

Political Communication Report Summer 2026 - Issue 33

“Political Communication Research under Data Constraints”

The Challenges of Working with Platform Data from Clean Room Environments¹

Axel Bruns, Digital Media Research Centre, Queensland University of Technology

Laura Vodden, Digital Media Research Centre, Queensland University of Technology

Modes of researcher access to digital trace data from social media and other digital platforms have undergone several changes over the decades. What Puschmann (2019) called the “Wild West” of relatively permissive data access via Application Programming Interfaces (APIs) – primarily designed to support third-party application development but also highly useful for research purposes – ended with the “APIcalypse” (Bruns, 2019) in the aftermath of the Cambridge Analytica scandal. Platform providers such as Meta and Twitter used the scandal as an excuse to decommission or severely curtail their APIs, ostensibly to prevent abuse. However, API closures also frustrate critical, independent, public-interest scrutiny by researchers. This is especially concerning given that platforms have been increasingly criticised for their lack of decisive action against ills including mis- and disinformation, hate speech, and other abuse (Bruns, 2018).

Such criticism has intensified since the mis- and disinformation infodemic (United Nations, 2020) accompanying the COVID-19 pandemic, the “enshittification” (Doctorow, 2025) of Twitter under Elon Musk, and the role of social media-based propaganda in democratic backsliding trends observed globally. New regulatory interventions – most prominently, the European Union’s Digital Services Act and its provisions under Article 40 (Giglietto & Puschmann, 2026) – now require major online platforms to provide comprehensive data access to researchers investigating “systemic risks” to democratic societies. Departing from conventional API models, a growing number of platforms now provide such access via ‘clean room’ facilities whose fit for purpose has yet to be comprehensively assessed.

¹ Copyright © 2026 Axel Bruns and Laura Vodden. Licensed under the Creative Commons Attribution Non-commercial No Derivatives (by-nc-nd). Available at <http://politicalcommunication.org>.

Clean Rooms in Principle and Practice

The clean room data access model was created in response to concerns that research data provided by platforms for access via conventional APIs may be stored, shared, and reused in ways that extend beyond the terms and conditions set in API access agreements, and by data users beyond the researchers who had been formally authorised to do so. In other words, once retrieved from traditional APIs, platforms no longer have any effective control over what happens with datasets, and rely on researchers' adherence to applicable ethical, privacy, and legal rules. Such concerns *are* valid, but have also been overemphasised by platforms in their attempts to justify the closure of API access models. For example, while data sharing amongst academics has constituted a legal grey zone ever since APIs became widely available, genuine negative effects from this are exceptionally rare. Furthermore, the public sharing of source datasets is strongly encouraged by research funders and publishers to ensure the transparency, rigour, and replicability of scholarly work, and the legality of deeply restrictive platform prohibitions against data sharing is disputable. Additionally, while research ethics require the protection of ordinary users who made and then deleted ill-advised posts, this concern must also be balanced against the need to document key platform activities for subsequent analysis and as a part of communicative history. From Donald Trump's infamous 'covfefe' tweet through the accounts of major political and other actors to the disinformation and propaganda posts made in the course of domestic and foreign influence operations, it is essential to preserve such content for posterity even if it is deleted from the platform itself (Weller et al., 2026).

Whatever its actual merits, the argument that platforms must retain tighter control over their (or more appropriately, their users') data – which implies that researchers themselves cannot be trusted to handle such datasets appropriately – has been central to the introduction of clean room environments as an alternative data access model. Such clean room models no longer permit researchers to retrieve platform data at scale for storage and analysis on their own computers or systems (or if they do, restrict this to a small subset of all available data, and/or to the export of aggregate communication statistics only). Rather than such data downloads from the API to researchers' own computers, they instead provide a platform-hosted research environment that is disconnected from the open Web to avoid any data leakages – the 'clean room' – that researchers log on to and within which they are allowed to interact with the platform data, often in strictly prescribed fashion and with a limited set of analytical tools as provided by the platform.

Of these clean room environments, the Meta Content Library (MCL), covering Meta platforms Facebook, Instagram, and Threads, is the most prominent to date. The MCL replaced CrowdTangle, previously a third-party data access tool for Facebook and Instagram which Meta acquired in 2016 (Newton, 2016) and which provided large-scale downloads of posts from selected public pages and groups on Facebook and public profiles on Instagram via Web-based and API facilities. Following CrowdTangle's decommissioning in August 2024 (Grevy

Gotfredsen & Dowling, 2024), and a beta trial phase for selected MCL users, approved applicants were granted full access in November 2024 – coincidentally resulting in substantial limitations to research data access to Meta’s platforms precisely during the final phase of the 2024 US presidential election.

The MCL has undergone several substantial changes since its initial release. Access was provided at first via a clean room environment hosted by the Social Media Archive (SOMAR), a facility housed at the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. To access the system, users first connected remotely to a virtual Windows machine, then connected from there to a virtual Linux machine, started up a Jupyter Notebook instance within that environment, and there could finally write Python or R scripts to access the MCL data hosted on Amazon Web Services (AWS) database servers.

