CDT Europe Comments on Delegated Regulation on data access provided for in the Digital Services Act

Introduction

Providing independent researchers with access to data is a vital component of increased transparency and will be critical to better understanding how systemic risks concretely impact society and to identifying which mitigation measures will be appropriate. The delegated act will need to adopt an iterative approach through which appropriate mechanisms for fundamental rights safeguards and independent oversight are established from the outset, whilst methodologies for operationalising these aspects remain adaptable to the circumstances. Consistent consultation with experts across academia and civil society will also need to be envisaged so that mechanisms can be consistently improved in this regard.

CDT Europe therefore recommends that in the development of this delegated act, priority is placed on: ***the establishment of an independent intermediary body to assist in implementation and oversight; developing a transparent and accessible process for the vetting of researchers; adopting a tiered approach to data access; and ensuring essential fundamental rights safeguards are in place from the outset***.

Independent Third Party Body: Additional Oversight for Effective Research Development & Transparency

As a primary recommendation, the delegated act should prioritise the establishment of an independent intermediary advisory body, with its own legal personality and allocated annual financial resourcing, which can provide support and additional oversight for the implementation and functioning of Article 40. Many of the pertinent questions on privacy protections, ensuring researcher independence and access, and preventing corporate capture or government overreach have been raised by the adoption of this provision. Addressing such reflections will require consistent consultation with experts familiar with similar mechanisms or people with experience in research generally and platform research specifically. The most effective way to ensure this expertise is readily available and consistently informs what will undoubtedly need to be an iterative process is through the establishment of an independent intermediary body as proposed by the EDMO Code.

There are several already foreseen aspects of implementation in which independent expertise would be valuable in addressing potential enforcement bottlenecks. For example, such a body could provide needed advice in the vetting process and the process of informing decisions about what data should be provided and by what methods. Experts such as those who contributed to the European Digital Media Observatory analysis should be consulted to provide insight on how to ensure coherence between GDPR and the DSA. Similar expertise will also be consulted regularly on issues such as data security and privacy protections necessary to ensure that platform data is handled responsibly and that its use for research does not invade individual users' privacy. More crucially, this body will be able to independently assess compliance of researchers and platforms with Article 40 and its delegated acts as well as ensure and assess consistency across Member State DSCs in their implementation and enforcement of the provision.

Data Access Needs for Addressing Systemic Risks

Article 40 could allow a wide range of research to occur on assessing systemic risks and evaluating mitigation measures. This includes projects that evaluate the prevalence and reach of specific risk-posing content, reverse engineer recommender systems to uncover latent patterns, and track the specific impacts of interface and algorithmic product changes. Given the scope of the systemic risks identified in Article 34 and the novel nature in EU-wide mandated research into these risks, there are examples of existing research that can be drawn upon to identify what types of data and analyses researchers may require and develop. For more detailed and long-term research, there are certain types of data that are of great interest to researchers conducting research on systemic risks and mitigation measures, such as:

- Baseline data about the overall volume of data on the service, e.g., the overall number of posts on a service, which is necessary to facilitate effective sampling techniques and wider confidence in results. Knowing that 10 million posts per month contain hate speech cannot reveal the prevalence of hate speech on a service if researchers do not know the total number of posts per month.
- Advertising data. i.e., the content of ads, and other data such as full ad targeting data, information about who purchased or viewed an ad, and data on ad expenditures.
- Social networks or social graph data, i.e., data that shows how users of a social network are connected to each other. For example, researchers have used this data to uncover trends in social media group membership, and how disinformation spreads within and across groups.
- Engagement data, i.e. information about users' engagement with posts through their reactions or comments. (Similar to what CrowdTangle provides.)
- Data on content moderation. Specifically, researchers are interested in data about individual, specific enforcement decisions, which allows them to investigate both platforms' policies and the accuracy, fairness, transparency, and efficacy of a service's content moderation decisions. However, disclosing the specific content a user has posted raises specific privacy concerns, discussed below.
- Data on ranking and recommendation algorithms, i.e. data about how a service generates a candidate pool for recommendation and then ranks it. Specifically, researchers are interested in the relative strengths of different inputs into recommendation algorithms, the ends they attempt to maximise, and how those ends are balanced against one another. Many researchers are interested in using this data to examine the consequences of algorithmic ranking and recommendation of content, and especially whether services treat particular types of content or users disproportionately and the impact of that disproportionate treatment.

