EDMO
European Digital Media Observatory

European Digital Media Observatory

Report on the EDMO DSA Data Access Pilot

December 2024

Prepared and submitted in response to the European Commission's call for feedback on the draft Delegated Regulation addressing technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data pursuant to Article 40 of the Digital Services Act.

10 December 2024

In October 2023 the European Digital Media Observatory (EDMO) initiated a pilot project designed to test potential processes and procedures for implementation of researcher data access provisions under Article 40 of the EU Digital Services Act. The following offers a report on the key outcomes, takeaways, and recommendations from the pilot project.

The report is designed, in part, to provide feedback on the Article 40 draft Delegated Act, which the European Commission published on 29 October 2024; the findings from the pilot expressly inform our reactions to the draft.

At the highest level, the pilot produced **four key takeaways:**

1. **Researchers require mechanisms for updating their applications, including information relevant to GDPR obligations.**

2. **Data provision processes are as important to the integrity of Article 40 as are researcher vetting procedures.**

3. **An independent intermediary can help significantly streamline vetting processes, but timelines must be realistic.**

4. **There will be a significant learning curve for all parties.**

On the basis of these takeaways, the report offers a series of recommendations to both the European Commission and Digital Services Coordinators.

However, by providing a detailed description of the pilot, we hope that the report will also prove informative for all stakeholders involved in the implementation of Article 40.

Rebekah Tromble
EDMO Advisory Council
George Washington University

Claes de Vreese
EDMO Executive Board
University of Amsterdam

Report on the EDMO DSA Data Access Pilot

Overview

In October 2023 the European Digital Media Observatory (EDMO) initiated a pilot project designed to test potential processes and procedures for implementation of researcher data access provisions under Articles 40.4 through 40.11, as well as Article 40.13, of the EU Digital Services Act (DSA). These articles permit Digital Services Coordinators (DSCs) of establishment to make reasoned requests under which Very Large Online Platforms (VLOPs) and Very Large Online Search Engines (VLOSEs) must provide data to vetted researchers who seek to contribute "to the detection, identification and understanding of systemic risks in the Union…and to the assessment of the adequacy, efficiency and impacts of the risk mitigation measures" (DSA 40.4).

The pilot involved researchers from France and the Netherlands, two VLOPs, (proto-) DSCs from France, the Netherlands, and Ireland, and Data Protection Authorities in all three countries. As the pilot got underway, regulators were beginning to draft a delegated act for Article 40 and were considering other guidelines that might be needed to support vetted researcher access. EDMO's pilot was therefore designed to help inform this thinking and to aid regulators, researchers, and platforms alike in identifying opportunities and challenges that might arise as the researcher data access system is put into place.

The following report lays out further details regarding how the pilot proceeded and offers key lessons learned from the pilot.

Background

In May 2022, the European Digital Media Observatory (EDMO) released a report and draft GDPR Code of Conduct on Platform-to-Researcher Data Access. The result of a year-long multi-stakeholder working group engagement, the draft Code of Conduct lays out procedures and guidance for both platforms and researchers to legally and ethically share and process data for research purposes. Though the EDMO draft Code was published before the DSA was finalized, the draft Code offers extensive details and guidance that EDMO believed could be instructive for further specifying how vetted researcher data access processes might work in practice under the DSA.

However, still in draft form, the Code had not been put to the test. Undertaking a pilot would therefore allow EDMO and its partners to simultaneously assess the strengths and weaknesses of the draft Code and its potential for application to the DSA.

Following months of preparatory work, the pilot launched in October 2023 with the participation of:

- Meta and TikTok;
- Researchers from the Institut Mines-Télécom (France) and the University of Amsterdam (the Netherlands);
- (Then proto-) Digital Services Coordinators from the French *Autorité de régulation de la communication audiovisuelle et numérique* (ARCOM),

Dutch A*utoriteit Consument & Markt* (ACM), and Irish *Coimisiún na Meán* (CNAM);

- Digital Protection Authorities from the French *Commission Nationale de l'Informatique et des Libertés* (CNIL), Dutch *Autoriteit Persoonsgegevens* (AP), and Irish *Data Protection Commission* (IDPC);
- The French data repository organisation *Centre d'accès sécurisé aux données* (CASD);
- A scholar who acted as the lead of a "mock" independent intermediary body (IIB) that solicited external expert reviews related to (a) data protection considerations, (b) ethical concerns, and (c) the scientific merits of the researchers' data access proposals.

The pilot was chaired by Dr. Rebekah Tromble, member of the EDMO Advisory Council and Director of the Institute for Data, Democracy & Politics (IDDP) at George Washington University. IDDP staff managed the day-to-day operations of the pilot.

