CSMaP | CENTER FOR SOCIAL MEDIA AND POLITICS

May 23, 2023

Margrethe Vestager
Executive Vice-President
A Europe Fit for the Digital Age
European Commission
Rue de la Loi / Westraat 200
1049 Brussels, Belgium

Thierry Breton
Commissioner
Internal Market
European Commission
Rue de la Loi / Westraat 200
1049 Brussels, Belgium

Dear Ms. Vestager and Mr. Breton:

Thank you for the opportunity to provide input on the Delegated Regulation on data access provided for in the Digital Services Act (DSA).

As you know, social media and digital technology have transformed society and presented urgent challenges to democratic governance. We are grateful for the European Commission's leadership, demonstrated through the DSA, in responding to this new technological environment.

We lead NYU's Center for Social Media and Politics, 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 data is the foundation of our work. Our team of academic researchers use these data to study how the ever-shifting online environment impacts politics, policy, and democracy.

But from the beginning, because social media companies tightly control the data necessary to study the platforms' impact, we've been conducting our research with one hand tied behind our backs.

The DSA's data access provisions have the potential to revolutionize research in this field and create an international standard for other nations to follow, which is why we're grateful for the opportunity to weigh in on this important topic.

Our comments will focus on three broad areas: (1) the importance of independent research, (2) the best data access mechanisms for research, and (3) what research we can conduct with robust access to data.

1. The Importance of Independent Research

Public, independent research is critical to understanding the risks posed by our new online information environment. These risks range from the spread of mis/disinformation to the impact of hate speech and illegal content to the influence of foreign actors in democratic elections. Currently, most of the data necessary to study this topic is controlled by platforms themselves, and they have little incentive to share it with researchers like us.

Requiring very large online platforms (VLOPs) to share data for independent research is critical. Just as critical is setting up mechanisms to ensure that research remains independent.

Under Article 40, Section 4, researchers must submit a research proposal identifying the type data they need and for what purpose. The Commission must ensure the process for reviewing those proposals remains independent and free of platform influence. If not, platforms could easily reject proposals for fear the results will paint them in a bad light, potentially undermining urgent topics of inquiry that have societal and scholarly value but that are most damaging to platform interests.

While there will of course be issues regarding the feasibility of providing different types of data on which the platforms will need to be consulted, it is crucial that the *questions to be answered*, and the privacy concerns surrounding how to answer those questions, should be assessed by an independent body, using a peer review mechanism of outside experts. Peer review serves as a vital mechanism for ensuring independence in academic research by providing an impartial evaluation of the work, identifying conflicts of interest, mitigating biases, promoting transparency and replicability, and protecting against fraudulent or unethical research practices. It helps maintain the integrity and reliability of the research process, which is crucial for the advancement of knowledge.

The Commission should use those same principles when determining the application and review process for research proposals under Article 40.

2. The Best Data Access Mechanisms for Research

Researchers' ability to study a secure and trustworthy online sphere requires rich digital data about user and platform data. Although the following is by no means an exhaustive list, in general the data we need for research falls into a few broad categories:

- *Exposure data:* What posts appear in people's timelines when they're actually using platforms.
- *Engagement data:* What posts people have clicked, shared, liked, etc.
- *Recommendation data:* What content the platforms recommend that people view and which accounts, pages, and/or groups they recommend users follow.

- *Network data:* Friend/follower relationships.
- *Content moderation data:* A catalog of actions taken against a post or account (if it's flagged for review or removed, for example).

What's more, there should be standard protocols in place for platforms to disclose the full scope of data available, including internal product experimentation and "AB testing," ad model input data, content moderation-relevant data, and other relevant data that are not public-facing.

a. Uniform Structures, Flexibility for the Future

The current data environment is incredibly fractured. For example, few platforms have publicly available application program interfaces (APIs), which provide the type of detailed information researchers need for large-scale analyses. Historically, Twitter has had the most publicly accessible API, which led to a flowering of academic research using Twitter data. Recently, however, under Elon Musk's leadership, Twitter began charging exorbitant fees for API access, essentially shuttering most public interest research about the platform. YouTube has an API, but its limits make it difficult for use in research. And while TikTok is reportedly working on a researcher API, it is unclear how robust it will be.

