UCD Centre for Digital Policy
Ionad um Bheartas Digiteach UCD

UCD Centre for Digital Policy
Newman Building
University College Dublin
Belfield
Dublin 4
Ireland

www.digitalpolicy.ie

+353 1 716 7777 | +353 85 7492363
naoise.mcnally@ucdconnect.ie

31 May 2023

Response to the Call for Evidence for
Delegated Regulation on data access provided for in the Digital Services Act

Submission on behalf of the UCD Centre for Digital Policy, University College Dublin, Ireland

Author: Naoise McNally

Key Recommendations

- **Data required to conduct meaningful research:** Data to be accessed for research under the DSA must encompass a broad range of information held by VLOPs/VLOSEs including technical/functional documentation, data used in the design and assessment of algorithmic systems, and information in the form of institutional memory held by internal stakeholders, in addition to the more commonly considered datasets of content and user data collected by the platforms.

- **Establishment of an Independent Intermediary Body:** Such a body would provide expert guidance and technical support to the DSCs in a range of functions including creating and managing data documentation such as codebooks and dictionaries; create guidelines for researchers and platforms and vetting research requests.

- **Coordinated Approach and Adaptive Governance:** An integrated approach involving the European Board for Digital Services, the DSCs and an independent intermediary body, within an adaptive governance framework that is flexible and responsive to the evolving technological landscape.

- **Standardised and accessible vetting process:** A single standard registration and vetting process operated collaboratively by the DSCs and independent body would streamline the process, reducing costs and duplication of effort while increasing efficiency and transparency.

- **Tiered Data Access Model with Progressive Obligations:** A progressive or tiered approach to data access would offer multiple mechanisms for data access requests with

increasing obligations according to the level of complexity and risk associated with the data requested.

- **Mechanisms for Transparency and Accountability:** Measures such as reporting on access requests, appeal and dispute resolution mechanisms, clarity on funding of the intermediary body are required to ensure the proper functioning of the framework

- **Adequate Funding:** A collaborative approach to resourcing in terms of funding and specialised staffing will be essential, especially in the context that the majority of VLOPs/VLOSEs are located in one of the smallest member-states: Ireland.

About the UCD Centre for Digital Policy

The members of the **UCD Centre for Digital Policy** believe that policy making and evaluation must be deliberative, emergent, and iterative, with sociocultural values at their core. Such an ambitious agenda will require working with stakeholders and beneficiaries to develop effective and evidence-based formal and informal regulation and institutional digital policies, maintain such policies over time, and foreground urgent issues of sustainability, equity, and human rights. The members of the centre draw on interdisciplinary methods from computing, law, design, human rights, and social science to create policy, amplify positive effects on society, especially vulnerable citizens, and study policymaking across technologies and sectors.

About the Author

Naoise McNally is a PhD Researcher in Algorithmic Governance at Science Foundation Ireland's Centre for Research Training in Machine Learning (ML-Labs) and the School of Information and Communication Studies (ICS) at University College Dublin, and a member of the UCD Centre for Digital Policy.

Introduction

We welcome the opportunity to contribute to the development of the framework for data access under Article 40 of the Digital Services Act (DSA). This provision of the DSA offers a unique and unparalleled opportunity to gain insight into the dynamics at play within online platforms and how these systems affect society across a range of issues.

Researchers have long been stymied in their attempts to examine the effects of online platforms. The data held by platforms was considered a proprietary resource with strict controls over access. A contentious and often adversarial dynamic has evolved. Platforms have imposed unilateral limits on the types and kinds of data provided; unilaterally curtailed or cut access[1], offered tiered and preferential access to specific researchers and institutions[2] creating questions of potential capture[3]; imposed artificial constraints[4] and limits to publication[5], impeding rigorous independent research[6] and creating asymmetry in terms of the topics and subjects of research[7].

Article 40 of the DSA provides a unique opportunity to reset the system, and establish a new direction for research into digital intermediaries globally. The DSA represents a paradigm shift in terms of reducing information asymmetry in understanding the dynamics and effects of digital intermediaries and offering pathways to obtain knowledge about how platforms work[8].

An important example is the question of the effect of social media usage on teen mental health. Despite being the cause of much concern and debate for many years, the research agenda has been limited in no small part by lack of access to data, an issue highlighted in the recent research review conducted by the American Psychological Association[9]: *"the data required to make cause-and-effect conclusions are challenging to collect and/or may be available within technology companies, but have not been made accessible to independent scientists."*

The desire with the DSA is to set a gold standard, so this opportunity should not be squandered. Researchers have been set a task weighty with responsibility: identifying and investigating systemic risks and risk assessment compliance. They will need extensive access to data and support to fulfil this mission - to produce independent, rigorous, replicable research that can shine light on the issues and possible solutions.

