Peer Review Information: Nature Human Behaviour thanks Cornelius Puschmann, Amelia Acker, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Platform-controlled social media APIs threaten Open Science

Brittany I. Davidson[1], *Darja Wischerath*[1*], *Daniel Racek*[2*], *Douglas A. Parry*[3*], *Emily Godwin*[1], *Joanne Hinds*[1], *Dirk van der Linden*[4], *Jonathan F. Roscoe*[5], *Laura Ayravainen*[1], *& Alicia G. Cork*[6]

[1] School of Management, University of Bath, Bath, UK
[2] Department of Statistics, Ludwig Maximilians Universität München, Munich, Germany
[3] Department of Information Science, Stellenbosch University, Stellenbosch, South Africa
[4] Department of Computer and Information Sciences, Northumbria University, Newcastle, UK
[5]BT Applied Research, Ipswich, UK
[1] Department of Psychology, University of Bath, Bath, UK

Corresponding author: Brittany I. Davidson, bid23@bath.ac.uk
* These authors contributed equally

STANDFIRST

Social media data enable insights into human behavior. Researchers can access these data via platform-provided Application Programming Interfaces (APIs), but these come with

restrictive usage-terms that mean studies cannot be reproduced or replicated. Platform-owned APIs hinder access, transparency, and scientific knowledge.

Social media (SM) data hold tremendous value for studying behavioral patterns over time and across contexts at individual, group, and population levels[1,2]. For example, these data can be used to examine where conflict is likely to occur, where to allocate aid in the event of natural disasters, how online polarization or misinformation is impacting voting patterns. SM data are thus relevant to a broad range of disciplines in the social and behavioural sciences.

Because SM data are constantly changing as users interact and platforms alter the structure of feeds and interactions, it becomes ever more important for researchers to engage with Open Science (OS) practices to ensure that work is reproducible (reusing the same data and methods provides the same results), and replicable (using the same methods on different data produces comparable results). Reproducibility and replicability are essential to ensure knowledge produced about human behavior is robust, valid, and credible.

Open Science principles discourage academic misconduct (e.g., misreporting of data, problematic methods, improper documentation of results). While this is important for all scientific endeavors, is of importance for SM research due to the power imbalance between academic researchers and SM platforms. Moreover, given the relevance of SM data for a host of important domains (e.g., politics, mental health, misinformation), it is imperative that findings using SM data are trustworthy and can be relied upon to inform policy.

SM data can be gathered in a variety of ways (Fig. 1), with some facilitated by SM platforms and others falling outside of the methods officially mandated by the SM platform Terms of Use (henceforth: Terms). Because of the inconsistency in Terms between platforms and the changes that platforms make to their Terms, researchers face substantial ambiguity in how they can collect, store, and disseminate SM data. Further, many of these Terms restrict the extent to which SM data can be shared with other researchers, undermining research transparency and hindering the verification of prior results. Compounding this, recent platform changes have removed many widely used data-collection routes previously essential for SM research, inhibiting the replication of prior findings with new data. Here, drawing on our experiences working with SM data, we shed light on impediments to research transparency associated with platform-controlled access to SM data.

[insert Figure 1 here]

Data, APIs, and Terms in Flux

SM data and the Terms that govern researchers' access to this data are not static, with both users and platforms able to effect changes that alter the data available to be (re)collected from the platform. Users have the capability to remove or edit data, whilst platforms control many of the routes used by researchers to access data. Researchers are therefore at the mercy of any changes that platforms make to their data-access APIs and Terms governing this access.

Fundamentally, SM data consist of the digital traces that users provide via their engagement and interactions on the platform. Over time, this data shifts and erodes due to (a) the structural changes that platforms make to the interaction features available to users and (b) the ephemeral nature of the action-records as users alter or delete their content and profiles or change their privacy settings. Illustrating this, Pfeffer et al.[3] found that after one year, less than 70% of original tweets were still available, decreasing to ~54% after three years. This can impact some content more than others: political campaigns have extremely high proportions of tweet and user decay (missingness over time)[4], which has implications for reproducing results especially when data sharing is restricted (see below).