Pilot procedure

Pre-Launch Activities

Dataset Selection and Codebook Preparation

Though the pilot was designed to adhere as closely as possible to the steps and procedures already laid out in the text of Article 40 of the DSA, in order to ensure participation by all relevant stakeholders, certain parts of the pilot necessarily deviated from what might be expected once Article 40 is fully enacted. Most notably, EDMO representatives discussed and came to an agreement with the two participating VLOPs regarding the datasets that the VLOPs would provide before the pilot launched*.* As such, the researchers were artificially constrained to research questions that could be answered by the specific datasets that the VLOPs agreed to provide.

During discussions with the VLOPs about datasets for the pilot, EDMO representatives laid out two key parameters:

1. Datasets should allow for meaningful scientific exploration of questions with policy significance and relevance to the DSA.
2. Datasets should comprise individual-level, not aggregated, data, allowing all parties to examine important data protection standards and procedures.

Following these discussions, Meta agreed to provide a dataset built on its existing URL Shares dataset. For the pilot, the company agreed to provide pseudonymized individual-level data, rather than the differentially-private aggregated data contained in the URL Shares dataset. This individual-level data could be used to more robustly and accurately examine correlations between (a) various user characteristics and (b) views and engagement with Facebook posts containing links to external websites (including, e.g., websites known to traffic in hate speech, disinformation, etc.).

TikTok in turn agreed to provide a dataset containing pseudonymized information about individual-level user content moderation reports, as well as the outcomes of those reports. These data could be used to study (a) the relationship between user characteristics and report type and volume, (b) how TikTok responds to different types of user reports, and (c) whether report outcomes differ based on user characteristics.

Both companies created codebooks for their datasets and provided these codebooks to EDMO and the participating researchers before the researchers prepared their data access requests. After receiving the codebooks, EDMO representatives and the researchers provided feedback to the platforms, requesting clarifications about certain variables in the codebooks and asking for additional information to help contextualize the data. (For further discussion of the codebooks, see Key Takeaway #1 below.)

Data Repository and Secure Processing Environment

After discussions with EDMO representatives, the platforms chose to provide these datasets to the researchers via the *Centre d'accès sécurisé aux données* (CASD), a secure data access center based in France that has a long track record of enabling researchers to legally and securely analyze sensitive data from a variety of sectors, including business, finance, education, and health. CASD makes data available to researchers via remote access to its secure processing environment, sometimes referred to as a "data cleanroom" or "data enclave." As neither dataset included sensitive category data, all data were pseudonymized using randomly generated, irrevocably hashed user IDs, and other data protection measures were applied to the data before they were transferred outside of the companies, a secure processing environment was likely unnecessary for this use case.

Following guidance in the EDMO draft Code of Conduct, the proposed processing activities for these data would likely be assessed as "medium risk," for which the draft Code advises relying on a limited-access API, rather than a secure processing environment, as the latter imposes significant constraints on researchers' analyses. However, EDMO representatives and the researchers agreed that, as part of the pilot, it would be useful to assess whether a third-party secure processing environment might be a reasonable alternative to relying exclusively on secure processing environments created and maintained by the platforms themselves. (For further discussion of the data intermediary and data access mechanisms, see Key Takeaway #4.)

Data Access Application Preparation

Article 40.8 of the DSA specifies seven conditions that researchers must meet to be vetted. Quoting directly from the text, researchers must demonstrate that:

a. *they are affiliated to a research organisation as defined in Article 2, point (1), of Directive (EU) 2019/790;*
b. *they are independent from commercial interests;*
c. *their application discloses the funding of the research;*
d. *they are capable of fulfilling the specific data security and confidentiality requirements corresponding to each request and to protect personal data,*

and they describe in their request the appropriate technical and organisational measures that they have put in place to this end;

e. *their application demonstrates that their access to the data and the time frames requested are necessary for, and proportionate to, the purposes of their research, and that the expected results of that research will contribute to the purposes laid down in paragraph 4;*

f. *the planned research activities will be carried out for the purposes laid down in paragraph 4;*

g. *they have committed themselves to making their research results publicly available free of charge, within a reasonable period after the completion of the research, subject to the rights and interests of the recipients of the service concerned, in accordance with Regulation (EU) 2016/679.*

Following guidance laid out in the EDMO draft Code of Conduct on Platform-to-Researcher Data Access, supplemented with information provided by the researchers about common practices for conflict of interest, funding, employment, and similar disclosures and verification, EDMO representatives and IDDP staff prepared a set of template forms for researchers to document that they meet the above requirements. The templates comprised three core elements:

1. Documentation attesting to the researchers' affiliation with a research organisation as defined in Article 2, point (1), of Directive (EU) 2019/790;
2. Statements certifying that the researchers were free from conflicts of interest;
3. A Data Needs and Management Plan containing:
    a. An overview of the research to be conducted;
    b. A data protection risk assessment;
    c. A proposed set of organisational and technical safeguards to be put in place to mitigate identified data protection risks;
    d. An analysis of ethical considerations that extend beyond data protection;
    e. A discussion of publication plans; and
    f. A disclosure of the source(s) of funding for the research.

