DCU Institute of Future Media, Democracy and Society

DCU Anti-Bullying Centre

EDMO Ireland

We submit this feedback on behalf of two research institutes at Dublin City University - the Institute of Future Media, Democracy and Society (DCU FuJo) and the DCU Anti-Bullying Centre (ABC) - as well as the EDMO Ireland hub, which is coordinated by DCU FuJo.

Our researchers assess a range of systemic risks, including cyberbullying, online hate and extremism, disinformation, electoral integrity, gender-based violence, and other forms of online harm. We investigate these areas to better understand the issues and develop effective strategies to mitigate their impact. We also have experience in assessing platform compliance with the EU Code of Practice on Disinformation for the national media regulator.

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

22nd May 2023

Kirsty Park
Eileen Culloty
Kanishk Verma
Tijana Milosevic
Tetyana Lokot
Nhung Dinh
Alan Smeaton
Jane Suiter

DCU Institute of Future Media, Democracy and Society

The Institute of Future Media, Democracy and Society (FuJo) is a research centre located in DCU's School of Communications. FuJo's multidisciplinary research investigates how to counter digital pathologies including disinformation and digital hate; how to enhance public participation through democratic innovations; and how to secure the sustainability of high-quality journalism. www.fujomedia.eu

DCU Anti-Bullying Centre

DCU Anti-Bullying Centre is a University designated research centre located in DCU's Institute of Education. It is home to scholars with a global reputation as leaders in the fields of bullying and digital safety. The Centre hosts the UNESCO Chair on Bullying and Cyberbullying.The Centre also hosts the national anti-bullying website www.tacklebullying.ie.

EDMO Ireland

DCU FuJo coordinates the Ireland hub of the European Digital Media Observatory. It is part-financed by the European Union to monitor and analyse disinformation; conduct factchecks and investigations; develop media literacy resources; assess and inform policy; advance tools to detect and analyse disinformation; conduct research; and increase capacity among the community of Irish stakeholders. https://edmohub.ie

Data access needs:

a) What types of data, metadata, data governance documentation and other information about data and how it is used can be useful to DSC's for the purpose of monitoring and assessing compliance and for vetted researchers for conducting research related to systemic risks and mitigation measures?

Algorithmic transparency data

- Research on online harms is limited due to a lack of data showing how users come across information through content delivery algorithms in timelines, ads, and recommender systems.
- It's crucial to understand how harmful content is ranked and given priority by platform algorithms, based on engagement metrics. This includes examining factors that influence the algorithm's decision-making process, such as demographics, user interactions, visibility, and whether harmful content is amplified or suppressed.
- Logs of algorithmic changes - including, for example, any data and metadata on disclosures to users, tweaks in the algorithms, how users deal with switching between algorithmic/chronological timelines - would be useful to assess compliance and to support research on systemic risks and mitigation measures.
- Increased transparency regarding the use and training of algorithms would help researchers understand their strengths, potential biases, and areas for improvement.
- Such transparency would also foster greater trust and collaboration between researchers, platforms, and the broader community dedicated to mitigating the impact of harmful online content.
- Research on algorithmic accountability would also contribute to a better understanding of how company algorithms shape user experience and agency online and how users become aware of and realise their rights for holding companies accountable for their algorithms.

User Experience and Feedback data

- Collecting user feedback and experiences regarding encounters with harmful content can provide valuable insights into the effectiveness of content moderation systems and other mitigation measures. Understanding user perspectives helps identify areas where algorithms may fall short and aids in refining detection and mitigation strategies.
- Providing additional insights into underlying algorithms can facilitate a more informed discourse among researchers, policymakers, and stakeholders. This understanding can lead to the development of improved systems, increased

accountability, and proactive measures to address potential biases or unintended consequences.

- This knowledge is crucial for implementing targeted interventions, developing effective content moderation policies, and improving algorithmic systems to prioritise user safety and well-being.

