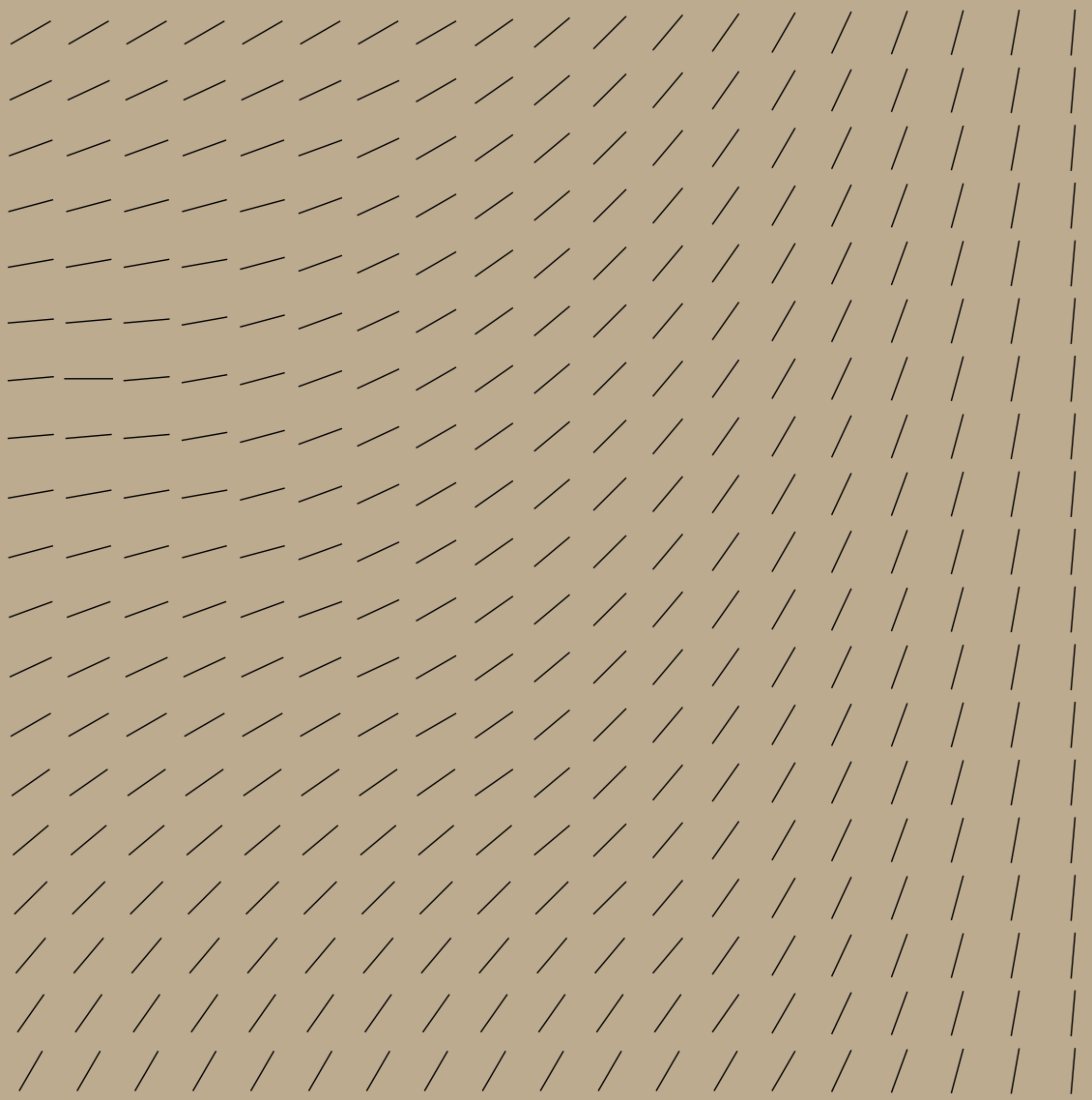
weizenbaum institut

ens European New School of Digital Studies



OCTOBER 2023

Weizenbaum Institute for the Networked Society

What the Scientific Community Needs from Data Access under Art. 40 DSA

20 Points on Infrastructures, Participation, Transparency, and Funding

Imprint

What the Scientific Community Needs from Data Access under Art. 40 DSA. 20 Points on Infrastructures, Participation, Transparency, and Funding, Weizenbaum Policy Paper 8, October 2023

