weizenbaum institut

ens European New School of Digital Studies

MAY 2023

Ulrike Klinger, Jakob Ohme

Delegated Regulation on Data Access Provided for the Digital Services Act

Response to the Call for Evidence DG CNECT-CNECT F2 by the European Commission

Imprint

AUTHORS:

Ulrike Klinger \ Europa-Universität Viadrina \ klinger@europa-uni.de

Jakob Ohme \ Weizenbaum Institute for the Networked Society \ jakob.ohme@weizenbaum-institut.de

PUBLISHER:

Weizenbaum Institute for the Networked Society –
The German Internet Institute
Hardenbergstraße 32 \ 10623 Berlin \ Tel.: +49 30 700141-001
info@weizenbaum-institut.de \ www.weizenbaum-institut.de

DOI: 10.34669/WI.WPP/7

ISSN: 2940-8490

Berlin, May 2023

For inquiries regarding this paper please contact Robert Peter at: robert.peter@weizenbaum-institut.de

The Weizenbaum-Institut e.V. adheres to the Code of Conduct of the EU Transparency Register, register ID 194885445254-71

This work has been funded by the Federal Ministry of Education and Research of Germany (BMBF) (grant no.: 16DII121, 16DII122, 16DII123, 16DII124, 16DII125, 16DII126, 16DII127, 16DII128 – “Deutsches Internet-Institut”).

About the Weizenbaum Institute

The Weizenbaum Institute – The German Internet Institute analyzes, evaluates and shapes relevant aspects of the digitally networked society. With basic, interdisciplinary and problem-oriented research projects on the ethical, legal, economic, political and social dimensions of digital change and the exploration of concrete solutions, it provides politicians, business and civil society with evidence- and value-based options for action in order to shape digitalization in a sustainable, self-determined and responsible manner. The Institute is supported by a research network from Berlin and Brandenburg, which includes Freie Universität Berlin, Humboldt-Universität zu Berlin, Technische Universität Berlin, Berlin University of the Arts, and the University of Potsdam, as well as the Fraunhofer Institute for Open Communication Systems (FOKUS) and the WZB Berlin Social Science Center. The Weizenbaum Institute is funded by the German Federal Ministry of Education and Research (BMBF) and the State of Berlin. For more information, visit www.weizenbaum-institut.de.

About the authors

Ulrike Klinger is Professor for Digital Democracy at the European New School of Digital Studies, European University Viadrina, Frankfurt (Oder).
Contact: klinger@europa-uni.de

Jakob Ohme is Research Group Lead of the “Digital News Dynamics” group at the Weizenbaum Institute for the Networked Society, Berlin.
Contact: jakob.ohme@weizenbaum-institute.de

I. Summary

1. *Data access needs*: We identify three types of data access needs to study and mitigate systemic risks: 1) non-private communication data (e.g., user connections and interaction, distributed content data, content exposure data, and content engagement data); 2) user account metadata (e.g. profile information, group memberships, and 3) data governance documentation (i.e., platform governance decision that can affect user-to-user or user-to-content connections). Researchers must be enabled to conduct research on publicly communicating actors (e.g., users, political parties, NGO's), distributed content (e.g., emotionality, hate speech), targeted and reached audiences (e.g., exposure, diffusion patterns), platforms themselves (e.g., platform-use patterns, media diets, affordances), and media effects (e.g., willingness to political participation, depression, anxiety), particularly in times when these communication acts are essential for democratic processes (e.g., opinion formation during electoral campaigns).

2. *Data access application and procedure:* Data access requests should be handled by DSCs via online forms that adapt to the type of request and data and should follow the logic: the more sensitive and encompassing data requests are, the more information is necessary. DSCs should make decisions about substantial quality (i.e., the study of systemic risk and mitigation measures, see Article 34(1) and Article 35), ethical and technical conditions by themselves. A peer-review process (i.e., by objective referees such as uninvolved scholars from the same discipline) is deemed impractical. DSCs should develop cross-national assessments that are guided by data access formats and data sensitivity. We suggest a tiered process dependent on sensitivity of the requested data and access method. Ideally, DSCs develop a general assessment and list of data types to be made available by platforms and a risk assessment on each data type before first access requests are filed. Participating researchers should sign data use agreements that ensure how they uphold high standards of data security and use data they have obtained only for the purpose and duration specified in the description of their research project. Data protection measures used in the vetting process can also include the request of data handling plans and consent forms for data-handling agreements, signed off by the institutional

review boards of a researcher's institution. The added value of an independent advisory mechanism (IAM) would be to provide professional guidance, oversight, mediation, and scrutiny of the data access request process. Specifically, the independent advisory mechanism should be executed by an intermediary body governed by academic and civil society researchers (more detailed information is provided in the attached full response).