Platform governance and moderation data

- Platforms are increasingly relying on automated moderation and pre-screening systems and while there are valid concerns about disseminating this information publicly, shedding more light on the algorithmic processes employed in these systems can aid in identifying potential algorithmic and systemic biases as well as in assessing the effectiveness of such systems in preventing harm.
- Meta's OPT model[1] and Google's Perspective API[2] , are valuable resources for researchers to evaluate their effectiveness in identifying toxic textual content. However, there is limited information regarding specific aspects of these systems:
    - It is unclear whether these systems are utilised for moderating other forms of online harm - online hate, cyberbullying.
    - It is unclear to the extent to which these systems are specifically trained and applied for addressing online harm
    - It is important to ascertain whether the training datasets incorporate a wide range of harmful behaviour and contextual nuances, enabling such models to effectively identify and classify various types of online harm.
- While platform policies are often clear and publicly available data about enforcement of these policies can be vague and the data provided in transparency reports can vary. Therefore, to gain a deeper understanding of compliance and potential systemic risks associated with content labelled as bullying/hate speech/extremist, it would be beneficial to include detailed information on types of content (profiles / audio-visual posts / comments) takedowns carried out for violating community guidelines.
- Insight on internal platform decision making processes and applications of policy can also be useful. For instance, how does a platform process and decide upon a request to takedown content from a government official or a major company? What combination of technology and human assessment is used in combatting ToS violations? Understanding these types of decisions is essential to understanding how companies comply with their obligations and how they are mitigating systemic risks.

---

[1] https://ai.facebook.com/blog/democratizing-access-to-large-scale-language-models-with-opt-175b/

[2] https://perspectiveapi.com/

Content and engagement data

- Access to meaningful breakdowns of content and engagement data facilitates an understanding of how much harmful content is available on a platform and allows for multi-method and multi-disciplinary analyses, which will inform research about essential areas of systemic risk. This includes content related to topics such as suicide, self-harm, cyberbullying, elections, online hate, grooming, disinformation and others.
- Data which allows researchers to assess the outcomes of using various search terms can be useful in understanding how easily harmful online content can be accessed, particularly when searching for specific hashtags and their variations commonly used by vulnerable demographics like teenagers.
- Data revealing the number of views, shares, and other visible forms of engagement provide insights into the reach and spread of harmful online content. Tracking these metrics can help quantify the level of exposure and potential impact on users.
- It would also be useful to understand how engagement data is influenced by algorithms, for instance, if content gains popularity in a certain context is it more likely to be promoted to other users, and are algorithms trained to take additional precautions if the subject matter or hashtags used are associated with areas of systemic risk e.g. discussing vaccines or using language associated with suicide.
- Additionally, social media content which is removed from platforms often becomes inaccessible to researchers who may need access to such data to conduct study phenomena such as online hate or disinformation or to assess moderation decisions. Access to deleted content should therefore be made accessible in some instances.
- It is also notable that some platforms have been taking actions which are reducing the opportunities for researchers to access content and engagement data. For instance, Meta has drastically reduced the team working on Crowdtangle and has not confirmed whether it plans to continue offering access to the tool, while Twitter made changes to its API programmes that have introduced excessive costs to academic researchers. This highlights the importance of effective measures to ensure that the DSA operates as intended, particularly if platforms are reluctant to provide such access.

Multi modal datasets

- In addition to the inclusive dataset for fairness [1] shared by Meta AI, the recent availability of multimodal data from Facebook [2,3] made accessible to research through collaboration with Harvard University, holds significant potential for advancing research and analysis. This data encompasses various aspects including demographics of individuals who interacted with web pages shared on Facebook and

aggregated interactions with Facebook and Instagram posts from public pages, groups or people.

- Access to such data and metadata empowers researchers to gain valuable insights into user behaviour, engagement patterns, and the impact of online content. For example, researchers can investigate how different demographics interact with and respond to specific types of content.
- However the types of comprehensive datasets currently available from VLOPS are limited in both scope and subject area. There is a need for greater transparency and data sharing initiatives across the digital landscape, enabling researchers to conduct comparative studies and gain a more comprehensive understanding of online phenomena.

Insights from the Code of Practice on Disinformation

While a limited number of VLOPs and VLOSEs are signatories to the Strengthened Code of Practice on Disinformation, the types of data requests under the Code provide an example of data requests that could be made under the DSA, including to those who are not signatories to the Code. The Code includes specific requests around policies and enforcement. For instance, explanations of systems and procedures to ensure policy enforcement, details and metrics surrounding any appeals processes, or metrics surrounding enforcement at a Member State level. Additionally, the ongoing work on structural indicators for the Code will provide additional insight into the types of questions that researchers might ask of platforms and the types of data required. It would be useful for the Commission to examine the structure of the QRE's and SLI's contained with the measures of the Code, the type of data provided by platforms in their transparency reports and the ongoing discussions surrounding data access taking place between signatories and other stakeholders in the Permanent Taskforce.