Each of these sections is explained in additional detail below.

Researcher Affiliation and Organisational Qualification

To complete this section, researchers were asked to provide:

1. An organisational mission statement;
2. Documentation of the organisation's not-for-profit status;
3. Evidence of accreditation as an academic institution from an EU Member State approved source; and

4. A statement, signed by an institutional signatory, attesting to the researchers' affiliation with the qualifying organisation.

Requests 1-3 were satisfied with public links to documents available on the research organisations' websites. Note that #3 above would not apply to a non-academic institution. Digital Services Coordinators should consider what other forms of documentation might help verify the authenticity of non-academic research organisations.

Conflict of Interest Documentation

Researchers signed conflict of interest forms that confirmed:

1. A lack of employment conflicts;
2. A lack of meaningful conflicts due to financial interests;
3. A lack of engagement with adversarial state actors (be they sanctioned countries or any nation state's defence, intelligence, or law-enforcement agencies);
4. Commitment to ensure compliance with these requirements for all staff engaging on a project; and
5. Commitment to rapidly notify the relevant Digital Services Coordinators if any of the above changed during the project.

Data Needs and Management Plan

The Data Needs and Management Plan (DNMP) template was modeled on guidance provided in the EDMO draft Code of Conduct on Platform-to-Researcher Data Access, with modifications designed to meet the unique requirements of researchers requesting data access under Article 40.4 of the DSA.

The DNMP template was designed to be an *initial application* document. This has two important implications. First, the template assumes as little as possible about the exact *records* or *rows* of underlying data. This is necessary because, in most instances, at the time of application, the specific data or datasets that researchers are requesting would not yet be created, organized, and/or documented in a data codebook. Thus, researchers would not yet know the volume of data (i.e., number of observations) that would be made available. And in many instances, researchers would only be able to speculate about the *columns* or *variables* that might be available.

In short, when initially submitting an application, researchers are unlikely to be able to provide fine-grained details about the data they would expect to receive. As such, the DNMP template asks researchers to consider the rough *shape* of the data they would be accessing – thinking through the implications of accessing particular categories of data, etc. In this way, the DNMP template bears similarities to a GDPR Data Protection Impact Assessment (DPIA).

However, researchers will not be able to finalize a traditional DPIA unless and until they know the more precise contours of the data that will in fact be made available to them. (For more on this issue, see Key Takeaway #1 below.)

The DNMP template also covers more information than a traditional DPIA. This points to the second important implication of the DNMP template's design: In an attempt to avoid duplicating information and in order to support streamlined application and review processes, the DNMP template weaves together a number of related but separate purposes. That is, it offers a format in which researchers can discuss the purposes and merits of their research, the proportionality of their data request, and other details that are required under DSA Article 40, while linking these considerations directly to data protection risk assessments and analyses of appropriate data protection safeguards. (For further analysis regarding the advantages and disadvantages of this approach, see Key Takeaway #4 below.)

The DNMP template therefore includes six sections.

Research Overview

The research overview requests information about:

1. The research team, including the researchers and their institutions;
2. A description of the proposed research, including:
    a. The objectives and the DSA systemic risks (DSA Art. 34) being assessed;
    b. Specific research questions; and
    c. Hypotheses, if relevant.
3. The specific data the researcher would need from the platform in order to conduct the research, including:
    a. Specific data points/variables required;
    b. Detailed analysis of the personal data required, along with a justification of its necessity made by tying the personal data back to one or more research questions (data minimization);
    c. Any additional data to be added or combined to the requested platform data for the purposes of the proposed research;
    d. Descriptions of the intermediate and output data, provided to ensure that if additional personal data, including special category data, are being inferred through the research, appropriate safeguards will be applied;
    e. Justification of the "proportionality" of such data needs, as per DSA Article 40.8(e).
4. The proposed research methods, with sufficient detail provided such that an expert reviewer could assess if the proposed methods, applied to the requested data, could, in fact, answer the proposed research questions.

Risk Assessment

This section asks researchers to assess the risks of the proposed research along the two axes proposed in the EDMO draft Code of Conduct on Platform-to-Researcher Data Access. The draft Code explains the logic behind the risk assessment framework in detail, but in sum it asks researchers to assess the risks presented by data inputs, outputs, and research processes based on: (1) the reasonable expectations of data subjects and (2) the potential impact on data subjects' rights and freedoms.

Organisational and Technical Safeguards

Following the risk assessment procedures laid out in the EDMO draft Code of Conduct results in categorising the proposed research as either "high," "medium," or "low" risk. The draft Code, in turn, offers recommendations as to the organisational and technical safeguards to apply under each category. Based on this approach, this section of the DNMP asks researchers to discuss the organisational and technical safeguards that they would and could implement and to propose a specific mechanism for accessing the data (e.g., direct dataset transfer, limited-access API, secure processing environment, or on-platform researcher sandbox (for surveys and experiments)).

