Improving Data Access for Researchers in the Data Services Act

Dergacheva, Daria; Katzenbach, Christian; Schwemer, Sebastian Felix; Quintais, João

Publication date: 2023

Citation for published version (APA):
We submit this feedback in response to the Call for Evidence on the Delegated Regulation on data access provided for in the Digital Services Act (DSA).

We are four researchers based in different research institutes in Germany, Denmark and The Netherlands. This feedback is based on our own individual research, as well as our joint research in the context of the Horizon 2020 ReCreating Europe project\(^1\), where we conducted research on the impact of copyright content moderation by online platforms. One of the key conclusions

\(^1\) See [https://www.recreating.eu/stakeholders/wp6-intermediaries/](https://www.recreating.eu/stakeholders/wp6-intermediaries/).
of our research in that project was the need for robust data access rules to enable adequate study of platforms’ content moderation practices.\footnote{Quintais, João Pedro and Katzenbach, Christian and Schwemer, Sebastian Felix and Dergacheva, Daria and Riis, Thomas and Mezei, Péter and Harkai, István, Copyright Content Moderation in the EU: Conclusions and Recommendations (March 1, 2022). reCreating Europe Report (March 2022), Available at SSRN: https://ssrn.com/abstract=440342.}

Article 40 DSA is a crucial provision to operationalize the regulation’s risk mitigation provisions vis-à-vis very large online platforms (VLOPs) and very large search engines (VLOSEs). In essence, Article 40 DSA enables data access to Digital Services Coordinators (DSCs) or the Commission, “vetted researchers”\footnote{See Article 40(8) lit. a-g DSA.} and other researchers\footnote{Article 40(12) DSA.}, provided certain conditions are met. Our submission is predominantly concerned with the data access for vetted researchers and researchers in relation to VLOPs.

**Vetted Researchers that may benefit from Article 40**

Article 40(4) DSA, in conjunction with Article 40(8) DSA on data access and scrutiny, state that researchers can be granted the status of “vetted researchers” for the “sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union (…) and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures (…)” put in place for VLOPs and VLOSEs.\footnote{See Evaluation and Recommendations of WP6 on Intermediaries in the H2020 ReCreating Europe project: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4380345 and https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4403423.}

Among the conditions imposed for DSCs to grant the status of “vetted researchers”, Article 40(8) DSA states that said researchers must be affiliated to a “research organisation” as defined in Article 2(1), of Directive (EU) 2019/790 on Copyright in the Digital Single Market (CDSM Directive). Article 2(1) of this Directive provides a complex definition of “research organisation” as

“a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research:

(a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or

(b) pursuant to a public interest mission recognised by a Member State;
in such a way that the access to the results generated by such scientific research cannot be enjoyed on a preferential basis by an undertaking that exercises a decisive influence upon such organisation”.

Recitals 8 and 11 to 15 CDSM Directive offer in varied ways and to different degrees some aid in the interpretation of the concept of “research organisation”, which plays a key role inter alia in the regime of text and data mining (Articles 3 and 4 CDSM Directive). In this context, questions have arisen as to whether the concept of “research organisation” requires a geographical limitation of the vetted researcher to a researcher that carries out research from or is affiliated to an institution based in the EU.

Along with other colleagues⁶, it is our view that such a limitation is not justified. Rather, the reference to a research organisation should be construed as a requirement that the organisation in question, if outside the EU, conducts scientific research (1) on a not-for-profit basis or (2) reinvests all the profits in its scientific research, and (3) generally provides non-discriminatory access to that research.

Beyond this, the requirements applicable to vetted researchers are mostly substantive (e.g. purpose of research and its alignment with the contribution) and formal (independence, funding disclosure, security, confidentiality, etc) as spelled out in Articles 40(4) and (8) DSA. As noted by Husovec⁷, the only other geographic constraint refers to the purpose of the research being tied to systemic risks in the EU, which does however have no bearing on the location of the research organisation. Importantly, this constraint should not act as a limitation to the researchers examining non-EU data as a means to effectively carry out the research, since such data may be crucial to for instance draw comparisons and provide control variables. Article 40 DSA already provides safeguards against disproportionate or excessive requests, so it would be unreasonable and undesirable to impose such a limitation on vetted researchers.

**Data access: needs and applications, institution-building and procedure**

Data access demands the highest levels of data security and infrastructure facilities on the side of researchers and their institutions, which can incur prohibitive costs and thus hamper research into platforms. Finding practical and fair solutions as well as best practices for data access that are not only accessible to researchers at elite and perfectly-equipped institutions is therefore a key challenge for policy and research. This challenge could be

---

⁷ Id.
supported, for example, by providing **best practices** as well as **capacity building** both on national and EU-level.

