instances," that can communicate with each other. This federation of servers allows for a more diverse and customizable social media experience, as each instance can set its own rules and moderation policies while still enabling its users to connect and interact with the wider network.

Elon Musk's acquisition of Twitter in 2022 catalyzed a significant migration of users to alternative social platforms. Driven largely by concerns over changes in content moderation policies and the overall direction of the platform under new leadership, people migrated to other platforms such as Mastodon and Bluesky. Notably, the federated social network platform Mastodon gained roughly 500,000 users within ten days of Musk's Twitter takeover on October 27, 2022. As of March 2023, Mastodon had over 10 million registered users, achieving a 300% increase within the next 5 months. Another federated platform, Bluesky, which launched in February 2023 amassed 1 million downloads worldwide in about 5 months.

The rise of federated social media platforms has sparked both positive and negative views. On the positive side, these decentralized social media introduce a vision for a diversified social web, suggesting more ways it might evolve. It encourages competition among different providers, likely improving the user experience. What's more, the decentralized structure can help fight against risks seen in centralized platforms, such as issues with leadership that may influence the whole platform.

However, there are concerns too. First, since federated platforms are open-source, it's easier for their codes to be used for harmful purposes, as seen when the alt-tech platform Gab reused Mastodon's code as a way of circumventing Google and Apple's ban. Moreover, content moderation is a bigger challenge for these new platforms. Established platforms, with their longer experience, might have more effective systems in place for this purpose.

The shift from centralized to federated social media offers researchers in digital media unprecedented opportunities to explore its implications across diverse fields. However, it also presents significant challenges for scholars in these fields. For instance, the adoption of federated social networks has led to users splitting across multiple platforms, requiring researchers to gather data from a broader array of sources to gain a comprehensive understanding of online behaviors and events. This fragmentation means that traditional methods of data collection and analysis may no longer suffice, pushing researchers to develop innovative methodologies and tools that can navigate the complexities of the "fediverse."

Generative AI

The concept of generative AI—defined as computational tools that are used to generate content (text, audio, images, or video)—has existed for many decades. However, the development of deep learning and neural networks have led to an explosion in the accessibility and use of generative AI through third-party applications like GPT-3.5 (also known as ChatGPT), Midjourney, and Sora. Given the increasingly widespread use of these tools around the world, it is no wonder that 2023 was described as a groundbreaking year for AI by tech companies like Google and Microsoft, as well as news organizations such as the Associated Press.

When analyzing how generative AI has been adopted by people, researchers have found that generative AI has been used as a new way to curate or source information, such as to look up a word or to edit their code. But user adoption of generative AI has not spread as quickly as awareness about it. According to a Pew Research Center survey, only 18 percent of Americans have used ChatGPT; however, nearly 60 percent of Americans have heard of it. While this percentage is likely to increase as generative AI becomes more ubiquitous in society, the economy, and digital media content production, there is an important caveat: people are also concerned about how generative AI can impact daily life. For example, researchers have raised alarms that AI could potentially be used by malicious actors to produce large quantities of disinformation in the runup to the 2024 elections.

The ongoing research on digital media data has produced a mixed results regarding how generative AI can be utilized within scholarship. Some researchers are more optimistic, arguing that generative AI can make machine learning and computational methods more accessible or efficient. Few-shot learning, which aims to teach a generative AI tool to label data based on a small sample of labels, and zero-shot learning, which uses a generative AI tool to label data with no manual labels, have become increasingly popular methods in the fields of machine learning, natural language processing, and computational social sciences. Generative AI can also be used to help researchers learn programming skills, increasing access to these methods. As Chris Bail notes in a recent publication, "ChatGPT user[s], for example, can ask the model to explain what is happening in a single line of code, or how a function operates."

However, others have also noted that the use of generative AI in digital media data research is also trading accessibility and transparency or reproducibility. In other words, while generative AI can make computational methods more accessible, the lack of open information about these for-profit models is a big vulnerability, particularly for open science. Using LLMs for classification is a black box, which shares the same features as other black-box classifiers with respect to the challenges of using them for research. Additionally, using LLMs to generate synthetic samples may not yield similar results as real data with real people.