Further Ethical Considerations

While there is a great deal of overlap between data protection considerations and the ethical concerns typically addressed in human subjects research, the latter is broader beyond the former. The DNMP template therefore includes a separate section that asks researchers to answer a series of questions drawn from the Association of Internet Research (AoIR) Internet Research Ethics 3.0 guidelines:

1. What are the potential benefits associated with this study? Who benefits and how?
2. Do you plan to seek consent from the data subjects in your study? If so, how? If not, why not? Will there be bystanders in your dataset (people whose personal information is in the data but they were not the user that originally posted/authored the content [a friend in an influencer's video but the friend is not an influencer])
3. Who are the subjects of your study? Are there any vulnerable populations (pregnant women, children, prisoners)?
4. Could the findings of your study harm a particular community?
5. Will your data set include data subjects outside the EU?
6. Does your study include deception?
7. Does your study offer participants incentives? If so, how are those incentives determined and distributed?
8. What is the culture (typical use, affordances and norms) on the platform you are studying? How do your research questions and methods interface with this culture?
9. Does your study require the use of labelers or task work (Mechanical Turk or similar)? If so, how did you determine fair enumeration?

10. Does your study include a review or discussion of dangerous content? What therapeutic counter-measures will be available to your research team?
11. Are you using methods that require large amounts of computing power (deep machine learning models, generative models, etc)? Why are these methods required over less compute-intensive methods? Are there actions your team is taking to reduce the project's impact on the environment?

Publication Plans

This section asks researchers to offer a publication plan, as public access of research findings is a requirement of DSA Article 40.8(g).

Project Funding

This section asks researchers to disclose the funding source for this specific research project, as required in DSA Article 40.8(c). In the pilot, researchers disclosed that they were receiving a stipend from EDMO for their participation and shared the amount of the stipend.

Pilot Activities

Working together, researchers from the Institut Mines-Télécom and the University of Amsterdam prepared two data access requests. The request for TikTok data was led by the University of Amsterdam, and the request for data from Meta was led by the Institut Mines-Télécom. In the first official step of the pilot, the lead researcher from the University of Amsterdam submitted the request to the Dutch A*utoriteit Consument & Markt* (ACM) on behalf of all of the researchers, and the lead researcher from the Institut Mines-Télécom submitted the request to the French *Autorité de régulation de la communication audiovisuelle et numérique* (ARCOM). The pilot then proceeded as follows:

1. The local DSCs verified that all required documents were provided. In one instance, the local DSC noted and requested a missing document.
2. The local DSCs then passed the application materials to a scholar acting as the head of a mock independent intermediary body (IIB). DSA Article 40.13 envisions the possibility of DSCs relying on "intermediary mechanisms" to provide advisory opinions as part of the researcher vetting and other data access processes. For the pilot, the mock IIB organized independent expert evaluations of
    a. data protection considerations;
    b. ethical concerns; and
    c. the scientific merits of the researchers' applications.

   This process was in line with preparations being made by EDMO's Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms.

3. Several review processes then occurred simultaneously:
    a. The mock IIB solicited the independent expert reviews described above;
    b. The local DSCs conducted their own reviews, focusing in particular on requirements under DSA Article 40.8(a-c).
    c. The local DPAs also reviewed the application materials, focusing on data protection considerations. Though review by DPAs is not required under DSA Article 40, as part of the pilot, all parties were interested in receiving feedback on data protection issues.
4. The mock IIB provided favorable advisory opinion letters to the local DSCs, sending copies to the researchers.
5. Drawing on the IIB advisory opinions and their own internal reviews, the local DSCs each reached favorable decisions, informing the researchers and IIB.
6. The local DSCs forwarded the application materials, along with their recommendations, to the DSC of Establishment, the Irish Coimisiún na Meán.
7. The Coimisiún na Meán conducted its own review.

At this stage of the pilot, several complications arose that caused significant delays in and changes to the intended final stages of the exercise. DPA capacity limitations; staffing changes, plus other inopportune timing at the Coimisiún na Meán; logistical complications between the researchers at the University of Amsterdam and CASD; and disagreements over contracts between the researchers and TikTok meant that the Coimisiún na Meán ultimately did not issue mock reasoned requests, as intended, and the researchers did not receive access to the TikTok data. However, the researchers did receive access to the data from Meta. Moreover, as the pilot’s very purpose was to uncover potential obstacles in data access processes and procedures under DSA Article 40, even these failures provided important lessons for all parties involved.

We discuss the most important lessons and insights drawn from the pilot below.

Key Takeaways

#1 Researchers Require Mechanisms for Updating Their Applications, Including Information Relevant to GDPR Obligations

There are a number of reasons researchers will need to update their applications over time, including after “vetted” status has been granted. For instance, researchers may add or remove personnel from a project. When a data access request is first submitted, funding for the research may be uncertain, or researchers may secure new or additional funding once data access has been granted. Researchers may move institutions, and so on.