AUTHORS:
Ulrike Klinger \ European University Viadrina \ klinger@europa-uni.de
Jakob Ohme \ Weizenbaum Institute \ jakob.ohme@weizenbaum-institut.de

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

EDITORIAL TEAM:
Moritz Buchner \ Weizenbaum Institute
Christina Dinar \ Leibniz-Institut für Medienforschung | HBI
Robert Peter \ Weizenbaum Institute
Kaltrina Shala \ Weizenbaum Institute

DOI: 10.34669/WI.WPP/8.2
ISSN: 2940-8490

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

This policy paper has been funded by the Stiftung Mercator.

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

The Weizenbaum Institute is funded by the Federal Ministry of Education and Research of Germany (BMBF).

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-institut.de

About this paper

The paper underwent an iterative development process stemming from two workshops held in March and June 2023. These workshops were organized by the authors and supported by Stiftung Mercator. Approximately 70 scientists and platform researchers from various fields actively contributed to the paper's development. Concurrently, informal discussions occurred with scientists hailing from different European Union nations, as well as regulatory authorities in Germany. In response to the EU Commission's Call for Evidence issued in May 2023, the authors submitted an interim status report, see Klinger & Ohme (2023). *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*. https://doi.org/10.34669/WI.WPP/7

What the Scientific Community Needs from Data Access under Art. 40 DSA

20 Points on Infrastructures, Participation, Transparency, and Funding[1]

To ensure free and objective research on social media platforms and their impact on systemic risks to the economy and society, reliable and independent access to platform data with high data quality is essential.

Article 40 of the Digital Services Act (DSA) creates for the first time a clear regulation that grants science independence from individual platforms and improved **data quality**, thus ensuring that socially relevant aspects of digitization can be investigated appropriately, consistently, and **independently**. It makes it possible to respond more quickly and accurately to new issues and developments in an evidence-based manner, thus contributing to a **fair, digital public sphere** that considers societal risks and opportunities.

This policy paper aims to inform the expected Delegated Act of the EU Commission[2] as well as the legislative process for the German Digital Services Act (Digitale Dienste Gesetz) and to formulate necessities from the perspective of platform researchers. This perspective is of utmost importance, as research on systemic risks depends on the expertise of scientific actors. In the implementation and further legislative process at national and European level, the following 20 points are particularly important from a scientific perspective:

Who should have access?

1. **Enable equal access for platform researchers**: Research on digitization and its impacts has been characterized by unequal access to resources and data. The goal must be to grant all platform researchers (if accredited) equal access to data. This will lead to greater diversity and independence in research. One-stop shops and research data centers can support this.

---

[1] We acknowledge the work of the contributors who helped shaping this paper with their valuable inputs and comments: Alexander Wehde, Andrea Lorenz, Daniela Stockmann, Erik Tuchtfeld, Johannes Breuer, Judith Möller, Julia Niemann-Lenz, Katharina Kaesling, Katharina Kleinen-von Königslöw, Katrin Weller, Marie-Therese Sekwenz, Mario Haim, Matthias C. Kettemann, Matthias Spielkamp, Mia Berg, Michael Meyer-Resende, Philipp Darius, Philipp Lorenz-Spreen, Pia Sombetzki, Richard Kuchta, Simon Munzert, Sonja Schimmler, Valerie Hase.

[2] Cf. the detailed proposals on data access in Klinger & Ohme (2023). 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. https://doi.org/10.34669/WI.WPP/7

2. **Allow access to non-EU researchers and journalists**: The DSA must not lead to the compartmentalization of European science. Therefore, access for non-European researchers in cooperation with European researchers must be possible. The access criteria described in Art. 40 para. 8 lit. b DSA should be interpreted broadly for access according to DSA 40 para. 12 for journalists who act in the sense of public education and information.

How can fair access work?