3. *Data access formats and involvement of researchers:* To facilitate use in academic research, access to VLOP and VLOSE data should be technically tiered in a measured fashion that balances legitimate research interests and ease of use with aspects of data security and privacy concerns. Two chief modes of access that have in the past proven their worth are web application programming interfaces specifically for research (research-oriented APIs, or R-APIs, such as Twitter's soon-to-be-defunct Academic Track or CrowdTangle) and virtual lab environments (or VLEs, e.g., Meta's Facebook Open Research & Transparency, or FORT, based on the popular JupyterLab). Note that we consider both approaches to be equally important (more detailed information is provided in the attached full response).

4. *Access to publicly available data:* The implementation of Article 40(12) should be prioritized as this will cover a great number of research questions scholars are working on, while the resources and developments necessary are minor. It will also prevent a gap in current platform research, as the implementation and first successful access requests under Article 40(4) will take more time and create a time gap where platform accountability is not yet in place.

II. Full Response

1. Data access needs:

a) What types of data, metadata, data governance documentation and other information about data and how it is used can be useful to DSC's for the purpose of monitoring and *assessing compliance and for vetted researchers for conducting research related to systemic risks and mitigation measures?*

The first data access need is public communication data. This means 1) user connections and interaction (e.g., follower/friendship connections, likes, replies between users); 2) content data (e.g., text of posts and comments, images, URLs, hashtags, image descriptions), 3) content exposure (e.g., individual feed-content and sorting), and 4) content engagement (number of views, likes, and comments). The second data access needs regards user account metadata. This means self-reported public info, group memberships (e.g., communities/sub-fora), activity logs (e.g., receiving of paid advertisement). Both types of data should be time stamped and available retrospectively as well as continuously. The third access need is data governance documentation. This means the disclosure of strategic changes in algorithmic selection mechanisms, AB testing and any platform governance decision that can affect user-to-user or user-to-content connections.

Researchers might conduct research on communicating actors (e.g., users, political parties, NGO’s), distributed content (e.g., emotionality, hate speech), targeted and reached audiences (e.g., exposure, diffusion patterns), platforms themselves (e.g., platform-use patterns, media diets, affordances), and media effects (e.g., willingness to political participation, depression, anxiety).

Systemic risks that can be studied with this type of data are: (1) long-term observational studies on various risk factors, like prevalence of misinformation, degree of social group formation/radicalization, prevalence of hate speech, diversity of news exposure, all-over long-time scales to be able to see trends (due to a lack of offline baseline). (2) Studies on A/B test or natural experiments of new feature roll-outs (e.g., new sorting algorithm) and their (causal) influence on the above described risk factors, such as sharing of misinformation, group formation and radicalization, narrowing of individual news-feed diversity etc. (3) Comparative studies between different samples (e.g., different countries or other groups) for assessing heterogeneities in the outcomes and risk factors, including survey research on political attitudes, participation etc., linking to social-media behaviour and exposure. (4) Analysis of political elites, media, and foreign influence operations (including geolocation data, focusing on their organic reach) that aim at influencing the above risk factors.

2. Data access application and procedure:

a) Digital Services Coordinators (DSCs) in the Member States will play a key role in assessing researchers' applications and they will act as intermediaries with the platforms. How should the application process be designed in practice? How can the vetting process ensure efficient exchanges between researchers and platform providers?