However, while conducting the pilot, an even bigger concern became apparent: *the difficulty researchers face in assessing data protection risks and proposing appropriate safeguards before the full contours of the data are known, data access mechanisms have been selected, and other protections such as contracts are in place*. Pilot participants began referring to this as the data request "chicken and egg dilemma."

Take, for example, a researcher's request for data on user age. Platforms typically have at least two types of user age data:

1. User-reported age
2. Algorithmically-inferred age

User-reported age is notoriously inaccurate. Many users do not provide this information, and many others provide false information. Algorithmically-inferred age data tends to be much more accurate, and, thus, in many instances, researchers would prefer the latter.

Presuming that a platform has such data, a researcher might request the inferential age data, bucketed by inferred birth year, and perform a risk assessment on the basis of the assumption that they will receive this data. Because the requested data would be relatively fine-grained–making re-identification more likely–and because users are less likely to expect that such data would be available to researchers (i.e., because few users even know that it exists), the researcher might propose relatively strict organisational and technical data safeguards, perhaps even proposing that the data be analysed in a secure processing environment. A secure processing environment would limit some of the techniques and analyses that the researcher could run, but the researcher recognises that this is appropriate, given the risk inherent in the data processing activities.

However, imagine that, after receiving a reasoned request for the algorithmically-inferred age data, the VLOP in question responds by saying that, while they do have inferred age data, it is not nearly as fine-grained as the researcher requested. Instead, the company uses very broad age categories–under 18, 18-34, 35-50, 51-65, and 65+. If the DSC verifies this to be true and the researcher accepts this substitute, the researcher may also want to update the data protection risk assessment and proposed safeguards, as the risk of re-identification may be reduced and a secure processing environment, with its attendant constraints on the researcher's analysis, no longer necessary.

Of course this scenario might also run in the opposite direction. That is, a researcher might request data that seems relatively low risk, only to learn that available data do in fact present more significant risks–e.g., because the data may leak sensitive category information or the applicable data are so sparse that the re-identification risks are higher than anticipated.

As part of the pilot, researchers received data codebooks from both Meta and TikTok. As such, the researchers knew what variables were available to them, how fine-grained each variable was, and what data protection techniques had already been applied to the data. For example, both platforms provided individual-level user/user report data but pseudonymized the data by randomly generating irreversibly hashed user IDs. Variables containing user characteristics (e.g., location or age) were provided in rather wide buckets or bands (e.g., country-level and 10-year age ranges).

This was a best-case scenario. In the early stages of Article 40 implementation, researchers will likely have to make many guesses–perhaps informed and reasonable guesses, but guesses nonetheless–about what data could be requested and how it might be made available. Robust platform data inventories and codebooks will develop over time. Indeed, even with direct, productive cooperation between the researchers and platforms that participated in the pilot, it took several rounds of discussion and feedback to achieve the level of detail needed for the codebooks to support the researchers' data protection assessments.

During the pilot, researchers also had the advantage of knowing that they would access the data via a secure processing environment, and they had access to substantial documentation regarding the technical specifications and other procedural safeguards involved in accessing the data through CASD.

And yet, even this amount of foreknowledge was insufficient to fully complete standard data protection impact assessments as part of the data request applications themselves. Because the platforms had not yet prepared the datasets in question, the researchers lacked essential information about the underlying data–in particular how many observations there would be, as well as how many unique users would be found in the dataset. Researchers also lacked information about the contractual terms under which the platforms would be providing data to CASD, and were therefore missing important information about the terms of data storage and deletion and additional requirements placed on CASD relevant to data protection.

The researchers also lacked finalized contracts between themselves and their institutions, on the one hand, and CASD and the platforms, on the other hand. (For further discussion of the contracts, see Key Takeaway #2 below.) Finally, because the researchers would receive pseudonymised data, the researchers were not in a position to help data subjects manage their rights under the GDPR. Only the creators of the datasets–the platforms themselves–were in such a position, but the researchers had no means by which to guarantee that the platforms would do so.

All of the above information is needed for researchers to fully exercise their obligations under GDPR; yet, when the initial data access requests were submitted, there was no way for the researchers to gather this information. In other words, without knowing precisely what data are available, in what form, and through what mechanisms, researchers will need to update their data protection risks assessments and proposed safeguards once the final contours of the data and access mechanisms are known.

Recommendations

Data Inventories

Recital 6 of the draft Delegated Act for DSA Article 40 states the following:

> *To help applicant researchers to design effective research projects and to reduce the administrative burden of the data access process, data providers should provide an overview of the data inventory of their services easily accessible online, including indications on the data and data*

structures available, and where possible, indicate suggested modalities for accessing them.