---

[1] Bastos, M., & Walker, S. T. (2018). Facebook's data lockdown is a disaster for academic researchers. *The Conversation*, https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533

[2] Preferential access to proprietary data held by platforms is discussed further in: Urman, A., Smirnov, I., & Lasser, J. (2023)The right to audit and power asymmetries in algorithm auditing. arXiv preprint arXiv:2302.08301. Notable examples include Raj Chetty (Harvard) who was given preferential access to Facebook data and went on to publish two major studies in Nature (Social capital I: Measurement and associations with economic mobility. *Nature*, *608*(7921), 108–121. https://doi.org/10.1038/s41586-022-04996-4; Social capital II: Determinants of economic connectedness. *Nature*, *608*(7921), 122–134. https://doi.org/10.1038/s41586-022-04997-3). In the field of Economics, Johannes Stroebel (NYU) published a number of influential papers on a diverse range of subjects including Covid-19 spread and housing (e.g. Kuchler et al., 2022; Bailey et al, 2018) with the use of proprietary data provided by Facebook.

[3] Discussion of the issue of academic capture due to preferential access to data is explored in Zingales, L. (2013) 'Preventing Economists' Capture' in Daniel Carpenter and David A. Moss (eds.) *Preventing Regulatory Capture: Special Interest Influence and How to Limit it* (CUP 2013), 124.

[4] Young, M., Katell, M., & Krafft, P. M. (2022). Confronting Power and Corporate Capture at the FAccT Conference. 2022 ACM Conference on Fairness, Accountability, and Transparency, 1375–1386. https://doi.org/10.1145/3531146.3533194

[5] Privileged access to data such as through the Social Science One initiative include a "first read" clause, ensuring platforms review before publication, offering an opportunity to object or offer feedback to researchers.

[6] Timberg, C. (2021). "Facebook made a big mistake in data it provided to researchers, undermining academic work." *Washington Post.*https://www.washingtonpost.com/technology/2021/09/10/facebook-error-data-social-scientists/

[7] See footnote 2 - Urman et al (2023)

[8] Geese, A. (2023). 'Why the DSA Could Save Us From the Rise of Authoritarian Regimes' in Joris van Hoboken, João Pedro Quintais, Naomi Appelman, Ronan Fahy, Ilaria Buri & Marlene Straub (ed.s) *Putting the DSA into Practice*. Verfassungs Books.

[9] American Psychological Association. (2023). *Health Advisory on Social Media Use in Adolescence.* https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use

The following submission offers a range of observations and recommendations relating to the questions put forward by the European Commission in the Call for Evidence. The report is structured into distinct sections to address key aspects of the data access framework. The first section delves into the concept of an adaptive governance framework, emphasising the significance of coordination and cooperation between stakeholders and regulators. The subsequent section focuses on capacity building and the necessity for adequate resourcing to ensure effective implementation. In the following sections, we explore the intricacies of data management, highlighting best practices and potential challenges. The issue of publicly accessible data is examined separately, underscoring the importance of making data available to the public in a transparent and accessible manner. Finally, the section relating to vetted research access, proposes a tiered system of access and associated obligations, alongside issues around data grants, archiving and dispute resolution. A brief concluding remark synthesises the key findings and recommendations, providing a concise overview of the submission's main points.

1. Adaptive & Integrated Governance Framework

Advocate an integrated approach using a standardised system for request, access and appeal, coordinated initially through the European Digital Services Board (EDSB), encompassing the Digital Services Coordinators (DSC) of member states and an Independent Intermediary Body, which would support the DSCs by providing and undertaking key governance tasks. Additional support could be sought from the newly-formed European Centre for Algorithmic Transparency (ECAT)[10] and the wider Joint Research Centre (JRC) of the European Commission. This integrated system would establish a cohesive framework, allowing for the pooling and optimization of limited resources, avoiding duplication, increasing speed, and enhancing efficiency. Furthermore, this approach promotes the sharing of knowledge and best practices, fostering capacity building across the system and reducing dependency on specific member states' DSCs.

Adopting an adaptive governance approach, emphasising flexibility, collaboration, and continuous learning to address emerging issues is key to developing a successful long-term data access framework. This approach encourages ongoing dialogue and collaboration among key stakeholders, including the platforms, research community, DSCs, civil society organisations and users.

1.1. *Independent intermediary body*

An independent intermediary body, as advanced under the Report by the European Digital Media Observatory (EDMO) Working Group[11], comprising a diverse range of expert researchers and technologists is an essential requirement for building a robust, independent, and transparent data access framework. An organisation with depth of technical knowledge and research expertise will be essential, particularly in the initial stages of development. Beyond the foundational tasks of creating guidelines, establishing standardised data dictionaries (section 2.2) and facilitating the creation of initial datasets (section 4.2), this organisation could fulfil a diverse range of specialised support functions to the DSCs, as well as overseeing governance and standards.

