Call for evidence on the Delegated Regulation on data access provided for in the Digital Services Act - Summary & analysis

Leerssen, P.

Publication date
2023

Document Version
Final published version

Call for evidence on the Delegated Regulation on data access provided for in the Digital Services Act

Summary & analysis

Dr. Paddy Leerssen, University of Amsterdam[1]

I. Introduction

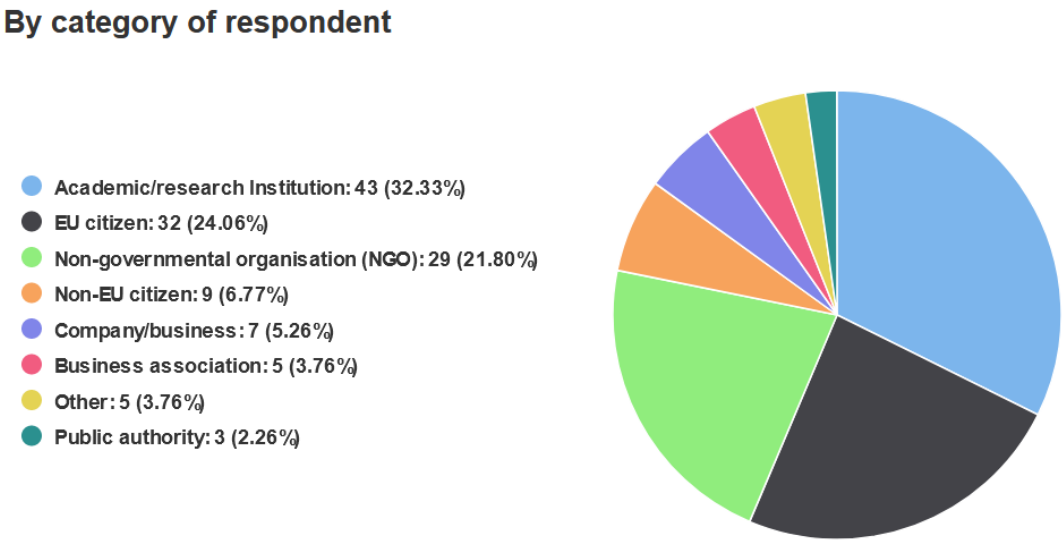
This report provides a summary and analysis of the feedback received in response to European Commission's call for evidence in the context of the Delegated Regulation on data access provided for in the Digital Services Act. It starts by briefly reviewing statistics on the consultation respondents, followed by a detailed assessment of the substance of their responses. It closes with a list of supplementary readings, selected from the works cited in the consultation responses.

II. Who replied to the call for evidence?

The call for evidence ran from 25 April 2023 until 31 May 2023. In total, 133 valid responses were received.

A breakdown of the respondents per category is reproduced below:

[1] The author thanks Eline d'Hoore for her research assistance.



Source: European Commission.[2]

The most common countries amongst respondents were the United States (29), Slovakia (22), Germany (15), Belgium (11) and the Netherlands (7).[3]

III. Main findings of the call for evidence

This report's analysis will be structured around the questions posed in the call for evidence, which are reproduced in full below:

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 DSCs for the purpose of monitoring and assessing compliance and for vetted researchers for conducting research related to systemic risks and mitigation measures?*
   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?*

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.*

[2] Delegated Regulation on data access provided for in the Digital Services Act.
[3] A full overview can be found on the EC website: Delegated Regulation on data access provided for in the Digital Services Act.

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?*

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

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?*

Before proceeding, a brief note on referencing is also in order. First, citations to relevant responses are not presented as exhaustive of all relevant responses, but rather as *indicative* and *illustrative*. Second, for the sake of consistency and clarity, responses are cited by reference to the individual or organisation listed in the response form, though it should be noted that the response is *not necessarily made on behalf* of the organisation named here. For universities in particular, responses may originate from individual researchers or research groups within the organisation, and may not necessarily be representative of the organisation as a whole. More detailed information on the authorship for each response can be found via the hyperlinks inserted in each footnote.

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 DSCs for the purpose of monitoring and assessing compliance and for vetted researchers for conducting research related to systemic risks and mitigation measures?*

The responses mention many different categories of data, from which several themes emerge:

- **Data related to users, accounts, and pages,** e.g. profile information; group memberships; friend/follower relationships networks; individual-level content exposure and engagement histories; associated profiling and labelling;
- **Data related to content,** e.g. individual post content; interaction metadata such as comments, engagements, impressions rates; associated tags and labels; monetization status;
- **Data related to content recommendations,** e.g. technical documentation on algorithmic ranking systems, including data used to personalise recommendations; data on recommended content outcomes; user interaction with recommended content; usage data for algorithmic ranking controls and settings;
- **Data related to ad targeting and profiling,** e.g. technical documentation on algorithmic targeting systems, including user data used to profile types/market segments; data on advertising outcomes and payments; usage data for ad targeting controls and settings; and
- **Data related to content moderation and governance**, e.g. technical documentation on (algorithmic) moderation systems and processes; archives or repositories documenting moderated content and/or accounts; item-level, disaggregated data on moderation actions, appeals rates, effects.[4]

Most respondents' data needs fall into one or more of these categories, though they are conceptualised in different groupings and with different terminology. For instance, the Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access (hereinafter: 'EDMO Researchers') refer to *post-related* data, *user-related* data and *content moderation* data, whereas the Weizenbaum Institute refers to *communication* data, *user account* (meta)data, and *data governance documentation*.[5] Comparable views are shared by other respondents including Amsterdam School of Communication Research (ASCoR), Arcom, Institute for Strategic Dialogue (ISD), Centre for Democracy &

[4] Stiftung Neue Verantwortung e. V. (SNV) refers to Twitter's compliance API, which "distinguishes between different statuses like "deleted" (meaning that the tweet or user account has been deleted) "deactivated" (meaning that the tweet or user account has been deactivated) "scrub_geo" (meaning that the geo-information associated with the tweet or user has been removed), "protected" (meaning that the account from which the tweet originated has become private) and "suspended" (meaning the account from which the tweet originated has been suspended)". More detailed information along these lines might also include other moderation actions (e.g. fact-checks, labels and interstitials, demotions and delistings) and other metadata such as moderation ground(s), nature of decision-making, appeals status, and so forth.

[5] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. Weizenbaum Institute for the Networked Society Berlin.

Technology (CDT), Dublin City University (DCU) and the Slovak Council for Media Services.[6]

It should be clear that the above categories are not entirely separate but rather interrelated. For instance, information about content *recommendations* or amplifications would include the types of *content* being recommended as well engagement patterns from *users* interacting with that content, as well as *moderation* actions such as demotion enacted against that content. Similarly, information about content or accounts might also include moderation information, such as whether the content has been demonetized or fact-checked, as well as user engagement history. Emphases will differ between research topics, projects and methodologies, as will be discussed further under 1B below.

Researchers emphasise the need for **historical, longitudinal** access – allowing researchers to trace patterns over time – but also **real-time access**.[7] To enable this, there is widespread support amongst researchers on the need for **automated access via APIs**. For instance, EDMO Researchers recommend the development of real-time and historical APIs for post-related data, user-related data and for content moderation.[8] NYU's Center for Social Media and Politics and Dublin City University recommend the development of multi-modal datasets, spanning different platform ecosystems and therefore necessitating some degree of standardisation across them.[9]

Besides APIs, other disclosure methods are also proposed. Relatively straightforward requests might be handled through simple databases (CSV-formats) or even text files. For especially sensitive queries, respondents mention various **secure access formats.** Methods that are mentioned in this space include clean rooms (virtual or physical) and virtual lab environments, data vaults, sandboxes, and remote query execution.[10] Few detailed technical standards are submitted on these concepts, and OpenMined notes that the industry has not yet coalesced on a consistent terminology for such solutions.[11] What these techniques generally have in common, however, is that they permit researchers to derive insights from third party datasets without copying the raw data to their local machines.[12] An example mentioned by several researchers as a potentially valuable blueprint is Facebook's Open Research and Transparency (FORT) environment, as well as

[6] Amsterdam School of Communication Research (ASCoR). Arcom.Institute for Strategic Dialogue. Centre for Democracy & Technology, Europe Office. Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Council for Media Services (Slovakia).

[7] e.g. Stanford Internet Observatory ("There are two modes of operation that could be considered: historical and real-time. Historical queries would allow searches back in time, and real-time would be a non-stop stream of events matching particular rules (such as Twitter's PowerTrack). One question to iron out is how to perform research on data that has been deleted in such a way that it complies with other EU regulations. By and large, Twitter's former API offerings are a good model to base future work off of, though other platforms more focused on multimedia content could offer additional contents such as the identifiers of soundtracks to video content or transcription.")

[8] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access.

[9] Standardisation, for the sake of comparability, is also supported by various other respondents including Stiftung Neue Verantwortung e. V. (SNV). This matter is also related to the development of common definitions or vocabularies, which is discussed further in Section 3.C.

[10] OpenMined (also citing relevant policy reports from the Royal Society, the United Nations, and the United States Government).

[11] Ibid.

[12] Ibid.

the now-defunct JupyterLab.[13] These secure methods can also be more complex, costly and also restrictive for research purposes. Most researchers therefore recommend a **tiered access system**, with conventional API access handling the majority of requests and the most restrictive access methods reserved only for the most sensitive requests.[14]

The sensitivity of disclosed data can also be mitigated through privacy-enhancing technologies (PETs).[15] At a minimum, anonymisation/pseudonymisation are endorsed as essential safeguards. Going further, some parties also propose more restrictive methods such as differential privacy and k-anonymity. These matters are discussed under sections 2.B and 2.C below.

Across these issues, there is a recurring emphasis amongst researchers on the importance of **technical documentation** from providers of VLOPs and VLOSEs (hereinafter referred to collectively as 'VLOs') providing context to the data which they disclose. Researchers require technical documentation in order to make effective use of APIs and other automated tools, and also to make sense of the data which VLOs provide. Such documentation can address *inter alia* how the data is collected and preprocessed (e.g. sampling, anonymization methods); relevant variables in the dataset; how these variables are defined and calculated; code examples to assist users; and an accessible point of contact for researchers.[16]

Besides technical documentation, some respondents express a more general interest in **internal documentation** of platforms, and more broadly in the facilitation of **qualitative research methods.** Such internal documentation might include, for instance, policies related to content moderation decision-making processes; worker instructions or training; internal research into the performance of algorithmic systems or user control features; or advertising and monetisation payments (e.g. to and from known sources of

[13] Stanford Internet Observatory. Weizenbaum Institute for the Networked Society Berlin. NYU's Center for Social Media and Politics ("Under FORT, researchers are given access to a sandbox environment, where they can search, filter, and conduct analysis. Researchers can export results of their findings, but not export any of the actual data. This tool can serve as a model for sandboxed environments that can provide multi-platform data for researchers").