These examples provide some insight, however it must be stressed that the potential intersectional manifestation of the systemic risks means that access to data will need to be flexible enough to allow researchers to request a wide variety of kinds of data on a case-by-case basis. In other words, the process should allow researchers to request whatever data they believe is necessary to conduct their specific research. 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.

This is not to say that researchers should necessarily be given access to whatever data they request, as some data may be too sensitive in that regard. However for researchers who may not have prior experience with such data access it would be important to establish what data is available in order for requests to be more precise. Researchers will also need to understand how this data is generated or compiled, as it may impact how they interpret the data. As a primary step, online platforms and search engines within the scope of the provision should be required to publish code books that describe the data they hold and can provide to researchers. Ideally the code books would use common terminology to describe the data. Common language could be developed by the independent intermediary body, and would require significant input from VLOPs and VLOSEs. In addition, because there will be differences between how VLOPs and VLOSEs operate and collect and maintain data, it may not be possible to have entirely common language for all relevant terms.

Similarly, when data is shared, companies should also provide detailed explanations of how they define those data types. If a company shares the number of posts occurring monthly on its platform, it should also define what constitutes a post — do they include reposts or deleted posts? — to allow researchers to better understand how they can or cannot incorporate individual data points into their research. Covered platforms and search engines should also be prevented from overly collapsing categories of data together. For instance, a VLOP could offer a statistic on “engagements” that includes favourites, bookmarks, and reshares. Covered entities should offer as specific categories as possible and allow researchers themselves to collapse them together if they see fit.

Publicly Available Data: Article 40(12)

Article 40(12) clearly establishes a mandated framework for real-time access to data and the delegated act must clarify the circumstances under which researchers will be given access to data that is available as it is generated, or that is at least relatively recent. Recent trends to restrict access to real-time data by platforms requiring exorbitant access fees or imposing harsh rate limits must be addressed. Real-time data is important for research on systemic risks and mitigation measures addressing current events, like mis- and disinformation concerning elections or public health issues, wars and military conflicts, and natural disasters. Providing researchers who meet the conditions of Article 40(12) with access to real-time publicly available data will ensure that a broader spectrum of researchers can contribute to research regarding systemic risks.

The delegated acts must define “publicly accessible.” Because access to this data will be broader than other categories of data, we suggest defining “publicly accessible” in this context to mean data available on the open internet and accessible to anyone without restriction. This would include, for example, content that is available to all users of a service and not otherwise restricted, assuming that anyone can create an account. It would not include, for example, content posted in a closed group on a service, to which users must request access.

As with the rest of Article 40, it is important that Article 40(12) not become an avenue that could be exploited by law enforcement for unjustified or abusive requests for data access. Making it easier for law enforcement to gather even publicly available data could increase its ability to surveil activists, protestors, dissidents, journalists, and members of minority racial, ethnic, or religious communities. Unjustified surveillance based on public data still raises significant privacy and human rights concerns. As a result, the delegated acts should make clear that personnel at law enforcement agencies, organisations that are acting on behalf of law enforcement agencies (such as contractors or organisations funded by law enforcement), or organisations acting in a law enforcement capacity or with a purpose similar to law enforcement should not be granted access to data under Article 40(12). In addition, researchers should be permitted to access data under Article 40(12) only pursuant to data sharing agreements with platforms that include binding obligations on the researchers not to share the data they access with any other party.

Data Access Formats: Technical Considerations

APIs have been powerful tools for researchers whose research depends on accessing platform data in bulk and many say they find APIs to be a robust, comprehensive, and easy way of accessing platform data. However, researchers have also raised concerns about the completeness and accuracy of data from APIs, as well as caps on the amount of data that researchers can access. Vetted researchers should be permitted to obtain data through APIs, and the delegated acts should require that the data provided be accurate and complete. Caps on data collection should be considered on a case-by-case basis, and should be permitted only if allowing greater collection is not feasible or would damage individual users’ privacy.

Static data sets can also be very useful tools for researchers. They take less technical knowhow and investment to use, can provide large volumes of information at low cost both to platforms and researchers, and do not need to be continually evaluated for privacy and security concerns in the same way. For static datasets to be useful, companies need to be forthcoming about their completeness and how they develop and filter them. Since static datasets do not update overtime, they are also limited in how useful they are for research on timely topics or longitudinal studies.

Virtual or physical clean rooms should also be used where feasible. A virtual clean room is a digital environment that would "permit researchers to import their own data, perform research analyses, and export the results of their analyses," while preventing them from exporting the social media data itself.[1] A physical clean room is a space to which physical access is restricted and "data analysis takes place on designated machines that are secured by encryption and disconnected from the internet."[2] Virtual or physical clean rooms would allow a researcher to access, inspect, and analyse data, but not possess it. To ensure the independence of research, the platform's ability to control how researchers use and analyse the data within the clean room environment should be limited to only what is necessary to ensure the security of the data.