3. **Access must be free of charge and independent**: The current monetary exploitation of X's (formerly known as Twitter) API access clearly shows how important free access is for making research egalitarian and diverse. The DSA clearly puts the onus on the platforms to provide data access for science; the platforms should also bear the costs for this.
4. **Comprehensive data access for accredited researchers**: The following information should be available for vetted researchers as defined in Art. 40 para. 4, 8 DSA:
    - Information on data structures, sample (representativeness), and available platform data, e.g., as a list of variables, collected metadata, and operationalizations at data access
    - Information about recent or soon-planned changes to the data structures and the available data
    - Information about the exact processing, aggregation, and anonymization of the raw data
    - Data that includes internal weighting of data features for algorithmic sorting/selection and other internal decisions such as A/B testing or new platform functionality
    - Platform-specific, publicly accessible data in machine-readable form, in real-time and historically, and available at any time
    - Platform-specific, personal data in machine-readable form, historically, only after a successful vetting process of researchers and, if applicable, consent of the users (for personal, non-public data)
    - Cross-platform data prepared in a standardized, machine-readable form

5. **Custom-fit access modes**[3]: Access to historical and real-time data (steaming APIs) should be as complete as possible via data interfaces (APIs) provided free of charge. Different levels of data sensitivity[4] should offer differently formulated accesses and procedures[5]. At high sensitivity levels in so-called data clean rooms[6], lower level in virtual laboratory environments (VLEs)[7]. Other possibilities for access modes are closed collections, ad libraries, or data donations, which can also be integrated via APIs or other interfaces, for example. For experimental studies and algorithmic audits, it should be possible to set up so-called sock puppets (i.e., online identities explicitly created for research purposes).
6. **Flexible access design**: Art. 40 para. 8 DSA so far only provides for individual access, i.e., by individual researchers. However, the application procedure should be more flexible. An institutional anchoring with a longer term and the possibility of adding new team members would be optimal. Data access should also be usable for academic teaching as well as early-career researchers who ensure the future of platform research. A digital self-disclosure of the research project's independence from commercial interests (DSA 40. para. 8, b) and funding of the research project (DSA 40. para. 8, c) should be sufficient for the approval process under DSA 40. para. 12. Conditions DSA 40. para. 8, d) (data security and confidentiality requirements) and e) (access necessary and proportionate) disproportionately restrict access, as this concerns publicly accessible data. It must be clarified to what extent or how collected data can be shared (e.g., for replication studies) or reused (e.g., in consortia).

How should the approval process be designed?

7. **Conduct of the approval process preferably by the local Digital Services Coordinator (DSC)**: the approval process should be conducted in a timely manner (within 7 days), ideally by (or in close coordination with) the DSC of the applicant's home country, to allow for relief of busy DSCs (likely in Ireland). An appeal mechanism to this process should be offered through an independent advisory board attached to the local DSC (see draft bill for the German Digitale Dienste Gesetz, para. 22), due to greater knowledge of national institutions and processes.

---

[3] See note 1.
[4] See recital 51 GDPR https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:32016R0679&qid=1694877783373
[5] See the Hertie School Data Science Lab's detailed proposal, Implementing Data Acess of the DSA, https://hertie-school-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 accessed 04/24/2023.
[6] Data clean rooms are safe, secure environments where personally identifiable information (PII) is cleaned and processed so that it can be made available for a variety of data analysis purposes.
[7] The virtual laboratory environment is an interactive environment for creating and performing simulated experiments and analyses.