[14] e.g. Weizenbaum Institute for the Networked Society Berlin. Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. A book chapter by Wood et, submitted by The Data Co-Ops Project provides some rubrics on the relative sensitivity of different requests in the context of differential privacy. See: Wood, A., Altman, M., Nissim, K., Vadhan, S. (2020), "Designing Access with Differential Privacy", in: Shawn, C., Dhaliwal, I., Sautmann, A., and Vilhuber, L. (eds), Handbook on Administrative Data for Research and Evidence-Based Policy, Cambridge MA: Abdul Latif Jameel Policy Action Lab (p. 207-208). Further discussion on the sensitivity of different forms of personal data can also be found in: EDMO, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access.

[15] Lujain Ibrahim et al. (Oxford Internet Institute).

[16] e.g. Amsterdam School of Communication Research (ASCoR). Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Ludwig-Maximilians-University Munich (Germany). Weizenbaum Institute for the Networked Society Berlin. GDR Internet IA et Société – CNRS.

disinformation).[17] Others, going further, also propose that researchers be given the opportunity to interview relevant platform employees.[18]

Several researchers also indicate the importance of **experimental** research, where the goal is not merely to *observe* the service but to *test* it through interventions. In particular, **A/B-testing** is a common form of experimental research, in which different versions of a service feature are deployed (simultaneously) within subsections of the user base. Such experimental methodologies can be important in order to develop causal theories about platform behaviour and to better understand the impact of specific policies or design choices, for instance in the context of UX-design or content ranking.[19] Responses refer to 'experiments' in various ways, which are worth clarifying here:

- **Platform-initiated experiments**: Several researchers express an interest in gaining access to internal platform testing results. This is related to the above point about access to internal documents. Leveraging insights or data from these existing tests, some argue, may be less burdensome on platforms than requiring them to carry out additional tests.[20]
- **Researcher-initiated experiments**: Other researchers foresee the need to request their own A/B-testing from platforms. This would further expand the scope and ambition of potential research – not only into platforms' own design choices, but also to evaluate novel ideas for risk mitigation and trustworthy design.
- **Natural experiments:** This concept refers to observational datasets which, as a matter of circumstance, enable a relatively straightforward comparison on the effects of a specific variable. Natural experiments are therefore still an observational method; they consist simply in the selection of opportune case studies, and should not be confused with 'experiments' in the more literal sense. Since they do not always control for other relevant variables, natural experiments do not typically sustain such fine-grained causal analysis as actual A/B-testing.

Platform- and researcher- initiated experiments therefore raise the question whether Article 40 requires services to cooperate not only in the *disclosure* of data but also in the *generation* or *collection* of relevant data. The same question arises in the context of **data donation**, which is another method mentioned by several respondents.[21] Here, user participants consent to submit their data to a research project. This can be arranged independently of the platform, but several researcher respondents argue that platforms should support these methods by providing infrastructure to allow users to download

[17] Institute for Strategic Dialogue. Arcom. Platform Governance, Media and Technology Lab (University of Bremen). Lujain Ibrahim (Oxford Internet Institute). GDR Internet IA et Société – CNRS. Universidade Catolica Portuguesa on behalf of Fair MusE. 5Rights Foundation. NYU's Center for Social Media and Politics.

[18] Platform Governance, Media and Technology Lab (University of Bremen). Centre for Information and Innovation Law (CIIR), Faculty of Law, University of Copenhagen.

[19] Academic Researchers from the Massachusetts Institute of Technology, Harvard University and Northeastern University.

[20] The response from Global Partnership on AI, in particular, discusses internal A/B-testing in detail and how insights from such work can be made public in a privacy-compliant manner.

[21] University of Amsterdam (Valkenburg). Ludwig-Maximilians-University Munich (Germany). Institute for Strategic Dialogue. NYU's Center for Social Media and Politics. Amsterdam School of Communication Research (ASCoR) ('citing datadonation.eu').

their data and/or for researchers to obtain consent from participants. These methods may make use of browser plugins or existing data takeout features or **'data download packages' (DDPs).**[22] Patti Valkenburg (University of Amsterdam)'s response discusses problems of self-regulatory DPPs and how these could be improved, and NYU's Center for Social Media and Politics discusses how facilities for user content and data donation on certain platforms have deteriorated over time.[23] ASCoR proposes the creation of data donation pipelines for obtaining consent and downloading data. SNV suggests that such infrastructure should offer access to the same data that personal data that subjects are entitled to request under Article 15(3) GDPR.[24] (To clarify: Independent data donation projects, which proceed without support from the platform through solutions such as browser plugins, are often mentioned in the context of data scraping and the debate around Article 40(12), which I return to under Section 4 below.)

Several industry submissions request that the Delegated Act include a list of data categories which are excluded from the scope of Article 40. This point is discussed further in Section 2C below.

Finally, a large number of researcher respondents mention the importance of independent collection methods such as scraping or sock puppet audits. This matter is discussed under Section 4 below.

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?*

[22] University of Amsterdam (Valkenburg) provides feedback on TikTok's existing DPPs and how those could be improved.

[23] NYU's Center for Social Media and Politics ('Easier Data Donations: All VLOPs should be required to have an easy way for users to download their data and donate it for scientific research. A core focus of our research at CSMaP is to pair offline political opinions and behaviour to online activity. We conduct surveys asking respondents their views on certain political topics, how they voted, etc. We then ask them to donate their social media and other online data. This combination of surveys and digital trace data allows us to draw connections between what they see online and what they do offline, and vice versa. For example, our Bilingual Election Monitor project uses this method to explore the attitudes of Spanish-speaking social media users in the U.S. Currently, downloading data is a cumbersome process. In the past, Facebook had a process to allow users to quickly give researchers access to their data. We were able to use data donated in this way to show that older people were more likely to share low quality news on Facebook. The DSA should make it easier for people to do this by requiring platforms to create a mechanism similar to Facebook's prior system — using the same uniform standards outlined above. In addition to aiding scientific research, it will also empower users to have more ownership over their digital data and decide whether it should be used for the public good.'). See also Women in AI Austria ('In our view, it is furthermore important that researchers using qualitative methods (e.g. digital ethnography) can rely on established procedures for obtaining consent, i.e. by asking the communities/fora/group chats/etc. for their consent. This practice should be retained and strengthened in the sense that when researchers obtain consent from a group active on a platform to conduct qualitative research using publicly available (or mixed) data, platforms cannot prevent the use of this data for research purposes.')

[24] Stiftung Neue Verantwortung e. V. (SNV).

Respondents expressed research interests in a variety of **domains**. The research topics and domains mentioned include:

- Political communications (e.g. polarisation, online elections and campaign);[25]
- Disinformation;[26]
- Hate speech;[27]
- Mental health and the protection of minors;[28]
- Advertising and consumer protection.[29]

Notably, many researcher commentaries tend to focus not so much on specific domains, harms or risks as much as they do on specific **platform *systems***, such as content recommendation and amplification, content moderation, and targeted advertising. Recommender systems and algorithmic amplification in particular are the subject of many responses, as is advertising. A recurring theme is the need to study how recommender systems work in relation to individual users or what Lewandowsky et al. refer to as 'human-algorithm entanglement' – how interactions between user and the platform recursively shape each-other and create feedback loops.[30] To this end, researchers are interested in studying what the Data Co-Op Project describes as the 'outgoing and incoming vectors' of recommender systems; the information they distribute *to* users alongside the information they collect *from* users to personalise those offerings.[31] Given this importance of user signals and inputs in the recommending process, Stiftung Neue Verantwortung states the need for 'user simulation' facilities through which researchers can 'automatically simulate users and user inputs to be able to obtain data on what actual users experience by interacting with the platform'.[32] Stiftung Neue Verantwortung also provide an example of such user-algorithm interaction research, in the context of mental

---

[25] e.g. Global Partnership on Artificial Intelligence (Social Media Governance project). Institute for Strategic Dialogue. CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley. Amsterdam School of Communication Research (ASCoR). Universidade Catolica Portuguesa on behalf of Fair MusE. Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Trust Lab; NYU's Center for Social Media and Politics. vera.ai (EU Horizon Europe Research and Innovation Project). European Digital Media Observatory (Task Force on Disinformation on the War in Ukraine). Julia Angwin. Weizenbaum Institute for the Networked Society Berlin.

[26] Global Disinformation Index. EU DisinfoLab. Institute for Strategic Dialogue. NORDIS - Nordic Observatory for Digital Media and Information Disorder. Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. Goodly Labs. Avaaz.

[27] Analyse & Tal. Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Weizenbaum Institute for the Networked Society Berlin. NYU's Center for Social Media and Politics.

[28] University of Amsterdam (Valkenburg). Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. 5Rights Foundation.

[29] Verbraucherzentrale Bundersverband. University of Amsterdam (Milan, Agosti and Beraldo). ALLAI. Panoptykon Foundation.

[30] Ayelet Gordon-Tapiero, Alexandra Wood & Katrina Ligett, "The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization," Proceedings of the 2nd ACM Symposium on Computer Science and Law (CSLAW'22) (2022), https://dl.acm.org/doi/10.1145/3511265.3550450; included as appendix to The Data Co-Ops Project. See also Lujain Ibrahim (Oxford Internet Institute) on the use of user control settings. Global Partnership on Artificial Intelligence (Social Media Governance project), citing Jiang, R., Chiappa, S., Lattimore, T., György, A., & Kohli, P. (2019). Degenerate feedback loops in recommender systems. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 383–390).

[31] The Data Co-Ops Project.