The initial stages of implementation are likely to see a significant interest among the academic community in accessing data to conduct studies under Article 40 of the DSA. An independent body

[10] ECAT https://algorithmic-transparency.ec.europa.eu/index_en

[11] EDMO (2022) Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

would be invaluable in assisting DSCs with high volumes of data access requests. Clear and timely communication between platforms and researchers will be important as the program launches with questions about what data is available, and to determine which studies qualify. Within the vetted researcher data access request system (discussed further in section 5), the independent body could facilitate a "peer-review" system for evaluating the most complex and novel research proposals.

Finance of the independent intermediary body

The Strengthened Code of Practice on Disinformation[12] serves as the basis for the recommendation of an independent intermediary body by the EDMO working group. Commitment 27 of the code requires that the signatories (platforms) provide support and funding for such an independent third-party body. If the VLOPs/VLOSEs subject to the obligations of the DSA are to be funders of such a body, then strict governance protocols will be required. Funding should be independently ring-fenced and guaranteed.

1.2. *DSCs*

As the DSC of origin for 12 of the 19 currently designated VLOPs/VLOSEs[13], Ireland will play a pivotal role in the success of the data access framework. However, as one of the smallest member-states it lacks the technical resources and depth of expertise to manage the workload of facilitating the large volume of research requests likely to ensue once the system is operational. There is a strong possibility that without adequate consideration, Ireland could become a bottleneck in the system for data access. Ireland has form in this regard, as evidenced by the issues surrounding GDPR enforcement[14]. This possibility was clearly understood in the drafting of the DSA, which differs considerably from those of the GDPR in terms of joint investigation and the involvement of the Commission in investigations. Data access under Article 40 was however, not given the same consideration, and the country of origin principle applies. This puts the weight of data access request regime on the shoulders of Coimisiún na Meán, the newly-created Media Commission in Ireland, which will hold responsibility for the DSC.

It is therefore crucial that an integrated approach involving an intermediary body and the other DSCs, with the support of ECAT and EDSB is developed to manage the pipeline of requests effectively. Adequate funding, as well as sufficient technical expertise should be mandated across the DSC members. This is particularly important in the case of Ireland, but may require counterbalancing through the network given the uneven distribution of platforms geographically. By addressing funding concerns, implementing mandates for specific skill sets, and enhancing the scientific competence within the DSCs, an integrated approach involving all stakeholders can ensure that Ireland can effectively fulfil its pivotal role in the data access system and avoid creating a bottleneck within a critical element of the DSA accountability framework.

2. Capacity Building

In collaboration with the DSCs, the independent intermediary body should lead the development of a community of practice, encouraging collaboration, knowledge sharing, and collective learning.

---

[12] European Commission. 2022. “The Strengthened Code of Practice on Disinformation”. *https://ec.europa.eu/newsroom/dae/redirection/document/87585*

[13] Country of origin for all VLOPs/VLOSEs is not available at time of writing

[14] ICCL. (2021). Europe’s enforcement paralysis: report on the enforcement capacity of data protection authorities. https://www.iccl.ie/news/2021-gdpr-report/

2.1. *Training & Support*

The limits, and often, lack of direct access to platform data has hindered research agenda and methodological development. The deployment of advanced research methods, such as causal enquiry or experimentation, was impossible, and has severely stunted research in a wide range of areas of particular risk such as teen mental health, suicide ideation, addiction and other behavioural issues. Platforms should be required to provide training on each available dataset, including demonstrations on tool use, case studies, and outlining limitations. Training should be given on how to submit applications effectively and the best practices for ensuring success.

2.2. *Documentation: Dictionaries, Codebooks, Usage Terms*

Detailed codebooks for each platform's data will be essential for researchers to effectively access data. Such codebooks must use standardised language and provide sufficient detail to describe different but corresponding elements or actions within systems consistently across platforms. The development of a comprehensive data dictionary which spans all platforms will be a fundamental step in creating effective codebooks. The Independent intermediary body can play a vital role in this process - overseeing the development of a suitable data dictionary to form the basis of the codebooks, and overseeing these in turn. This should be created as part of the initial efforts in documenting data collected and creation of standardised datasets (section 4.2). By taking on this extensive effort, the intermediary body can facilitate the access and utilisation of data across platforms more efficiently. Protocols must also be put in place to ensure this process is fulfilled as new forms of data or datasets are developed. This ongoing process will help maintain the consistency and relevance of the codebooks and data dictionaries.

As research progresses, a repository of published research should be assembled. This resource could provide guidance on the data's potential uses and limitations, and offer exemplars for best practice. In addition, this repository could serve as a central repository for information on the datasets available, providing specifications on data usage, citation and publication guidelines as well as limitations on publication of specific datasets.

2.3. *Funding*

It is crucial that all institutions responsible for the evaluation of applications under 40(4) have sufficient capacity, in terms of funding and staffing resources, to carry out the processes required to manage the pipeline of requests in a timely manner.

