Stanford | Internet Observatory
Cyber Policy Center

May 25, 2023

Director General Roberto Viola
Directorate-General for Communications Networks, Content and Technology
European Commission
1049 Bruxelles/Brussel, Belgium

Re: Delegated Regulation on data access provided for in the Digital Services Act

Roberto Viola and DG Connect,

Thank you for the opportunity to provide written evidence on facilitating data access for researchers under Article 40 of the Digital Services Act (DSA). The Stanford Internet Observatory is a cross-disciplinary program of research, teaching, and policy engagement for the study of abuse in current information technologies, with a focus on social media and a global scope.

The following comments provide key considerations for the European Commission from our team's experience with data sharing arrangements for social media research, including technical considerations for receiving and processing data. We also raise a common question for non-EU researchers about whether and how they can make use of the mechanisms that will become available for researcher access to social media data under the DSA.

Article 40 of the DSA can unlock a wide field of research, enabling researchers to pursue new studies and generate important findings that advance our understanding of the role social media plays in societal issues, including but not limited to, mental health, child safety, and responses to natural disasters. We commend the Commission for its important work and for this call for evidence on implementing researcher access to data.

Executive Summary

Based on extensive experience with online platform programs for research access, the Stanford Internet Observatory offers background and recommendations for generating impactful research and addressing common issues with data sharing arrangements. The primary challenges researchers will face with platform data sharing under Article 40 are understanding the data and information that can be requested and ensuring comprehensive data is received in accessible formats.

An independent intermediary body of experts is needed to review requests, facilitate data sharing, and provide clear guidance to designated industry and qualified researchers on how to request and facilitate data sharing.

Multiple mechanisms for data access will be needed to address different types of research questions. Standard methods should be developed for common types of access permissible under Article 40 with considerations for accessibility, efficiency, and the security and privacy of shared data and information. We suggest prioritizing historical and real-time APIs and access to historical data sets of public content. A virtual cleanroom can provide vetted researchers access to non-public data.

The Commission, Digital Services Coordinators, or a delegated authority should establish standard formatting and minimum requirements for what data is provided, including accompanying media where relevant. The Commission or an independent delegated authority should also audit company data sharing processes for quality assurance to ensure all relevant and requested data is provided to researchers in a secure, but accessible manner.

Finally, we encourage the development of guidance on access for non-EU residents who may have affiliations with qualified EU organizations. Global collaboration that adheres to strict privacy and data protection rules and expectations can unlock the black box of insights waiting to be tapped into from the designated online platforms and qualified field of researchers.

Introduction to Data Sharing Arrangements

Data sharing arrangements under Article 40 should draw upon past lessons from platform agreements with researchers, with an initial focus on developing clear expectations and guidance for qualified researchers and designated industry partners to apply for and facilitate access. Guidance on implementation should consider future innovations in the online social and search space, and allow for iteration once the program is launched.

Since our research program was established in 2019, the Stanford Internet Observatory (SIO) has had extensive engagement with social media companies' researcher access programs. Platforms that provide data do so through individually negotiated data use agreements with individual researchers or institutions, or through a more publicly accessible application programming interface (API). Similar to the provisions in Article 40, some industry pilots, such as the Twitter Moderation Research Consortium (TMRC), offered data access to defined data sets released on a regular basis.[1] Researchers applied for access to the TMRC by verifying their affiliation with an academic, journalistic, or nonprofit institution; demonstrating technical competency to ingest, process, and analyze the data; describing their non-commercial use case for the data; and attesting to their data security practices. Once admitted to the program, researchers signed a data use agreement with Twitter.

This model sought to address many potential concerns about vetting researchers and ensuring data privacy. One drawback to such a model is that it can limit data access to well-resourced research groups with the expertise and infrastructure to process complex data structures. It can also index too heavily on individual researchers over a research team. Strengths of such a model are that it reviews researchers for their qualifications and requires them to demonstrate their data security practices.

Researcher access programs that address these issues take several forms including: API access to historical or real-time public data, generation of data sets on specific topics, ad libraries or similar content databases, and dashboards with real-time public content and performance metrics. Access to public information is broadly available to users while researcher-specific APIs and data sets require additional vetting and agreements.

Research API Access

One of the most valuable research assets from social media platforms are APIs that provide real time or historic data on public posts.[2][3][4] APIs provide access to platform data — such as the text of posts, date and time of content creation, and details about user engagement — and enable researchers to query the platform content for all data relevant to a research question. Researchers can conduct regular structured queries of platform data to build out a data set that helps to answer their research question. APIs can be powerful tools for researchers to understand real-time information flows and prevailing online discourse.