[32] Stiftung Neue Verantwortung e. V. (SNV).

health, based on journalists' research into 'rabbit holes' of depressive content on certain platforms using automated bot inputs.[33]

More specific ideas for research projects with relevance for systemic risks from the call for evidence include the following:

- Global Partnership on AI offers a detailed proposal on the study of recommender systems leveraging access to internal platform A/B-testing, and how this might fuel the dissemination of harmful content including political misinformation, extremism, or domains promoting eating disorders. Their proposal draws inspiration from a self-regulatory study based on voluntary Twitter access partnerships with Huszár et al. 2020.[34]
- The Stanford Internet Observatory discusses various research projects conducted via the self-regulatory Twitter Moderation Research Consortium (TMRC), primarily concerning state actor influence networks .[35]
- The Panoptykon Foundation describes lessons learned from their "Algorithms of Trauma" research into harmful health-related advertising, and data needs for future followups.[36]
- Lujain Ibrahim et al of the Oxford Internet Institute offer a detailed discussion of user controls in particular, pertaining to recommender and moderation systems, investigating their usage rates as well as effects.[37]
- Dublin City University lists possible research into the protection of minors (evaluating the effectiveness of child rights assessments and mitigation measures, e.g. via survey methods; studies on the efficacy of age-assurance measures; and compliance with proactive AI-based cyberbullying moderation) as well as content governance (e.g. on freedom of expression in relation to government and other third party takedown requests; research into the effectiveness of pre-screening processes for content such as CSAM; and research into sexual violence and harassment).[38]
- ISD lists numerous indicative research questions to underscore the breadth and diversity of possible research topics (listed in footnote).[39]

---

[33] Stiftung Neue Verantwortung e. V. (SNV) (also citing the Wall Street Journal).

[34] Global Partnership on Artificial Intelligence (Social Media Governance project), citing: Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). *Algorithmic amplification of politics on Twitter.* Proceedings of the National Academy of Sciences, 119(1). doi: 10.1073/pnas.2025334119.

[35] To wit ● My Heart Belongs to Kashmir (September 2022) ● Unheard Voice (August 2022) ● The New Copyright Trolls: How a Twitter Network Used Copyright Complaints to Harass Tanzanian Activists (December 2021).

[36] Panoptykon Foundation.

[37] Lujain Ibrahim (Oxford Internet Institute) (requesting "sharing data on the nature, usage, and effectiveness of human-algorithm interactions—in particular user controls pertaining to recommender & moderation systems […] In addition to engagement data and recommendation data, additional data needed to study user controls may for instance include: a. Documentation of implemented user controls and the outcomes they lead to when used; b. Usage data of user controls & feedback signals; c. Data and metadata for content flagged by users through controls, and the resulting automated decisions").

[38] Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub.

[39] Institute for Strategic Dialogue. ('• "What is the prevalence of content that could be classified as "incitement to hatred" under the German penal code on Facebook? • How many views did video clips of RT and Sputnik broadcasting activities receive on YouTube one month prior and one month after the Russian invasion of Ukraine? • How effective are warning labels from independent fact-checkers or authoritative sources in reducing the spread of misinformation on Twitter? • What types of users are more likely to be exposed to hate speech

- UC Berkeley emphasises the importance of experimental research (as discussed in Section 1A) in order to address causal questions, which are often central in risk assessment and mitigation.[40]
- EU DisinfoLab lists several areas of interest and also refers to several past projects as possible sources of inspiration, including research into the prevalence of climate change disinformation via comments under social media posts; their own investigation into the Russian 'Doppelganger' project which disseminated false articles, links and so on; and their own investigation of YouTube Disinformation Entrepreneurs, which traced revenue streams for pro-Russian disinformation channels.[41]
- Snap Inc. suggests that the effectiveness of relevant performance metrics used in relevant reporting under the DSA would be useful to have researched.[42]

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?*

Many respondents emphasise the importance of **timeliness** in the data access procedure.[43] There are concerns that overly complex or protracted proceedings would deter researchers and lead to the procedure being underutilised. Several respondents request that the Delegated Act provide greater clarity as to what constitutes a 'reasonable period' for the handling of requests.[44] Industry respondents support a case-by-case

across different platforms? • Do moderation decisions about what content is allowed on a platform affect some user groups disproportionately? • Are Instagram's 'Explore' page algorithms systematically amplifying the visibility of cyberabuse content? • What is the proportion of so-called 'superusers' that show hyperactive and abusive behaviour on Facebook? How can we measure the effect of 'superusers' on algorithmic feeds? • Are high-profile users treated preferentially in content moderation processes? • Are TikTok's algorithms intentionally demoting Black Lives Matter activists, i.e., reducing how frequently their videos appear on the 'For You' feed? • Are users able to silence others through the misuse of moderation tools or through systemic harassment designed to censor certain viewpoints?).

[40] CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley.

[41] EU DisinfoLab.

[42] Snap Inc.

[43] e.g. Ministry for Industry, Business and Financial Affairs (Denmark). Trust Lab. Institute for Strategic Dialogue. CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley. AI Forensics. The Mozilla Foundation. European Digital Media Observatory (Task Force on Disinformation on the War in Ukraine). Centre for Democracy & Technology, Europe Office. University of Amsterdam (Van Drunen & Noroozian). University of Amsterdam (Borra, Peeters & Rieder). NORDIS - Nordic Observatory for Digital Media and Information Disorder.Institute for Research on the Information Environment; Princeton University.

[44] e.g. Ministry for Industry, Business and Financial Affairs (Denmark) ('due regard to the difficulty in retrieving the relevant data on one hand and the deadlines for and timeliness of the research on the other hand.')

assessment.[45] Google's response proposes several factors that might be taken into account in this assessment: "the sensitivity or potential sensitivity of the data (which may warrant additional investigation and implementation of security measures, such as aggregation or anonymisation, and/or additional review and data validation by the VLOP or VLOSE before providing the data); the volume of the data and the range of data types requested (which would increase the time needed to collect, validate and present the data in a safe manner); practical hurdles in identifying, collecting and transferring the data; and the total volume of requests which have been submitted to (and are pending for) the same provider, and the resources needed to respond to those requests".[46] Booking.com requests that the Delegated Act install a minimum period of one month, with possibilities for extension pursuant to reasoned and timely request.[47]

As regards timing, researchers also warn of challenges related to **funding**. Article 40(8) requires researchers to demonstrate that they possess the funding necessary to carry out research, but researchers object that it may be difficult to obtain research funding for such projects before data access is secured.[48] Several solutions are suggested for this 'Catch-22' scenario: the University of Amsterdam (Borra, Peeters and Rieder) recommends that the vetting process be aligned with application procedures for ERC funding instruments.[49] The Centre on Regulation in Europe (CERRE) proposes the development of dedicated 'DSA Grants' to fund research in this space.[50] Arcom recommends that funding allocations are decided on by other bodies than DSCs, in order to safeguard researcher independence.[51]

Besides timeliness, researchers attach great importance to **independence**. To preserve the autonomy and objectivity of academic research, many researchers wish to ensure that providers of platforms are not put in a position to decide over research projects or applications.[52] CNRS observes that academic freedom is enshrined as a fundamental right under Article 13 of the EU Charter on Fundamental Rights, as well as being protected under many national constitutional regimes, including that of France.

To protect their independence, researchers offer several suggestions. The most common proposal is an emphasis on **peer review** during the application process, so that questions of methodology and research ethics are assessed by other independent academics.[53] Most of these responses propose that the **independent advisory mechanism** be involved in organising these peer review tasks (discussed further in Section 2(e) below). The Weizenbaum Institute differs on the issue of peer review, warning that this process often

[45] Computer & Communications Industry Association (CCIA Europe). DOT Europe. Google.
[46] Google.
[47] Booking.com.
[48] University of Amsterdam (Borra, Peeters & Rieder).
[49] University of Amsterdam (Borra, Peeters & Rieder).
[50] Centre on Regulation in Europe (CERRE).
[51] Arcom. ('Ces mesures de renforcement ne sauraient être mises en place sans les financements correspondants. Les CSN ne peuvent néanmoins pas intervenir dans l'attribution de ces financements afin de veiller à l'indépendance de la recherche et d'éviter les conflits d'intérêt.')
[52] e.g. Stiftung Neue Verantwortung e. V. (SNV). Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Institute for Strategic Dialogue. Coalition for Independent Technology Research. CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley. EU DisinfoLab. Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. RESET. University of Michigan Center for Social Media Responsibility.
[53] Ibid.

leads to delays. They prefer to leave the DSC responsible for assessing research applications, although they do agree that their work should be supported by an independent advisory body.[54]

Peer review safeguards independence but also **expertise**. There is a concern that DSCs acting alone will struggle to assess research proposals on aspects such as methodology, technical safeguards and research ethics. Besides involving peer reviewers and other independent advisory mechanisms, respondents therefore also emphasise the importance of staffing and capacity-building at DSCs.[55] SNV and Dublin City University propose that DSCs install dedicated **data units** with specialised skills in the handling of data access requests.[56]

Researchers as well as industry representatives both mention the importance of occasions for **dialogue** between providers of platforms, DSCs and researchers throughout the access process. Due to the complexity of the subject matter, formulating actionable data access requests may often require coordination with or feedback from the disclosing party.[57] Industry responses request the opportunity for early-stage dialogue with researchers and DSCs prior to formal applications. Researchers also emphasise opportunities for dialogue with DSCs, as well as points of contact at relevant platforms (often associated with technical documentation, see 1a above).[58] Researchers' views here are not entirely uniform, and tensions arise with the demand for independence. They tend to emphasise dialogue with DSCs rather than direct dialogue with platforms, and some respondents wish to minimise platforms' role in the application process.[59] The University of Helsinki recommends that researchers be given the opportunity to remain anonymous vis-à-vis the platform throughout the process, in order to minimise the risk of gatekeeping or chilling effects.[60]

Several other suggestions are made to enable speedy and efficient handling of requests:

There is broad support for **transparency** in the application process, amongst researchers as well as industry. Specifically, many respondents propose that every researcher application and DSC request be published.[61] This type of transparency is supported on at

---

[54] Weizenbaum Institute for the Networked Society Berlin.

[55] Stiftung Neue Verantwortung e. V. (SNV). Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Centre for Democracy & Technology, Europe Office. NORDIS - Nordic Observatory for Digital Media and Information Disorder. 5Rights Foundation.

[56] Stiftung Neue Verantwortung e. V. (SNV) (also citing: Julian Jaursch, Here Is Why Digital Services Coordinators Should Establish Strong Research and Data Units.) Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub.

[57] e.g. Booking.com. CCIA Europe. DOT Europe. Google. Coalition for Independent Technology Research. Mozilla Open Source Audit Tooling (OAT) Project. Daphne Keller.

[58] Weizenbaum Institute for the Networked Society Berlin. Coalition for Independent Technology Research. Verbraucherzentrale Bundersverband.. Arcom. Sciences Po médialab.