This cumbersome and – especially from outside the United States – exceptionally slow ‘Virtual Data Enclave’ (VDE) has since been complemented and increasingly replaced by an alternative access framework, the ‘Secure Research Environment’ (SRE): streamlining the access process, the SRE uses the Amazon Workplaces Secure Browser plugin to connect directly to a virtual browser within the user’s own Web browser, and launches a Jupyter Notebook instance where researchers can again write Python or R code to query and process data from Meta’s AWS servers. While the SRE represents a vast improvement to the usability of the system, it addresses concerns about the handling of deleted posts and comments by mandatorily deleting all accessed data from the working directory on the first of each calendar month; this severely obstructs longitudinal studies.

Key Limitations of Clean Room Environments

Clean room environments like the SRE and VDE share an underlying philosophy of data provision and processing that severely limits both users and uses. Previously, API-provided datasets, once accessed, could be processed and analysed by researchers not only programmatically, but also interactively and visually using industry-standard data analytics tools such as Tableau or Power BI. The limitation to working with Jupyter Notebooks functionally excludes researchers without the necessary coding skills (or support staff) from accessing these data; this is deeply problematic since much of the most critical research on social media platforms and practices originates from disciplines like media studies, communication studies, and political communication whose researchers, at least historically, have had far less formal programming training than their colleagues in computer science.

Even where such hurdles in developing scripts for data access and processing can be overcome, data analysis and interpretation remain severely curtailed by the limited set of analytical tools available within clean room environments. It is usually possible to install standard data processing packages from the Python Package Index (PyPI) or Comprehensive R Archive

Network (CRAN). More specialised social media analytics tools (as developed by individual researchers for their own or the broader research community's use) are sometimes permitted, at the discretion of the clean room provider and subject to internal review. Researchers may also be able to copy and paste publicly-released data processing scripts – such as Giglietto's (2025) scripts for assessing the impact of Meta newsfeed policies on engagement with politicians' posts – into the clean room environment, but for more expansive and complex workflows this process quickly becomes inefficient and error-prone. Overall, the range of computational tools available to clean room users is highly restricted.

More generally, this emphasis on the computational processing of social media data privileges mostly quantitative analytical approaches despite the superior insights generated by mixed-methods approaches to social media data analysis. Clean room environments are not usually set up to support qualitative research components. Even small subsets of data usually cannot be exported, coded outside the clean room, and then reimported into the environment for subsequent analysis. They do not provide manual qualitative data coding tools and facilities such as the industry-standard NVivo or MaxQDA, or even basic spreadsheet editors. Additionally, the inability to access the open Web from within the clean room prevents now-emerging uses of commercial Large Language Models (LLMs) to assist qualitative data coding. In some clean room setups, it will be possible to install open-source LLMs, via package repositories or direct upload, but this too is limited by the hard drive and memory space allocations available within the virtual environment.

Finally, in addition to privileging quantitative over qualitative or mixed-methods approaches to data analysis and interpretation, the enclosure of research processes within exclusive clean room spaces hosted by platform providers also creates substantial hurdles to any cross-platform or comparative analysis, which is especially problematic as the communicative environments provided by social media platforms have diversified considerably, in part as a consequence of the decline of Twitter (Bruns, 2026). By their very design, clean rooms exist in isolation; data from one clean room environment cannot be transferred to and combined with data from another clean room. And even where data from one platform were retrieved via conventional APIs or other means, the upload of such datasets to another platform's clean room environment may be explicitly prohibited by the API's terms of service. One solution, then, is for researchers to conduct their analyses in parallel across different clean rooms and other research environments, and to export and combine only the final aggregate outcomes for comparison and correlation. On top of multiplying the workload needed to carry out such parallelised research, any resulting cross-platform analysis will inevitably be restricted to very high-level, abstract observations.

Conclusion: Researching the Walled Garden, from within the Walled Garden

While the Meta Content Library is perhaps the most significant clean room for social media data access at present, this model is not limited to Meta alone; it is not even limited to social media data access. Notably, the TDM Studio service offered by media content database ProQuest embraces a similar clean room philosophy to enable its users to computationally process large datasets of news media articles and similar content without exporting the data. Here, the motivation for the clean room model appears to be a concern about large-scale copyright violations, rather than about privacy (since ProQuest provides only professionally produced, published articles, licenced from commercial and public-service publishers). Here, too, however, clean rooms privilege some users and uses over others, in ways that shape, channel, and constrain the research that can be done with this platform.

We raise these issues not to dismiss the clean room concept altogether: the ethical, moral, legal, privacy, and copyright concerns that underpin the platforms' justifications for these approaches should be, and are, taken seriously by researchers as they access, process, analyse, interpret, and publish results from their data.

