# Researcher Access to Platform Data: European Developments

Mathias Vermeulen

**Abstract.** This commentary examines the scope of the data-sharing regime under the Digital Services Act, as well as which researchers will be able to access data under the framework, and using what process. It then evaluates the guidance and commitments contained in the European Union's Code of Practice on Disinformation and the European Digital Media Observatory's Code of Conduct, including how these instruments relate to one another and operate within the broader regime.

The European Union (EU) is setting up a ground-breaking new regime that will fundamentally change platforms' incentives to allow researchers to access platform data, thereby creating new avenues for third parties to scrutinize the work of Trust and Safety professionals. This regime consists of three important initiatives.

First, the **Digital Services Act** (DSA) requires Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSE) to provide data to researchers on request from a regulator.[1] The act stipulates conditions for a regulator to vet researchers and research proposals, and proposes different access regimes for sensitive and non-sensitive data. In a significant departure from current practice, companies will no longer be the final arbiter in deciding which entities obtain access to data, and do not have the authority to decide for which research purposes data can be shared. A regulator can use this data to assess the company's compliance with obligations under the DSA.

Second, a number of companies (Google Search, YouTube, Twitter, Microsoft Bing, Linkedin, Meta, Instagram, and TikTok) have promised to make data available to enable research on disinformation under the **EU's Code of Practice on Disinformation**.[2] In theory this is a voluntary commitment, but in practice the EU will consider adherence to these commitments when assessing a company's compliance with the DSA's obligations, specifically on risk mitigation measures.

A third initiative, led by the **European Digital Media Observatory (EDMO)**, developed a draft Code of Conduct under the EU's General Data Protection Regulation (GDPR), which specifies how platform-to-researcher data access might be achieved in compliance

1. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
2. https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation



doi:10.54501/jots.v1i4.84

with Europe's most stringent privacy regime.[3] The report of the EDMO Working Group on Platfrom-to-Researcher Data Access develops the contours of an independent, third-party intermediary body that could vet researchers and research proposals, and evaluate the codebooks and datasets made available by platforms. The same companies that have committed to making data available to researchers under the Code of Practice on Disinformation have also agreed to support the set-up of such a body.[4]

This commentary examine these three initiatives, first examining the scope of the DSA's data-sharing regime and who will be able to access data under the framework, and using what process. It then evaluates the guidance and commitments contained in the EU's Code of Practice on Disinformation and the EDMO Code of Conduct, including how these instruments relate to one another and operate within the broader regime. Together, these initiatives explore some of the questions Professor Persily raised in the inaugural issue of the *Journal of Online Trust and Safety*:[5] to which companies should such a regulatory regime apply? Who should have access? To what data should they have access? And how should such access be regulated to protect both user privacy and research integrity?

## 1 EU-Mandated Data-sharing Regime in the Digital Services Act

The DSA imposes a new set of due diligence obligations on VLOPs and VLOSEs, covering companies that have more than 45 million monthly active users in the EU, regardless of their size or turnover. Article 34 requires these companies to identify, analyze, and assess any systemic risks associated with the design, functioning, or use of their services. It lists four broad categories of risk:

- The extent to which a company disseminates content, as defined by the laws of EU member states
- Actual or foreseeable negative effects on a range of fundamental rights, including freedom of expression, the right to privacy, the prohibition of discrimination, and the rights of the child
- Actual or foreseeable negative effects on civic discourse, electoral processes, public security, gender-based violence, public health, or minors
- Serious negative consequences for users' physical and mental well-being

DSA Article 35 stipulates that a company must take appropriate measures to mitigate these risks. The act recognizes that the effectiveness of measures will vary by company, and suggests a small number of potential measures that companies could implement such as changing content moderation policies, recommender systems, or ad delivery systems. Risk assessments and risk mitigation measures are both subject to independent audits.

The DSA's access to data regime for vetted researchers is a crucial component of its broader transparency and accountability measures. Under Article 40, a VLOP or VLOSE must give vetted researchers access to data upon the request of a regulator for the

3. https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf
4. https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation
5. https://tsjournal.org/index.php/jots/article/view/22/11

sole purpose of conducting research on the "detection, identification and understanding of systemic risks in the EU" (as set out in Article 34), and to the "assessment of the adequacy, efficiency and impacts of the risk mitigation measures" pursuant to Article 35.[6]

**In the DSA context, researchers can be considered both pathfinders and quasi-auditors**: they can spot new and emerging risks that may not have been covered by a company's risk assessment report, and assess whether self-regulatory initiatives to mitigate specific risks have been effective in practice. As such, their research will make a crucial contribution to independent auditors and the European Commission as they assess a company's adherence to its DSA obligations. Valid research questions could include:

- To what extent are YouTube's recommendation algorithms amplifying COVID-19 misinformation?
- Does the use of Instagram by teenage girls lead to depressive symptoms, social anxiety, or body image concerns?
- Is TikTok disproportionately removing content from Black creators or LGBTIQ content?
- Is LinkedIn's system of selecting and presenting job ads to people discriminatory in nature?
- Is Meta's content moderation system robust enough to prevent exposure to hate speech in Dutch or Bulgarian?