[59] e.g. NYU's Center for Social Media and Politics.

[60] University of Helsinki ('The anonymity of the researcher(s) to the VLOP VLOS should be ensured for independent and impartial analysis of the data'.)

[61] Weizenbaum Institute for the Networked Society Berlin. Arcom. Centre for Democracy & Technology, Europe Office. CCIA Europe ("All decisions made by DSCs within the framework of the data access regimes should be detailed, reasoned, and public (subject to redaction for the protection of users' and businesses' rights), so as to enable challenge by providers.'').

least two grounds. First, it can contribute to speedy and efficient exchanges by helping researchers to understand how other (successful and unsuccessful) applications are designed. Second, it can help to ensure accountability and legitimacy of the procedure, by allowing third parties to check how Article 40 is being implemented in practice.

Relatedly, several researchers propose that efficiency can be enhanced by **streamlining applications** for **commonly-requested data** and for **repeat players**. Designing an effective research API may entail up-front investments, but once it is in place the marginal costs of granting access to additional researchers are comparatively minor. By clearly establishing the eligibility requirements for specific APIs or other infrastructures, they could become accessible to a large number of researchers.[62] This streamlined access would also be beneficial for purposes of **replicability**, so that third party researchers may reassess the quality of previous research projects.[63] Other researchers propose a comparable streamlining for the vetting of repeat players; after successfully demonstrating their capacities as researchers in one request, a second request would not necessarily need to repeat each point in full detail - only those which are particular to the new request. EU DisinfoLab, for instance, refer to this as a **two-track system**, with one track for vetting research proposals and another for vetting organisations, which then maintain the status for a longer term.[64] Similarly, Stiftung Neue Verantwortung (SNV) proposes that researchers and institutions should be able to apply as such, independent of separate search projects, for purposes of exploratory research (a concept discussed further in Section 3d below).[65] Google's submission, by contrast, insists that the 'delegated act should also specify that DSCs should not rely on representations made by researchers in prior applications, regardless of whether those prior applications were granted'.[66] This would in effect require *de novo* review of each request, and would likely rule out such streamlining.

Given the complexity of handling applications, especially at the initial stage before standardised infrastructures are developed, several parties propose a **phased rollout** in which resources are initially concentrated on a small number of high-priority projects.[67] On that basis, best practices could be refined iteratively, gradually expanding data access in terms of scope and depth.

---

[62] e.g. Centre for Democracy & Technology, Europe Office. CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley. Amsterdam School of Communication Research (ASCoR). Universidade Catolica Portuguesa on behalf of Fair MusE. Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access.

[63] e.g. Stanford Internet Observatory. Amsterdam School of Communication Research (ASCoR).

[64] EU DisinfoLab. See also European Digital Media Observatory (Task Force on Disinformation on the War in Ukraine) ('To minimise administrative overheads, once researchers are cleared for data access, they shouldn't have to re-apply on a per-project basis').

[65] Stiftung Neue Verantwortung e. V. (SNV).

[66] Google.

[67] e.g. Stanford Internet Observatory. CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley. Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. Brandon Silverman. Arcom. Snap Inc.

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?*

Many respondents, from industry as well as the research community, emphasise the need for consistency and predictability in the vetting process. Greater clarity is desired from the delegated act on the vetting requirements listed in Article 40(8), in terms of their substantive and evidentiary requirements. Respondents suggest various means to **standardise applications**: (1) standardised application forms, (2) standardised data access agreements, and (3) standardised non-disclosure agreements. EDMO has developed a model data access agreement.[68]

The Denmark Ministry for Industry, Business and Financial Affairs has submitted a study on **best practices** for researcher vetting, which includes an analysis of vetting procedures for administrative and genomic data.[69] Naomi Shiffman provides a list of information that researchers should disclose as part of the vetting procedure, based on her experience in overseeing the vetting of researchers at CrowdTangle.[70] In addition, Shiffman proposes that guidance can be provided by way of **criteria grids,** which include examples of what would qualify, what would not qualify, and what occupies the grey area (examples Shiffman gives include 'the RAND Corporation, a military-affiliated research institution such as a Naval academy, or the national university of an authoritarian government'. According to Shiffman, 'this will be a common type of applicant for DSCs, and they should have a consistent approach to dealing with them'.) Furthermore, these criteria grids could also list 'common "red flags" (for example, no published research), and green flags (for example, an institution whose research consistently includes robust methods sections and lists the limitations of its own findings)'.[71]

Regarding **(a) affiliation with a research organisation,** questions are raised inter alia about the nature of eligible 'affiliations', as well as the status of non-university researchers, non-EU organisations and researchers.

Regarding the concept of affiliation, Booking.com highlights the possibility of multiple affiliations and requests the delegated act to clarify that researchers should disclose 'any and all affiliations as part of their application, to provide VLOs with visibility regarding the research organisations with whom findings may be shared'.[72] Reset Tech asks for clarification on two hypotheticals: '(1) A journalist who is based outside the EU signs a

[68] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access (citing their model data access agreement).

[69] Social Observatory for Disinformation and Social Media Analysis, Evaluating Safe space solution including data management and processing setup - Modelling academic social media data safe spaces based on administrative and genomic data management From unit level to access level.

[70] Naomi Shiffman ('As part of the vetting process, researchers should be required to share: ● Name, location, and institutional affiliation; ● Image of national identification card to prove identity; ● Major sources of funding; ● Links to previously published research ● Overview of proposed research process, including possible research questions, exploratory research requirements, a description of the type of data they hope to obtain from the platform, potential alternative data types')

[71] Naomi Shiffman.

[72] Booking.com.

contract and enters a data sharing agreement with an EU-based research organisation as part of a broader research consortium. (2) A US journalist who is on sabbatical for two years and works remotely, but full time, for a European research organisation.'[73]

Non-academic respondents including NGOs and media organisations advocate for the importance of enabling these **non-university researchers** to participate as vetted researchers.[74] Witness proposes that the Delegated Act should clarify that non-peer reviewed research outputs also fall within the scope of Article 40.[75] Some industry respondents suggest that a track record of peer reviewed publications be included as evidentiary requirement for vetting.[76] For Witness, this would have the potential to exclude non-university researchers, who may not prioritise conventional peer review publications.[77]

Several respondents advocate that the vetting process should be accessible to **non-EU researchers**.[78] Stanford Internet Observatory requests guidance on 'access for non-EU residents who may have affiliations with qualified EU organisations.'[79] Going further, legal analysis by Martin Husovec, by the University of Copenhagen and by the University of Bremen argue, based on relevant language in the Copyright in the Digital Single Market Directive, that the concept of a 'research organisation' is without territorial restrictions.[80] These respondents do note that the territorial scope of Article 40 is confined to the study of 'systemic risks in the EU', a point which I discuss further under 2D below.

Regarding **commercial independence** (40(8)(b)) and **funding transparency** (40(8)(c)), many respondents request clarity on the substantive and evidentiary requirements. DOT Europe proposes that commercial independence requirements should exclude researchers, including academics, who have major projects funded by competitors – taking into account not only the research at issue in the vetting application, but also 'the relevant wider funding of the researcher and/or the research institution of which s/he forms part.'[81] To this end, the researcher application should disclose, according to DOT Europe, 'the source, nature, scope and conditions of such funding, going beyond the research in question and extending to the funding of the wider research of the researcher or institution)'.[82] Booking.com requests that the terms of data access expressly prohibit any commercial usage of the data or insights derived therefrom, and, to this end, 'historic, existing or potential future relationships with other online platforms, competitors or other stakeholders which may pose risks concerning the independence or confidentiality of the research activity'.[83] EU DisinfoLab argues for a lenient interpretation as regards

[73] RESET.
[74] e.g. Witness. EU DisinfoLab. Institute for Strategic Dialogue.
[75] Witness.
[76] e.g. Google.
[77] Witness.
[78] e.g. Centre for Democracy & Technology, Europe Office. The Data Co-Ops Project.
[79] Stanford Internet Observatory.
[80] Martin Husovec. Platform Governance, Media and Technology Lab (University of Bremen). See also: The Data Co-Ops Project. RESET.
[81] DOT Europe.
[82] see also Snap Inc. ('It should be a critical element of the application process for researchers to demonstrate that they are not affiliated with competitor or activist organisations.')
[83] Booking.com.

private funding, since CSOs tend to rely on private funding for much of their activity.[84] They also request that funding information should be shared exclusively with the vetting body and not with the public at large, for fear that this information could be used to 'harass civil society'.[85] There may be tensions here with the more widely-voiced proposal, discussed in Section 2A, that vetting applications and decisions be made public.

CDT requests that the delegated act close any loopholes which might enable law enforcement entities to make use of Article 40.[86]

The following sections discuss in further detail issues related to subsequent vetting criteria: Section 2C discusses data security, confidentiality and technical and organisational measures (see Article 40(8)(d)). Section 2D discusses purpose limitation and proportionality (see Article 40(8)(e) and (f)).