Applications should be submitted digitally to DSCs via a standardised form that includes a series of questions on planned research, data requirements, specification of systemic risks studies, as well as contact information and affiliation of researchers. Application forms should ask for specific ethical considerations taken and data protection measures in place. DSCs should orient themselves on existing schemes (TwitterAcademicAPI, Facebook Crowdtangle, YouTube Researcher Program). DSCs should make decisions about substantial quality (i.e., study of systemic risk and mitigation measures, see Article 34(1) and Article 35), ethical and technical conditions by themselves. A peer-review process (i.e., by objective referees such as uninvolved scholars from the same discipline) is deemed impractical, due to the long duration of the process, biases in decision making and the unpaid labour involved for researchers, as known from academic publication processes.

The DSCs should pre-determine the data protection measures necessary per data type, so researchers know beforehand which requirements they must fulfil. Vetting of appropriate data protection measures by the DSCs should take precedence into account, so that repeated requests from researchers from the same institution can be processed more quickly based on previous applications and showcased protection measures. For ethical considerations, DSCs can collaborate with existing institutional review boards (IRBs) of research institutions and agree on terms of evaluation. Designated IRBs that DSCs collaborate with should also be accessible to researchers from institutions without an existing IRB. In order to avoid bias or gaming and to safeguard the autonomy of scientific enquiries we stress that it should not be the platforms themselves to decide about the content and quality of the research projects (and researchers) for which access is requested. It should be considered whether research questions are shared with platforms at all or if DSCs only forward access requests, to avoid platforms adapting their response towards the research questions asked.

DSCs could motivate applicants to optionally provide details on how to open data for similar requests also for colleagues. This would pertain an infrastructure managed by the DSCs that lists successful data access requests and makes them available in line with the FAIR principles, so that scholars with similar interests can access data after a certain 'grace period' that allows initial requesting researchers to work with the data. A grace period deemed reasonable is 12–24 months after reception of data. Such infrastructure should be available on a European level to foster cross-country research.

Should data clean rooms be one access mode to highly sensitive data, national DSCs should be responsible for hosting such clean rooms in each country so researchers can get access to data or collaborate with research institutions in developing those (or use existing ones).

b) Article 40(8) exhaustively defines criteria for vetting researchers. How can a consistent assessment across DSCs be ensured, while still taking into consideration the specificities of each request?

DSCs should develop cross-national assessment procedures that are guided by data access formats and data sensitivity. We suggest a tiered process dependent on sensitivity of data requested and access method (see 3. and Darius and Stockmann, 2023). There should be unique standards which type of data has what level of sensitivity and can be accessed via what kind of access method. While data sensitivity is a continuum, a (adaptable) category system can help to develop standards across DSCs. Ideally, DSCs develop a general assessment and list data types to be made available by platforms and a risk assessment on each data type before first access requests are filed. A matrix cross-tabulating data sensitivity with data access methods can help to develop consistent standards across DSCs.

c) What additional provisions or specifications could be useful to help balance the new data access rights and the protection of users' and business' rights, e.g. related to data protection, confidential information, including trade secrets, and security?

Participating researchers should sign data use agreements that ensure they uphold high standards of data security and use data they have obtained only for the purpose and duration specified in the description of their research project.

Data protection measures used in the vetting process can also include the request of data handling plans and consent forms for data-handling agreements, signed off by the institutional review boards of a researcher's institution. This creates a second layer of security as institutions and not only individual researchers are per current data protection regulations under GDPR required by follow up on data access plans, including the deletion of raw data after a research project has been finished. While open science standards should be followed where possible, cases exist where the study and mitigation of systemic risk justify non-open data practices.

For certain types of study designs (e.g., the integration of user digital trace data with survey data that is necessary to study how platform communication can affect individuals), additional informed consent from the users studied should be gathered. Mechanisms for this should be developed, for example, including lists of names of users into data access requests of who researchers acquired informed consent for study participation beforehand.

d) What kind of safeguards can be put in place to assure that data gathered under Article 40 is used for the purposes envisaged and to minimise the risk of abuses?

Safeguards can be annual reporting of vetted researchers on the progress of the research project and data security and confidentiality requirements, data access only to lab members listed on the access request or authorised via a project access list afterwards, and permanent bans to access in cases of serious abuses.

e) Article 40(13) introduces the possibility of an independent advisory mechanisms to support the management of data access requests and vetting of researchers. What would be the added value of such a mechanism?

The added value of an independent advisory mechanism (IAM) would be to provide professional guidance, oversight, mediation, and scrutiny of the data access request process.

