CSMaP | CENTER FOR SOCIAL MEDIA AND POLITICS

Dear Mr. Roberto Viola,

Thank you for the opportunity to give input on the Delegated Regulation on access to non-public data provided for in the Digital Services Act (DSA). We are grateful for the European Commission's continued commitment to transparency and accountability by mandating that very large online platforms (VLOPs) and very large online search engines (VLOSEs) provide researchers access to data for scientific studies investigating systemic risks in the European Union.

NYU's Center for Social Media and Politics (CSMaP) is an academic institute that works to strengthen democracy by conducting rigorous research, advancing evidence-based public policy, and training the next generation of scholars. Social media is core to our work, as our research teams use data from platforms to study how the online information environment impacts politics, policy, and democracy. Since social media companies tightly control the data vital to study platform impacts, researchers' use of these rich digital datasets on users and platforms is scarce, especially as access to social media data continues to shrink. With limited data access, it becomes increasingly challenging for researchers to study the social media ecosystem and inform the mass public and policy community.

The DSA is critical to ensuring that VLOPs and VLOSEs share data for independent research. However, just as crucial are the data sharing structures and mechanisms to ensure that this rich digital data is standardized and made accessible to research teams. Drawing on our experiences working with large-scale datasets to conduct policy-relevant research, our comments on the Delegated Regulation focus on four areas concerned with these structures and mechanisms for data sharing: (1) streamlining the application and response communication process, (2) improving the documentation of Data Inventories, (3) clarifying components of the data access application, and (4) clarifying the data security and protection measures necessary to work with platform data. We close by looking ahead to the future as the digital environment evolves.

1. Streamlining the Application and Response Communication Process

The DSA's Delegated Acts have set the global standard for data access to protect public interest research and we commend the Commission on establishing a transparent application process that is accessible to global and interdisciplinary research teams. Recital 4 and Article 7 provide structured guidelines that clearly establish expectations for applicants. We know that research can be lengthy, with many applications and approvals throughout the process, and therefore, we appreciate that there is a centralized data access portal to standardize the process for all applicants. In addition, we appreciate that there is a published timeline for responses — five working days for an initial response and 21 working days for a full determination.

While this structure should streamline the application length, we are concerned that a short review period for complex applications may result in more rejections or mediations. Therefore, in cases where the applicant's request for data is exceptionally complex, we suggest allowing additional time (no more than 60 working days) to ensure a more comprehensive review. While this recommendation does lengthen the research timeline, we suggest that this only be applied on a limited case-by-case basis and that the rationale and length of the extended review time is still communicated within 21 working days.

2. Improving the Documentation of Data Inventories

To build ambitious research agendas, teams of researchers must know what data is available for each platform to formulate research questions and prepare data requests. However, the current system does not offer clear data inventories or guidelines to platforms, leaving researchers to prepare and submit requests that may or may not be aligned with the available data. Not only will this slow the research process, but it will also drain the resources of the Commission in reviewing and re-reviewing applications. A clear solution is well-documented data inventories.

Based on our experiences collecting and accessing datasets over the past decade, we are acutely aware of the importance of well-documented data inventories. Recital 6 requires VLOPs to provide an "inventory of their services easily accessible online including indications of the data and data structures." However, available data types, data structures, and data documentation differ across platforms. We are concerned that the current lack of clear, specific guidelines for the platform data inventories will lead to inconsistent stated inventories across VLOPs. When designing research applications, researchers may need to make *a priori* judgments about presumed available or unavailable data. This may unduly limit the scope of ambitious research agendas and render the requests for data access counterproductive — for researchers, platforms, and the Commission itself. As an important next step, we recommend that the Commission sets baseline expectations by providing guidelines for VLOPs to develop more transparent data inventories. Inventories should include defined data fields, data time ranges, metadata, and note any missingness from datasets.

Some of these guidelines may be straightforward, such as establishing uniform documentation structures and consistently defined data fields and variable names across platforms. A benefit of this standardization is that it may reduce the technical barriers to conducting cross-platform research by structuring data at the same unit of analysis. As we have seen more and more in 2024, the social media ecosystem has continued to rapidly fracture. Therefore, to enable cross-platform research, data inventories can be standardized by using uniform column headers where possible (such as standard variable names for data types) and defining measures (such as a view, impression, or engagement). This prioritization of designing data inventories to ensure direct comparability can aid researchers in studying issues of key concern and systemic risk, such as the direction and spread of misinformation and the influence of foreign interference in democratic elections.