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?*

Respondents mention both technical and legal safeguards that can minimise risks and help to balance relevant rights. Many submissions, including EDMO Researchers', propose a tiered approach in which the stringency of safeguards is determined based on the risks associated with this data.[87]

Regarding **data protection and privacy**, the most extensive guidance in this space is offered by the EDMO's working report for a Code of Conduct on researcher access, which is referenced in their response and is also endorsed by several other respondents.[88]

A minimal technical safeguard for managing privacy risks is **anonymisation/pseudonymisation**, as a default measure for disclosed data. Re-identification of individuals remains a risk despite anonymization, and the use of various **privacy-enhancing technologies (PETs)** is possible to mitigate this risk, including differential privacy, k-anonymity and synthetic data.[89] K-anonymity, in short, reduces the chance of singling out individuals in a dataset by either excising specific values or replacing them with broader ranges. Differential privacy is a more rigorous and restrictive method for avoiding reidentification, which revolves, in short, around introducing minor errors ('noise') into the dataset.[90] There is disagreement, however, as to the merits of

[84] EU DisinfoLab.
[85] EU DisinfoLab.
[86] Centre for Democracy & Technology, Europe Office.
[87] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access.
[88] EDMO, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access.
[89] Centre on Regulation in Europe (CERRE). Goodly Labs.
[90] The Data Co-Ops Project.

these methods. For differential privacy and synthetic data, researchers have warned that they may hamper and/or distort meaningful findings.[91] Work by EDMO has highlighted problems with differential privacy in a self-regulatory context.[92] Work by Altman et al (2022) submitted to the consultation questions the efficacy of k-anonymity.[93]

In sum, further research into the relative merits of these methods may be necessary, and their applicability may differ on a case-by-case basis depending on researcher needs and the nature of the data involved. The Data Co-Ops project has submitted a visualisation plotting this balancing of access restrictions versus data sensitivity, which is taken from the context of government data reuse.[94]

Other privacy-protective technical restrictions that are mentioned include clean rooms (virtual or physical), data vaults, sandboxes, and/or virtual lab environments. As discussed in Section 1A above, these methods focus not on perturbation of relevant data but rather on maintaining exclusive stewardship and avoiding duplication onto researcher machines. CDT also requests that providers of platforms and researchers 'put in place pragmatic mechanisms for upholding the spirit of transparency and for accomplishing notification of data subjects that their data has been shared with researchers', pursuant to Article 21 GDPR.[95] (It is worth noting that derogations to this right may apply pursuant to the exceptions for public interest, scientific or historical research purposes or statistical purposes, pursuant to Article 89 GDPR and Article 21(6) GDPR.[96])

In addition to technical restrictions, many respondents also mention **data management plans** or **data access agreements** which may take the form of contractual commitments, including commitments not to use the data for purposes outside the proposed research plan, and specifically to refrain from attempts to re-identify individuals from anonymised/pseudonymised sources. A best practice mentioned by researchers as well as industry respondents is the **activity logging** of researcher queries, whether through APIs or other access methods such as data vaults.[97] In this way, the disclosing provider can monitor for problematic and potentially abusive activity, and better enforce relevant data access agreements.

---

[91] Lujain Ibrahim et al. (Oxford Internet Institute) cite criticism of differential privacy and synthetic data solutions, stating that 'researchers have expressed strong concerns that the validity of research findings may be altered by privacy-preserving techniques, distorting statistical inferences and increasing disparities in outcomes for racial minorities.' Citing e.g.: Hauer, M.E. and Santos-Lozada, A.R., 2021. Differential privacy in the 2020 census will distort COVID-19 rates. Socius, 7, p.2378023121994014. Santos-Lozada, A.R., Howard, J.T. and Verdery, A.M., 2020. How differential privacy will affect our understanding of health disparities in disparities in the United States. Proceedings of the National Academy of Sciences, 117(24), pp.13405-13412 . Privacy Loss in Apple's Implementation of Differential Privacy on MacOS 10.12. J. Tang, A. Korolova, X. Bai, X. Wang, and X. Wang. CoRR (2017). Stadler, T. and Troncoso, C., 2022. Why the search for a privacy-preserving data sharing mechanism is failing. Nature Computational Science, 2(4), pp.208-210.

[92] EDMO, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access.

[93] The Data Co-Ops Project.

[94] The Data Co-Ops Project (reproducing a figure from Micah Altman, Alexandra Wood, David R. O'Brien, Salil Vadhan & Urs Gasser, "Towards a Modern Approach to Privacy-Aware Government Data Releases," 30 *Berkeley Technology Law Journal* 1967 (2015).)

[95] Centre for Democracy & Technology, Europe Office.

[96] For further analysis on this point, see EDMO, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access.

[97] e.g. CCIA. The Data Co-Ops Project.

Regarding business confidentiality, trade secrets and security, there are not many technical solutions proposed. Responses from industry as well as researchers tend to highlight **non-disclosure agreements (NDAs)** as a key means of managing risk in this space. To enforce these agreements, certain technical measures are also mentioned. Activity logging, mentioned previously, can help flag potential risks to business confidentiality and security. CCIA proposes procedures for **pre-publication review** to assess compliance with access terms and to avoid abuse.[98]

Concerning the substance of disclosures, there are diverging interpretations as to how far protection reaches for confidential information, including trade secrets. Some industry respondents request that confidential information including trade secrets should be exempted entirely from disclosure (in apparent contrast to the statutory language of Article 40(5), which refers to 'significant vulnerabilities').[99] DOT Europe insists that trade secrets 'should ordinarily be out of scope for vetted researcher data access requests, save in exceptional cases, and even then, subject to enhanced safeguards'.[100] Other industry respondents, such as Snap Inc, do not seek such a categorical exemption for trade secrets, and acknowledge that 'platforms may have to hand-over business critical information'.[101] Researchers generally advocate for a more restrictive interpretation of trade secrets limitations, expressing concerns that an overly broad interpretation will be abused to unduly limit the potential scope of Article 40. For CERRE, '[e]ven if a certain dataset is highly commercially sensitive and trade secret protected, we believe it should still be possible to mandate its disclosure to a researcher under Article 40(4) of the DSA'.[102] Such an analysis, for CERRE, is to be made on a case-by-case basis, weighing all relevant rights and interests involved, including the sensitivity of the data, applicable safeguards, and the urgency and importance of the systemic risk research being conducted.[103] The Denmark Ministry for Industry, Business and Financial Affairs similarly advocates that trade secrets be excluded only 'under exceptional circumstances'.[104] In addition, CERRE and EDMO Researchers both propose to clarify that the burden of proof to invoke these limitations must rest with the providers invoking them, rather than with researchers.[105] Snap Inc, by contrast, instead proposes that the protection of business' rights be the 'default position'.[106]

Security, as a ground for platforms to refuse access, received the least attention in the consultation responses. The most detailed treatment is by CERRE, particularly in the in-depth report underlying their short-form submission.[107] They propose a 'threat modelling' approach, taking into account what data are being disclosed, to whom, and what data these parties already possess.[108] In addition, they distinguish between **data security** (concerns related to the unauthorised accessing of data during the access process) and **systems security** (concerns related to the unauthorised use of the service/system

[98] CCIA.
[99] e.g. Google.
[100] DOT Europe.
[101] Snap Inc.
[102] Centre on Regulation in Europe (CERRE).
[103] Ibid.
[104] Ministry for Industry, Business and Financial Affairs (Denmark).
[105] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access.
[106] Snap Inc.
[107] CERRE report 'Access to Data and Algorithms: For an Effective DMA and DSA Implementation'.
[108] Centre on Regulation in Europe (CERRE).

enabled by the disclosure of data). Data security concerns might include, for instance, the use of inappropriate or outdated encryption protocols or authorisation mechanisms when disclosing data to researchers. System security concerns, by contrast, might include the divulging to researchers of security-sensitive information such as user passwords. Related to systems security, DOT Europe proposes 'particularly careful scrutiny over any application seeking access to data relevant to the operation of the controls environment of any given VLOP/VLOSE, given the potential for loss of data to compromise the integrity of such controls.'[109]

Across all these issues, the legal-procedural question arises whether the delegated act should prescribe specific categories of data as either eligible or ineligible for data access requests. Some respondents seem to prefer a more flexible, case-by-case analysis, whereas others call for clarification as regards specific types of data. For instance, Google requests that 'the delegated act should provide guidance about the types of data that are excluded from Article 40(4) requests, such as 'confidential information, by providing an example list of excluded data, including trade secrets and competitively sensitive information, source code, privileged data, contractually protected data, data necessary for the security or integrity of the platform, machine learning or other algorithmic model coefficients, and internal documentation (e.g. decisions or memos)'.[110] CDT has submitted a comparable list.[111] On substance, as discussed, this request is already in tension with demands that trade secrets should only lead to a refusal of access in exceptional circumstances. But at a procedural level, this proposal also illustrates how the delegated act could be more or less specific in detailing Article 40's substantive scope. The Slovak Council for Media Services proposes that platforms publish their own non-exhaustive lists, of (meta) data they will not disclose under the DSA regime (art 40 (5))', as well as 'provide reasons for these decisions and specify alternative sources of data (art 40 (6))'.[112] In addition to black lists of *ineligible* data, the delegated act may in theory also introduce (non-exhaustive) white lists of *eligible* data,[113] or grey lists creating rebuttable presumptions either for or against eligibility. CDT and VZBV both emphasise that such lists would need to be non-exhaustive, to allow sufficient flexibility, nuance and adaptability in light of changing conditions.[114]

---

[109] DOT Europe.

[110] Google.

[111] Centre for Democracy & Technology, Europe Office ('certain data may be too sensitive to ever share with researchers: personally identifiable biometric information; precise geospatial information; personally identifiable information about children under the age of 13; information revealing an individual's physical or mental health diagnosis; log-in credentials' information identifying an individual's sexual orientation or sexual behaviour; and phone or text logs, photos, audio recordings, or videos, maintained for private use by an individual. Moreover, the sharing of user-generated content, whether posted publicly or privately, can present unique risks of revealing this kind of sensitive information, if the user has included it in the text, image, or video that they have posted. It may not be possible for platforms to adequately mask this information when sharing content with researchers, so content data may necessarily need to be shared in data clean rooms or otherwise under heightened privacy and security measures.')

[112] Council for Media Services (Slovakia).

[113] Many researcher respondents include such non-exhaustive lists, and these are synthesized in Section 1A above.

[114] Verbraucherzentrale Bundesverband ('In principle, it is common ground that it is not possible to list exhaustively what data may be useful. In particular, it is important to bear in mind that functionalities are constantly changing. For example, formats such as shorts or rocks [sic] are new queried features. These should, of course, also be included, so that no new blind spots are created for researchers.'). Centre for Democracy &

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?*

Many relevant technical and legal safeguards are already outlined in the previous section. As regards **research purposes** and how these are defined and enforced, respondents also expressed differing views. Article 40(4) limits data access requests 'for the sole purpose of conducting research that contributes to the detection, identification and understanding of systemic risks in the Union, as set out pursuant to Article 34(1), and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures pursuant to Article 35'. Relatedly, Article 40(8)(f) requires that 'the planned research activities will be carried out for the purposes laid down in paragraph 4;' and (e) that 'their access to the data and the time frames requested are necessary for, and *proportionate to*, the *purposes* of their research, and that the expected results of that research will contribute to the purposes laid down in paragraph 4'. There is evidence of diverging interpretations on these points in the call for evidence.