Application and Vetting Procedures; Practical Recommendations

There are several aspects that will need to be considered when establishing the initial practical framework for the functioning of this provision. The delegated act should appropriately outline the specificities of some of these mechanisms to ensure the development of vital research can commence as soon as the provision allows.

Data access application and procedure:

A primary consideration should be given to a standardised application process, used across DSCs, that requires all applicants to provide the same information under the same vetting criteria. There should be a reasonable deadline by which DSCs commit to reviewing applications and issuing a decision, and consideration should be given to an expedited track for applicants seeking to study a time-sensitive matter, such as systemic risks related to an upcoming election. The application process should be conducted as transparently as possible. Applications should be made public (subject to redactions, if they contain confidential information) and the DSCs' determination should be announced publicly. The public should have the ability to easily track the number of applications, who is making the applications, the subject matter of the applications, and the DSCs' determinations on applications.

DSC staff should include individuals with relevant expertise to evaluate the importance or legitimacy of requests for data access for research purposes who are tasked with reviewing applications. If this is not possible, consultation with the independent intermediary body should be conducted. Platforms should be given the opportunity to comment on applications before the DSCs approves or denies the application, but limited in the type of feedback they can provide. For example comments about whether the data requested exists and is feasible to provide; whether providing the data would violate the privacy interests of users or the platform or undermine the security of their service, and how those harms can be mitigated; and whether the researchers' technical and organisational measures for ensuring data security and confidentiality are sufficient (and, if not, what technical and organisational measures should be required). The

[1] EDMO Code at Part II, para. 6.3.1.
[2] EDMO Code at Part II, para. 6.3.2.

research applicant in turn should be given an opportunity to respond to the platform's comments and/or modify their application.

Criteria for Vetting Researchers

The delegated act should specifically outline that the vetting procedure should be conducted using a standard application across all relevant Digital Services Coordinators, ensuring coherence between the information requested by all applicants. The application process should be conducted transparently and should provide enough appropriate information to the public in order to allow third parties to assess whether DSCs are acting consistently. Once again, the independent intermediary body should be engaged to assist in the vetting process across all DSCs and can aid in the conducting periodic reviews in which a number of applications are randomly selected and assigned to a different Member State DSC (or DSC of establishment) to review, to see if they come to the same determination as the original DSC.

Balancing Data Access with Privacy Protections

The Commission must ensure that individual user privacy is not compromised by researcher access to platform data under Article 40; this means ensuring researchers understand their obligations under GDPR. As explained in more detail in the Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access (the "EDMO Code"), the GDPR recognises the importance of scientific research and includes both exceptions and limits on data processing for research purposes. These obligations can be respected whilst not hindering the development of vital research to better understand systemic risks. The aforementioned independent intermediary body would play an important role here in ensuring GDPR compliance for research and protecting platform users' privacy.

Alongside this, the Commission should adopt a tiered approach to data access, in which researchers should be permitted to access the most sensitive data, that could have the largest impact on data subjects' rights and freedoms, only if they can demonstrate an equally high level of ability to protect the data. Researchers' plans for keeping data confidential and their research methodology must be carefully scrutinised in the vetting process to ensure they are sufficient and that the researcher is capable of implementing them. However, researchers or vetted institutions will have varying degrees of access to internal mechanisms for data protection and resources. Therefore the use of virtual or physical clean rooms is another way to achieve this method of access, but it is important that the social media company's ability to control how researchers use and analyse the data within the clean room environment is limited to only what is necessary to ensure the security of the data. The delegated acts should also provide for independent auditing of these clean rooms and procedures to ensure that companies are not manipulating data, and require platforms to preserve (consistent with the GDPR) data provided to one researcher and make it available to subsequent researchers seeking to replicate the research.

It is important to note however that certain data may be too sensitive and impactful to share with researchers under any circumstances, or at least not without the individual data subject's affirmative informed consent. This may include one-to-one direct messages; 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.

With this in mind, the delegated acts will need to address data retention, i.e., how long must covered platforms keep data so that it can be requested by researchers? On the one hand, requiring platforms to keep data for longer periods of time may enable more research, especially historical research; on the other hand, data retention poses risks to user privacy and periodically deleting data is an important way to protect users from misuse of their data or data falling into the hands of bad actors.