Other guidelines may be more difficult, such as accounting for content excluded from a social media dataset, generally known as moderated content. This content can appear in many forms, including posts moderated by the platform or posts deleted by a user. This data — both the moderated content itself and the associated metadata, such as the time flagging and removal from the platform — is valuable to understanding the prevalence of harmful or toxic text and images on a platform and to evaluating the efficacy of content moderation policies. However, we are concerned that this moderated content will be difficult to obtain depending on a platform's data storage policies. For example, it is not evident to researchers if deleted content is retained in perpetuity or stored in data warehouse logs. This means that researchers will not know what data is available when designing research projects and applying for data access.

In addition, while DSA requires VLOPs to publicly justify content moderation actions in a "statement of reasons," Recital 12 suggests that "data related to content moderation and governance" may include reasonable non-public data that researchers can request. This leaves open the question of whether there is additional non-public data on content moderation policies and enforcement decisions that researchers might need to request. One such example includes Meta's now known "break the glass" measures implemented during and after the U.S. 2020 elections, the specifics of which were not publicly communicated at the time. These sorts of emergency platform actions aimed at slowing the spread of viral content and increasing moderation during critical moments like elections are often not made public and provide valuable insight into platform data and policies, not only in the U.S. but globally. Clarifying what supplementary non-public data may exist, and establishing a baseline data inventory for VLOPs for reporting moderated content, is a productive next step in ensuring that researchers are sufficiently aware of what non-public data can reasonably be requested.

Finally, it is critical that these guidelines for data inventories are flexible for the future. For example, future data fields might account for AI-generated and labeled content. Even the past year has shown considerable shifts in the digital information environment and we can expect these shifts to continue into the future. Recital 6 also ought to include expectations for periodical updates by VLOPs and for occasional opportunities for the Digital Services Coordinator to amend data inventory guidelines.

3. Clarifying Components of the Data Access Application

The aforementioned data inventories are a key component in informing researcher applications. Recital 12, and Article 40.4 altogether, pertain to non-public data that must “speak to the necessity and proportionality of the data request” and demonstrate that the data is not “available through other sources including public data access.” This recital leaves open questions for potential research applicants, such as, what is meant by “proportionality” and “necessity”?

Given the inherent unknowns associated with requesting non-public data, clarifying the expectations of “proportional” requests is crucial. The Recital should outline how the DSCs measure “proportionality” by offering benchmarks or a baseline metric for sampling. When providing such metrics to researchers, DSCs should be in consultation with our community to determine appropriate sample sizes for specific types of research, such as projects exploring user interactions on a specific platform versus studying the spread of a specific instance of hate speech across multiple platforms. Currently, it is unclear whether proportionality means the scope of the dataset is explicitly tied to the research objectives outlined in a data application or other factors, such as the sensitivity of the data request. Therefore, providing a benchmark or a baseline metric for particular types of studies would help researchers determine what volumes of data are appropriate for their specific research project. Additionally, we recommend researchers be afforded flexibility in the amount of data they can request as their studies progress, especially when initial analyses are drawn and more data is required for further exploration or a new direction on a project is taken.

Better documented data inventories, as outlined in Recital 6, will be a valuable resource and a critical first step for researchers preparing applications. These inventories would direct researchers to data points that are private and allow us to request access to these in our proposals. We also recommend that the Commission provides additional clarity on what is considered “necessary.” At present, we understand that this modifier refers to the study of systemic risks, which must be explained and justified in the application itself. Overall, the lack of clarity and expectations for components of the data access application will make it challenging for researchers to be precise about the data they hope to access — both because the full scope of what data is available is unknown and how it must be justified is unclear.

Given the highly specialized knowledge needed to anticipate what is included in a dataset and in what format the data is structured, we further suggest that the Commission or a supporting body provide a consulting mechanism to assist researchers with understanding data availability and their applications before submission. While we are aware of additional responsibilities this may impose, a limited consultation or “office hours” for applicants may ultimately improve the quality and success of the data request process.

4. Clarifying the Data Security and Protection Measures Necessary to Work with Platform Data

We applaud the Commission for including Recital 13 since researchers must uphold confidentiality, security, and data protection when analyzing user data. While the Recital includes sufficient examples of safeguards, such as data access agreements, non-disclosure agreements, and organizational measures, the Commission can strengthen the law's usability for researchers by providing additional definitional clarity and guidance from the DSC on risk assessments and the proposed safeguarding measures.