8. **Distinguishing data sensitivity for access requests**: In the approval process, a distinction should be made between access to medium- and highly-sensitive data as defined in Art. 40 para. 4 DSA and prioritized access to publicly available data as defined in Art. 40 para. 12 DSA.
9. **Prioritize publicly available data**: Publicly accessible data from the platforms must be made available at all times and without a vetting process, according to Art. 40 (12) DSA. This refers to data not restricted by the user and can be viewed by freely browsing the platform. Technically, this means, above all, the provision of real-time access through APIs (data interfaces) and dashboards. A definition of semi-public data is needed, e.g., "private" Facebook or Telegram groups with thousands of members. Art. 40(12) DSA should be implemented as a priority (also concerning upcoming EU and US elections) and will already cover a large part of scientific data needs. This requires early, clear guidance from the EU Commission on the implementation of Art. 40(12).
10. **Low-threshold, modular application procedure**: For simplification and standardization, we propose a modular principle in which data and usage interests, as well as data protection measured, are queried in a standardized manner. Users should only need to specify general research interests, similar to the former Twitter Academic API, and also allow exploratory research. Information about what specific data/variables are available should be easily accessible. In addition to German, the application should be available at least in English to enable inclusive data access. Access should be granted via login to an online tool where data from multiple platforms can be accessed simultaneously, also to ensure much-needed cross-platform research on systemic risks.
11. **Expand infrastructures and resources, build competence through close exchange**: Uniform and clear infrastructure requirements must be formulated for research institutions so that appropriate preparations can be initiated, such as a secure server infrastructure. There should be contact persons for questions and support in access for those who do not have the necessary technical skills. There is a need for targeted research funding programs at the European and national levels that explore systemic risks based on requested platform data. These highly specialized and resource-intensive projects cannot be carried out under existing funding programs.
12. **Involvement of the scientific community in peer review processes**: Applications that have already been submitted in this way in a similar form should be reviewed primarily for formalities, ethics, and feasibility of data sharing before direct, timely approval is given. Applications for which there is no precedent should be reviewed more comprehensively, in terms of content, in a peer review system (and may also be reviewed only formally in the future). This reduces review cases and resources. Adequate compensation for researchers involved in the peer review process should be guaranteed.

How can data quality be safeguarded?

13. **Advisory board structure for the DSC**: The planned, independent advisory board of the German DSC should advise on overarching strategic issues with regard to the implementation of the DSA and allow the inclusion of scientific issues. In order to fulfill these tasks, researchers who can demonstrate expertise in the field of empirical research with platform data should be represented on the advisory board. Duplicate structures with a planned Intermediary Advisory Mechanism according to Art. 40(13) DSA should be avoided.
14. **Conditions for the quality of data preparation**: The preparation of the data should be the main responsibility of the platforms and the form of preparation should meet uniform standards in order to meet the listed requirements. The catalog of interfaces and data points should not be defined unilaterally by the platforms but should be determined in a standardized manner through a process independent of the platforms and involving the researchers (e.g., independent advisory mechanism (IAM) to support data sharing under Art. 40(13) DSA). Standardization is necessary to ensure that platform-comparative and cross-platform studies are methodologically valid[8]. This also applies to the provision of text, image/video and audio to enable platform comparisons. Direct forwarding of research questions to the platforms should be prevented to exclude any influence on the part of the platforms.
15. **Documentation and transparency of variables, measurements, and data collection**: Information about when collected data and measurements from platforms were changed over time should be retrievable. Research needs more transparency about metrics, e.g., what does a “like” or “views” mean on different platforms? As uniform as possible variable naming and metadata structure should be strived for – also for “public” data (Art. 40(12) DSA)- here the opportunity for cross-platform simplification and harmonization of indicators exists.
16. **Right to scraping for scientific quality assurance**: Indispensable for scientifically sound quality control of data is the “right to scrap”. Without the right to scrap, the data access offered by platforms cannot be validated. Incomplete data sets, such as those repeatedly provided by platforms in the past (e.g., the Facebook Ad Archive), are hardly useful for research and, in the worst case, are harmful because they falsify the results. Data donations by users are also essential for research and quality control. The possibilities for this should be expanded and simplified.

[8] Cf. also item 1 in https://algorithmwatch.org/en/wp-content/uploads/2023/05/Open_letter_DSA.pdf

17. **Platform data quality reports**: Platforms should publish regular reports that include aspects of the data quality provided (e.g., validation by external bodies, own data cleansing initiatives, completeness of data, etc.).
18. **Insight into the platform data pool for research purposes, including non-European**: Art. 40 para. 4 DSA is intended to provide access to the entire platform data pool, as systemic risks can only be identified if there is clarity about the totality of all data available to the platforms. Viewing and pre-structuring of existing platform data should be done centrally by the DSCs. Non-European platform data must be accessible, e.g., in the context of manipulated social media data and disinformation. The market location principle (*lex loci solutionis*) must be applied here.[9]
19. **Oversight**: An authority must be appointed that can independently check the quality of the data preparation (possibly an Intermediary Advisory Mechanism, which would have to be provided with extra resources for this purpose, cf. Klinger & Ohme). Highly qualified data scientists are needed to assess the data provided by the platforms in terms of quality and potential uses, who can provide a professional assessment, including on the open question of what the data queries of approved researchers may even contain in terms of societal risks. In the event of poor data quality, a short-term and transparent (complaint) procedure should be available for readjustment; in case of doubt, fine procedures should also be able to follow.
20. **Serious penalty mechanisms are good, but incentive structures are better**: Deliberate delays in data access by a platform should result in penalty proceedings. However, we recommend not only to work with serious sanction mechanisms, but also to create incentives for platforms in the medium term by providing sufficient resources for data access (e.g., special compliance). If the platforms refuse to grant access, research institutions should be able to proceed by means of an action for failure to act or a refusal counterclaim, so that mechanisms and corresponding resources should also be provided for this.