On the one hand, industry respondents advocate a relatively strict interpretation of this purpose limitation and proportionality. CCIA warns that 'overly broad data requests from vetted researchers (only to decide at a later stage which subset is relevant) would go beyond the intended scope of this provision' and advocates that 'data access requests should be strictly limited to the specific data requested'. CCIA, DOT Europe and Booking.com all wish to exclude 'fishing expeditions' from the scope of Article 40.[115] To address this, they demand that researcher applications clearly state what data is requested, and how it helps to contribute to the purposes mentioned by Article 40. Going further, Snap Inc proposes that research questions should be limited to those aligned with research objectives set by the European Commission.[116]

Researchers, by contrast, tend to view an overly restrictive purpose limitation as problematic. Many of them emphasise the importance of **exploratory research**, which does not necessarily focus on testing specific hypotheses but rather on generating new theories or hypotheses. This type of research is considered important for platform ecosystems precisely due to a lack of data access, which leads to a situation in which, as per EDMO Researchers' submission, 'we often don't know what we don't know'.[117]

Technology, Europe Office ('It will be difficult, and counterintuitive to the iterative nature of the due diligence obligation of the DSA for the delegated act to develop a closed list of the types of data, metadata, documentation, and other information that researchers might need in advance.')

[115] CCIA. DOT Europe. Booking.com.

[116] Snap Inc. ('The EC should develop an overarching EU-wide research strategy on the systemic risks and mitigation measures to ensure consistent assessment of the applications. Only those which are aligned with these research objectives and can demonstrate contribution to them should be allowed.')

[117] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. See also University of Michigan Center for Social Media Responsibility. Stiftung Neue Verantwortung e. V. (SNV). Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's

Exploratory analysis can therefore help to uncover unanticipated risks, and formulate more targeted research strategies for anticipated risks.[118] In this research context, the concept of 'fishing expeditions' may therefore be inapposite, since it originates from the context of more narrowly-targeted criminal investigations (where inquiries are typically limited to specific grounds of suspicion for a particular offence). An approach based on avoiding 'fishing expeditions' therefore seems to be at odds with researchers' demands for exploratory and inductive analysis.

Some respondents view Article 40(12) in particular as an important avenue for exploratory research and monitoring.[119]

Calls to enable exploratory research also connect to the demand for automated, API-based disclosure explored in the sections above. Referring to existing platform practices, Stiftung Neue Verantwortung proposes that researchers be accredited not on specific hypotheses but on research *interests*, so that specific researchers or organisations can be accredited for longer periods allowing for extended and iteratively refined research projects.[120] In this way, so long as their overarching goals are clearly defined and well-motivated, purpose limitation can be maintained while allowing for broadly-scoped and exploratory research.

CCIA proposes that researchers must be able to demonstrate that the data requested is not available through other means, such as the DSA's other transparency disclosure mechanisms.[121] Additional evidentiary burdens such as these may clash with researchers' demands, discussed previously, for a speedy, efficient and accessible access procedure.

Respondents also address the **geographical scope** of research purposes. Several researcher respondents mention the importance of **comparative research** (which can refer to comparisons between platforms or between countries),[122] and some expressly request that this include comparisons with non-EU countries. These non-EU countries are deemed important in at least three ways: First, non-EU countries can serve as a control group to better understand EU samples.[123] Second, risks impacting the EU may originate outside the EU (for instance, per Reset Tech, when 'candidates' outside the EU 'incite violence against a minority, and that minority has a significant diaspora in the EU').[124] Third, non-EU countries may provide important case studies of significant risks (e.g. incitement and/or facilitation of violence in Myanmar and India¸ election-related violence in the United States and Brazil).[125] In many cases, such systemic risks may not

Anti-Bullying Centre (ABC) - EDMO Ireland hub. Amsterdam School of Communication Research (ASCoR). CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley. The Mozilla Foundation.

[118] University of Michigan Center for Social Media Responsibility.

[119] Martin Husovec. RESET. Council for Media Services (Slovakia).

[120] Stiftung Neue Verantwortung e. V. (SNV).

[121] CCIA.

[122] e.g. Universidade Catolica Portuguesa on behalf of Fair MusE. Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Weizenbaum Institute for the Networked Society Berlin.

[123] The Data Co-Ops Project. Martin Husovec. Platform Governance, Media and Technology Lab (University of Bremen).

[124] RESET. See also: Coalition for Independent Technology Research. Platform Governance, Media and Technology Lab (University of Bremen). Centre for Information and Innovation Law (CIIR), Faculty of Law, University of Copenhagen.

[125] Lujain Ibrahim (Oxford Internet Institute). Rappler.

(yet) have manifested (at a significant scale) in the EU, and in-depth research is only possible via non-EU case studies.

Several respondents also discuss penalties and liabilities for the misuse of data.[126] Many respondents mention **contractual penalties** as part of data access agreements or NDAs.[127] Penalties mentioned by CCIA include 'restrictions on future access, exclusion from future vetted-researcher status, exclusion from future EU funding, and in last recourse, fines'.[128] Snap Inc sees a role for the European Commission here, ('the EC should establish procedures for revoking an individual's vetted researcher status and other sanctions on those who fail to meet their obligations and research standards. In this respect, the Commission might look to application processes for access to data for clinical research').[129] In other submissions it is not always specified who would be tasked with enforcing these penalties. The response from Annenberg Public Policy Center, University of Pennsylvania requests that '[a]ny enforcement of violations of the review process or sanctions would remain the province of government authorities'.[130]

Some respondents also address (briefly) the issue of **liability**. CCIA proposes that '[i]n case of a data breach (e.g. loss or unauthorised disclosure) or other misuses, the delegated act should clarify that the respective researcher or DSC will be held liable, when necessary, for the data they have requested. The list of damage should cover at least users' privacy breaches and harm to providers (e.g. financial losses or competitive harm).'[131] DOT Europe makes a similar proposal.[132] In addition, European Tech Alliance and DOT Europe request that VLOs be exempted from any liability related to the misuse of their data by researchers, including potential data breaches.

Also relevant for liability determinations is the question of **GDPR controllership** for personal data processing related to the data access process. For breaches of the GDPR, liability turns on which party (or parties) qualify as data controller(s). Google proposes that the delegated act should clarify that Article 40 disclosures operate on the basis of independent and not joint controllership, with reference to EDPB Guidelines 07/2020.[133] EDMO Researchers, by contrast, have proposed that controllership should be determined on a case-by-case basis.[134]

Finally, liability is also discussed in the specific context of data scraping under Article 40(12) – an issue which is explored further in Section 4 below.

---

[126] Booking.com. vera.ai (EU Horizon Europe Research and Innovation Project).. CCIA Europe. DOT Europe. Google.

[127] Trust Lab. Arcom.

[128] CCIA Europe. See also DOT Europe.

[129] Snap Inc.

[130] Annenberg Public Policy Center, University of Pennsylvania.

[131] CCIA Europe.

[132] DOT Europe.

[133] Google (citing European Data Protection Board, Guidelines 07/2020 on the concepts of controller and processor in the GDPR).

[134] See Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access (n.b.: This report refers to the context of voluntary disclosure rather than statutorily-mandated disclosure, as in the case of Article 40 DSA and as discussed in the aforementioned EDPB Guidelines).

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 independent advisory mechanism (IAM) enjoys broad support amongst respondents, from researchers as well as industry representatives. As discussed in section 2B above, an independent advisory body staffed by researchers, legal experts and ethicists is supported for reasons of *independence* and *expertise*. In particular, respondents foresee the IAM playing an important role in the **peer review of researcher applications,** on such points as methodology and research ethics.[135] EDMO Researchers propose that this work take the form of advisory opinions to DSCs, as well as accrediting other organisations as qualified to undertake vetting and review.[136] There are numerous expressions of support for the ongoing work of the EDMO Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms. Sciences Po's submission goes further than mere advisory opinions and proposes that the IAMs role in the vetting procedure 'should not just be consultative but decisionmaking'.[137] Others propose that the IAM be involved upon request by the DSC.[138] GDR Internet IA, by contrast, argues that certain aspects of researcher vetting should be the exclusive domain of IAM-led peer review with a view to preserving academic freedom.[139] Snap Inc, by contrast, oppose an independent advisory mechanism on the grounds that Article 40 framework 'is already quite complex' and this 'would make the process even more cumbersome'.[140]

Some responses also suggest that the IAM be involved in **other tasks**, including: producing non-binding guidance or standards,[141] coordinating action across Member States and fostering collaboration,[142] monitoring compliance,[143] and dispute resolution and the handling complaints and appeals.[144]

Responses also differ as regarding the **composition** of this advisory body. Many respondents refer to 'experts' without much further specification, but this leaves ambiguities. Industry respondents propose that the body include platform representatives.[145] Many researchers insist on a body staffed by researchers, and oppose the inclusion of platform representatives, either in explicit terms or implicitly through

[135] The School of International and Public Affairs (SIPA) at Columbia University. Opsci. CCIA Europe

[136] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access.

[137] Sciences Po médialab.

[138] TikTok Technology Limited.

[139] GDR Internet IA et Société – CNRS.

[140] Snap Inc.

[141] Academic Researcher Members of the EDMO Working Group on Platform-to-Researcher Data Access. Trust Lab. Council for Media Services (Slovakia). Stanford Internet Observatory.

[142] University of Helsinki. GDR Internet IA et Société – CNRS.

[143] Ministry of Industry, Business and Financial Affairs (Denmark).

[144] Booking.com. Trust Lab. University of Helsinki. GDR Internet IA et Société – CNRS.

[145] Google ('professionals with adequate research, legal, privacy, and security expertise, including professionals with expertise and understanding of the perspectives of different stakeholders, to include users, providers, and researchers').

repeated emphasis on the primacy of an independent composition.[146] Besides researchers and industry representatives, other categories that are mentioned include specialised lawyers, ethicists and technologists.[147] Tiktok suggests that the body currently being established under Commitment 27 of the Code of Practice on Disinformation might also contribute in this context.[148]

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)?*

See the discussion on disclosure formats, methods and safeguards in sections 1A and 2B above.

A handful of respondents go into somewhat further detail by mentioning specific technical standards – listed below in footnote.[149] A common theme is a preference for JSON representation of data, to a lesser extent also CSV and/or XML are mentioned. Stanford Internet Observatory proposes that public data be accessed via HTTP or Web Socket.[150] Project Fair MusE mentions REST API as an industry-acknowledged best

[146] GDR Internet IA et Société – CNRS. Positions such as these focused on researcher freedom, suggest that the mechanism, or at least its decision-making roles, would lie solely with researchers. See also Arcom, calling for scrupulous attention to the independence and composition of the IAM, with a view to avoiding potential interference from service providers. See also The School of International and Public Affairs (SIPA) at Columbia University and University of Michigan Center for Social Media Responsibility mentioning the risk of 'capture' for the independent advisory body.