EDMO sees this as a crucial and welcome mechanism for addressing the data request "chicken and egg dilemma" described above. The more information researchers have about potentially available data at the outset, the more efficient and effective the data access processes will be for all parties.

EDMO recognizes that VLOPs' and VLOSEs' data inventories will necessarily evolve over time, requiring some patience from both researchers and regulators. Indeed, researchers will need to provide feedback on precisely what types of information are needed as part of the data inventories. However, creating such data inventories is not only feasible, EDMO considers it essential to the effective implementation of DSA Article 40. Most VLOPs and VLOSEs already have APIs that they make available to developers, commercial and other partners, and the data provided via these APIs are of course documented for those partners.

These APIs offer a natural starting point for platforms, regulators, and researchers alike to begin thinking about what data inventories created in response to the Delegated Act might look like. Indeed, before it was deprecated in 2023, Twitter/X had a best-in-class researcher API that contained data, and documentation for those data, based on researcher needs and feedback. EDMO recommends that all parties consider this as a model for the data inventories referenced in Recital 6.

Robust Reasoned Requests

Even with informative data inventories in place, the pilot demonstrated that researchers will still need additional information to fulfill their data protection obligations under the GDPR. This includes information contained in data sharing agreements and other contracts, as well as information about VLOPs' and VLOSEs' own data protection activities. As such, EDMO recommends that reasoned requests issued by a DSC of Establishment to one or more VLOPs/VLOSEs at minimum:

- Provide guidance on the contracts to be relied upon by platforms, research organisations, and, where applicable, data intermediaries. EDMO has created a model data sharing agreement that could be adapted for the DSA's purposes.
- Require that a VLOP/VLOSE provide relevant information about how it assists users in exercising their rights under the GDPR, updating researchers and regulators should these practices change.
- Where data intermediaries are used, require that a VLOP/VLOSE disclose contractual information relevant to researchers' GDPR obligations.

Application Procedures

Following insights gleaned from the pilot, EDMO recommends that the DSC of Establishment confer "vetted" status to a researcher on the basis of the best information available at the time of the initial application and allow researchers to revisit the data protection risk assessment, technical and organisational safeguards, and other elements of a traditional data protection impact assessment once full information about the data and access mechanisms are available. In many instances, additional or more constraining safeguards may be needed. However, researchers should also be permitted to propose less-constraining safeguards–particularly in regards to data access mechanisms–when appropriate. EDMO also recommends that researchers be provided with a mechanism–ideally via the DSA data access portal referenced in the Article 40 draft Delegated Act–to notify DSCs of changes to their research teams, funding, or other elements of their research plans.

#2 Data Provision Processes Are As Important to the Integrity of DSA Article 40 As Are Researcher Vetting Procedures

While many of the lessons learned during this pilot related to application reviews and researcher vetting processes, the most significant obstacles arose when planning and implementing processes related to data provision. Notably, three key concerns emerged related to (1) contracts, (2) the period of data availability, and (3) data quality.

Contracts

The previous section discussed problems that arose as part of the researcher vetting process because contractual terms were unknown when researchers submitted their initial applications. Yet even more fundamental issues arose as researchers and their institutions grappled with the contractual terms that were presented to them by the companies and CASD.

Interestingly, Meta and TikTok took very different approaches to these contracts. Meta, for its part, chose not to require contracts between itself and the researchers or their institutions. Instead, it entered into a contract with CASD that allowed CASD's data analysis agreements with the researchers' institutions to suffice. This offered a distinct advantage from the researchers' perspective, as they would simply be required to use CASD's standard agreements, some of which had already been signed by their respective institutions. Further, researchers preferred this mechanism from a political perspective, as it meant avoiding contractual relationships with an entity that was overwhelmingly more powerful than their own institutions.

However, the power imbalance was not fully resolved, as Meta refused, despite repeated requests, to share a copy of the contract between the company and CASD with EDMO representatives or the researchers. CASD and Meta each discussed the broad contours of the "triggering mechanism" that would allow CASD to provide the researchers with access to the data via its secure processing environment. (For the purposes of the pilot, this was a letter from the mock IIB attesting to the positive external reviews it received.) However, neither EDMO nor the researchers could confirm whether there were additional, undisclosed terms that might impact the researchers' data access, ability to analyse the data, right to download the results of their analyses, and so on.

Though TikTok also deposited its data with CASD, contrary to Meta, TikTok took the view that it would need data sharing agreements between themselves and researchers directly. While the company used EDMO's model data sharing agreement as a starting point, negotiating its terms to meet the needs of the pilot took numerous rounds of back-and-forth. Particular issues arose around questions regarding (a) what security requirements were to be put on researchers (vs CASD itself), (b) what the obligations would be of each party in the event of data breaches, and (c) what, if any, obligations researchers had if they were to create inferences that were special category data under the GDPR. Ultimately, at least one of the researchers did not feel comfortable signing this contract, and as a result, the researchers never received access to the TikTok data.