[1] Roth, Yoel. 2022. "The Twitter Moderation Research Consortium is now open to researchers." Twitter Blog. https://blog.twitter.com/en_us/topics/company/2022/twitter-moderation-research-consortium-open-researchers.

[2] Twitter. n.d. "Twitter API for Academic Research." Twitter Developer Platform. https://developer.twitter.com/en/products/twitter-api/academic-research.

[3] Meta. n.d. "FORT Researcher API." Facebook Open Research and Transparency. Accessed May 17, 2023. https://fort.fb.com/researcher-apis.

[4] YouTube. n.d. "How it Works." YouTube Researcher Program. https://research.youtube/how-it-works/.

The APIs are often designed to provide data relevant to commercial interests, such as surfacing all mentions of a keyword or URL on the platform. They typically only return content present on a site and will not show results for content that has been removed by either the user or the platform. Until recently, Twitter offered approved researchers free access to their rate-limited API, a comprehensive API that provided robust data endpoints to researchers, and was second only in rate limits to their Enterprise API, a product they sell to advertisers and other customers.[5]

Topical Data Sets

For specific topical research areas, such as those in which there are privacy or security concerns, archived data sets or platform-generated reports may provide useful information for researchers that an API could not support. For example, platforms investigate and remove large networks of accounts that violate their coordinated inauthentic behavior policies. This removed content can be archived in a structured data set. Researchers can apply for access to this archive to conduct research on these networks after they have been identified and removed by the platform.

One example is TMRC, of which SIO was a global partner.[6] The TMRC was an industry-leading effort to release Twitter data to outsiders for analysis while protecting user privacy. The data sets were overwhelmingly state actor influence networks, consisting primarily of inauthentic accounts tied to governments and mercenaries. The TMRC enabled outside analysts to understand how these groups were evolving their tactics for manipulating public discourse all around the world.[7][8]

A few examples of SIO reports made possible by the TMRC include:

- My Heart Belongs to Kashmir (September 2022)
- Unheard Voice (August 2022)
- The New Copyright Trolls: How a Twitter Network Used Copyright Complaints to Harass Tanzanian Activists (December 2021)

Research Data Sharing Challenges and Limitations

Successful data sharing relationships can lead to impactful research and policy outcomes. However, bad experiences, sometimes involving incomplete data, are prevalent. These issues generally fall under two categories that should be addressed by the Commission: trust and transparency, and setting clear expectations and guidelines between qualified researchers and designated platforms.

[5] Twitter. n.d. "Twitter API for Academic Research." Twitter Developer Platform. https://developer.twitter.com/en/products/twitter-api/academic-research.

[6] Roth, Yoel. 2022. "The Twitter Moderation Research Consortium is now open to researchers." Twitter Blog. https://blog.twitter.com/en_us/topics/company/2022/twitter-moderation-research-consortium-open-researchers.

[7] Roth, Yoel, and Vijaya Gadde. 2021. "Expanding access beyond information operations." Twitter Blog. https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations-.

[8] Dang, Sheila. 2023. "Twitter research group stall complicates compliance with new EU law." Reuters. https://www.reuters.com/article/twitter-moderation-insight-idCAKBN2U61IU.

Trust and Transparency

The DSA will enable important public transparency and researcher access to assess online interactions and potential online harms, as provided within the scope of Article 40.[9] Until now, researchers have depended on the goodwill and whim of private companies to study spaces of online discourse used extensively by the public. In some cases, when platforms set the terms and conditions for researcher access, those terms have been rooted in commercial agreements that don't adhere to research best practices.[10] Calls for regulating researcher access to social media data have stemmed in part from past collaborations between academic researchers and platforms to facilitate data sharing. Under one of these programs, known as Social Science One, Facebook delivered a data set that failed to include half of users in the United States.[11] In other cases, Twitter being the most recent, API access has been deprecated or revoked at the discretion of the platform, with no other viable option presented for continued access to that data.[12]

Academic researchers depend on accurate and comprehensive facilitation of their data requests, but have little means to understand what data or information is available, or to check the validity of the data they receive to conduct this research.[13] For this reason, and due to the competing incentives[14] between platforms and researchers requesting data that can unearth potential harms, clear expectations for platforms and researchers and independent or governmental oversight will be important to ensure the validity of research requests and the data and information provided to conduct studies.