[147] Martin Husovec. Council for Media Services (Slovakia).

[148] TikTok Technology Limited.

[149] Stanford Internet Observatory ("For public data, such as that required under Article 40.12, the easiest method for researchers to consume is an API exposed via HTTP or a WebSocket that allows for fetching of JSON representations of that data, using a rich set of search operators. There are two modes of operation that could be considered: historical and real-time. Historical queries would allow searches back in time, and real-time would be a non-stop stream of events matching particular rules (such as Twitter's PowerTrack). One question to iron out is how to perform research on data that has been deleted in such a way that it complies with other EU regulations. By and large, Twitter's former API offerings are a good model to base future work off of, though other platforms more focused on multimedia content could offer additional content such as the identifiers of soundtracks to video content or transcription. Amsterdam School of Communication Research (ASCoR) ('We see JSON-based APIs as the go-to solution that should be used for most of the data sharing mechanisms.') Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub ("as SSL/TLS for data transmission […] Comprehensive documentation with code examples and developers will assist researchers and developers to seamlessly integrate and utilise the data access interfaces. Data exchange formats could be standardised for sharing as currently indexed structured formats"). EU DisinfoLab ("CSV data works well, as well as GEXF format, which allows quite an easy mapping using Gephi. Moreover, having a common language would be positive and reflect a general trend towards standardisation that we see in FIMI (e.g., in the DISARM framework or the Kill Chain)").

[150] Weizenbaum Institute for the Networked Society Berlin

practice.[151] LMU Munich refers to ISO standards where relevant, such as ISO 8601 for dates.[152]

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

Many expect the independent advisory body, discussed in section 2E, to act as an important source of expertise in the Article 40 process. As discussed, adequate staffing at DSCs, for instance through the creation of dedicated 'data units', is also mentioned as an important form of capacity building. Dedicated points of contact within VLOs, discussed in section 1A, can also be seen as a relevant form of capacity building.

A more general theme amongst respondents is remuneration for **peer review activities**. Whether organised via an independent body or through other venues, peer review activities are time- and resource-intensive. Peer review processes which are not remunerated, such as for many academic publications and conferences, are often marred by delays and inefficiencies. Remuneration can help to ensure a speedy, effective process. In addition, adequate funding is seen as important to ensure independence of peer reviewing processes, and avoid capture by platforms or other parties.

Another theme is respondents advocating for dedicated **funding schemes for research making use of Article 40.**[153] CERRE proposes the creation of new '**DSA Research Grants'**, which would not only provide necessary funding but also serve as an initial assessment / prima facie evidence of eligibility under Article 40's vetting process.[154] It is worth repeating here, as discussed in Section 2A, that researchers risk being placed in a double-bind or 'Catch-22' scenario since many funding sources require them to be able to demonstrate access to the necessary data, whereas Article 40 seems to permit access only when they have *already secured* the necessary funding. Dedicated funding for relevant research would be one solution, alongside flexible vetting standards, to alleviate this problem. Coordination or integration with the ERC's activities on research funding is also mentioned as an option.[155] CDT calls for funding specifically on the issue of law

[151] Universidade Catolica Portuguesa on behalf of Fair MusE.

[152] Ludwig-Maximilians-University Munich (Germany).

[153] Centre on Regulation in Europe (CERRE). ALLAI. Carnegie Endowment for International Peace. Daphne Keller ('Informed investment in the underlying infrastructure to support sound research will be important, given the ultimately limited resources available to academics and other researchers. To the extent that the European Union itself can provide funding to build foundational tools and resources to support an array of future research projects, such spending will be a sound priority'). Platform Governance, Media and Technology Lab (University of Bremen). University of Copenhagen.

[154] Centre on Regulation in Europe (CERRE) ('To obtain these grants, applicants must explain in their applications which data they must access, how their research contributes to the detection or minimisation of systemic risks in the EU, and how they plan to comply with all the requirements of Article 40(8). Obtaining a DSA Research Grant would provide researchers with strong prima facie evidence that they have passed the vetting process, so that their requests should be authorised by the relevant DSC in an expedited time frame.'). The School of International and Public Affairs (SIPA) at Columbia University.

[155] Martin Husovec. University of Amsterdam (Borra, Peeters & Rieder).

enforcement requests for access to data, and the relationship between their activity and the introduction of Article 40 DSA.[156]

Finally, on capacity-building, researchers express a desire for venues and occasions for future knowledge-sharing and network-building events, such as through workshops or conferences, as well as opportunities for training in relevant skills such as data management.[157]

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?*

Opinion is somewhat divided as regards the merits of a common glossary or dictionary. Support for the proposal is broad but qualified in various ways.[158] Amongst its supporters, DCU in particular offers a detailed list of items it might address.[159] Several groups refer to FAIR (Findable, Accessible, Interoperable, and Reusable) principles as a relevant standard.[160] The independent advisory body is also mentioned as a body that might play a coordinating role here. The Weizenbaum Institute notes that this is a resource-intensive task which would require dedicated funding.[161]

Some are also critical, largely due to the dynamic and divergent nature of VLO data structures: their service designs and dataset structures differ (for instance, what qualifies as a 'post', an 'account' or a 'view') and are constantly changing.[162] Several respondents, including various industry respondents, therefore offer qualified support for a dictionary of limited scope, and stress the need to sufficiently address divergence between different

[156] Centre for Democracy & Technology, Europe Office.

[157] Google. Global Disinformation Index. University of Helsinki. Naomi Shiffman (suggesting '● Live trainings provided by platforms on each available data-set, including demonstrations on how to log into and query tooling, different use cases, and limitations. Ideally, DSCs will also provide training on how to use multiple datasets concurrently. ● Extensive documentation and examples of published research leveraging datasets ● Guidance on how to cite datasets and instructions for submitting data to journals for verification, as well as limitations on dataset publications ● Trainings by DSCs on submitting applications and best practices for thorough and successful applications')

[158] Dublin City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub. Trust Lab. Weizenbaum Institute for the Networked Society Berlin. Weizenbaum Institute for the Networked Society Berlin. Europe Technology Policy Committee of the Association for Computing Machinery. Amsterdam School of Communication Research (ASCoR). vera.ai (EU Horizon Europe Research and Innovation Project). NORDIS - Nordic Observatory for Digital Media and Information Disorder. GDR Internet IA et Société – CNRS. Universidade Catolica Portuguesa on behalf of Fair MusE. Women in AI Austria, Avaaz. ALLAI. Google. TikTok Technology Limited. CCIA Europe

[159] City University's Institute of Future Media, Democracy and Society (DCU FuJo) - Dublin City University's Anti-Bullying Centre (ABC) - EDMO Ireland hub.

[160] Amsterdam School of Communication Research (ASCoR). Weizenbaum Institute for the Networked Society Berlin.

[161] Weizenbaum Institute for the Networked Society Berlin.

[162] Sciences Po médialab; Stiftung Neue Verantwortung e. V. (SNV). CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley.

services.[163] For Stiftung Neue Verantwortung, a dictionary is not desirable so long as technical interoperability or standardisation has not yet been achieved, and similarly, Snap Inc considers the proposal premature and only to be considered after initial stress-testing of access requests.[164] Sciences Po rejects the proposal outright on the ground that '[a] precise common language already exists within the research and engineering communities concerned'.[165]

What remains a relatively stronger point of agreement amongst researchers is the need for clear definitions and technical documentations in relation to specific disclosures – as discussed in section 1A. Whereas it may be challenging to produce definitions that are workable across different service contexts, it appears relatively more feasible for services to state definitions for their own disclosures.[166]

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?*

In varying detail, many responses addressed the importance of Article 40(12) and access to publicly available data. This provision is seen as important by many researchers, and some expressly demand that it be treated as a priority.[167]

Among the issues raised were (1) possible disclosure formats and mechanisms, (2) procedures for exercising 40(12)'s access rights, and (3) the definition of publicly available data.

As with Article 40(4), many researchers express an interest in **online interfaces** as well as **APIs**. A user-friendly model that is mentioned repeatedly is Facebook's **CrowdTangle,** which provides engagement information about public posts in real-time via an online interface as well as through an API.[168] Two submissions by former CrowdTangle (executive) staff, Brandon Silverman and Naomi Shiffman, discuss best practices from this

[163] CCIA Europe. Google. TikTok Technology Limited. CITRIS Policy Lab & Goldman School of Public Policy, UC Berkeley.

[164] Stiftung Neue Verantwortung e. V. (SNV).

[165] Sciences Po médialab.

[166] Institute for Strategic Dialogue. Stiftung Neue Verantwortung e. V. (SNV), more generally Section 1A above.

[167] e.g. Weizenbaum Institute for the Networked Society Berlin (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.) Democracy Reporting International ('We see it as essential that the delegated act covers Art. 40.12 as most studies in this field are done on public data and will be done on the basis of this Article 40.12').

[168] Analyse & Tal. Weizenbaum Institute for the Networked Society Berlin. Coalition for Independent Technology Research. Centre for Democracy & Technology, Europe Office. Arcom.

model, focused respectively on the tool's design and its policies for researcher vetting.[169] There is broad support to enhance this model and replicate for other platforms.[170] Some respondents also note shortcomings and limitations in the present design of these tools.[171] Rappler, for instance, calls for 'a *better* public CrowdTangle' (emphasis added).[172]

Even more frequent are references to the importance of **independent collection methods** such as scraping or adversarial sock puppet auditing.[173] These methods are considered important because they are already an established practice, taking place at an appreciable scale, and because they allow researchers to independently verify other disclosures and claims made by the service under study.[174] Respondents refer to instances where platform scraping mechanisms detected discrepancies between information being shown to users and information being disclosed via dedicated researcher tools.[175]

Data scraping can be initiated by researchers without support or approval from platforms. Compared to dedicated disclosure mechanisms, questions therefore remain about its relationship to Article 40(12) and under which conditions or procedures such scraping might fall under this provision. Respondents make several proposals in this regard:

- **Terms of Service protections:** The most common demand is to clarify that researchers who satisfy the conditions of Article 40(12) are protected against contractual claims from platforms. Many VLOs prohibit scraping in their Terms and Service, and some have taken legal action against public interest research on this basis, resulting in chilling effects according to some researchers.[176] In addition to clarifying legal protections under Article 40(12), ISD proposes that platforms institute voluntary carveouts in their Terms of Service to permit protected research.[177]
- **Clarification on copyright:** Some respondents also raise the issue of copyright, and specifically the relationship to the text and data mining exemption in

[169] Naomi Shiffman, Brandon Silverman.