Period of Data Availability

The companies also differed regarding the length of time that they were willing to deposit the data with CASD and make it accessible to the researchers. Citing its consent decree with the US Federal Trade Commission, Meta initially proposed a 30-day timeframe. That is, it would contractually require CASD to terminate researchers' access to the data after just 30 days.

EDMO representatives adamantly pushed back on this. Not only was 30 days woefully inadequate to complete analyses, but EDMO representatives were aware that the company had received a public letter from the FTC warning them against using the consent decree to avoid or undermine transparency efforts, including "good-faith research in the public interest." Meta eventually acknowledged that the specific 30-day timeframe was not dictated by the consent decree and agreed to extend the data access period to 90 days. EDMO representatives and the researchers still felt this was insufficient. Indeed, the researchers were not able to complete the (already limited) analyses that were planned for the pilot. But it was the best EDMO representatives were able to achieve during negotiations with Meta for the pilot.

TikTok, in contrast, agreed to provide its data for 365 days. However, as noted above, researchers were ultimately unable to access the TikTok data.

Data Quality

From the start of the pilot, researchers were concerned about how to assess data completeness, accuracy, and other data quality issues, and they planned to focus their analyses on these questions. Without access to the data, researchers could not assess the quality of the TikTok data, and they ran out of time to complete their intended analyses of the Meta data. However, what checks they could complete suggested that some data were likely missing from the dataset they received. None of the parties involved in the pilot had considered possible remedies had this concern been confirmed.

Recommendations

Robust Reasoned Requests

As discussed above, EDMO recommends that the DSC of Establishment provide guidance in its reasoned request on the contracts to be relied upon by platforms, research organisations, and, where applicable, data intermediaries. And where data intermediaries are used, EDMO recommends that the reasoned request require that any portion of a contract between the VLOP/VLOSE and the data intermediary that might impact researchers' ability to compete their approved research.

In light of the data provision concerns that emerged in the pilot, EDMO is also particularly encouraged by the Article 15.3 in the draft Delegated Act, which states that “When providing access to data, data providers shall not impose archiving, storage, refresh and deletion requirements that hinder the research referred to in the reasoned request in any way.” However, as Article 15.3 could be read to apply only to data that researchers store or otherwise hold directly, EDMO believes that the language in Article 15.3 should be clarified to explicitly prohibit these restrictions being imposed, no matter where the data reside, including with data intermediaries.

Expanded Researcher Involvement in Amendment Requests and Mediation Procedures

Article 12 of the draft Delegated Act lays out procedures for amendment requests–i.e., requests made by a VLOP/VLOSE to amend a reasoned request. This includes proposals for alternative data. As specified in Article 12.5, if the DSC of Establishment determines that a change to the requested data is justified, the DSC of Establishment “may consult the principal researcher to enquire about the suitability of any alternative proposals submitted by the data provider for attaining the objectives of the research project proposed in the data access application.” EDMO concurs with this provision, but also recommends that Article 12 permit the DSC of Establishment to consult with researchers about any element of the reasoned request likely to have a material impact on the proposed research.

Furthermore, EDMO recommends that Article 13 of the draft Delegated Act be updated to allow researchers to request mediation proceedings, either before or after data provision. As the pilot demonstrates, the ability for researchers to request mediation is likely to be particularly important if the researchers identify data quality issues.

#3 An Independent Intermediary Can Help Significantly Streamline Vetting Processes, But Timelines Must Be Realistic

EDMO’s work to set up an independent intermediary body (IIB) shortly after the EDMO-led multi-stakeholder working group released its draft Code of Conduct on Platform-to-Researcher Data Access. In the accompanying report, members of the working group unanimously called for the creation of a new data access intermediary body that, among other core tasks, could facilitate independent expert reviews of research proposals and researcher qualifications in support of digital platform data access (under any regime, DSA or otherwise). Review processes overseen by an independent organisation would increase the independence of researchers and could help reduce the legal liabilities faced by platforms that had previously insisted on reviewing and approving researchers’ proposals before providing data access. Under the DSA, such an independent intermediary can help increase the capacity of regulators by facilitating interactions with, and advice from, experts.

The pilot bore this out. Working from a pool of contacts from around the world, the head of the pilot’s mock intermediary body was able to solicit reviews from experts in digital research, ethics, and data protection, and these reviewers delivered their evaluations in a timely fashion.

However, two relatively small issues that arose during the pilot are likely to be magnified as the DSA Article 40 system gets off the ground. To begin, fewer than half of the 15 experts that the mock IIB leader contacted were available to participate. This was not problematic for the pilot, as only a handful were needed. However, this reflects larger, more systemic issues inherent in peer review systems, as experts' time is stretched incredibly thin. (Note that this is one of the reasons that EDMO's planning for a real-world IIB includes compensating experts for their time.) Pulling from a global pool of experts, an IIB is more likely to successfully recruit reviewers than would local DSCs. But all parties must be realistic about these constraints.