Funding should be considered for the development of a research agenda addressing systemic risks already identified but requiring qualification and quantification. In addition, funding provisions will be required for the necessary data security infrastructure required to support work on sensitive data, as costs for such facilities will prove prohibitive to many institutions and research groups

3. Data Requirements

3.1. *Establishing a Common Understanding of the Data Available*

As noted in the EDMO working group report, a key issue facing the research community as it grapples with access to data offered by Article 40 of the DSA is the issue of "unknown unknowns": researchers have only a limited understanding of the vast amounts of data collected by the platforms[15]. This

[15] Shapiro, E. H., Sugarman, M., Bermejo, F., & Zuckerman, E. (2021). New Approaches to Platform Data Research https://www.netgainpartnership.org/resources/2021/2/25/new-approaches-to-platform-data-research

information asymmetry makes it difficult a priori to set out specific data needs across a wide variety of subjects and fields.

It is therefore crucial for platforms, at the earliest opportunity, to provide a detailed overview of what data is gathered and can be meaningfully used for research. A useful starting point for this process would be to mine their own internal research, from technical teams, product teams, content moderation, trust and safety teams among others, which would provide insight into the vast array of data available to be used for research. From this initial scoping exercise, standard protocols should be put in place for platforms to disclose the full scope of data available.

3.2. *Understanding Data Beyond APIs*

The purpose of providing data access to vetted researchers under the DSA is to facilitate investigation of systemic risk on a given platform, it is important to consider the broader scope of research required to understand such risks. As multi-faceted information platforms with many constituent elements, including algorithmic recommender systems, the risks come not only from the content flowing through the system and the behaviours of users within the system, but are also as a result of the affordances of the system itself and the algorithmic systems determining recommendations. System affordances encompass the design and implementation of the platform in terms of user interfaces and platform controls, for example infinite scrolling, privacy controls, forwarding limits etc. The algorithmic systems are varied but the mainstay is the recommender system which determines how content is ranked and therefore what is shown to the user (and what is not). To this end, 'data' as understood under the DSA must include a wide variety of types of information from quantitative data (large scale datasets on content and user activity on the platforms) to data in the form of documentation such as internal reports and model data.

3.3. *Geographical Scope of Data*

Systemic risks are identified by their trans-national[16] nature, and the internet by its nature is without borders. It is therefore essential that the scope of data not be restricted to that originating with EU users, but rather that the data requested be the required input to answer a research question, which affects EU citizens. The research data access framework must use this cross-national view to interpret and grant data requests in order that the research undertaken to understand systemic risks is as rigorous and effective as possible.

Cross-national or cross-jurisdictional research could provide an important source of data in terms of natural experiments relating to different uses and behaviour by different groups[17]. When paired with information about A/B tests or experiments undertaken by platforms on users in different geographies[18] it could shed light on algorithmic dynamics. In addition, such data could facilitate research on under-researched communities and differing effects in other cultures that will directly affect EU citizens. Finally, facilitating research on how systems operated in other jurisdictions[19] assists in deepening our understanding of potential risks to EU society.

---

[16] Renn, O., Lucas, K., Haas, A., & Jaeger, C. (2017). Things are different today: The challenge of global systemic risks. *Journal of Risk Research*, *22*(4), 401–415.

[17] Dunning, T. (2012). *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press.

[18] Instagram Starts Test To Hide Number of Likes Posts Receive for Users in 7 Countries. *Time Magazine.* https://time.com/5629705/instagram-removing-likes-test/

[19] A notably catastrophic example of the role played by platforms in creating risk and facilitating harm is the role of Facebook in the Rohynga Genocide in Myanmar see: UN Human Rights Council. (2018). *Report of the Independent International Fact-Finding Mission on Myanmar*, A/HRC/39/64.

3.4. *Types of Data Required for Research*

3.4.1. Quantitative Data

Quantitative datasets focussing on particular user behaviour and content dynamics are required by researchers to investigate a plethora of issues. Large scale or specifically targeted sets of such data have been out of reach for most researchers and could provide answers to questions long under consideration. Such datasets include user engagement data such as frequency, duration, time of day of usage; granular user data, while sensitive, is essential and includes (when necessary) age, gender, socioeconomic status, and geographic location, as well as interests and friend/affinity networks; content and the associated interaction data including posts, links, videos, images, likes/responses, shares, comments etc.

Systemic risks associated with new technological developments are proven not easily identifiable but require the passing of time and user adoption for the longer-term effects to come into view. Longitudinal data, therefore, is necessary to employ causal inference methods which can identify causation rather than correlation, which will be key to understanding systemic risks to individuals and society at large. An important example is the question of the effect of social media usage on teen mental health. Despite being the cause of much concern and debate for many years, the research agenda has been limited in no small part by lack of access to data, an issue highlighted in the recent research review conducted by the American Psychological Association[20].