Similarly, in order to provide for robust and effective data access, it would also be important to devote attention to **dispute resolution between platforms and researchers**, for example in instances, where there is disagreement regarding the requested data/information in order to avoid the necessity of strategic litigation for the purpose of effectuating Article 40 DSA.

Researchers need access to **readily-available research infrastructure**. Research projects often need to be delivered under high time pressure. Within a single research project it can be difficult to have the expertise and competence to operationalise the data access provision. Therefore we need **institutional efforts**, e.g. by creating structures or organizations for researchers to address those challenges. Such challenges demand institutional capacity to coordinate database creation and secure processes for data access.\(^8\) While there is increasing funding for research projects around the role and risks of platforms in society, there are few funding opportunities for such institutions. Yet, we need **sustainable public funding to facilitate sustaining coalitions and collaborations** between those observatories and for **institution-building in this field**.

In contrast to natural and engineering sciences, social sciences and legal scholarship have fallen behind in building and sustaining research infrastructures to study the challenges of our time. If this does not succeed, individual researchers will repeatedly need to go through these complex processes of data access on their own, bearing too many costs on all levels and failing to respond with robust expertise to the critical questions of our time.

At the moment, platform’s application programming interfaces (APIs) provide practically no data that could help to access content moderation or algorithmic recommendations. We have already indicated it our H2020 ReCreating Europe Report\(^9\) concerning the lack of meaningful data on copyright moderation. We believe the situation is the same with auditing moderation and algorithmic systems dealing with misinformation, hate speech, well-being and other kinds of societal issues that may translate into systemic risks under the DSA. There are numerous researchers in the EU and

---

\(^8\) Currently, there already do exist some smaller and bigger observatories and networks, such as: the social media observatory, Platform governance, media and technology lab at ZeMKI Bremen, which keeps the platform governance archive, as well the EU-funded European Digital Media Observatory focussing on misinformation, and the Platform Governance Research Network.

abroad that agree with this statement, including Sciences Po Shabayek et al., 2022\textsuperscript{10}; Bundtzen & Schwieter\textsuperscript{11} of German Institute for Strategic Dialogue (2022), Lurie et al (2023\textsuperscript{12}) of Stanford Internet Observatory Cyber Policy Center; Dergacheva & Katzenbach (2023, forthcoming), to name a few. In particular, the recommendation by Messer et al. (2023) has illustrated with examples – which we subscribe – the data that is missing\textsuperscript{13}.

In addition to these top-level remarks we note that in terms of empirical research on platforms, there are several layers of data access that we feel is urgently needed for the research community (cf. questions 1 a) and b); 3 a)).

**Data access formats and involvement of researchers**

We provide below a description of types of data required for analyzing societal risks by (1) computational methods in social science, (2) qualitative types of research, (3) experimental research.

*Data for analyzing societal risks by computational methods in social science*

- Platforms need to add more features for the academic API access of their data. This would ideally enable the provision of metadata from social media that could help researching various issues, such as copyright moderation practices, the spread of misinformation online, and hate speech content moderation. For example, in researching content moderation and spread of harmful content and misinformation, the needed data is often absent from the official APIs of platforms, even when there is some level of access provided to researchers. Some of the examples include data that has been removed by platforms, data that has been labeled, data that has been stopped from promotion, as well as data on how algorithms work in connection with harmful or misleading information.

  Currently, researchers come up with many kinds of research designs, which often do not provide satisfactory outcomes due to scarcity of data. Thus, there should exist research programs within platforms

\textsuperscript{10} Shaden Shabayek, Emmanuel Vincent, Héloïse Théro. Digital Platforms’ Governance: missing data & information to monitor, audit & investigate platforms’ misinformation interventions. 2022.

\textsuperscript{11} Sara Bundtzen, Christian Schwieter Access to Social Media Data for Public Interest Research: Lessons Learnt & Recommendations for Strengthening Initiatives in the EU and Beyond.

\textsuperscript{12} Lurie Emma, Bateyko, Dan & Schroeder, Frances (2023). TikTok just announced the data it’s willing to share. What’s missing? Stanford Internet Observatory Cyber policy Center Blog.

with APIs providing more data than currently available, including metadata on content moderation and algorithmic recommendations in its various forms and for various reasons.

- Research programs that platforms may run as a result of the DSA should include research access to moderated content for a certain period of time. For example, if there is research on content deleted due to copyright claims or misinformation, API metadata is often insufficient to audit the quality of such moderation. Thus, it should be possible for vetted researchers to access deleted data which could be stored, as a sample, by platforms for a certain amount of time for these purposes.