Ensuring Adequate Human Rights Safeguards

As highlighted, there are several potential aspects in which human rights safeguards will be necessary in order to protect against potential abuses of this provision. Given CDT's existing research however, our recommendations focus on establishing safeguards to minimise the risk of unjustified or abusive law enforcement surveillance using data gathered under Article 40.

EU law enforcement and other governmental agencies have a demonstrated interest in gathering data from social media, and, in some instances, this data gathering may arguably have similarities to research. This raises serious human rights concerns as law enforcement personnel have been shown to use social media data in the past for illegitimate purposes such as monitoring protestors, dissidents, and members of religious or racial minorities. More so, our analysis raises pertinent questions on the rights and safety of researchers who may be less able to resist unjustified law enforcement demands for user data than social media companies. Individual researchers may not know that they can bring legal challenges to attempts to compel access to their data, and they may not have the financial resources necessary to do so. In addition, some researchers may have close relationships with law enforcement agencies and wish to actively cooperate with them.

Succinctly, our research highlight that these safeguards could include:

- Not allowing law enforcement agencies or those affiliated with law enforcement agencies to qualify as vetted researchers under Article 40 and this must be specified within the delegated act as currently, neither GDPR Article 89 nor DSA Article 40 contain provisions that would necessarily prevent a law enforcement agency from qualifying as a "vetted researcher.” Concretely, the delegated acts should provide that vetted researchers may not be personnel at law enforcement agencies, organisations that are acting on behalf of law enforcement agencies (such as contractors or organisations funded by law enforcement), or organisations acting in a law enforcement capacity or with a purpose similar to law enforcement including immigration authorities.

- To the extent possible, providing vetted researchers with access to data through or at the social media company, rather than allowing the researcher to possess it to lessen the risk of unjustified law enforcement demands for data held by researchers. However In some cases, it may be necessary to allow researchers to possess data. The delegated acts should provide flexibility to allow a researcher to possess data if the researcher can demonstrate that possession of data is necessary to conduct their research, again reiterating the importance of the independent intermediary body.
- Requiring vetted researchers to destroy data they are permitted to possess after a certain time period or when their research has concluded. Requiring researchers to destroy data when it is no longer needed would ensure that researchers do not become a vast repository of social media data that law enforcement could attempt to exploit.
- Requiring data sharing agreements between platforms and vetted researchers to prohibit researchers from voluntarily sharing data with any other party unless legally obligated to do so in line with restrictions imposed by GDPR. To ensure this, researchers will need to uniformly understand their data protection obligations under GDPR. This risk can be mitigated by requiring the use of data sharing agreements, with binding obligations, between platforms and vetted researchers, assessed by the intermediary body, before data may be shared under Article 40
- Establishing additional transparency obligations for social media platforms and independent researchers. In this circumstance, and in line with GDPR Article 21, in order to exercise their right to object to the sharing of their data with law enforcement, a data subject must be aware of how their personal data is being used, making transparency from platforms and researchers vital. The delegated acts could mitigate this risk by requiring platforms and researchers to 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, to allow data subjects to exercise their right to object. This could include a system in which platforms provide an interface allowing a data subject to see at a glance which research projects (if any) their data has been shared with. In addition, if researchers cannot provide transparency to data subjects directly, they could make details of their personal data processing public on a dedicated website.
- Providing funding for additional research to understand whether providing data to vetted researchers will make law enforcement more aware of—and likely to demand access to—users’ social media data. Law enforcement agencies can make applications for compelled disclosure of data only if they are aware of the existence and availability of this data. Their applications are more likely to be successful the more they know about the data that might be available, as they will be better able to explain or document why the personal data could be relevant to an investigation. Future research should examine

whether law enforcement agencies are making more unjustified demands for social media data or are more successful in the demands they make following the implementation of DSA Article 40. The European Commission should provide funding to study this issue.

Additional Support Mechanisms for Researchers and DSC's

The knowledge capacity of researchers for their GDPR obligations and the protection of their own rights will need to be developed and the European Commission should provide resources for this in the early stages of enforcement of this provision. Researchers have also identified cost as a barrier to platform research. Even if platform data is available free of charge, there may be costs associated with using the data for research. For example, video and audio content, which are increasingly posted by both advertisers and users, often must be transcribed for research use. Because of the high costs of transcription, audio and video data are, in practice, inaccessible to many researchers. Providing additional funding for platform research is an important capacity building measure, as is providing support for the development of tools and data-analytics capabilities that researchers can employ.