Critically, none of the information shared in these APIs is uniform. Data content and structure differ greatly by platform, which means researchers must often undertake intense post-collection data processing in order to make any cross-platform comparisons. Some of these differences are easy to mitigate, e.g. using MM/DD/YYYY formats versus DD/MM/YYYY. Some are much more difficult, e.g. when one platform tracks impressions and another platform doesn't, or when platforms count impressions (or other important metrics) in different ways.

To address this problem, the DSA should require VLOPs to create uniform protocols, formats, and data models for sharing data for research, while maintaining flexibility for the future.

For example, when sharing data, column headers (i.e., variable names for different types of data) should be standardized. There should also be guidance provided to platforms related to defining measures — what's a view, what's an impression, what's engagement, etc.

It's also critical for these mechanisms to be flexible and adaptable for years to come. The social media platforms of today may not be the social media platforms of five years from now. The type of social media content could also shift. After all, while just a few years ago social media was primarily composed of text and image content, video now dominates. Flexibility includes creating a process to update these structures as needed in the future. The same independent body charged with reviewing research proposals could also help create, monitor, and update these standards.

Next, we will discuss examples of mechanisms to share both public data (under 40.12) and non-public data (under 40.4).

b. Example of Mechanisms for Public Data

Article 40, Section 12 says VLOPs shall give access to "real-time data, provided that the data is publicly accessible in their online interface by researchers." As discussed above, the best method for doing that is through a research API.

The requirements for each API will necessarily vary by platform — the information available for text- and image-based platforms like Facebook, Twitter, and Instagram will be different from video-based platforms like YouTube and TikTok, for example.

For public data, the best approach is for APIs to provide all information; however, if it's determined that providing data at this scale would be unfeasible, platforms should provide a large, random sample of content. Twitter, for example, had this until recently with its "Decahose API," which enabled researchers to access a 10 percent random sample of tweets every day. If there are certain types of data where a random sample is not possible, VLOPs should make it very clear what's available and how.

Another method researchers use for gathering public data is through scraping. By "scraping," we mean the process of loading publicly available web pages on one's own computer and then retaining the information contained in that webpage as it loads. Many social media sites do not have APIs, but they do have publicly available data. Researchers have designed programs to scrape and analyze this data on sites such as NextDoor, BitChute, and TikTok. But scraping often technically violates platforms' terms of service, putting researchers at risk.

While APIs are much preferred, there will still be instances where data scraping makes more sense for research purposes. Therefore, platforms should set up a way for researchers to safely and legally scrape social media data for research. In return, platforms should also get legal protection if these data are used in ways that violate privacy or for non-research purposes, including sale to a third-party (i.e. Facebook's Cambridge Analytica scandal). Further, providing legal protections for scraping can be a large impact policy change that ensures at least some access to data as larger, more robust methods of ensuring data access (including but not limited to APIs) are being developed.

c. Example of Mechanisms for Non-Public Data

Researching the impact of social media on society also often requires access to non-public data. For example, much of the work organizing anti-vaccine movements during the pandemic happened within private Facebook groups.

Article 40, Section 4 outlines the procedure for vetted researchers to gain access to non-public data for research. There are three ways the DSA could require VLOPs to do that.

1. *Restricted Access APIs:* This would be the same set up as a public API, except with non-public data. Streaming all data through an API, and making some public and some private, will better facilitate the research process. Researchers could then know which data points are private and request access to those data points in their proposals.