**Who will obtain access to data under the DSA framework, and how will this procedure work in practice**? Details on the procedures for vetting researchers and providing access to data will be specified in a 'delegated act'—a secondary piece of EU legislation that will be adopted in the next 18 months. However, the DSA lists a number of conditions that researchers need to fulfil in order to be vetted, and outlines a five-step process for requesting platform data.

**Step One**: Researchers file an application with the Digital Services Coordinator of Establishment, which is the regulator of the country in which a company has its headquarters in the EU.[7] The application must demonstrate that the researchers are affiliated with a research organization as defined in the EU's copyright legislation,[8] which is broader than a university and includes non-academic research institutes and civil society organizations that conduct "scientific research with the primary goal of supporting their public interest mission."[9] This definition would likely extend to consortia of researchers that include non-EU based researchers and journalists, as long as a European researcher is the main applicant. US or UK researchers who are visiting researchers at an EU-based university would be able to apply for access. Researchers would further need to disclose the funding source of the research and demonstrate they are independent of commercial interests.

The DSA's vetting procedure includes requirements specific to each data request. Researchers will need to:

- Describe the appropriate technical and organizational measures that will preserve data security and confidentiality requirements

6. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
7. Ibid.
8. https://eur-lex.europa.eu/eli/dir/2019/790/oj
9. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf

- Justify why the data are necessary for their research purpose, and how the research would contribute to understanding either the systemic risks in the EU (as defined by Article 34) or the adequacy, efficiency, and impacts of the risk mitigation measures (pursuant to Article 35)
- Commit to sharing their research results publicly and free of charge
- Suggest their ideal data access format (e.g. API, online database, json file), and by which date they would like to have the data

The DSA does not specify whether the researcher or organization needs to be vetted before a data request can be submitted, or whether both processes can take place simultaneously. The delegated act will likely offer more information on this.

**Step Two**: The Digital Services Coordinator of Establishment approves or rejects the application. Given that most VLOPs and VLOSEs are based in Ireland, an Irish regulator will likely be the ultimate arbiter of a vetting procedure. While researchers can simultaneously apply to their national regulator—which can give an opinion to the Irish Digital Services Coordinator—the Irish regulator will make the final decision. The DSA also specifies a procedure through which a researcher can lose their vetted status.

The DSA seems to acknowledge the challenges facing regulators in individual European Member States, which may lack the necessary context, skills, and knowledge to assess a variety of data access requests and research designs and methodologies, especially when researchers request access to sensitive datasets as protected by the GDPR. The act refers to a potential role of an "independent advisory mechanism" that can assist a regulator in vetting researchers and research proposals.[10] A recent EDMO report (see Section 3) provides guidance on such an independent mechanism, but the delegated act could shed more light on its mandate and functions.

**Steps Three and Four**: The Digital Service Coordinator submits a specific data request from a vetted researcher to the VLOP, which then—in **a fourth step**—has 15 days to respond to the regulator's request. The VLOP can then provide the data as requested or seek an amendment to the initial data access request if (1) it does not have access to the requested data or (2) providing access to the data will lead to significant vulnerabilities for the security of its service or the protection of confidential information, particularly trade secrets. The latter was the subject of substantial discussions during legislative negotiations. Members of the European Parliament were afraid it would offer companies a blanket excuse with which to refuse data requests. The DSA's final text states that a company's consideration regarding commercial interests "should not lead to a refusal to provide access to data necessary for the specific research objective."[11]

If a VLOP requests an amendment of the initial data access request, it will need to specify alternative means through which the data can be provided or suggest "other data which are appropriate and sufficient for the purpose of the request."[12] Trust and safety professionals will have an important role to play in this process.

**Step Five**: If the VLOP requests an amendment, the Digital Service Coordinator will confirm or decline the request within 15 days. Again, this raises questions about the extent to which the Digital Service Coordinator will be able to appropriately assess the legitimacy of the requests, and whether it may require external assistance.

10. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
11. Ibid.
12. Ibid.

**An additional mechanism, which is separate from the five-step process described above, is the DSA's access regime for "publicly accessible data,"** including real-time data, that contributes to the detection, identification, and understanding of systemic risks in the EU pursuant to Article 34. The DSA clarifies that this data refers to aggregated interactions from public pages, public groups, or public figures, including impression and engagement data such as the number of reactions, shares, and comments from recipients of the service. Informally known as the 'CrowdTangle provision,' companies are expected to give vetted researchers access to this type of data "without undue delay." It is unclear how this procedure will function. The delegated act may provide further detail.