The Recital refers to "appropriate access modalities" — a method of sharing data with end users — as one safeguard form, but it does not clarify what qualifies as "appropriate." A widely used modality CSMaP uses when working with shared data is virtual clean room environments, which provide highly controlled settings for analyzing data while minimizing security and privacy risks for sensitive datasets. Already, clean room software offers unique security features, including built-in de-identification capabilities for sensitive data, encryption protocols, secure file transfer mechanisms, and more. One example of what a secure computing environment for accessing sensitive data could look like is the Facebook Open Research and Transparency (FORT) Researcher Platform, which provides a secure way for qualified users to access privacy-protected Facebook and Instagram data. Another example is the United States Federal Statistical Research Data Centers (FSRDCs). As an added benefit, these environments frequently comply with data security safeguards required by university Institutional Review Boards, allowing researchers to work with data in an environment that meets multiple data security compliance mechanisms. However, the current Recital does not make it clear which security measures are sufficient under what circumstances and may benefit from standardizing what an "appropriate" clean room environment or other modalities are. By clarifying the Recital's language, the Commission will help establish a greater understanding of the privacy and security requirements in Article 40.4, reducing any additional burden on the researchers as they prepare their data access applications.

Lastly, Recital 13 should detail how DSCs will evaluate the adequacy of researchers' proposed safeguards relative to the risks identified in a requested dataset. We recommend that the Commission supply the criteria or benchmarks the DSCs will use to assess the outlined legal, organizational, and technical safeguards researchers are implementing to protect users' data. These criteria can also ensure a shared understanding among member states of the DSC role, ensuring consistency in their risk assessments, and helping researchers better understand compliance expectations. Such measures should also help make the application process more efficient for both the Commissioners and researchers.

5. Looking Ahead

While our comments focus on recommendations for Article 40.4, it is also critical to look to the future. The DSA is the world's first major transparency legislation aiming to hold online platforms and other intermediary services accountable. To ensure the law's long-term success, the Act must adapt as the online information ecosystem continues to fragment.

While Article 40.4 requires Very Large Online Platforms and Very Large Online Search Engines to provide access to their data for public interest research needs, the online information ecosystem is increasingly fragmenting. The social media landscape is no longer dominated by a few large, legacy platforms like X/Twitter, Facebook, and Instagram. Newer, broadcast-style entertainment platforms, such as Twitch, private messaging apps like WhatsApp, and other niche platforms, such as Gab and Rumble, have risen in their use over the last few years. Some of these platforms, perhaps most notably Telegram, are not considered VLOPs at the moment. This shift challenges the transparency efforts of the DSA since activity is increasingly occurring outside of the VLOPs and VLOSEs' thresholds. As the online information environment becomes more fragmented, the sources of potential systemic risk will likely no longer be limited to VLOPs and VLOSEs.

To address this incoming issue, future versions of the DSA should consider extending data access provisions to these emerging platforms. At CSMaP we are already building a new research infrastructure to explore these new online environments, including projects collecting data to understand how politics is talked about on WhatsApp and smaller niche platforms, such as BlueSky, Gab, Better, and Rumble. Additionally, we are also one of the founding consortium members of the Accelerator infrastructure project, which can serve as a model for creating policy-relevant research by building shared infrastructure to support data collection, analysis, and tool development to better understand today's information environment. Furthermore, we recommend the Commission consider conducting periodic reviews of the DSA to reassess whether new criteria should be added alongside the current VLOP or VLOSE threshold, ensuring the legislation reflects the emerging platforms in the digital ecosystem.

The DSA is a strong step in the right direction to creating a safer and more transparent online information environment. However, future efforts must ensure the law adapts to the ever-changing information landscape by incorporating our forward-looking measures. Taken together with our Recital recommendations, the DSA will continue to support researchers' commitment to studying online harms for years to come.

Sincerely,

Sarah Graham
Research Manager
NYU's Center for Social Media and Politics

Solomon Messing
Research Associate Professor
NYU's Center for Social Media and Politics

Lama Mohammed
Tech Policy Fellow
NYU's Center for Social Media and Politics

Jonathan Nagler
Co-Director
NYU's Center for Social Media and Politics

Joshua A. Tucker
Co-Director
NYU's Center for Social Media and Politics

Zeve Sanderson
Executive Director
NYU's Center for Social Media and Politics