It is likely that data missingness will increase as platforms enact policies that call for the removal of inactive accounts. This will further reduce the extent to which prior findings can be reproduced using the original data collection procedures. For example, X (formerly Twitter), announced (via an Elon Musk tweet) that they will be '*purging accounts that have had no activity at all for several years,*'. Although it was commented that tweets would be archived, no further information was provided. Google similarly announced that it will start deleting Google (and associated YouTube) accounts that have been inactive for two years from December 2023. As these changes are likely to result in the removal of large swathes of SM data, they will have a substantial impact on the reproducibility of findings drawing on older datasets, thus impacting digital archiving and preservation[5].

Not only does SM data change and disappear but, troublingly, the data-access APIs themselves are constantly changed and updated. These changes are frequently undocumented and poorly communicated[6]. Changes to APIs can include the addition of new fields to gather data previously unavailable or, in some cases, the removal of existing fields or changes in functionality. Updates can also include changes in the way metrics are calculated. This means that even if researchers shared their code to query an API, someone re-running it to re-gather the data may find that the code does not reproduce the same results as those generated by the original researcher. For example, the Reddit API once provided the raw number of upvotes and downvotes per post/comment, however this functionality was later removed, with only aggregate scores remaining. While these scores are derived from up- and down-votes, the individual values are now unavailable through the API. This impacts any attempts to reproduce or replicate prior research, as the new 'score' metric is not transparently comparable with the 'upvote-downvote count' which, if available, could be useful to understand the popularity/virality of content, user behavioral patterns, or (mis/dis)information spread. Unfortunately, platforms typically do not document these changes. This highlights how crucial it is for APIs to have up-to-date and transparent changelogs with their documentation.

Alongside changes to the data and the APIs, researchers must also be aware, firstly, of the restrictions that the Terms imply and, secondly, of changes to the Terms by which data can be collected, stored, and processed[7]. Many SM providers state that it is the researcher's responsibility to keep up to date with the Terms (e.g., TikTok: '*However, it remains your sole responsibility to review these Research API Terms from time to time to view any such changes*'). Terms are not fixed for a given platform and often differ between platforms. For example, SM platforms tend to adopt different data ownership models (user-owned vs. platform-owned), which can impact the viability of data collection routes. For instance, Reddit deems all user-generated content as user-owned, meaning that data donation would not breach their Terms (in theory), whereas TikTok forbids any data collection outside of their API, which is presently only available in the US and Europe, eliminating any attempts at reproduction by researchers (or reviewers) outside of these regions.

Raw Data Sharing Restrictions

Sharing the original data underlying the findings reported in a study is critical for facilitating reproduction. Unfortunately, alongside the transparency-risks associated with evolving datasets, APIs, and Terms, many platforms restrict the extent to which researchers can share the raw data collected via their APIs[7]. These restrictions undermine the extent to which findings using SM data can be reproduced as researchers cannot rerun analyses on the original data. For example, since 2016, X has restricted the sharing of raw platform data collected via its API, with subsequent versions of the Terms indicating that researchers were only allowed to share Tweet and user IDs (unique identifiers allocated to each tweet/user). In mid-2023, these Terms were revised to allow the sharing of 50,000 raw tweets a day between two researchers with an upper limit of 1.5M tweets. X backtracked again in August 2023, and stated: '*Academic researchers are permitted to distribute an unlimited number of Tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research*'. While the lifting of this restriction, in theory, enables the sharing of SM data for research purposes (e.g., collaboration, verification, reproducibility), in practice it remains restrictive as researchers can only share Tweet/User IDs and not the raw data. To collaborate/verify/reproduce results, IDs need rehydration (recollecting raw data from IDs via the API). This brings challenges with dynamic data and relies on third-parties paying for API access[1]. Currently, X's API is both expensive and restrictive regarding data collection, sharing, and thus impeding replication attempts, especially with large datasets.

These Terms are therefore in direct conflict with reproducibility and replicability because (a) researchers cannot openly share raw datasets and (b) a complete rehydration is not possible due to data deletion from users, potential field changes and updates to the API. This has a disproportionate impact on the extent to which older datasets can be reproduced and highlights the direct impact of constant changes in API Terms on research[5]. See Table 1 for more examples.

Other Terms essentially restrict data sharing by virtue of the compliance processes that researchers must follow. For example, TikTok has restrictions in place on the use of their data specifically in relation to users who remove or change their content (e.g., account, posts, engagement) in extremely short timeframes. The Terms state: '*You agree to regularly refresh TikTok Research API Data at least every fifteen (15) days, and delete data that is not available from the TikTok Research API at the time of each refresh.*' This is problematic, as it can cause research results to become unstable wherein results would likely fluctuate with each data refresh. Further, researchers would need to perform substantial amounts of additional work (recollecting data every 15 days), which would be especially challenging when working with large datasets.

