Response to the call for feedback on a planned Delegated Regulation on data access provided for in the Digital Services Act (DSA)

Amsterdam School of Communication Research

Publication date
2023

Document Version
Final published version

UNIVERSITY OF AMSTERDAM
Amsterdam School of Communication Research

Response to the call for feedback on a planned Delegated Regulation on data access provided for in the Digital Services Act (DSA)

28 May 2023

About ASCoR

The Amsterdam School of Communication Research (ASCoR) is the largest institute of its kind in Europe, and one of the largest Communication Science research institutes worldwide. Its research focuses on the production, content, and effects of media for individuals, groups, and society. ASCoR's research has an inherently broad focus and deals with a range of topics addressing crucial societal issues, such as (mental) health, aging, education, sustainability, diversity, mis- and disinformation, digital inequality, legitimacy of (public and private) organizations, and journalistic freedom. ASCoR researchers apply a variety of methods to conduct rigorous academic research on these challenges, being at the forefront of new computational methods and analysis of digital trace data within the discipline of Communication Science. More information about the institute can be found at https://ascor.uva.nl/

About this response

This response was initiated and edited by dr. Theo Araujo and dr. Damian Trilling. It was elaborated with input from dr. Anne Kroon, dr. Joanna Strycharz, and dr. Marthe Möller, and several other colleagues within ASCoR and in the other institutes listed in the next section. We acknowledge the rich discussion with other institutes and inspiration provided by other responses to this consultation, which resonate with many of the points we list below.

We also highlight that these answers are meant to illustrate the needs from research on communication – and the digital society – and to serve as a conversation starter and evidence for these discussions. We are very much willing to continue contributing to this dialogue, and to provide more detail and further information where needed. Questions can be sent to ASCoR.

Expressions of Support

The views expressed in this document are also relevant for the following institutes at the University of Amsterdam:

Amsterdam Institute for Social Science Research
The Amsterdam Institute for Social Science Research (AISSR) brings together six social science disciplines at the University of Amsterdam. With circa 350 staff members, it is one of largest social science institutes in the world. The AISSR stimulates researchers to develop research programs that are at the forefront of scholarly endeavour and achieve a better understanding of social issues. This pursuit of scholarly excellence and societal relevance contributes to the training and emergence of a future generation of researchers. https://aissr.uva.nl/

The Social and Behavioural Data Science Centre, University of Amsterdam
The Social and Behavioural Data Science Centre is the organizational hub to coordinate data science related activities at the University of Amsterdam. The centre includes methodological specialists for the four domains of the Faculty of Social and Behavioural Sciences: Social Sciences, Psychology, Child Development and Education, and Communication Science. The Social and Behavioural Data Science Centre covers all activities at the intersection of computer science, mathematics, and statistics, and at least one of the substantive disciplines of the social and behavioural sciences. The centre supports researchers working on data science organizationally, financially, and through consultancy activities. https://sobedsc.uva.nl/

The Amsterdam Institute for Humanities Research, University of Amsterdam
The Amsterdam Institute for Humanities Research (AIHR) is a research institute covering research in all important subdisciplines of the humanities, organized in five research schools: the Amsterdam Centre for Language and Communication, the Amsterdam School for Cultural Analysis, the Amsterdam School for Heritage, Memory and Material Culture, the Amsterdam School for Historical Studies and the Amsterdam School for Regional, Transnational and European Studies. AIHR collaborates with the Institute for Logic, Language and Computation, an interfaculty research institute of the Faculty of Humanities and the Faculty of Science and hosts seven Humanities Labs that conduct innovative research and serve as a hub for expertise on and training in digital methods and infrastructure for the study of human culture. https://aihr.uva.nl/

Executive Summary

We welcome the Commission's focus on how to improve the process for researcher access to platform data. We hope that these efforts lead to a privacy-friendly, fair, rigorous, and transparent process for data sharing with researchers, replacing the often arbitrary, opaque, and incomplete data access systems offered by some (but not all) digital platforms. Only through this renewed process can we be sure that independent and rigorous (academic) research can investigate the systemic risks within the digital society.