[170] e.g. EU DisinfoLab

[171] Lujain Ibrahim (Oxford Internet Institute) ("In 2021, academics discovered systematic gaps in Crowdtangle transparency data that Meta was providing to academics and regulators.") citing: [Bobrowsky, M, 2021. Facebook Disables Access for NYU Research Into Political-Ad Targeting. WSJ. Matias, J.N., 2023. Humans and algorithms work together—so study them together. Nature, 617(7960), pp.248-251.]

[172] Rappler (also claiming 'We have a design we could contribute for a data platform that would give tiered access, depending on purpose. It would take roughly six months to build')

[173] Julia Angwin. Brandon Silverman. Martin Husovec. Stiftung Neue Verantwortung e. V. (SNV). University of Amsterdam (Borra, Peeters & Rieder). EU DisinfoLab. Daphne Keller. Platform Governance, Media and Technology Lab (University of Bremen). Centre for Information and Innovation Law (CIIR), Faculty of Law, University of Copenhagen. Universidade Catolica Portuguesa on behalf of Fair MusE. Avaaz. The Data Co-Ops Project. Institute for Strategic Dialogue. Centre for Democracy & Technology, Europe Office. Amsterdam School of Communication Research (ASCoR). AI Forensics. AlgorithmWatch. NYU's Center for Social Media and Politics provides a definition of scraping ('By "scraping," we mean the process of loading publicly available web pages on one's own computer and then retaining the information contained in that webpage as it loads. Many social media sites do not have APIs, but they do have publicly available data.').

[174] EU DisinfoLab. AI Forensics. Mozilla Open Source Audit Tooling (OAT) Project.

[175] AI Forensics.

[176] e.g. Institute for Strategic Dialogue. The Mozilla Foundation. Amsterdam School of Communication Research (ASCoR). Platform Governance, Media and Technology Lab (University of Bremen). Global Disinformation Index.

[177] Institute for Strategic Dialogue.

Article 3 of the CDSM Directive.[178] Mozilla requests that this relationship be clarified, and the University of Bremen and the University of Copenhagen propose that "[o]ne possible way to ensure the smooth interplay 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 CDSM Directive".[179]

- **Lifting of technical restrictions:** Besides litigation, VLOs also restrict scraping through technical measures such as blocking and rate limiting. Some respondents propose that Article 40(12) should also protect researchers against such technical restrictions.[180]
- **Broader immunities:** Some respondents go further in requesting broader immunities for other applicable regimes that might deprive scraping of its effectiveness, for instance immunities against claims from platforms or against civil liability.[181] To this end, ISD also proposes that Member States should establish additional legal protections. Exemptions or immunities related to data protection law are not requested; several researchers note that researchers should remain responsible for GDPR compliance in their scraping practices.[182]

Amongst industry respondents, Google also proposes that, '[i]n some cases, compliance may be most appropriately achieved by allowing or supporting direct scraping by researchers' (though 'in other cases, API access may be more appropriate').[183] TikTok, by contrast, asks for clarification that 'that the practice of 'data scraping' is not permitted nor encouraged under the DSA given the real risk of data protection and privacy harms to users'.[184] Booking.com requests that 'that it should be up to the VLOP to determine the appropriate means and interfaces made available to grant access to publicly available data' and that '[t]he VLOP should remain in full control over the manner in which access to the relevant data is provided, and the technical limitations and security measures associated therewith'.[185]

Process

Some respondents also discuss questions of process related to Article 40(12). Tiktok requests clarification that requests by researchers for publicly available data should be assessed by platforms (rather than, for instance, DSCs as in the case of Article 40(4)), as well as several other clarifications, including further clarification on the relationship of Article 40(12) DSA to the GDPR, ideally in the form of guidelines by the European Data Protection Board.[186] Several of these responses appear to be premised on the idea that

[178] Platform Governance, Media and Technology Lab (University of Bremen). The Mozilla Foundation.

[179] The Mozilla Foundation. Platform Governance, Media and Technology Lab (University of Bremen). Centre for Information and Innovation Law (CIIR), Faculty of Law, University of Copenhagen.

[180] Martin Husovec. Daphne Keller. Institute for Strategic Dialogue.

[181] Platform Governance, Media and Technology Lab (University of Bremen). Centre for Information and Innovation Law (CIIR), Faculty of Law, University of Copenhagen. RESET. Institute for Strategic Dialogue.

[182] e.g. Daphne Keller.

[183] Google.

[184] TikTok Technology Limited.

[185] Booking.com

[186] TikTok Technology Limited. ('● Confirm that the obligation is on platforms to assess requests by researchers for public data. That guidance might also better specify how the platforms should assess if researchers comply with Article 40(8)(b-e). In particular, it might indicate that platforms can consider relevant industry codes, such

researchers would be required to undergo some kind of application process before making use of Article 40(12), rather than being entitled to its benefits *qualitate qua*, by direct operation of law. Other contributions stress the importance of 'adversarial' collection or collection without permission by the platform, which they see as an important supplementary method to verify permissioned access results and prevent 'audit washing'.[187] Daphne Keller therefore argues that 'the Delegated Act should not create new gatekeepers with power to limit scraping, or authority' and that regulators, though they may offer guidance or presumption, 'should not be vested with power to prevent, impede, or become necessary sources of permission for other scraping-based research'.[188] CCIA requests that the vetting process under this paragraph be further clarified.[189] Snap Inc. cautions against overly detailed specifications on this complex topic and proposes a phased, iterative rollout (see also Section 1a above).[190] A similar point is made by Daphne Keller, who advises against overly detailed guidance if it might unduly constrain other potentially fruitful solutions ('Article 40.12 is an effective backstop to other DSA provisions precisely because it is open-ended and flexible for unforeseen future uses. It should remain so').[191] Martin Husovec suggests that further guidance on the protection of scraping under Article 40(12) might be adopted as an EC Guidance or a Code of Conduct.[192]

Publicly available data

The concept of 'publicly available data' is interpreted in different ways. Some respondents offer a broad definition focusing on technical access affordances. For Reset, publicly available data is 'any data that could be realistically accessed by a hypothetically interested member of the public [e.g. by making an account'. ISD and Mozilla propose slightly narrower definitions in that they would exclude data which requires an account to access.[193] Narrower still are the definitions proposed by Brandon Silverman, based on policies at CrowdTangle regarding *meaningfully public* data, which focus not only on technical access but also take into account other contextual cues such as the account type and size, and special considerations for public figures such as elected officials and

as the draft EDMO Code when conducting this assessment; ● Confirm that platforms can ask researchers to sign data sharing agreements (as per the draft EDMO Code) in order to demonstrate that appropriate data security and confidentiality measures are in place; ● Address the interaction with this obligation and the platform's obligations under the GDPR. As set out above, ideally this would be addressed in guidelines from the European Data Protection Board; ● Confirm that this obligation only extends to data already publicly available from the platform; ● Acknowledge that the nature and scope of this obligation is not completely clear and that there is an expectation that platforms and researchers will co-operate in good faith to resolve any issues that arise in relation to the provision of this data; ● Confirm that the practice of 'data scraping' is not permitted nor encouraged under the DSA given the real risk of data protection and privacy harms to users'.)

[187] AI Forensics ("Broader researcher data access, coupled with adversarial data collection, enables what we refer to as adversarial audits or third-party audits. These audits are conducted by independent auditors without official data access or collaboration with the platforms, and they are crucial for bringing systemic harms to public attention and preventing platforms from engaging in audit-washing"). See also: University of Amsterdam (Milan, Agosti and Beraldo). Coalition for Independent Technology Research. Daphne Keller.

[188] Daphne Keller.

[189] CCIA.

[190] Snap Inc.

[191] Daphne Keller.

[192] Martin Husovec.

[193] Institute for Strategic Dialogue.

media outlets.[194] GDR requests that access under Article 40(12) 'should include the possibility to provide access to old (but not erased and still published) data, and not just to a stream of new data that is being posted.[195]

V. Further reading

Below is a selection of further readings cited in responses to the call for evidence:

- Altman, M., Wood, A., O'Brien, D., Vadhan, S. & Gasser, U. (2015), "Towards a Modern Approach to Privacy-Aware Government Data Releases," 30 *Berkeley Technology Law Journal* 1967 (2015).)
- CERRE, 'Access to Data and Algorithms: For an Effective DMA and DSA Implementation'.
- EDMO, Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access.
- European Data Protection Board, Guidelines 07/2020 on the concepts of controller and processor in the GDPR.
- Gordon-Tapiero, A., Wood A.m & Ligett, K., (2022) "The Case for Establishing a Collective Perspective to Address the Harms of Platform Personalization," Proceedings of the 2nd ACM Symposium on Computer Science and Law (CSLAW'22) https://dl.acm.org/doi/10.1145/3511265.3550450.
- Huszár, F., Ktena, S. I., O'Brien, C., Belli, L., Schlaikjer, A., & Hardt, M. (2022). "Algorithmic amplification of politics on Twitter", Proceedings of the National Academy of Sciences 119(1). https://doi.org/10.1073/pnas.2025334119 .
- Social Observatory for Disinformation and Social Media Analysis, Evaluating Safe space solution including data management and processing setup - Modelling academic social media data safe spaces based on administrative and genomic data management From unit level to access level.
- Stanford Internet Observatory, My Heart Belongs to Kashmir (September 2022)
- Stanford Internet Observatory, Unheard Voice (August 2022)
- Stanford Internet Observatory, The New Copyright Trolls: How a Twitter Network Used Copyright Complaints to Harass Tanzanian Activists (December 2021).
- The Royal Society, From privacy to partnership: The role of privacy enhancing technologies in data governance and collaborative analysis.

[194] Brandon Silverman.

[195] GDR Internet IA et Société – CNRS.

- United Nations Committee of Experts on Big Data and Data Science for Official Statistics, *The United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*.

- United States National Science and Technology Council, *National Strategy to Advance Privacy-Preserving Data Sharing and Analytics*.

- Wood, A., Altman, M., Nissim, K., Vadhan, S. (2020), "Designing Access with Differential Privacy", in: Shawn, C., Dhaliwal, I., Sautmann, A., and Vilhuber, L. (eds), *Handbook on Administrative Data for Research and Evidence-Based Policy*, Cambridge MA: Abdul Latif Jameel Policy Action Lab.