In addition to restrictions on the sharing of raw data, platform Terms can also impact the reuse of work using their platform data. For instance, TikTok states: '*After you publish any Research outputs, you agree that TikTok will have free and unlimited access to and use of your publication and Research outputs.'*, noting that researchers must send all research outputs to TikTok (August 2023). This is potentially problematic, as Terms like these could conflict with publisher agreements, alongside employer Terms relating to reuse of employer name and IP.

Final Thoughts

Platform-controlled APIs can threaten the reproducibility and replicability of SM research. The perspectives provided in this comment are grounded in our experiences using these APIs in our research and, while we have studied the Terms that bound this conduct, we do not and cannot provide legal advice. This is a complex area with dynamic policy and regulatory events, and specific legal counsel may be required to guide each study. Notably, at the time of writing, various regulatory bodies are considering whether and how to compel large online platforms to provide data access for research purposes[8]. While this has the potential to address some of the challenges for reproducibility and replicability, other challenges inherent to the data will remain, and the implementation of any policy reform will face substantial regulatory, institutional, platform, and infrastructural challenges. We also acknowledge that while we are encouraging transparency and data sharing, there are numerous ethical and privacy challenges that come with sharing SM data[1,5,9]. Making responsible decisions is a complex process that remains an open question within this landscape.

Whilst we have used a handful of SM platforms as illustrative examples, we have argued that, broadly, these data collection routes and the Terms that govern their use pose substantial restrictions that not only threaten the transparency of our research but, more fundamentally, risk restricting the advancement of our knowledge on human behaviour[10,11]. Specifically, we have highlighted challenges arising due to (a) the evolving nature of platform data, APIs, and Terms, and (b) the restrictions that platforms place on how data can be accessed, stored, processed, and shared. Alongside these elements, over the preceding years, a growing number of platforms have either removed their data-access APIs, restricted the nature and amount of data available through the API, or placed their APIs behind exorbitant paywalls. Despite the challenges, the constant changes to these APIs will continue to place restrictions on attempts to reproduce and replicate prior SM research and, in doing so, hinder scientific progress.

References

1. Walker, S., Mercea, D. & Bastos, M. *Information, Communication & Society* **22**, (2019).
2. Lazer, D. et al. *Nature* **595**, (2021).
3. Pfeffer, J. et al. *International AAAI Conference on Web and Social Media* **14**, (2023)
4. Bastos, M. *American Behavioral Scientist* **65**, (2021).
5. Acker, A. & Kreisberg, A. *Arch Sci* **20**, (2020).
6. John, N. A. & Nissenbaum, A. *The Information Society* **35**, (2019).
7. Freelon, D. *Political Communication* **35**, (2018).
8. Leerssen, P., Heldt, A. & Kettemann, M. C. *Digital Communication Research* **12**, (2023)
9. Sen, I., Flöck, F., Weller, K., Weiß, B. & Wagner, C. *Public Opinion Quarterly* **85**, (2021).
10. Bruns, A. *Information, Communication & Society* **22**, (2019).
11. De Vreese, C. & Tromble, R. *Political Communication* **40**, (2023)
12. Ohme, J. et al. *Communication Methods and Measures*, (2023)
13. Krotov, V. & Silva, L. *AMCIS* **17**, (2018).
14. Fiesler, C., Beard, N. & Keegan, B. C. *ICWSM* **14**, (2020).

Competing interests

Funding: This work was part-funded by the Engineering and Physical Sciences Research Council (EPSRC Grant Ref: EP/W522090/1) as a PhD studentship to Darja Wischerath as an EPSRC iCASE with Brittany Davidson and Jonathan Roscoe. This work also was part-funded by the Engineering and Physical Sciences Research Council (EPSRC Grant Ref: EP/SO22465/1) as a PhD studentship to Emily Godwin via the CDT in Cybersecurity TIPS-at-scale. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authorship Contributions

Conceptualization: BID
Writing Original Draft: BID, DW, DR, DAP
Writing Original Revision: BID, DR, DAP, JH
Writing Review & Editing: BID, DR, DAP, DW, EG, JR, JH, DL, AGC, LA
Writing Review & Editing Revision: BID, JH, DR, EG, DAP, AGC, DL, LA

Figure 1. Infographic of common routes for SM data access on LHS, RHS includes a description of each route and the types of data one typically obtains from each data access route.