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

In this section we have highlighted a number of examples of the types of research into systematic risks our researchers hope to conduct under the DSA.

Compliance Measures:

- **Rights of the Child Rights Impact Assessments**[3] - It is currently unclear how companies work with children to develop/design the safety and reporting tools and to evaluate the effectiveness of safety tools such as reporting tools with children.[4] We currently know that companies work with children to develop these tools but any further information is considered proprietary. To that end, it would be important from the perspective of children's rights to ensure that Child Rights Impact Assessments are conducted by companies prior to implementing their safety tools or products, as they call them. Independent research could then involve children (via quantitative and qualitative methods) in the process of evaluating the effectiveness of these tools and also as to their impact on children's rights, such as privacy, participation, expression. This could be done by conducting a survey, for example, in a specific country, with children of a certain age, inquiring into the effectiveness of the reporting tools. This could also be applied with respect to age-verification (e.g. how many 12-year-olds are using a platform if the digital age of consent is 13 in the given country and is the company in compliance with Article 8 of the GDPR in terms of how it handles their data?)
- **Effectiveness of age-assurance measures** - A number of companies partner with third party providers such as Yoti[5], but it would be helpful to understand better how effective this process is. Inquiring about the specific methods that companies use when someone signs up on the platform and subsequent checks in terms of verifying images, etc. If a child has been detected to be under the digital age of consent but was allowed to be on the platform, how does the company handle the data collected up to that point if parental consent had not been collected (if data processing was based on consent). This concerns GDPR application, but it has been framed as an online safety issue in the public discourse (not allowing under 13 or 16-year olds to be on platforms). It is important to be able to evaluate how companies address under-age use and subsequent personal data management.
- **Compliance with proactive AI-based cyberbullying moderation** - It would be helpful to have access to data that would allow researchers to assess the effectiveness of AI models used by the companies to address cyberbullying. This means access to AI models themselves and to anonymised data that the models were trained and tested on (provided that this is possible for the given context both ethically and according to

---

[3] Mukherjee, S., Pothong, K., & Livingstone, S. (2021). Child Rights Impact Assessment: A tool to realise child rights in the digital environment. London: 5Rights Foundation. Retrieved from: https://digitalfuturescommission.org.uk/wp-content/uploads/2021/03/CRIA-Report.pdf

[4] Milosevic, T., Verma, K., Carter, M., Vigil, S., Laffan, D., Davis, B., & O'Higgins Norman, J. (2023). Effectiveness of Artificial Intelligence–Based Cyberbullying Interventions From Youth Perspective. *Social Media+ Society*, *9*(1), 20563051221147325; Milosevic, T., Van Royen, K., & Davis, B. (2022). Artificial intelligence to address cyberbullying, harassment and abuse: new directions in the midst of complexity. *International journal of bullying prevention*, *4*(1), 1-5;

[5] https://www.yoti.com/

data protection). For example if a model was trained on online hate data but is being applied to cyberbullying detection–what are some of the shortcomings?

Systemic Risk Mitigation:

Freedom of information and expression - Such work, particularly in authoritarian regimes, requires data about government / third-party requests for user/account information and data about takedown or blocking requests from state/non-state actors. Equally important are records of how companies dealt with these requests, how many were satisfied, and how many denied. Some social media companies share some of this data in their transparency reports, but these are mostly voluntary, and having systematic access to this data would be much more useful.

Research on disclosures relating to external takedowns and internal takedowns would inform freedom of information research and would allow insights into company compliance with such requests in democratic and authoritarian settings. It would also provide insights into how companies' own ToS triangulate with external judicial/extrajudicial requests and where there may be room for exploitation of company ToS or local legislation by rogue actors or repressive regimes.

Effectiveness of pre-screening of content - This applies to harmful online content but also the detection of illegal content such as CSAM before it is posted. For example, to our knowledge, YouTube pre-screens their videos upon upload and some recently released tools (CSAI match and Content Safety API) provide some insight into this process, but overall there is much unknown about the pre-screening process, technologies and effectiveness.

Sexual violence and harrassment - For such a sensitive topic, it is relatively difficult to distinguish between disinformation/harmful information and the trustful content helping citizens to speak up about taboo issues. Access to social media data, specifically content and engagement data, allows researchers to gain insights into public opinions, attitudes, and experiences which can lead to a deeper understanding of societal issues, such as sexual violence and harassment, and help identify patterns, trends, and underlying causes. This information can inform the development of evidence-based policies, interventions, and support services to address these issues more effectively as well as informing policy decisions made by platforms.