Specifically, the independent advisory mechanism should be executed by an intermediary body governed by academic and civil society researchers. Given the immense workload and responsibility, this should not be unpaid labour. An initial IAM (such as the board of the European Digital Media Observatory) should develop standards on how members are selected and incentivized.

Specifically, the IAM

- Can mediate and suggest solutions in case of disagreement. The DSC should first- and foremost be responsible to keep platforms accountable to deliver on granted access requests. Only in cases where disagreements cannot be resolved, the IAM can function as mediator and suggest a (potentially binding) solution, should platforms and DSCs agree on this procedure.

- Can help determine the risk-level of data and projects and thereby support the DSCs, as this might change over time.

- Should be responsible for independent data quality and validity checks and developing quality assurance standards with DSCs and platforms.

- Help define what suitable access methods and data formats are that platforms should deliver, as this might change over time.

- Represent and communicate researchers’ perspectives towards DSCs, the Board of DSCs and platforms.

3. Data access formats and involvement of researchers:

a) What technical specifications could be considered for data access interfaces, which takes into account security, data protection, ease of use, accessibility, and responsiveness (e.g., APIs, data vaults and other machine-readable data exchange formats)?

To facilitate use in academic research, access to VLOP and VLOSE data should be technically tiered in a measured fashion that balances legitimate research interests and ease of use with aspects of data security and privacy concerns. Two chief modes of access that have in the past proven their worth are web application programming interfaces specifically for research (research-oriented APIs, or R-APIs, such as Twitter’s soon-to-be-defunct Academic Track or CrowdTangle) and virtual lab environments (or VLEs, for example Meta's Facebook Open Research & Transparency, or FORT, based on the popular JupyterLab). Note that we consider both approaches to be equally important in the tiered logic of research data access described above, and that we do not consider more restrictive modes of access, such as physical data clean rooms, though they may be the best option for extremely sensitive data. To allow for the broadest possible researcher access to platform data, physical data clean rooms should only be the last option for a small number of studies.

Research-oriented APIs (R-APIs)

Functionality:

- Structured way of accessing VLOP/VLOSE data
- Query different API endpoints with a set of parameters
- Let researcher receive data in easy to work with formats (e.g., csv, json)
- Access simplified through open source packages and languages (e.g., R, Python), and intermediary data transferring protocols (cURL)

Data level:

- Low to medium sensitivity (see also 4.)

Main advantages:

- Well established approach, known by researchers and platforms
- flexible, structured, and efficient way of accessing data
- data analyses on the local machines

Main disadvantages:

- Unsystematic data hoarding possible
- Data protection not sufficient enough for personally identifiable data, e.g., restricted to specific data types

Virtual Lab Environments (VLEs)

Functionality:

- Controlled environment to access social media platform data
- Hosted by a third-party data provider (could be instated by intermediary bodies or DSCs)
- Allow researchers to access data from multiple platforms in one place
- Resources for development and maintenance should be provided by platforms
- Examples are Meta's Facebook Open Research & Transparency, or FORT, based on the popular JupyterLab

Data level:

- Data considered highly sensitive
- Data that is personally identifiable
- Data on platform-internal processes

Main advantages:

- Access data from multiple VLOPs in one place
- Strong processing and analysis capabilities
- Controlled environments with increased data protection
- Aid reproducibility by keeping data analysis pipelines in a single place
- Can be quickly available, if prioritized

Main disadvantages:

- More challenging to use than R-APIs, standards have not emerged
- Constrain analyses by limiting researchers to languages and packages available or installable within the VLE
- Obfuscate errors in data or code by constituting a walled-off environment (external scrutiny needs to be introduced, e.g., validity checks by IAM)
- Require investments on the part of the platforms

b) What capacity building measures could be considered for the research community to take advantage of the opportunities provided by Article 40?

Capacity building measures can involve scholarly societies on national and European level. It is important to involve all disciplines who (potentially will) work with platform data. Resources for workshops or online learning material are necessary so the IAM, the DSCs, involved researchers, or other experts teach on how to write successful access requests. Resources need to be invested to keep data access stable over time to allow for longitudinal research designs which substantiate causal claims otherwise impossible. Platforms should provide documentation that allows researchers to assess the tool. For example, currently, much documentation only mentions a name to describe categories or variables, but no definition of the concept is provided. Measurement is not explained or too obscure to understand. Examples are helpful. A contact should be provided for questions and questions should be answered by the platforms.

c) Would it be desirable and feasible to establish a common and precise language for DSCs, vetted researchers, VLOPs and VLOSEs to use when communicating about data access, e.g. by formulating a standard data dictionary and/or business glossary? How might this be implemented?