Scholarly Solidarity

The last year in social media research has made apparent the necessity of scholarly solidarity. Since our last report in 2023, researchers have lost access to data from a myriad of platforms. The impact of these data restrictions will not be evenly felt. Researchers with funding will be able to weather the storm, either through purchasing data or taking on more expensive data collection procedures, such as empaneling participants and collecting digital trace data. Under-resourced researchers, and particularly junior researchers, non-academic researchers, and researchers in smaller institutions, will face additional challenges collecting data, as many new research programs require institutional agreements like Data Use Agreements (DUAs) or support letters from faculty.

Additionally, researchers face growing attacks from outside the field. Recent legal threats against scholars doing vital research in the public interest are making studying the information environment more challenging, both personally and professionally. For example, X (Twitter) sued the Center for Countering Digital Hate in 2023 after they published a report about hate speech on the platform. However, this case was dismissed in early 2024 as a blatant attempt to silence researchers who aim to conduct analysis on Twitter's platform. Additionally, amidst increasing polarization in the United States, researchers have been targeted by congressional inquiries into their research, resulting in a broader politicization of important research about the digital environment.

New groups like the Coalition for Independent Technology Research offer a promising path forward. In 2023, after Twitter shut down, the Coalition provided mutual aid to over 50 research projects that were threatened as a result of Twitter shutting down the Academic API. Moreover, the Coalition has provided public support to researchers facing threats because of their research. These events bring to light the increasing need for scholarly solidarity. We echo calls for scholarly solidarity for researchers studying digital media in this report, encouraging scholars to support one another as we navigate ongoing challenges in the discipline.

The Future: What to Anticipate in 2024

Digital media research is in a state of strong uncertainty. Social media platforms have been changing rapidly and those changes have been coupled with a dramatic restricting of researcher access. While some platforms have launched special APIs geared toward academic researchers, the process for acquiring access is often slow and bureaucratic while the resulting data access can be minimal. Furthermore, as we've seen with the removal of Twitter's Academic API track and the imminent sunsetting of Meta's CrowdTangle API, academic access exists at the will of the platform and can be removed at any time. While regulations such as the European Union's Digital Services Act (DSA) show promise as a strategy for protecting researchers' ability to conduct vital work monitoring the digital ecosystem, it remains unclear how widespread and enforceable such regulations will be. At the same time, the rise of AI-empowered disinformation, user shifts to federated platforms, and elections and events happening worldwide make the study of online messages and behaviors more critical than ever.

So, what should researchers expect in 2024 and what can they do to prepare? While we like to optimistically hope for the best, we recommend researchers realistically plan for the worst. The coming year promises to be a pivotal one in determining the future of researcher data access. Platforms are testing the waters for a new age of restricted data access–seeing how little they can give away and how many bureaucratic hurdles can be erected to interfere. Researchers should be prepared to advocate against these restrictions through legal, political, and public relations strategies. Researchers should document the social media data they use and work to articulate to a general audience what is lost when we are unable to study these platforms. We know that democracy hangs in the balance, but we must ensure that others know the stakes as well.

Furthermore, researchers should make contingency plans in the event of loss of data access. Researchers should think specifically about the research questions they are asking, the data needed to study those questions, and develop strategies to pursue those questions while maintaining the high ethical standards of respect for persons, beneficence, and justice. For example, data donation is one strategy that is gaining popularity. Under this approach, consenting subjects download their own social media data and share that information with researchers (though under the increasing litigiousness of platforms, this method may yield risks later as well). Researchers may need to pursue other strategies as well, such as data cooperatives, direct partnerships with a platform, or even data scraping in cases where that would be ethically appropriate.

We expect 2024 to be a tumultuous year, but there is one thing that remains clear: digital media is an integral part of the modern information landscape. Platforms, and data access, will continue to change. However, the need to study, understand, and document these platforms, and the wealth of content they house, will remain. To meet this moment, researchers must advocate, express their contributions to a general audience, and get creative in how they study these critical elements of modern life. This includes leveraging collective efforts through coalitions like the Center for Independent Tech Research, internet observatories, and general-use archives.