3.4.2. Documentation and Qualitative Data

In order to understand these systems and the risks associated with them, it is necessary to have a broader and deeper understanding of the mechanics, priorities and underlying logic underpinning the systems. In addition to quantitative data relating to content and user behaviour on the platforms, it is essential that data access process encompasses other important forms of data, such as documentation and qualitative data for social science research. Documentary data could be in the form of internal reports, research outputs, risk assessments and impact assessments[21] and other such material. Qualitative data takes the form of access to internal stakeholders involved in decision making and development of systems. Such interviews have taken place under strict limitations, as often employees are subject to NDAs[22].

3.4.3. Data relating to Algorithmic Systems

In terms of research on the algorithmic decision-making systems, a different set of data may be required, including understanding the systems' processing architecture and machine learning models. External investigations which focus on input/output studies have been necessary heretofore but insufficient in terms of deeply investigating risks and harms.

Assessing and investigating risks within and relating to online platforms requires a wide range of data on the accuracy, functioning and testing of algorithmic systems including, but not limited to, training

---

[20] American Psychological Association. (2023). *Health Advisory on Social Media Use in Adolescence.* https://www.apa.org/topics/social-media-internet/health-advisory-adolescent-social-media-use

[21] Selbst, A. D. (2021). An Institutional View of Algorithmic Impact Assessments. *Harvard Journal of Law & Technology. 35*(1)

[22] A good example and discussion of the difficulty of obtaining information from the internal stakeholders in platforms is: Gillespie, T. (2022). Do Not Recommend? Reduction as a Form of Content Moderation. Social Media + Society, 8(3). https://doi.org/10.1177/20563051221117552

data, model parameters, advertising model input data[23]evaluation data, A/B testing information[24] and internal product experiment documentation. The code of algorithms could also be interpreted as accessible under the DSA.[25] To facilitate this sensitive research, intra-system APIs, configured to respect privacy and security concerns of the platforms should be given consideration.

In the absence of clarity on what data is available (section 3.2) it is impossible to tell what else could be used for research, but other examples of other forms of data include:

- Data relating to content moderation processes and practices
- Fact-checking data
- Network threat and integrity data

This list is not exhaustive but reflects some of the variety of potential research paths to identify and qualify systemic risks associated with the platforms.

3.5. *Quality, Accuracy and Integrity of Data*

Ensuring the quality, accuracy, and integrity of data is crucial in research. To achieve this, measures for quality assurance should be implemented, guaranteeing that the data provided in response to researcher requests is complete, replicable, and auditable. Past experience shows that even with platforms like Twitter, which previously offered extensive access and well-developed documentation, there have been significant disparities in data quality[26]. Assurance of data accuracy and completeness in accordance with the data access request is essential. Such datasets should be subject to external audit and breaches of compliance considered as serious infractions. It is also essential to have a clear notification process and sufficient lead times for any changes in format to allow sufficient adaptation and prevent disruptions in the research process.

3.6. *Archiving of Data Requested for Research Replication*

Reproducibility is a core principle of the research process. It is therefore imperative that specific datasets, produced by platforms in response to requests by researchers, are archived and made available for future research, both for replication studies and for different methodological approaches. An option for the archive process, especially of highly sensitive data, is that such datasets may be held in trust by by the independent intermediary body[27] or by the platforms, acting as information fiduciaries[28].

---

[23] Huh, J., & Malthouse, E. C. (2020). Advancing computational advertising: Conceptualization of the field and future directions. *Journal of Advertising*, *49*(4), 367-376.

[24] An example of internal researchers at VLOP outlining the challenges and opportunities of A/B tests: Xu, Y., Chen, N., Fernandez, A., Sinno, O., & Bhasin, A. (2015, August). From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 2227-2236).

[25] Edelson, L., Graef, I., & Lancieri, F. (2023). Access to data and algorithms: For an effective DMA and DSA implementation. Centre on regulation in Europe (CERRE).
https://cerre.eu/publications/access-to-data-and-algorithms-for-an-effective-dma-and-dsa-implementation/

[26] Tromble, R., Storz, A., & Stockmann, D. (2017). We don't know what we don't know: When and how the use of Twitter's public APIs biases scientific inference. *Social Science Research Network*.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3079927

[27] Suggested in the EDMO Working Paper Report (note 11)

[28] Platforms as information fiduciaries sees them as custodians rather than owners of the data. Balkin, J. M. (2015). Information fiduciaries and the first amendment. *UCDL Rev.*, *49*, 1183.

4. Publicly Accessible Data

It is likely that the system will be overwhelmed with applications in the first months of implementation. In anticipation of this likelihood, a tiered data access framework is recommended (set out in detail in section 5). The base level of such a system would be access points for publicly available data, including real-time data, as set down in Article 40(12).

4.1. *Public Access Data: Dashboards & APIs*

Each platform should provide a user-friendly dashboards with real-time public content, and performance metrics, these would function as first-stop resources for exploratory research, especially useful for fields which require real-time and trending information such as disinformation studies and networked behaviour.