Measures should be taken to ensure data provided in response to researcher requests is complete, replicable, and auditable. If a platform provides a random sample of 10,000 data points to one researcher and 10,000 different points to another they would not be able to compare or verify their results, and they would depend on the platform's provided explanation for how that sample was chosen rather than being able to structure their own sample.

The broad scope of data and information available to researchers under Article 40 may be overwhelming at first to both researchers and platforms. This challenge has been referred to as an "unknown unknowns" problem in which researchers do not know what data or information platforms can make available in order to

[9] DiResta, Renée, Laura Edelson, Brendan Nyhan, and Ethan Zuckerman. 2022. "It's Time to Open the Black Box of Social Media." Scientific American, April 28, 2022. https://www.scientificamerican.com/article/its-time-to-open-the-black-box-of-social-media/.

[10] Bak-Coleman, Joe. 2023. "TikTok's API Guidelines Are a Minefield for Researchers." Tech Policy Press, February 22, 2023. https://techpolicy.press/tiktoks-api-guidelines-are-a-minefield-for-researchers/.

[11] Timberg, Craig. 2021. "Facebook admits it bungled data it shared with social scientists." The Washington Post, September 10, 2021. https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/.

[12] Mehta, Ivan. 2023, "Twitter's restrictive API may leave researchers out in the cold." TechCrunch, February 14, 2023. https://techcrunch.com/2023/02/14/twitters-restrictive-api-may-leave-researchers-out-in-the-cold/

[13] Lurie, Emma, Dan Bateyko, and Frances Schroeder. 2023. "TikTok just announced the data it's willing to share. What's missing?" Stanford Internet Observatory. https://cyber.fsi.stanford.edu/io/news/tiktok-just-announced-data-its-willing-share-whats-missing.

[14] Nonnecke, Brandie, and Camille Carlton. 2022. "EU and US legislation seek to open up digital platform data." Science 375, no. 6581 (February): 610-612. DOI: 10.1126/science.abl8537.

develop their research questions.[15] As New York University postdoctoral researcher Laura Edelson and colleagues outline, "Recital 96 lists certain types of data that can be accessed through the provisions of Article 40, namely: data on the accuracy, functioning and testing of algorithmic systems for content moderation, training data and, apparently, even the code of algorithms."[16] Designated platforms will also face challenges in determining the data and information that is available across their extensive systems. Further complicating the issue of cataloging, platforms often define the data and systems they use in company-specific terms. Understanding what is available, cataloging and developing methods for retrieval, and setting common definitions and formatting will require an extensive effort, which an independent intermediary body would be well positioned to address.

Clear Expectations and Guidelines

In considering how to facilitate vetted researcher requests for data, the Commission should develop a process that provides researchers with guidelines for developing research questions and submitting requests. Establishing that process will require collaboration with industry to outline which data and mechanisms are preferred and available.

Expectations and guidelines should be specific, and should cover a variety of data sharing mechanisms that meet researcher needs to access and analyze unique data types. The guidelines should be regularly updated to account for future technologies, affordances, tools, and data available on covered platforms and search engines.

Recommendations for Data Sharing Arrangements

Based on the above considerations, we recommend: 1) the development of an independent intermediary body to facilitate and provide guidance to industry and researchers, 2) multiple mechanisms for data access, 3) clear guidelines for data sharing formats and minimum requirements, and 4) a mechanism for auditing data sharing processes for quality assurance.

Independent intermediary body

We support the development of an independent intermediary body to develop clear guidelines for researchers and industry, and to prepare for the significant demand for data access in the research community. As recommended by an EDMO-led, multi-stakeholder working group, this body would evaluate research questions, provide guidance on expectations to designated platforms and vetted researchers, and facilitate data sharing between designated platforms and qualified researchers.[17] An independent body will also be important to assist Digital Service Coordinators (DSCs) with reviewing an expected high volume of

[15] Shapiro, Elizabeth H., Michael Sugarman, Fernando Bermejo, and Ethan Zuckerman. 2021. "New Approaches to Platform Data Research." NetGain Partnership. https://www.netgainpartnership.org/resources/2021/2/25/new-approaches-to-platform-data-research.

[16] Edelson, Laura, Inge Graef, and Filippo Lancieri. 2023. "Access to Data and Algorithms: For an Effective DMA and DSA Implementation." CERRE. https://cerre.eu/publications/access-to-data-and-algorithms-for-an-effective-dma-and-dsa-implementation/.