Yes, a common language and dictionary would be helpful. This is a resource-intensive process that should be funded sufficiently (e.g., through a call by the Commission) and involve experts in building epistemic structures, such as scholars from linguistic and computer science. Coordination could be a task for the IAM, with broad outreach to research communities. Open science practices should support establishing a common and interdisciplinary understanding of terms. Documentation should follow industry standards and be machine readable as well as human readable.

4. Access to publicly available data:

a) Not only vetted researchers will have greater opportunities for accessing data, all researchers meeting the conditions set out in Article 40(12) will be able to get direct access to publicly available data. What processes and mechanisms could be put in place to facilitate this access in your view?

The implementation of Article 40(12) should be prioritized as this will cover a great number of research questions scholars are working on, while the resources and developments necessary are minor. It will also prevent a gap in current platform research, as the implementation and first successful access requests under Article 40(4) will take more time and create a time gap where platform accountability is not yet in place. All VLOPs and VLOEs should provide API access to public data. Specifically in the case of Twitter, which is about to discontinue academic access, the Academic API should be kept open and free of charge to researchers to avoid gaps in ongoing data collection. The vetting process for Academic API access is considered feasible and can be a standard for developing access strategies to other VLOP and VLOSE APIs. In addition, and to increase the reach and usage of access, Dashboard solutions (following the example of Crowdtangle) should be an additional access mode guaranteed to platform researchers through Article 40(12). APIs with anonymized data access should be mandatory and platforms should harmonise terminologies and data structure where possible and appropriate. VLOPs/VLOSEs could host a shared API platform through which researchers can access data.

III. References

Darius, P., & Stockmann, D. (2023). Implementing Data Access of the Digital Services Act. Accessible via: https://hertieschool-f4e6.kxcdn.com/fileadmin/2_Research/2_Research_directory/Research_Centres/Centre_for_Digital_Governance/5_Papers/Implementing_Data_Access_Darius_Stockmann_2023.pdf (last access: May 22, 2023).

Pasquetto, I. V., Swire-Thompson, B., et al. (2020). Tackling misinformation: What researchers could do with social media data. The Harvard Kennedy School Misinformation Review. Accessible via: https://misinforeview.hks.harvard.edu/article/tackling-misinformation-what-researchers-could-do-with-social-media-data/ (last access: May 22, 2023).

List of Signatories

Dr. Jakob Ohme, Weizenbaum-Institut für die vernetzte Gesellschaft

Prof. Dr. Ulrike Klinger, European New School of Digital Studies, Europa-Universität Viadrina, Frankfurt (Oder)

Prof. Dr. Cornelius Puschmann, ZeMKI, Universität Bremen

Prof. Daniela Stockmann, PhD, Center for Digital Governance, Hertie School of Governance

Dr. Sonja Schimmler, Weizenbaum-Institut für die vernetzte Gesellschaft

Dr. Martin Degeling, Stiftung Neue Verantwortung

Dr. Anna-Katharina Meßmer, Stiftung Neue Verantwortung

Dr. Julian Jaursch, Stiftung Neue Verantwortung

Prof. Dr. Mario Haim, Institut für Kommunikationswissenschaft und Medienforschung, LMU München

Prof. Dr. Jakob Jünger, Institut für Kommunikationswissenschaft, Universität Münster

Joschka Selinger, Gesellschaft für Freiheitsrechte e.V.

Philipp Darius, Center for Digital Governance, Hertie School of Governance

Dr. Philipp Lorenz-Spreen, Max Planck Institute for Human Development

Christian Strippel, Weizenbaum-Institut für die vernetzte Gesellschaft

Prof. Dr. Judith Möller, Leibniz-Institut für Medienforschung | Hans-Bredow-Institut

Dr. Julia Levasier, Bayerisches Forschungsinstitut für digitale Transformation