APIs should provide real-time access to public data, including both historical and current information, and deleted or restricted content, with an accompanying broad range of metadata. User-friendly interfaces are crucial for the successful use of APIs by researchers in less technical fields, but APIs should also be configured to enable the development of packages (e.g. R, stata, gephi) to facilitate further research. They should provide data in a machine-readable format, such as JSON. The use of APIs should be free of charge, there should be no cost associated with accessing the public data they provide. There should be no restrictions on the number of API calls or requests that can be made, quotas should only be applied if queries disrupt the service quality.

4.2. *Standardised Datasets*

To facilitate exploratory research and provide readily accessible resources to ease the likely flood of data access requests, a set of standardised datasets could be made available by the platforms on a pilot-project basis, subject to review. Such datasets could include: random samples (e.g., 1% and 10%) of public posts or datasets on specific topics or with specific geographical/socio-demographic focus. The initial datasets should be negotiated with a set of researcher-stakeholders to prioritise the most 'in-demand' data within the research community, set the requirements and assure usefulness.

It should be noted that previous attempts at creating 'readymade' datasets for research have been markedly unsuccessful. The Social Science One[29] initiative was hailed as the beginning of a new era in platform-research partnership[30]. The initiative limped through more than a year overdue, and the datasets provided were severely compromised from a research perspective[31] across a number of dimensions including the arbitrary limits on content shared on the platform, lack of granular user data, and privacy-preserving obfuscation of the underlying dataset. Using privacy-preserving or enhancing techniques such as synthetic data or methods to make human data "anonymously" available to researchers such as "differential privacy", should be approached with caution. Researchers have expressed strong concerns that the validity of research findings may be altered by privacy-preserving techniques[32].

The independent intermediary body could liaise with the platforms to facilitate the process of preparing standardised dataset. However it will need to be empowered to avoid issues noted from

[29] https://socialscience.one/

[30] The vision and proposed approach is explored in King, G., & Persily, N. (2019). A new model for industry–academic partnerships. *PS:Political Science & Politics*, 1–7. https://doi.org/10.1017/S1049096519001021

[31] Hegelich, S. (2020). Facebook needs to share more with researchers. *Nature*, *579*(7800), 473-474. https://www.nature.com/articles/d41586-020-00828-5

[32] Hauer, M.E. and Santos-Lozada, A.R., 2021. Differential privacy in the 2020 census will distort COVID-19 rates. *Socius*, *7*, p.2378023121994014.

previous experience including ensuring the datasets include diverse samples reflecting the EU population, complete datasets, and not just specific subset of data from limited time points chosen by the platform.

Finally, such datasets should not be considered a solution in and of themselves. They represent a sandbox for exploratory research and to exercise methodological approaches or to develop research ideas. Testing this data and documenting the process could serve as a useful testing ground, uncovering operational chokepoints and deducing reasonable timeframes. The learnings derived in this pilot phase could help develop the template for managing future requests efficiently and expediently.

4.3. *External Data Collection Methods*

In the current context of limited data access, various research methods have been developed to study platforms from an external perspective (e.g. sock-puppeting and scraping). The introduction of a new framework for platform-provided data does not eliminate the need for adversarial research approaches. Under the GDPR the platforms are required to protect user data, including preventing scraping, however a carve out should be made to facilitate external research methods under Article 40(12). The provisions for publicly accessible data should be clarified to provide legal safeguards for researchers who engage in external research methods to access public data. These are urgently required as platforms have increasingly sought to prohibit such data collection through both technical and legal means[33][34].

5. Vetted Research Access

The vetting process for researchers should be standardised across the EU, transparent and entirely independent of the platforms. This could be centrally managed by the Independent intermediary body in conjunction with the DSCs in member states. While Article 40(8) deals with the question of vetting researchers and assessing their specific request for data as a combined process, in practice this process could be broken down into several elements and distinct stages to maximise efficiency and minimise duplication. A single standard registration process would serve as the first step providing access for publicly available data. A tiered access system would offer access to data based on a progressive set of application requirements depending on the sensitivity, complexity and novelty of the data requested.

5.1. *Single Standard Registration*

At the base level, a single point of registration should be created for all researchers wishing to access the publicly available data. This registration system and database could be managed by the independent advisory body, jointly by the DSCs or by the European Commission through the European Board for Digital Services. This system could function as the primary level of verification and vetting of researchers across all DSCs. This approach would streamline the process and create a foundation for more detailed data access requests.

[33] Edelson, L. and McCoy, D. (10 August, 2021). 'We Research Misinformation on Facebook. It Just Disabled Our Accounts'. *The New York Times.* https://www.nytimes.com/2021/08/ 10/opinion/facebook-misinformation.html.
[34] Brandom, R. (August 13, 2021). Facebook shut down German research on Instagram algorithm, researchers say. *The Verge*. https://www.theverge.com/2021/8/13/22623354/facebook-instagram-algorithm-watch-research-legal-threat

It should be a public register - detailing the researcher's institution affiliation, ORCID id, research interests and outputs[35], to facilitate reporting (Section 5.7) and collaboration across disciplines and research projects. Implementing a rigorous and homogeneous approach at this level, would create the foundations for a streamlined and efficient system for more detailed data access requests of platforms.