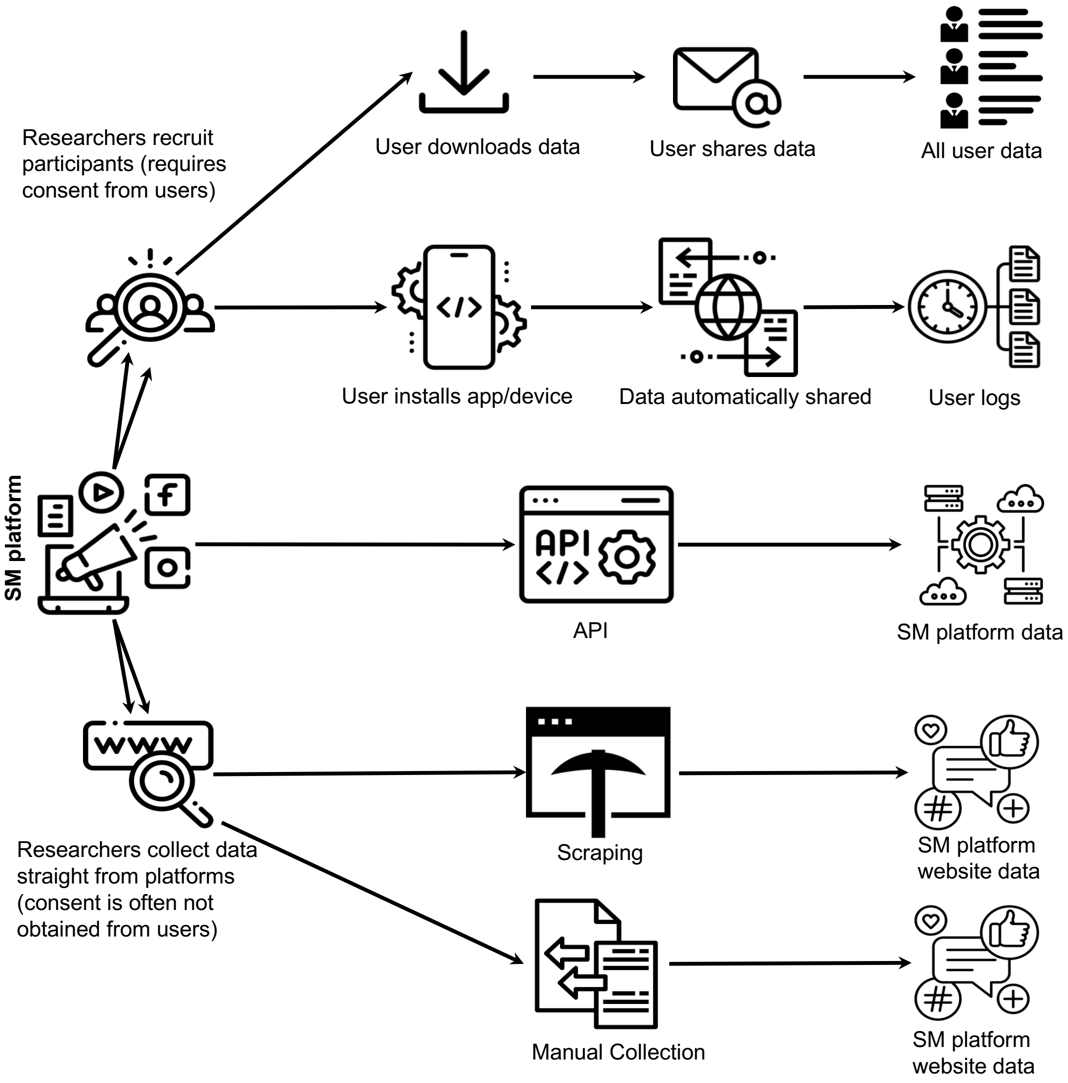
Table 1. Illustrative Examples of Terms that Impact Open Science