There is a lot more data which requires a more qualitative approach and is currently not accessible. It is important that both DSCs and vetted researchers use mixed methods to assess systemic risks of VLOPs content moderation. Although qualitative data research is not mentioned specifically in the DSA at the moment, it needs to be taken into account.

**Data for analyzing societal risks by qualitative types of research**

In addition to such quantitative data, there is a high need for data access from the perspective of qualitative social science research, in order to better understand and evaluate the operation and risks of platforms. Access to this type of data and process is still only superficially defined and developed in the DSA. At the very least, the following data is needed to assess systemic risks:

- Access to description of the decision-making processes on algorithmic systems (VLOPs do this kind of assessments in order to develop their recommender systems, for example);
- Providing access for qualitative interviews ("points of contacts") – for research interviews with staff involved in decision making and development of systems within VLOPs;
- Providing access to documentation to understand the labor conditions of moderators as well as related issue.

**Data for analyzing societal risks by experimental research**

From the perspective of social sciences and in relation to experimental research, at least the following data would be necessary to assess systemic risks:

- Providing DSCs and vetted researchers with the results of experiments on algorithmic amplification and content moderation that VLOPs perform within their own research units. This may include, for
instance, data on the effect of recommender systems on the spread of misinformation or hate speech etc, data on automated copyright or hate speech moderation, data on demonetization and visibility measures (for example, in order to assess if platforms still work with “shadow banning” or variations thereof), and data on virality of information and its amplification in the VLOP’s services.

- Possible cooperation of VLOPs and experimental researchers on algorithmic amplification which has the potential to cause systemic risks.

Access to publicly available data

Article 40(12) enables researchers that do not qualify as “vetted researchers” to directly access publicly available data, subject to certain requirements. This is an important provision as it enables a broader level of access beyond the category of “vetted researchers”, thereby mitigating the risk that the latter function as gatekeepers for research data.

As argued by Keller in her submission14, Article 40(12) DSA supports the extraction of data and information from publicly available web pages (web harvesting or scrapping), including through APIs.

In our view, one avenue to facilitate this access could be to incentivize or mandate (to the extent possible) the addition of data to traditional API access by VLOPs, thereby extending its availability beyond the realm of vetted researchers to include for instance civil society organisations, independent researchers, etc.

To achieve the aim of facilitating access under Article 40(12) DSA, it is also important to insulate and immunize the activities it enables (such as web harvesting or scrapping) from strict interpretations of other applicable regimes that would deprive it of its effectiveness and have a chilling effect on research.15 In particular, many of these activities would arguably follow under the definition of “text and data mining” (TDM) in Article 2(2) CDSM Directive, which covers “any automated analytical technique aimed at analysing text and data in digital form in order to generate information which includes but is not limited to patterns, trends and correlations”. If so, then such research activities under Article 40(12) DSA might trigger the application of the regime for TDM for the purposes of scientific purposes in Article 3 CDSM Directive. One possible way to ensure the smooth interplay

14 The point is made in some detail by Keller, Daphne. Feedback reference F3422727, https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act/F3422727_en
15 Id.
of both provisions is to at the very least clarify explicitly that researchers that meet the requirements of Article 40(12) DSA are also presumed to meet the requirements of Article 3 DSM Directive.

To be clear, this does not mean that existing restrictions in other legal instruments, such as the General Data Protection Regulation (GDPR), should not apply to web harvesting or scrapping. Rather, the aim of the Delegated Regulation in this respect should be to preserve the scope of data access to researchers under Article 40(12) DSA by (1) avoiding that existing rules are interpreted to unduly reduce its scope and (2) not imposing additional restrictions or requirements in in the Delegated Regulation that have a similar effect.

In this context it is also important to underline that existing data repositories and archives, such as, e.g. the Internet Archive\textsuperscript{16} or the Lumen Database run by Berkman Klein Center for Internet & Society at Harvard University\textsuperscript{17}, fulfill important functions for society at-large as well as for research purposes. Data access provisions such as Article 40 DSA should only be seen as compliment to – rather than as alternative to – these existing archives, which need to be preserved.

**Concluding remarks**

Neither platforms, their technology, nor researchers studying these phenomena are static. Therefore, it is important that Article 40 DSA will be approached with enough flexibility\textsuperscript{18} also by national DSCs in order to effectively fulfill the legislators’ intention.

We look forward to seeing suggestions from this Call for evidence reflected in a delegated Act by the European Commission according to Article 40(13) DSA. We encourage the European Commission to – where legislative intervention is not possible or sensible – continue to look into the facilitation of stakeholders.

\textsuperscript{16} https://archive.org.
\textsuperscript{17} https://www.lumendatabase.org.
\textsuperscript{18} This includes explorative studies which help map what kind of data meaningfully can contribute to the study of platforms; changing research environments etc.