5.2. *Tiered Access & Progressive Obligations*

Implementing a progressive set of application requirements depending on the sensitivity, complexity and novelty of the data requested would set the system up to effectively funnel applications to the appropriate level of review. Proposals could be subject to a prioritisation review in certain cases with time-sensitivity (e.g. relating to an upcoming election) and an expedited process available for urgent requests under the crisis protocol. A progressive list of requirements would thereby serve to streamline the application process.

At the base level, access to public APIs and prepared datasets should be open to all vetted researchers registered with the registration system noted above. Data requests for specific datasets would follow a standardised application system, which at a minimum would contain IRB approval, detailed research proposal and methodology. A tiered system of application could set out different levels of information required according to levels of data sensitivity, security requirements, risk, complexity and novelty of data request etc. For example, requests involving the highest level of sensitive data, the process could require additional information such as: identity verification for all researcher participants; security protocols/proof of ability to secure data, but these would not be required at lesser levels within the system.

Criteria for assessment of applications should be transparent and publicly available within the resources provided by the Independent body and DSCs (outlined in section 2). In the interest of equality of access, and to avoid previous preferential treatment of particular researchers, institutions or research topics, certain criteria should not be prioritised. Examples include: seniority, topic of research, citizenship, previous publication record, or research funding.

Peer-review should be incorporated as an element of the vetting process, but due to its demanding nature in terms of time and resources, implementing it uniformly for all access requests within the already constrained research sector would considerably impede the process, placing an unnecessary burden on the system. It should therefore only apply to certain tiers of novel and complex requests that require intensive resourcing from the platform.

Given the history of platforms providing privileged access to specific institutions and researchers, certain safeguards should be considered in order to ensure the independence and integrity of the data access system. Data access requests when forwarded to the platforms should be done in a "blind process" i.e. the details of the request do not include identifiers associated with the researchers/research team. In addition, Article 40(8)(c) sets out a requirement to disclose the research funding in the application. Likewise, this information should not be included in the data access request forwarded to the platforms, but revealed exclusively to the vetting body - the independent intermediary body and DSCs.

5.3. *Broad and Equal Access*

[35]A subset of information could be displayed publicly, with the majority of information held for access only by DSCs and intermediary body.

Ensuring the data access framework developed under the DSA provides an open and fair opportunity for diverse researchers from various backgrounds, institutions, disciplines, and regions, including those based outside the European Union member states, is of utmost importance. Provisions should be made for supervised students and international research teams, acknowledging different levels of risk and complexity.

Access to data should be granted to non-EU resident researchers, affiliated with a research organisation[36] in accordance with the definition as set out in Article 2(1) of the Copyright Directive. Given the nature of research projects, especially those funded by EU institutions, to have multi-institution and multinational partners, the data access framework must consider how best to manage access requests to international research teams with researchers from different institutions and countries, avoiding unnecessary complexity, duplication and delay. In this context, there will also be a requirement for a mechanism for adding researchers, EU and non-EU-based, to existing projects. PhD students, working under a supervisor should have the ability to make an application, with the support of that supervisor and their institution. Limiting access only to faculty members, limits the scope and depth of research, as much cutting edge research on platforms is being done at PhD level.

Beyond the academic research community, research organisations (NGOs, civil society, journalists, fact-checkers and independent specialists) have played a central role in identifying and investigating many of the risks and harms associated with platforms[37]. The provisions of Article 40(8) (a) and (b) appear to preclude specific data access requests from such researchers, thereby inhibiting their contribution to the risk assessment and mitigation framework. Article 40(12) therefore is essential in enabling their continued contribution, and must be prioritised within the access framework. Non-academic researchers should be afforded access to the public APIs and standardised research sets, by offering a special designation within the single central registration system (section 5.1)

5.4. *Public Database of Research Requests*

A public database of data access requests submitted to platforms should be established. This could follow examples such similar initiatives such as the American Economic Association's (AEA) Randomised Control Registry[38] or the World Health Organisation's International Traditional Medicine Clinical Trial Registry[39]. It could include information such as the researcher and institution who made the request; the platform; details of the data requested; purpose of research; risk to be assessed; the adjudicating DSC; status: whether the grant was submitted to the platform; whether the request was granted or denied; grounds for denial (if applicable); appeal status if applicable; timestamps for different stages of the process should be included. The register would also include details on previously granted data access requests and provide information on where to access data archives (section 3.6).