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?

We echo the view of the DSA Observatory[6] that an efficient process requires DSCs to develop their own research units with the capacity to assess research needs, the credibility of researchers, and the relevance of researchers' proposals. If they lack this capacity, DSCs are likely to be overwhelmed and the process will be severely hampered.

Relatedly, we note that the process will be more efficient if researchers in a given area cooperate and coordinate to prioritise research and share knowledge about the process. This would generate better insights into systemic risks across Europe and prevent DSCs from being overwhelmed with duplicate requests and poorly conceived requests. In the area of disinformation, for example, EDMO is well placed to facilitate a degree of prioritisation and coordination among researchers.

DSCs should develop a EU-wide approach to vetting researchers that includes credibility checks, but does not discriminate against researchers not affiliated with traditional HEIs/Research Centres, so as not to disadvantage precariously employed junior researchers or those working in independent research organisations. This would also require an application process that asks for just enough information, but not too much so as to create an equal playing field.

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?

Cross-country consultations are required to ensure DSCs in all Member States understand the diversity of the researcher environment in each country and the specificities of each national research environment, so that common ground can be found for assessing both the credibility of the researcher and the validity of their request. This need further supports the case for the establishment of research units within DSCs DSCs may need to invite additional expertise to assess requests in specific areas (e.g., CVT/CVE, gender-based violence online, etc.). The vetting mechanism must be transparent and have clear criteria for assessment.

---

[6] https://dsa-observatory.eu/2023/03/10/here-is-why-digital-services-coordinators-should-establish-strong-research-and-data-units/

There may be benefit to subjecting this part of the process to an expert peer-review in some form.

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?

Additional provisions may include obligatory training in data protection law for researchers applying to access data.

It might be helpful to convene an event with researchers from various disciplines who have accessed companies' APIs before for a variety of topics, and also researchers who have an interest in doing so but have not in the past, prior to designing the request for access application form. This would be done to collect feedback as to the data security and confidentiality requirements that researchers need to ensure to receive access to data (DSA Article 40, section 8, d).

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?

Legal obligations may need to be specified for researchers accessing particularly sensitive data, but with publicly available data this may not be much of an issue. Requesting evidence of institutional ethics approval and evidence of research outputs would be appropriate.

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 DSC holds the authority to approve or reject a researcher's data access request, however, there is a potential risk that financial motivations, in the case of a DSC in which a country benefits from being a country of establishment, or the motivation to limit data access requests, in the case of a platform, could influence the fairness and integrity of a DSC's role as an intermediary.

This is particularly true in the case of Ireland, as the Irish DSC is likely to process the majority of requests towards VLOPs and VLOSEs under the DSA, which will require a large amount of resources. It is essential that the application and procedure process is designed in such a way that it minimises reliance upon the staffing, expertise and management process of an individual DSC.

Additionally, it will be crucial to ensure efficient resource allocation and expertise within DSCs. To effectively fulfil its role as an intermediary, a DSC must possess specialised knowledge in areas such as systemic risk, platform governance, data science, and internet regulation. It is essential for a DSC to continuously update this knowledge in order to remain relevant and adaptable in the rapidly evolving landscape of digital services and the associated risks.

Having an independent advisory mechanism (or at least a partial structure) would ensure more rigorous peer review of requests and could allow the opportunity to draw from a pool of relevant experts for specific disciplines/cases. It would also provide for a more centralised process which reduces the level of workload required by an individual DSC and ensures that the process remains transparent and fair.

We also strongly echo the call for DSC's to incorporate a research and data unit to ensure they have the sufficient knowledge and skills to both process research requests and to handle compliance duties with platform data.

Data access formats and involvement of researchers:

a) What technical specifications could be considered for data access interfaces, which takes into account security, data protection, ease of use, accessibility, and responsiveness (e.g. APIs, data vaults and other machine-readable data exchange formats)?

To ensure authorised access and protect sensitive data, VLOPs & VLOSEs should implement robust authentication and authorisation mechanisms like 0Auth or API keys while providing access to data and AI models. Having a dedicated account/user ID for each researcher would be useful to ensure security/track specific access. In addition to complying with GDPR to safeguard user privacy, use of encryption techniques such as SSL/TLS for data transmission and secure storage could be considered for data access interfaces. To optimise the usability and integration of data access interfaces, consider designing intuitive APIs with clear documentation which should have clear naming conventions, parameters, and response structures. 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[7]. While APIs and machine readable formats have become a gold standard, accessibility should also extend to researchers not familiar with more hi-tech data formats, so data should be made available in a variety of ways.