In response to the Call for Feedback, we outline below a set of suggestions and recommendations. These recommendations are based on experiences from our members in conducting academic research in the digital society, adopting a range of methods to collect and analyse digital trace data from the (very large) digital platforms that influence the daily lives of millions of citizens in the EU. They are also inspired by discussions with other institutes.

Our key recommendations are:

1. **The data provided must be complete and reliable**, i.e., reflective of the complete experience of an individual within a digital platform. This means not only inclusive of content *produced* or reproduced by an individual, but also content to which an individual is *exposed*, including (automated) recommendations, and how an individual is *profiled* by platforms. Achieving complete and reliable data requires transparency from the platforms regarding their data collection and algorithmic processes.
2. **The data must be accessible and usable**, i.e., it should be provided in machine-readable form, through easily accessible APIs, and accompanied by clear and updated documentation describing all variables, and of all relevant metadata. In particular, it is important that the APIs have generous rate limits. This means that the maximum amount of data that an individual researcher can access must not be capped, and that the data needed for meaningful analyses can be retrieved within a feasible time frame.
3. **The access process must be independent**, i.e., we strongly support the notion of setting up an independent body to facilitate access by researchers to this type of data, to help keep the DSC's efforts at a manageable level and eliminate dependencies on decision-making by platforms as to whom access is provided, or what type of data is made available.
4. **The process for researcher access to data must support both confirmatory and exploratory research**, i.e., the process must be devised in such a way that researchers can submit requests for a specific research project and for academic institutions to have ongoing access to data in a way that enables much-needed inductive, bottom-up research on platform activity.
5. **The process for researcher access to data must be comprehensive**. The further specification of the DSA and of data access should include more methods than only the direct provision of data by platforms to researchers. Specifically, this means clarifying that scraping of public data is in accordance with the DSA - thus enabling the audit and enrichment of data provided directly by platforms -, and that experimentation with automated user accounts should be possible - thus enabling research on the effects and workings of algorithmic processes.

6. **Citizen participation must be facilitated and encouraged**, i.e., these efforts must support data donation initiatives, in which citizens are able to donate their data to academic research (making use of GDPR provisions facilitating data access), and help simplify this process (e.g., by establishing processes in which individuals can donate their data via one-click processes to vetted research institutions).

We detail our recommendations in the answers below and would be glad to further engage in this discussion.

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?

It is of utmost importance that data is reflective of the complete experience of an individual within a digital platform. This means not only inclusive of content *produced* by an individual, but also content to which an individual is *exposed*, including (automated) recommendations, and how an individual is *profiled* by platforms.

We therefore adhere to the principle that all data made available to an individual user (i.e., served to an end user client), created by an individual user (i.e., their activity), or created by the platform *about* an individual user (i.e., profiling) or served *for* that user (e.g., recommendations, or the content shown) must in principle be accessible to researchers, under clear, transparent and privacy-protecting conditions.

It is of crucial importance that this data is made available at individual level, to investigate and establish media effects. If one wants to study, for example, how disinformation spreads, then one needs to be able to do that on the level of an individual message. Similarly, aggregate statistics on how often something has been shared are not sufficiently informative - one needs to be able to reconstruct via which connections information is spread and what information each individual is exposed to online - and when. For example, to conduct network analyses, one needs to be able to reconstruct the full dissemination graph rather than a sample.

Documentation and transparency are also key: if metrics are provided (e.g., engagement scores) but it is unclear how they are calculated, then this does not help, because it cannot be established *what exactly* affects the outcome of interest. To study some of the risks mentioned (e.g., disinformation and hate speech), that data should contain features from three groups: (1) users, (2) content, and (3) structure (e.g., sharing patterns). Of course, measures to ensure privacy are to be taken for (1). It is also necessary to differentiate between (a) what is shown to users, and (b) what users interact with: we need data on both *in conjunction* as they both have effects on individuals. For example, to understand the role of a platform's algorithms on radicalization, one needs to know both what the platform shows to a specific user and what the user then does with the content. Regarding (2), it is important that exposure to all types of content - i.e., generated by users as well as sponsored/paid content (such as advertising or organizational content) - is accessible to researchers as both types of content have distinct effects on individuals and societal consequences.

b) What sort of analysis and research might DSCs and vetted researchers conduct for the purposes of monitoring and assessing compliance and conducting research related to systemic risks and mitigation measures?