Selected Terms (August 2023)	**Reproducibility**	**Replicability**
Reddit Noting Reddit states user content is owned by users, [one cannot] 'u*se the Data APIs to encourage [...] violation of third party rights (including using User Content to train a machine learning or AI model without the express permission of rightsholders in the applicable User Content)*'[a]	If data cannot be shared, researchers cannot reproduce original results from articles as they cannot recollect the same dataset. If the researchers had used ML/AI and did not share their trained models, then reproducing (retraining the model) violates Terms and thus reproducing the work is not possible.	Datasets may be replicated provided the API provides the same fields and the ways in which metrics are calculated remain the same. If the analysis used ML/AI then these analyses cannot be replicated as this violates current Terms.
X (formerly Twitter) '*Never derive or infer, or store derived or inferred, information about a Twitter user's: Health (including pregnancy), Negative financial status or condition, Political affiliation or beliefs, Racial or ethnic origin, Religious or philosophical affiliation or beliefs, Sex life or sexual orientation, Trade union membership, Alleged or actual commission of a crime*.' –however at an aggregate level, this is acceptable.	These restrictions mean any prior papers looking at any of these areas at an individual level cannot be reproduced.	These restrictions mean any prior papers looking at any of these areas at an individual level cannot be replicated.
TikTok *TikTok Research API Data shall not be kept for longer than is necessary for Research approved as part of your application. You agree to provide TikTok with written certification of data deletion upon TikTok's request.*	This means that data cannot be shared so the exact analysis cannot be reproduced. It is also vague as to what 'longer than necessary' means (e.g., end of analysis, publication, end of grant?). This is likely at odds with many university or funder data retention policies, too.	NA

LinkedIn [b] [You agree not to…] ‘*Sell, rent, lease, disclose, distribute, share (with the exception of making the Content available to Users through the Application), transfer, sublicense, communicate, or otherwise make available, any Content, directly or indirectly, to any third party (e.g. you may not sell access to an aggregated collection of Member profiles, the most relevant Members for a position, or any social activity, such as posts, likes, or shares by Members)*’	This means that data cannot be shared so the exact analysis cannot be reproduced.	NA

[a] This Term is vague and places researchers in a difficult position of not knowing what they can and cannot do, especially when terms such as ‘ML’ and ‘AI’ are incredibly broad and work at different scales (e.g., training a task-specific decision tree versus training a general large language model (LLM)). Furthermore, the wording relating to ‘training’ is ambiguous, and raises the question of whether researchers are free to apply other pre-trained ML models to Reddit data. For example, can researchers use a pre-trained model on Reddit data but not use the same data for training or tuning? This causes other issues, for example, the use of models that are not adapted to a specific SM dataset may negatively impact the models’ accuracy and the inferences that can be drawn.

[b] Note: LinkedIn also has strict Terms regarding storing data, where no data are allowed to be stored, unless you have explicit consent from Members.



Type of Data Collection	Types of Data Collected
Data Donation: A user-centric approach, where users share their data with researchers. Users download their SM data from platforms and share with researchers[12] or install a smartphone app that logs data that is shared with researchers (e.g., UsageLogger). This can be legally complex depending on SM platform Terms.	Typically, all user data (content, posts, direct messages, etc.). Number of users will be limited to the ethics forms, how many sign up, etc.
Tracking: A user-centric approach similar to data donation. Tracking involves capturing logs of data typically from web browsers, apps, or devices via plugins or apps and is automatically transferred to a research server. Contrary to data donations there are no user interactions or requirements for users to actively share their data[12]. This can be legally complex depending on SM platform Terms.	Typically, user logs (interactions with SM apps, other apps, typically no content from SM platform used). Number of users will be limited to the ethics forms, how many sign up, etc.
Application Programming Interface (API): APIs are often the only official way to access data and are tied to SM platform Terms of Service.	A variety of data as permitted from the SM platforms (e.g., post content, likes, images, URL links, language). Number of posts/users will be limited by the SM platform itself.
Scraper/Spider/Crawler: A tool created to gather data on websites (or as explicitly programmed) for data analysis[13]. These do not use an API and are often legally complex[14].	Publicly available data can be collected (e.g., post content, images, number of likes). Number of posts/users is large (can be limited for technical reasons, e.g., rate limits)
Manual Collection: An approach where researchers manually collect data (e.g., copying and pasting data from SM feeds). This can be legally complex depending on SM platform Terms.	Publicly available data can be collected (e.g., post content, images, number of likes). Number of posts/users is often limited due to time required for collection.