[7] https://developers.google.com/search/docs/appearance/structured-data/dataset

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

As per the opportunities provided by Article 40, VLOPs and VLOSEs should organise workshops, webinars, and training sessions to educate researchers about i) the process of requesting and accessing datasets and AI automated systems, ii) best practices and recommendations for researchers on how to effectively analyse and interpret the obtained data to assess the effectiveness of automated systems, iii) navigating and understanding the technical aspects of accessing and working with the data and the automated AI systems to facilitate data processing, analysis, ,and visualisation to derive meaningful insights from the available data. In addition, VLOPs and VLOSEs should foster collaboration and networking opportunities between researchers, civil society and regulatory authorities to facilitate knowledge exchange and sharing of experiences. Furthermore, VLOPs and VLOSEs should emphasise the importance of ethical consideration in research by encouraging researchers to adhere to ethical guidelines and principles, ensuring protection of user privacy and the responsible use of data. VLOPs and VLOSEs should establish feedback mechanisms to gather input from researchers regarding their experiences, challenges faced, and suggestions for improvement in the data access process, and continuously review and refine procedures and requirements based on the received feedback to streamline and enhance the research community's engagement with VLOP data. However, DSCs with adequate research expertise should be part of this process.

Capacity building measures can also include additional training in regulatory literacy; data protection in research context; responsible use of social media data/metadata in quantitative and qualitative research, including anonymisation of publicly available data. Additionally, researchers could also be trained as independent experts to evaluate research access requests - this would contribute to the creation of a community of practice and a growing pool of independent experts that could be used by DSCs.

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?

Implementing a standard data dictionary or business glossary would involve collaboration among relevant stakeholders, including policymakers, regulators, industry experts, and research communities. These stakeholders would work together to formulate a comprehensive set of agreed-upon terms, definitions, and specifications that are applicable to the data access processes, requirements, and parameters. This dictionary or glossary would serve as a reference point for all parties involved, promoting consistent and coherent communication. The standardised dictionary could include the following,

i) Data request types:
    1) Exploratory data request: A request made by a researcher to obtain data from VLOPs or VLOSEs for exploratory analysis and research purposes.
    2) Compliance monitoring data request: A request made by a regulatory authority or researcher to access data from VLOPs or VLOSEs for monitoring and ensuring compliance with the DSA Act.

ii) Data Access Agreement: The legal agreement between the researcher and the VLOP or VLOSES that outlines the terms, conditions, and limitations of accessing and using the requested data.

iii) Categories of Data:
    1) User profile data: Personal information and preferences of users, including demographic data, interests, and behavioural patterns.
    2) Content data: Text, images, videos, or other media shared or published by users on VLOPs or VLOSEs.

iv) Standard anonymisation techniques: Methods to protect individual privacy by adding noise or perturbation to data before it is, ensuring individual identities cannot be easily re-identified.

Regular updates and revisions of the data dictionary or business glossary would be necessary to keep pace with technological advancements, emerging data access practices, and evolving regulatory requirements. This collaborative approach to establishing a common language would promote effective communication, streamline data access processes, and facilitate a more harmonised and efficient ecosystem for research and compliance under the DSA Act.

To ensure transparency, data requests directed towards large platforms should be logged in a centralised public platform along with the status of which DSC dealt with the request and whether the request was granted, granted with an amendment or refused.

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?

It may be worth collecting all publicly available data in a universally accessible catalogue. If appropriate and if platforms are willing, certain datasets could be made available to be indexed by dataset search engines.

Even publicly available data could be presented in a controlled and structured manner, and additional guidance and training could be provided to researchers on how to access, download and use this kind of data in a responsible way. Such training / best practice handbook could include information about company ToS, EU privacy and data protection legislation, and best practices for ethical research using publicly available extant data in a variety of contexts (e.g., research on elections, research on conspiracy theories, research on gender-based violence, research on vulnerable groups).

A website similar to the signatory maintained website for the Code of Practice on Disinformation[8] could be useful to collect and display datasets, information relating to the data providers, the DSA and its obligations, as well as relevant information for researchers which is specified above.

[8] https://disinfocode.eu/