ASCoR scholars conduct academic research on a wide range of topics that address a set of (systemic) societal risks. This includes topics such as (mental) health, (digital and media) literacy, sustainability and energy transition, diversity, mis- and disinformation, polarization, digital inequality, the crisis in trust on (public and private) organizations, consumer empowerment, and challenges to journalistic freedom. In all those issues, information content and communication processes play a key role and turn out to be of vital importance. By having access to complete and reliable platform-related data, ASCoR researchers can conduct rigorous academic research in these topics (among many others) and contribute to the development of mitigation measures in the form of new academic knowledge, theory development, concrete, theory-driven empirically-tested interventions, and policy recommendations.

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?
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?
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?
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?
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?

We subscribe in principle to the recommendations by the European Digital Media Observatory (EDMO) and their draft code of conduct for platform-to-researcher data access[1].

In addition, we believe that the application process should be designed in a way that minimizes the role of platforms in the data access process. **We strongly welcome the idea of an independent body composed by public institutions, universities, and societal partners with domain-specific expertise as a way to manage these processes, and to help minimize the workload of DSCs in member states.**

We distinguish between three forms of access, and would welcome the creation of processes that accommodate these three scenarios:

1) **The vetting of specific research projects**, in a simple and rigorous process to allow for individual requests for specific data for clearly defined research aims to be granted.

[1] 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

2) **The vetting of research institutions**, ensuring that trusted institutions (e.g., universities, European Research Infrastructure Consortia) can have ongoing access to platform data to conduct regular monitoring of platform activity and/or systemic risks, as well as exploratory research. These trusted institutions must have clear processes in place to ensure rigorous, privacy-protecting research (e.g., Ethical Review Boards, Data Management Plans).

3) **The vetting of citizen data donation pipelines**[2], ensuring that digital platforms enable individuals to donate their own data to academic research using frameworks developed by researchers.

We also highlight the need to enable and support *both* confirmatory research (i.e., with a clearly delimited research question and causal model) and exploratory research (i.e., in which researchers may require data with a broader purpose, to detect patterns and inductively derive insights from platform and user activity). In addition, we subscribe to the principle outlined by the EDMO that the DSA *should not* serve as a reason for platforms to collect *even more data* or *increase surveillance* on individuals.

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 enable the execution of rigorous, reliable, and replicable research on systemic risks, the process must ensure that the data provided is machine-readable (e.g., JSON format) and well-documented - as outlined in the previous section. In addition, we see the need for:

1. **Ensuring that the workflow is efficient, real-time, and stable**. We see JSON-based APIs as the go-to solution that should be used for most of the data sharing mechanisms. In the vast majority of cases, APIs can be neatly integrated into researchers' workflows.

2. **Ensuring the adoption and enforcement of a common standard for data provision and documentation**. This includes the definition of a set of baseline variables that must be delivered by type of data and agreed upon standards for these variables. These common standards ensure, on the one hand, that research can be reliably done *across* platforms and, on the other hand, that the scientific community can develop common practices and standards for analysing these data.

   As a simple example, whenever providing content (e.g., a post, a video, etc.) for academic research, platforms should consistently deliver the same accompanying metadata (e.g., the timestamp when it was created, last updated etc.) and the same type of data (e.g., the text within the content or transcription as well, at a minimum, a link to the image or video

[2] For more details on data donation, see https://datadonation.eu/

file). While this might be considered straightforward, in practice every platform provides a (small) subset of what may be required, following often different formats (e.g., using a wild variety of incompatible timestamp formats) and, barring notable exceptions (e.g., Twitter at least until a few months ago), with minimal or no documentation.

3. **Enabling access for automated user accounts for research**. Given the need to study the impact of algorithmic processes *within* platforms as one of the major areas for systemic risks within communication, platforms must allow researchers to study their effects and workings using automated user accounts (e.g., to simulate specific user behaviour and understand the dynamics of algorithmic profiling and recommendations).

