/ Response to the European Commission's call for evidence for a Delegated Regulation on data access provided for in the Digital Services Act

The DSA must ensure public data for public interest research

May 2023

Our online environment today suffers from enormous information asymmetry: online platforms assemble information about us, while we know little about them. And while they share data with commercial third parties, the researchers who would hold them accountable to society or monitor societal concerns have had limited data access at best, and at worst, have faced technical barriers and legal threats. The voluntary data sharing between platforms and the research community has been fraught and fragile; research projects and essential watchdog efforts can crumble at a company's whim.

Historically, the terms of public data sharing have been set by the companies, not by their users or the public interest research community. Much of the data that is made publicly available by platforms is designed to serve advertisers and marketers, not to help the research community or to allow for scrutiny of systemic risks. On April 26th, the Mozilla Foundation and a group of civil society experts gathered to discuss the implementation of the Digital Services Act's public data sharing scheme in practice. This cohort combined technical expertise and years of experience in platform monitoring, data protection and human rights.

Together we put forward the following recommendations to the European Commission and to the designated platforms directly as they move to implement Article 40, paragraph 12, of the Digital Services Act.

1) Public data should be complete, comprehensive, and include historical data

- Even the best practices in data sharing are extremely limited, for instance omitting key metrics related to platform functionalities (like with Facebook's "Reels") and mitigations (like labels or fact-checks applied to content), or skewing towards certain languages or countries. Often, shared public data proves incomplete or inaccurate (ie: scattershot or poorly labelled ad archives).

- Publicly accessible data should be understood to include metadata and data that could have been captured historically over time. The term "publicly accessible", like the term "manifestly made public" suggests information that is accessible to any member of the public, without being required to create an account on the service to access the information. For the public interest purposes of Article 40.12, this term should include any information that would be accessible to a user of the platform with an account on the service, since these data are essential for monitoring and comparison.

- Regulators and researchers need to know what is in fact "publicly accessible in their online interface." What is publicly accessible in an online interface changes constantly. Keeping track of these changes as an external observer is impossible to do. Platforms should therefore be required to share a taxonomy of their publicly accessible data with researchers. Preferably they would use a common method for changes like a "changelog" that would also support reproducibility by creating documentation of what has changed and when.

2) Data must be usable, accessible and verifiable, for which multiple access methods are needed

Article

- Useability is a critical factor. Platforms must deliver data in a way that has real-world impact, meets researchers' needs, and fulfils the spirit of the data access and scrutiny obligation.

- API (Application Programming Interface)[1] access is a minimum viable method for permissioned access to public data. APIs can allow for cross-platform research and the creation of interfaces and tools suited to research projects.

- That said, multiple methods of access are needed. API access or any other single mode of permitted access should not preclude the use of other research methods necessary to ensure the integrity of the public data that platforms share formally, for instance through automated data collection (scraping), data donation or the use of unofficial APIs.

- Platforms should also provide visual interfaces themselves to facilitate research and cross-platform analysis and to empower a more general public interest research audience. Interfaces simplify access for research with similar purposes but where specific domain knowledge is needed (e.g. for election monitoring, an understanding of the local legal and political context is needed.) To be useful interface access must be timely, the data verifiable, and the interfaces well maintained.

[1] Notably, several of the designated VLOPs and VLOSEs do not have any form of public API.

3) Permissioned access must come on fair and reasonable terms

- Permissioned access should be free or at a nominal cost. Higher costs risk discouraging the use of the DSA's data access provisions and perpetuating inequity among less well-resourced research organisations. Both the DSA and the Code of Practice on Disinformation are explicit that platforms should not prevent or discourage good-faith research.[2]
- Restrictions and mitigation measures to address privacy concerns, for instance related to commingling of data (e.g. combining crowdsourced data with API data) or to sharing with third parties (like research partners), should be modelled on the GDPR.
- Researcher access requests under Article 40.12 should be approved on a researcher or research organisation basis and not project by project. Access to continuous, real-time data is necessary for exploratory research, and repeated vetting should not be applied to particular project-by-project research questions.
- A standardised data access request process should be considered for all VLOP/SEs.
- Approved researchers should receive sustained access appropriate to their research. For instance, an organisation monitoring a platform's protection of election integrity will need sustained access for the period necessary to assess the platform's integrity efforts. Once approved, they should retain access for a minimum of three years, and there should be a streamlined, expedited renewal process.
- Provision of access must be timely and transparently communicated, as any delay could compromise the research.
- The process for approving access requests must be transparent and with the possibility of appeal by researchers to an independent third party, such as the independent advisory body foreseen by the DSA Article 40.13.

4) Platforms must not hinder independent, public interest research

- Many platforms actively hinder research through their terms of service, through technical measures, or through intimidation and threats of legal action. In particular, platform efforts to prevent scraping have had a chilling effect on the researchers trying to hold them accountable. The irony of this is that many companies ignore the same data-gathering techniques when they are used by marketers or "social listening" tools that advance their bottom line. The DSA Article 40.12 should be understood as a safe harbour for research addressing systemic risks and should at a minimum require non-interference into public interest, GDPR-compliant research efforts in line with the rest of the provision.

[2] Code of Practice Measure 28.3.; DSA recital 98 "where data is publicly accessible, such providers should not prevent researchers meeting an appropriate subset of criteria from using this data for research purposes that contribute to the detection, identification and understanding of systemic risks"

5) Data sharing should include a diversity of researchers

- Research related to systemic risks in the European Union is conducted by researchers physically located within and outside of the EU. Currently, it is unclear whether researchers located outside of the EU will have access to data needed to conduct this research. This should be clarified to ensure they have access. The current lack of clarity could also make commonplace international research collaboration difficult by limiting the potential research partners of EU-based researchers.
- Access to public data should also be possible for journalists, who have historically played a role in holding these very companies to account for these very concerns and made impressive use of platforms' publically accessible data as part of their watchdog function.

In drafting these recommendations we understand well the responsibility accompanying privileged data access, in particular in relation to security and privacy. The past years have seen a flourishing of researcher efforts to align on best practices and to improve risk mitigation strategies. While working with this data will never be zero risk, the implementation of the Digital Services Act provides a much-needed, structured avenue to further align on best practices as a public interest research community.

Undersigned:

AlgorithmWatch
Mozilla Foundation
Amnesty International
AMO Association for International Affairs
Unfollow Everything
Institute for Strategic Dialogue (ISD)
Democracy Reporting International (DRI)
AI Forensics
Check First oy
Stiftung Neue Verantwortung (SNV)
Avaaz
MEMO 98
The Forum on Information and Democracy
Algorithmic Transparency Institute, National Conference on Citizenship
The Coalition for Independent Technology Research
The Institute for Data, Democracy & Politics, George Washington University Brandon Silverman, Former CEO & Co-Founder of CrowdTangle