[17] "Launch of the EDMO Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms." 2023. EDMO. https://edmo.eu/2023/05/15/launch-of-the-edmo-working-group-for-the-creation-of-an-independent-intermediary-body-to-support-research-on-digital-platforms/.

requests due to significant interest among the academic community in accessing data to conduct studies under Article 40 of the DSA.

The independent body should consist of qualified researchers and technical advisors with expertise in social media research. Members of the intermediary body should provide “peer-review” evaluation of requests, soliciting input from academics and other experts in the field. The body should also be empowered to consider circumstances such as evolving research questions or additional data needs. Clear and timely communication between platforms and researchers will be important as the program launches with questions about what data is available, and to determine which studies qualify.

Multiple mechanisms for data access

Multiple mechanisms for data access are needed to share and study unique types of data or information about platform design and algorithmic systems, and to ensure sensitive data is protected. Different types of data lend to distinct delivery mechanisms. Standard methods can be developed for common types of data and information that should be made accessible under Article 40 with considerations for accessibility, efficiency, and the security and privacy of shared data and information.

Depending on the public data needed for a study, researchers should have access to archived databases, historical APIs, and streaming APIs with real-time data. A well-audited virtual cleanroom should facilitate access to non-public data for vetted researchers. This risk-based approach to mechanisms for data access will provide researchers the data they need while balancing control in how to parse data with security measures to protect sensitive data. Similar considerations can be made for potentially sensitive information about design and internal records that platforms may need to provide. This can be adapted for future forms of data and information that may become available.

Guidelines for standard formatting and minimum requirements for data provided

The Commission should establish standard formatting and minimum requirements for what data is provided, including accompanying media where relevant. The Commission should also task an independent intermediary body with developing expectations for designated platforms to catalog the types of data and information that are available for research requests with shared nomenclature and for how to provide access to those materials.

Quality assurance auditing

Independent auditing of company processes for quality assurance is essential to academic research that relies on the assumption that all relevant and requested data is received. The Commission and DSCs should conduct audits or delegate that authority to independent auditors, or as a function of the independent intermediary body. Monitoring and auditing will help to ensure data requests and data sharing mechanisms deliver the entire data sets they are expected to generate for researchers and that data and information sources that fall under the authority of Article 40 are cataloged.

Technical specifications for research sharing and secure data access

Technical specifications for access should be dependent on needs and the sensitivity of the data or information that is being shared. Non-public data should have strict controls, but enable researchers to use their own tooling to process and analyze data, as possible. Public data should use common formats, common forms of access — such as through API tooling — and include important post and user details, such as text, media, links, engagement and view counts, metadata, and the public data of the content creator.

For public data, such as that required under Article 40.12, the easiest method for researchers to consume is an API exposed via HTTP or a WebSocket that allows for fetching of JSON representations of that data, using a rich set of search operators. There are two modes of operation that could be considered: historical and real-time. Historical queries would allow searches back in time, and real-time would be a non-stop stream of events matching particular rules (such as Twitter's PowerTrack). One question to iron out is how to perform research on data that has been deleted in such a way that it complies with other EU regulations. By and large, Twitter's former API offerings are a good model to base future work off of, though other platforms more focused on multimedia content could offer additional content such as the identifiers of soundtracks to video content or transcription.

Studying non-public data, which may be necessitated under Article 40.4, requires significantly more controls which likely necessitate a different method of access and delivery. Ideally, this data would be delivered in such a format that researchers would be able to leverage their existing analysis tools, rather than need to use tooling internal to the company.

Access for Non-EU Residents

An open question for Article 40 of the Digital Services Act is access to data for non-EU persons who may have affiliations with qualified EU organizations. We ask that the Commission consider the benefits of research by a global community of scholars on important societal issues under Article 40 of the DSA, so long as the requirements set in Article 40.8 and other researcher qualification components of the DSA and EU regulations are met, and their associated organizations adhere with data privacy and security law and expectations.

Sincerely,

Jeffrey T. Hancock
Faculty Director
Stanford Internet Observatory*

Renée DiResta
Research Manager
Stanford Internet Observatory*

John Perrino
Policy Analyst
Stanford Internet Observatory*

Elena Cryst
Deputy Director
Stanford Internet Observatory*

Daniel Bateyko
Special Project Manager
Stanford Internet Observatory*

David Thiel
Chief Technologist
Stanford Internet Observatory*

** Affiliation for identification purposes only. The views expressed are those of the authors and not of Stanford University.*