4. **Supporting pipelines for citizen data donation**. As per the GDPR, individuals already have the rights to access and download the data that digital platforms have on them. To advance our ability to perform much-needed research on systemic risks associated with communication, we see data donation (i.e., individuals donating their digital trace data to research[3]) as a crucial way to combine access to personal (digital) data and individual perceptions and behaviour (collected through surveys or interviews), importantly, with *informed consent* from participating individuals. As such, we believe it is essential to create mechanisms through which platforms enable individuals to donate their data in a straightforward manner to vetted institutions, using frameworks developed by researchers (e.g., OSD2F[4] and PORT[5]).

We also acknowledge the increasing knowledge within the research community on how to access, process and analyse different types of data[6]. Exploring how to integrate this knowledge and existing infrastructure can facilitate (secure) data access, and reuse initiatives already funded by the EU or national bodies.

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

In addition to the capacity measures outlined above, we see two main capacity building and/or support measures for the research community:

1. **Clarifying that scraping is in line with the DSA**.
   We see scraping - i.e., researchers collecting data directly from platforms via end-clients or other means - as a much-needed mechanism to ensure that data provided *by* platforms can be audited by being compared to data that can be extracted *from* platforms, and to

[3] For more details on data donation, see https://datadonation.eu/
[4] https://www.aup-online.com/content/journals/10.5117/CCR2022.2.001.ARAU
[5] https://doi.org/10.1016/j.patter.2022.100444
[6] We highlight, for example, https://sshopencloud.eu/ or initiatives funded by PDI-SSH in the Netherlands (https://pdi-ssh.nl/nl/home/) such as the Secure ANalysis Environment (SANE) initiative (https://odissei-data.nl/en/2022/02/secure-analysis-environment-sane-secure-data-for-social-sciences-and-humanities/)

enrich the data provided by platforms. We welcome the development of codes of conduct and further specification of the practice.

2. **Creating clear mechanisms and standards for data access and data sharing**.
We note three separate (and interrelated) developments that reinforce the need to create clear mechanisms and standards for data access and sharing.

First, we note that open science principles, which includes open data and FAIR data policies, are a key priority for the European Commission/European Research Council, national science foundations (e.g., NWO), general scientific advisory bodies and universities. As a research institute we subscribe to these principles. Along the same lines, more and more journals and funders require data to be made available for replication purposes or to extend other researchers' work. Furthermore, research transparency and integrity principles require that datasets are available for replication purposes.

Second, we see that *team science* is becoming the new norm in how we do research. This means that our researchers are increasingly taking part in larger teams, often collaborating across disciplinary and institutional boundaries. As such, access to research data is not restricted to the principal investigator of a project, but rather something that must be enabled for multiple project participants, both within one's own university (e.g., research assistants, collaborators) and in other institutes (e.g., other universities in case of national or international collaborations).

Third, we note that often for research based on social media data collected via APIs, it is often not possible to make the underlying data open access because platforms do not allow that in their terms of service. Yet, replicability is considered a cornerstone for the gathering of scientific evidence and is also increasingly demanded by funders. If the exact dataset cannot be redistributed or at least reconstructed by others, it makes research less replicable and leads to large downstream problems such as researchers shying away from conducting research on these data in the first place.

We therefore consider the following capacity building and support initiatives to ensure that **data access and sharing procedures are compatible with open science and FAIR principles**. This means:

a. **Clarifying under which conditions data can be shared**, with clear guidelines, manuals and instructions that can be used by data stewards, privacy officers and ethical review boards.

b. Ensuring that the data request process via DSC's or independent bodies (suggested above) supports **cross-institutional and multi-country applications**, being consistent across countries.

c. **Reconciling data access standards with open science and FAIR data principles**, ensuring that data can be shared publicly (or within the research community) unless there are privacy-related concerns for citizens. In case of privacy concerns, clarify that anonymized or pseudonymized datasets can be shared instead.

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?

It is considered desirable to establish and implement a standard data dictionary, which also aligns with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles–to emphasize standardization and clarity in data management.

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 standardized research access requests for research to access data. This process should be transparent and adhere to the minimal standards we suggest above for all types of data.