5.5. *Data Grants*

[36] Article 2, point (1), of Directive (EU) 2019/790 stipulates the definition of a research organisation but does not require the organisation to be based in an EU member state.
[37] Examples include: Angwin, J., Larson, J., Mattu, S. and Kirchner, L. (2016) "Machine Bias, There's software used across the country to predict future criminals. And it's biased against blacks." *Propublica.* https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; Faddoul, M., Chaslot, G. and Farid H. (2020) "A longitudinal analysis o f YouTube's promotion of conspiracy videos" (2020) Technical Report. https://github.com/youtube-dataset/conspiracy; Center for Countering Digital Hate. (2021). *The Disinformation Dozen*. https://counterhate.com/research/the-disinformation-dozen/
[38]American Economic Association (AEA) Randomised Control Registry: https://www.socialscienceregistry.org/
[39]World Health Organisation International Traditional Medicine Clinical Trial Registry: https://www.isrctn.com/

In the interest of academic best practice - rigour and replicability - grants of data access should also not come with any restrictions with regard to research usage (e.g. first review, reproduction, types of analysis). Data should be granted with due regard to adequate data security, but without onerously short retention periods such as currently stipulated by TikTok[40], which takes no account of the research timeline, which stretches from months to years. In this spirit, access should be offered for a sustained period appropriate to the research with a mechanism for extension and refinement of request, based on research outcomes. This refinement clause would offer a streamlined system, rather than have a research team undergo a full vetted access request process.

Per section 3.6 provision should be made for the archiving of data grants in order to facilitate replication, follow-on studies and further avenues of research.

5.6. *Dispute Resolution*

Article 40(5) of the legislation outlines the reasons why a platform might reject a data request, and Article 40(6) details the process for responding to and amending such a rejection. However, the provisions do not clearly delineate the procedures or consequences in situations where there is either non-compliance with these regulations or a disagreement between the researcher and the platform over the data requirements. This omission represents an important issue and requires serious consideration to avoid a serious deficiency in the governance framework.

In the case of a refusal to put forward a request to a platform by a DSC, a comprehensive explanation of the reasons should be provided. In the initial stages an appeal process will be necessary as the community of practice develops a common understanding of the data available, what is feasible operationally and Within the context of an adaptive framework, this should evolve into a set of grounds for request, refusal and appeal.

Article 40(5) requires platforms to reply to the request from the DSC within 15 days but a consequent procedure, in the case of a denial of request by a platform is not described i.e. there is a requirement for an explicit dispute resolution process, where a request is refused by the platform. a thorough description/statement of reasons for not providing said data.

5.7. *Transparency Reports on Research Access*

In line with the requirements of the DSA for transparency reporting (e.g. Art. 22(3) trusted flaggers, Art 24, Art. 15, Art. 35 Board reports on Risk mitigation). There should be a comprehensive annual report on data access. This should be derived from the public database of data access requests submitted to platforms (section 5.4). With information relating to the institution, location, subject areas and methods as well as the adjudicating DSC, such a database could facilitate reporting on data access to ensure transparency around questions and metrics such as speed of processing by DSCs; speed of processing of requests by platforms; levels/rates of grant and denials by different platforms. It could also serve the research community and policy makers more widely by revealing trends around research activities e.g. lack of research of certain risks and/or (over)concentration of research efforts on specific subject areas or on specific platforms.

[40] TikTok researcher API ToS requires researchers “to regularly refresh TikTok Research API Data at least every fifteen (15) days, and delete data that is not available from the TikTok Research API at the time of each refresh." https://www.tiktok.com/legal/page/global/terms-of-service-research-api/en

6. Conclusion

In conclusion, we want to emphasise the importance of a flexible approach in this delegated act. This approach will allow the structure to evolve and adapt to the needs of researchers. The key recommendations outlined in this report highlight the essential elements needed for such a framework. These elements include adopting an adaptive governance approach and promoting integrated efforts between the DSCs of member states and an independent intermediary body. It's important to ensure adequate resourcing for the intermediary body, DSCs, and researchers. Emphasis should also be on capacity building through standardisation of protocols and processes as well as knowledge sharing. To that end, a tiered data access model with progressive obligations has been proposed as a streamlined approach that can facilitate both exploratory and more targeted research requirements in a timely and efficient manner. Alongside this, we recommend strong protections for external adversarial data collection. The framework should be designed for access to non-EU resident researchers, and also provide access to publicly available data for NGOs, civil society, journalists and independent specialists. Transparency and accountability mechanisms should be in place, along with robust dispute resolution procedures. Finally, the definition of ‘data’ should be understood in the context of both social science and computer science research, to encompass a wide range of quantitative, qualitative and documentation data utilised by VLOPs/VLOSEs in developing, deploying and assessing their systems.

Ultimately, these recommendations seek to foster an environment that upholds the principles of good research practice—producing independent, rigorous, and reproducible research for public consumption. Underpinning this framework are the principles of transparency, equity and cooperation. For this system to work effectively, a new cooperative dynamic must replace the current adversarial dynamic which has developed over the past two decades. The recommendations, taken together, offer a roadmap for implementing a comprehensive and flexible approach to data access that caters to the diverse needs of researchers, while also ensuring the transparency, accountability, and robustness of the research produced.