Finally, the increased complexity of the clean room model also further widens the gap between those research teams who are able to navigate the cumbersome accreditation, access, (self-) training, data access, analysis, and results extraction and publication process and those who are not. This gap exists between interdisciplinary research teams at large, well-resourced universities and isolated, individual researchers at smaller institutions, but more systematically also between universities in the Global North and those in the Majority World. In concert with the broader platform data availability and fidelity issues that still persist for non-Global North, non-English-language contexts, this will continue to channel scholarly attention to the relatively well-covered WEIRD (Western, Educated, Industrialised, Rich, Democratic) and English-speaking nations, while contemporaneous developments in the Majority World that deserve at least as much attention are addressed much more rarely.

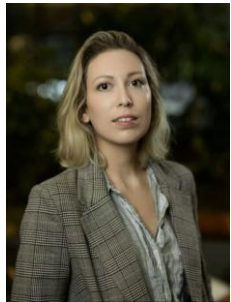
As the sole point of origin for authoritative data on what happens on their platforms, social media providers remain in a powerful position; while they are increasingly required by laws such as the EU's Digital Services Act to facilitate research data access, exactly how they do so still impacts substantially on whether such access is genuinely useful to researchers. Legislators and regulators must not only continue to pay close attention, therefore, to the base requirement that (some) access is provided, but also remain vigilant about whether and to what extent such access actually meets researchers' needs. That the European Union has already begun proceedings against Meta and TikTok for failing in these duties is encouraging (European Commission, 2025), but much more pressure must yet be brought on these and other platforms – not least also in jurisdictions beyond Europe.

References

- Bruns, A. (2018). Facebook Shuts the Gate after the Horse has Bolted, and Hurts Real Research in the Process. *Internet Policy Review*.
<https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786>
- Bruns, A. (2019). After the ‘APIcalypse’: Social Media Platforms and Their Fight against Critical Scholarly Research. *Information, Communication & Society*, 22(11), 1544–66.
<https://doi.org/10.1080/1369118X.2019.1637447>
- Bruns, A. (2026). The Death of Twitter and the Decline of Public Debate Online. *M/C Journal*, 29(2). <https://doi.org/10.5204/mcj.3247>
- Doctorow, Cory. *Enshittification: Why Everything Suddenly Got Worse and What to Do about It*. London: Verso, 2025.
- European Commission. (2025, 24 Oct.). Commission Preliminarily Finds TikTok and Meta in Breach of their Transparency Obligations under the Digital Services Act.
https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503
- Giglietto, F. (2025). “A Pretty Blunt Approach”: Meta’s Political Content Reduction Policy and Italian Parliamentarians’ Facebook Visibility. *SocArXiv*.
https://doi.org/10.31235/osf.io/8dqag_v2
- Giglietto, F., & Puschmann, C. (2026). From the Wild West to the Walled Garden: The Evolution of Twitter/X Data Access for Research. *M/C Journal*, 29(2).
<https://doi.org/10.5204/mcj.3257>
- Grevy Gotfredsen, S., & Dowling, K. (2024, 9 July). Meta Is Getting Rid of CrowdTangle—and Its Replacement Isn’t as Transparent or Accessible. *Columbia Journalism Review*.
https://www.cjr.org/tow_center/meta-is-getting-rid-of-crowdtangle.php
- Newton, C. (2016, 11 Nov.). Facebook Buys CrowdTangle, the Tool Publishers Use to Win the Internet. *The Verge*. <https://www.theverge.com/2016/11/11/13594338/facebook-acquires-crowdtangle>
- Puschmann, C. (2019). An End to the Wild West of Social Media Research: A Response to Axel Bruns. *Information, Communication & Society*, 22(11), 1582–9.
<https://www.tandfonline.com/doi/abs/10.1080/1369118X.2019.1646300>
- United Nations. (2020, 31 Mar.). UN Tackles ‘Infodemic’ of Misinformation and Cybercrime in COVID-19 Crisis. <https://www.un.org/en/un-coronavirus-communications-team/un-tackling-%E2%80%99infodemic%E2%80%99-misinformation-and-cybercrime-covid-19>
- Weller, K., Wang, Y., Peters, Y., & Gruber, J. B. (2026). Can 20 Years of Twitter Be Preserved? What Is Lost, What Remains, and What Is in Between. *M/C Journal*, 29(2).
<https://doi.org/10.5204/mcj.3253>



[Axel Bruns](#) is an Australian Laureate Fellow and Professor in the Digital Media Research Centre at Queensland University of Technology in Brisbane, Australia, and a Chief Investigator in the ARC Centre of Excellence for Automated Decision-Making and Society. His books include *Are Filter Bubbles Real?* (2019) and *Gatewatching and News Curation: Journalism, Social Media, and the Public Sphere* (2018), and the edited collections *Routledge Companion to Social Media and Politics* (2026 / 2016), *Digitizing Democracy* (2019), and *Twitter and Society* (2014). He served as President of the Association of Internet Researchers in 2017–19.



[Laura Vodden](#) is a data scientist and PhD candidate at the Digital Media Research Centre at Queensland University of Technology, with extensive experience working with social media APIs, as well as emerging cleanroom data environments such as the Meta Content Library and ProQuest TDM Studio. Laura's doctorate research develops methodological approaches to applying large language models to communications research, particularly for large-scale text and content analysis.