[9] Since the scope of the DSA applies according to the market location principle (regardless of the platform's place of establishment), non-European data from non-European platforms are also included, as long as these platforms offer services in the EU or to EU citizens.

Signatories (institutions)

Weizenbaum Institute for the Networked Society, Berlin

European New School, Europa-Universität Viadrina, Frankfurt (Oder)

AlgorithmWatch

Center for Advanced Internet Studies (CAIS)

D64 – Zentrum für Digitalen Fortschritt

Democracy Reporting International (DRI)

GESIS Leibniz-Institut für Sozialwissenschaften e.V.

Stiftung Neue Verantwortung (SNV)

Signatories (individuals)

Ulrike Klinger – Professorin, European New School, Europa-Universität Viadrina, Frankfurt (Oder)

Jakob Ohme – Head of Research Group, Weizenbaum Institute, Berlin

Alexander Wehde – Student Assistant, Forschungsstelle für Rechtsfragen neuer Technologien sowie Datenrecht (ForTech) e.V.

Andrea Lorenz – Research Associate, Universität Hamburg

Christoph Neuberger – Scientific Managing Director, Weizenbaum Institute, Berlin; Professor, Freie Universität Berlin

Daniela Stockmann – Professor, Center for Digital Governance, Hertie School, Berlin

Erik Tuchtfeld – Research Associate, Max-Planck-Institut für ausländisches öffentliches Recht und Völkerrecht, Heidelberg

Jan-Hendrik Passoth – Professor, European New School, Europa-Viadrina Universität Frankfurt (Oder)

Johannes Breuer – Senior Researcher, GESIS Leibniz-Institut für Sozialwissenschaften & Center for Advanced Internet Studies (CAIS)

Judith Möller – Professor, Leibniz-Institut für Medienforschung | Hans-Bredow-Institut, Universität Hamburg

Julia Niemann-Lenz – Senior Research Associate, Universität Hamburg

Katharina Kaesling – JProfessor, Technische Universität Dresden

Katharina Kleinen-von Königslöw – Professor, Universität Hamburg

Katrin Weller – Senior Researcher, GESIS Leibniz-Institut für Sozialwissenschaften

Marie-Therese Sekwenz – PhD Candidate, TU Delft/ Technology Policy and Management/ AI Futures Lab

Mario Haim – Professor, Ludwig-Maximilians-Universität, München

Matthias C. Kettemann – Professor, Universität Innsbruck, Alexander von Humboldt-Institut für Internet und Gesellschaft, Leibniz-Institut für Medienforschung | Hans-Bredow-Institut

Matthias Spielkamp – Co-founder and Executive Director, AlgorithmWatch

Mia Berg – Research Associate, Ruhr-Universität Bochum

Michael Meyer-Resende – Co-founder & Executive Director, Democracy Reporting International (CRI), Berlin

Philipp Darius – Postdoctoral Researcher, Center for Digital Governance, Hertie School, Berlin

Philipp Lorenz-Spreen – Research Scientist, Max Planck Institute for Human Development, Berlin

Pia Sombetzki – Policy & Advocacy Manager, AlgorithmWatch

Richard Kuchta – Analyst and Researcher, Democracy Reporting International

Simon Munzert – Professor, Data Science Lab, Hertie School, Berlin

Sonja Schimmler – Head of Research Group, Weizenbaum Institute, Berlin

Valerie Hase – Postdoctoral Researcher, Ludwig-Maximilians-Universität, München