Second, despite their deep expertise, reviewers brought in by the mock IIB struggled to understand the purpose of their reviews, as well as how to properly digest the materials. In short, this was a very new exercise for the reviewers. While there were some overlaps with standard review processes for scientific journal articles, the Data Needs and Management Plan looked nothing like a typical journal article. And though the head of the mock IIB provided extensive written guidelines, several of the reviewers still struggled to apply the requested standards and procedures. In the context of the pilot, this was easily manageable. The mock IIB leader was able to answer questions, receive feedback, and, when necessary, intervene to course correct. At scale, this will not be possible. Instead, based on the pilot, we believe that the IIB will need to develop more robust training materials for reviewers.

Recommendations

Each of these concerns points to the need for realistic expectations regarding the timelines for researcher vetting processes. Put simply, no matter who is responsible for organising advisory input, it will take time to identify and properly train expert reviewers. In the current draft of the Delegated Act, Article 7.2 gives the DSC of Establishment just 21 days (following confirmation that an application is complete) to review and provide a decision to the researcher.

Unfortunately, EDMO believes that even under the best conditions, this timeline is unrealistic. Though researchers certainly desire an efficient and timely review process, they also understand the constraints involved. On the basis of the pilot, and after having consulted with members of the research community, EDMO recommends a 60-day timeline for researcher vetting processes.

#4 There Will Be a Significant Learning Curve for All Stakeholders

Challenges that arose during the pilot's expert review processes point to another key takeaway: Everyone involved in the DSA data access regime will need to learn an unfamiliar and complex system.

For researchers, pieces of the process naturally align with standard procedures for creating research proposals. They will devise research questions and formulate hypotheses. They will consider what data they need to conduct the research and tie that to specific analytical methodologies. But for many, that's where the similarities will end.

Rather than developing a lengthy and scientifically robust research proposal, researchers will need to prepare a concise and efficient text that foregrounds

1. how the research will support the study of systemic risks in Europe,
2. data protection considerations, and
3. the proportionality of the data request.

This did not come naturally to researchers during the pilot. They were not yet at ease with DSA-specific terminology and expectations, and they (and the EDMO representatives) discovered that they did not necessarily think about "systemic risks" in the same way as the participating DSCs.

Similarly, some expert reviewers were looking for lengthy methodology sections and detailed discussions of the scientific merits of the proposal. Though the researchers certainly could have produced these, for the purposes of DSA Article 40, this was not necessary and would have significantly delayed the application writing and review processes.

It is also worth noting that the researchers who participated in the EDMO pilot came from relatively well-resourced institutions. They had significant in-house expertise (lawyers, data protection officers, technical support) to assist with various aspects of the pilot. This will not be the case for many other researchers who seek to make use of DSA Article 40 mechanisms.

We also found that the participating platforms–though genuinely eager and productive partners–struggled to help internal stakeholders understand this new, complex process. For Meta and TikTok, the pilot involved work with engineers and data scientists, project managers, policy personnel, and lawyers, among others. For example, many of those involved in producing datasets for the pilot did not understand the intended use case (research) well enough to realize when their decisions might adversely impact data quality or other factors relevant to sound research. Ultimately, it is VLOPs'/VLOSEs' (legally-mandated) responsibility to sort out these issues, the pilot left little doubt that this will present challenges.

Encouragingly, the staff of the participating DSCs were highly constructive partners who were willing and eager to jump straight into the deep end in this pilot. But even they were learning an unfamiliar system. And differences emerged among them about how to proceed. For example, one of the local DSCs intentionally chose to give the advisory opinions from the IIB significant weight, while the other local DSC performed more thorough, additional in-house reviews.

Moreover, as EDMO representatives and the various DSCs sought feedback from Data Protection Authorities, the DPAs' capacity limitations proved a significant obstacle. This was a new test case, even for the DPAs, and their inclination was to look for information in a familiar form–i.e., in traditional data protection impact assessment forms. However, as described above, there were a number of reasons why traditional DPIAs were not fit-for-purpose, and because all parties (but the regulators in particular) were so busy, it proved difficult to find time to sort through competing expectations and determine what changes might be needed to EDMO's templates in order to satisfy all parties.

Recommendations

The DSA Article 40 draft Delegated Act represents an important step forward in providing clarity to all stakeholders involved in the DSA researcher data access regime. However, it does not–and cannot–go nearly far enough to resolve the issues described just above. Instead, EDMO recommends that the Board of DSCs develop detailed, clear, and harmonized guidelines for each step of the process. Indeed, we hope that this report will offer some support in that direction.

EDMO also recommends that the Commission and Board of DSCs work with the independent intermediary body to develop training materials and regular workshops for various stakeholders.

EDMO

www.edmo.eu

The European Digital Media Observatory has
received funding from the European Union
under contract number LC-01935415