## 2 Semi-voluntary data sharing via the Code of Practice on Disinformation

The DSA framework requires companies to share data *reactively* (i.e., in response to a request from a regulator). However, a number of companies—including Google Search, YouTube, Twitter, TikTok, Microsoft Bing, Linkedin, Meta, and Instagram—have committed to *proactively* share data with researchers through the EU's Code of Practice on Disinformation.[13] Similar to the DSA, the Code of Practice accepts researchers from civil society organizations "whose primary goal is to conduct scientific research on a not-for-profit basis, pursuant to a public interest mission recognised by a Member State."[14]

The Code of Practice, while voluntary, is connected to the DSA; Article 45 of the DSA states that "adherence to—and compliance with—a given code of conduct may be considered as an appropriate risk mitigating measure."[15] **By proactively giving researchers access to data, a company can signal to a regulator that it approaches its due diligence obligations under DSA Article 34 seriously, thereby decreasing the risk of retaliatory action by the regulator. Yet such a measure does not provide a company with a free pass.** The DSA makes clear that participating in and implementing this code "should not in itself presume compliance."[16]

The relationship between the DSA and voluntary codes of conduct is deliberate, and was partly created to address the perceived failures of an earlier Code of Practice on Disinformation, which was purely self-regulatory.[17] Through the DSA, the EU seeks to incentivize platforms to share data on disinformation by linking adherence to a Code of Conduct or Code of Practice to a company's obligations.

Like the DSA, the Code of Practice distinguishes between access to public data and access to datasets that "require further scrutiny."[18] For the former, relevant signatories of the Code commit to "continuous, real-time or near real-time, searchable, stable access to non-personal data and anonymised, aggregated, or manifestly-made public data for research purposes on Disinformation through automated means such as APIs or other open and accessible technical solutions allowing the analysis of said data."[19] Signatories commit to provide public access to such information, including engagement

13. https://digital-strategy.ec.europa.eu/en/library/signatories-2022-strengthened-code-practice-disinformation
14. https://ec.europa.eu/newsroom/dae/redirection/document/87585
15. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
16. Ibid.
17. https://digital-strategy.ec.europa.eu/en/library/2018-code-practice-disinformation
18. https://ec.europa.eu/newsroom/dae/redirection/document/87585
19. Ibid.

and impressions (views) of content hosted by a company, with reasonable safeguards to address the risks of abuse (e.g., API policies prohibiting malicious or commercial uses). The code separates this type of public access from "real-time or near real-time, machine-readable access to non-personal data and anonymised, aggregated or manifestly made public data on their service for research purposes, such as accounts belonging to public figures such as elected official, news outlets and government accounts subject to an application process which is not overly cumbersome."[20] The Code borrowed the term "manifestly made public data" from the GDPR. It is currently unclear how the relevant signatories will interpret this notion in relation to their service, or what a "not overly cumbersome" application process could look like, although it would likely be impractical for companies and researchers to employ a procedure that is very different from the DSA's 'CrowdTangle' provision.

Companies were not ready to commit to anything in this code beyond what the DSA and its delegated act require. **Importantly, the relevant signatories did commit to "developing, funding, and cooperating with an independent, third-party body that can vet researchers and research proposals"—which could be seen as the same "independent advisory mechanism" mentioned in the DSA,[21] and elaborated in the EDMO report** (see Section 3). The code states explicitly that they would take into account "ongoing efforts" such as the EDMO proposal for a Code of Conduct on Access to Platform Data and commit to co-fund the development of an independent third-party body from 2022 onwards.[22]

Finally, companies committed to support good faith research on disinformation that involves their services "and will not take adversarial action against researcher users or accounts that undertake or participate in good-faith research into Disinformation."[23]

Despite its crucial link to the DSA, the ultimate purpose and finality of the research and the access to data regime in this context is different from that of the DSA. Under the Code of Practice of Disinformation, access can be granted for any research purpose on "disinformation," and is not limited to assessing platforms' roles as they address risks or take appropriate risk mitigation measures.[24]

## 3 Privacy-compliant data sharing with researchers

Platforms have at times invoked Europe's GDPR as a key obstacle preventing them from sharing data with independent researchers.[25] This argument rests on the assumption that the GDPR does not specify whether or how companies might share data, and, given the potentially significant penalties for violating the regulation, the platforms have argued that a conservative, risk-averse approach is warranted.