2. *Researcher Sandboxes:* Researcher sandboxes allow platforms to keep control of their data, while providing researchers access to search and analyze the data. An example of this is Facebook Open Research & Transparency (FORT), which "facilitates data sharing and the publication of independent research about Facebook's role in society, with the right privacy protections in place." Under FORT, researchers are given access to a sandbox environment, where they can search, filter, and conduct analysis. Researchers can export results of their findings, but not export any of the actual data. This tool can serve as a model for sandboxed environments that can provide multi-platform data for researchers.

3. *Easier Data Donations:* All VLOPs should be required to have an easy way for users to download their data and donate it for scientific research. A core focus of our research at CSMaP is to pair offline political opinions and behavior to online activity. We conduct surveys asking respondents their views on certain political topics, how they voted, etc. We then ask them to donate their social media and other online data. This combination of surveys and digital trace data allows us to draw connections between what they see online and what they do offline, and vice versa. For example, our Bilingual Election Monitor project uses this method to explore the attitudes of Spanish-speaking social media users in the U.S. Currently, downloading data is a cumbersome process. In the past, Facebook had a process to allow users to quickly give researchers access to their data. We were able to use data donated in this way to show that older people were more likely to share low quality news on Facebook. The DSA should make it easier for people to do this by requiring platforms to create a mechanism similar to Facebook's prior system — using the same uniform standards outlined above. In addition to aiding scientific research, it will also empower users to have more ownership over their digital data and decide whether it should be used for the public good.

3. What Research Can We Conduct With This Data?

Under the current data regime, researchers have been able to study several critical issues regarding the risks involving social media's impact on society, politics, and democracy. For example, research has provided insights into the recommendations of algorithmic systems, the patterns and effects of foreign influence, the relationship between social media and political behavior and beliefs, the prevalence of hate speech and harassment, and the efficacy of interventions.

Better access to data would allow researchers to probe even deeper. Specifically, increasing and standardizing data access will enhance research capabilities in four areas.

1. *Multimodal Research:* While the social media ecosystem was once dominated by a few large text-based platforms, it has transformed and fractured into a diverse ecosystem of multimodal platforms. Indeed, one of the key methods for identifying information quality online is still using text-based methods, such as fact-checks and domain-level ratings. However, in popular video dominant platforms, such as TikTok and YouTube, the size and complexity of these video data make collection and classification difficult relative to text data. Recent research has made advancements in video classification, but significant further work is needed to understand the impact of video content. Requiring VLOPs to share publicly available data in a uniform manner will greatly enhance our ability to analyze various forms of content.

2. *Multilingual Research:* While social media usage is quite high around the world, research has tended to focus on North America, and much of the work in global contexts still uses English language data. Given the linguistic diversity of the European Union, multilingual research is critical to detect, identify, and understand potential risks. The DSA's data access provisions can address this gap and provide researchers the information we need to study this landscape.

3. *Cross-platform research:* Put simply, people exist in multi-platform environments, and yet cross-platform research remains scarce compared to within-platform research. There are key technical challenges to solve, such as identifying users across platforms, collecting data at the same unit of analysis, and processing data to have standardized structures or metadata. As the social media ecosystem grows and fractures, developing methods for comparative and systemic cross-platform research will be especially important to gain a comprehensive picture of the public's information environment. Creating robust data sharing mechanisms for VLOPs would be an important first step in this direction.

4. *Generative AI:* The development of foundational models — which enable the generation of text, image, and video at scale — will undoubtedly impact the networked information ecosystem. These developments have the potential to facilitate the wide scale production of spam, harassment, and false information, much of which will be both cheaper to produce and more difficult to detect. There are already reports of AI-generated websites pretending to be reputable news sites. In the case of foreign influence campaigns, a key method for identifying accounts was the use of stock photos, grammatical errors in posts, and other behaviors that suggested automation. With the advancement of these foundational models, the outputs are more likely to appear human-like, introducing new challenges for both the public and researchers alike. Gaining access to platform data under the DSA will help us better track and understand the impact of the burgeoning AI landscape on the information ecosystem.

Sincerely,

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

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

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

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