To address this challenge, in May 2021 the EDMO established a working group comprising representatives from academia, platforms, and civil society to draft a Code of Conduct under Article 40 of the GDPR.[26] Its draft Code of Conduct published in May 2022 clarifies how platforms can provide data—and what steps researchers must take

20. https://ec.europa.eu/newsroom/dae/redirection/document/87585
21. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
22. https://ec.europa.eu/newsroom/dae/redirection/document/87585
23. Ibid.
24. Ibid.
25. https://knightcolumbia.org/content/the-keys-to-the-kingdom
26. https://edmo.eu/2021/08/30/launch-of-the-edmo-working-group-on-access-to-platform-data/

to protect it—under the GDPR.[27] **Though specific to the GDPR, the guidance laid out in this Code is applicable much more broadly: it provides practical approaches for platforms and researchers who seek to design and implement data access regimes around the world.**

The EDMO draft code consists of two broad sections. Part I provides legal guidance, specifying which GDPR requirements apply to data access and research processes and clarifying how platforms and researchers should implement those requirements. It elaborates on the GDPR's research exemptions; specifies the legal roles, responsibilities, and liabilities for both platforms and researchers; clarifies the security safeguards required from a data security perspective; and clarifies a company's transparency obligations vis-à-vis its platform users. Its key message is clear: platforms can share personal data, even sensitive personal data, with researchers in a GDPR-compliant way by implementing specific safeguards. It is also important to note that **from a GDPR perspective, it does not matter who performs the research using platform data. The main consideration is that a researcher/organization is equipped to properly protect the data it receives and processes**.

***Part II* includes operational guidance and standards for (a) evaluating the level of risk associated with accessing and analyzing specific data and (b) implementing appropriate technical and organizational safeguards based on the risk level.** This section offers a novel risk assessment framework that researchers can use to evaluate the level of risk involved in accessing and analyzing the data necessary for a specific research project. This framework is based on two considerations:[28]

1. Data subjects' reasonable expectations in relation to the "data processing" activity, including how private they might reasonably expect the data to remain, given the circumstances of its generation.
2. The data processing activity's potential impact on data subjects' rights and freedoms, including if the data or research outputs are misused.

In the EDMO framework, where data can be shared on the basis of a contract between a research institute and a platform, these attributes are mapped along a continuum from low to high risk, and the two dimensions are combined to form a risk framework with four quadrants. Depending on the outcome of this risk assessment, the EDMO Code highlights required and recommended technical and organizational safeguards that researchers and platforms must implement before sharing data. It requires researchers to develop plans and protocols for data storage (including specified retention periods and criteria), destruction (e.g., when unneeded or at the end of the data storage period), security, and access.[29]

Under the current status quo, platforms retain some responsibility to vet researchers and research proposals. For example, they are expected to assess the appropriateness of the safeguards proposed by a researcher before entering into a data-sharing agreement with their institution. Since this undermines research independence, the EDMO Working Group unanimously agreed that **an independent intermediary body should be created to certify that research proposals and proposed data safeguards comply with the EDMO Code of Conduct.** Streamlining these review and certification processes and housing them in an independent intermediary body would reduce the burdens placed on smaller, under-resourced universities and research institutions, thereby offering data access to a much more diverse pool of researchers. Moreover, an

27. https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf
28. Ibid.
29. Ibid.

independent intermediary could simultaneously review and certify that the platforms' datasets, codebooks, and technical systems adhere to the EDMO Code requirements. Given the DSA's explicit reference to such an intermediary body, and companies' commitment in the Code of Practice on Disinformation—under the principle of 'the polluter pays'—to co-fund its creation, such a body is very likely to be set up in the next 18 months.

## 4 Conclusion

The EU has laid the foundations for an ambitious new data access regime for researchers that consists of a mandated data-sharing regime, as established by the EU's Digital Services Act,[30] and a semi-voluntary data-sharing regime, as established by the EU's Code of Conduct on Disinformation.[31] The work of the EDMO Working Group provides a path forward for both regimes to operate in a privacy-compliant way.[32] Important elements of these three documents still need to be worked out in practice over the next 18 months, but together they provide a potential blueprint for other countries that seek to design and implement data access regimes.

## Authors

**Dr. Mathias Vermeulen** is public policy director at AWO, and an affiliated researcher at the Centre for Law, Science, Technology and Society at the Vrije Universiteit Brussel. mathias@awo.agency

## Acknowledgements

Many thanks to Rebekah Tromble, Aparna Surendra, and Louis Dejeu-Castang for their input in writing this commentary.

In the past 5 years I have received funding from Reset, Mozilla Foundation, the Knight First Amendment Institute at Columbia University, and George Washington University to study this topic. I have also provided expert advice on these three initiatives to all relevant regulators and legislators in the EU, including the European Commission, members of the European Parliament, and European governments. AWO drafted the EDMO Code on the basis of the EDMO Working Group's guidance and provided legal advice on applying the GDPR to the EDMO Working Group.

## Keywords

Access to data; Digital Services Act; GDPR; European Union; transparency.

30. https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/IMCO/DV/2022/09-12/p3-2020_0361COR01_EN.pdf
31. https://ec.europa.eu/newsroom/dae/redirection/document/87585
32. https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf