SciencesPo
MÉDIALAB

Delegated Regulation draft on data access provided for in the DSA

Feedbacks from médialab Sciences Po, France

December 2024, Paris
https://medialab.sciencespo.fr/
contact: Benjamin Ooghe-Tabanou benjamin.ooghe@sciencespo.fr

About médialab Sciences Po

médialab is an interdisciplinary research laboratory composed of sociologists, engineers and designers, conducting thematic and methodological research to investigate the role and impacts of digital technology in our societies using innovative participatory and digital methods.

Digital technology is transforming our societies on the whole and the production of scientific knowledge in particular. Through the multitude of data that it generates, the digital turn contributes to deepening our understanding of different social worlds, while offering at the same time investigative tools that spawn new forms of knowledge production through the modelling, visualization and interactive exploration of data. These research approaches are developed around a variety of themes revolving around systemic risks such as the diffusion of disinformation, fake news or hate speech on the digital public space, as well as the penetration and transformation of our societies by new artificial intelligence computation technology.

To empower such studies, data on and from Very Large Online Platforms (VLOPs) and Search Engines (VLOSEs) such as Google, YouTube, Facebook, X (Twitter), Instagram or TikTok have been crucial resources over the past decade, and the recent trend of restricting the means of access to that data over the past few years has raised a wide variety of concerns to ensure our capabilities to secure our ongoing and future work.

Global overview of the draft delegated regulation

In that regard, médialab researchers are potentially very interested in benefiting from the delegated regulation on data access provided for in the DSA and welcome this very important effort.

The proposed regulation draft is very promising and includes some crucial parts which should be preserved within future versions, including elements along some of the lines we already encouraged during earlier stages of the regulation drafting process.

We also believe that some parts of the proposed draft could sometimes be more precise, enriched or reformulated to better ensure the accessibility and usability of the communicated data, as well as the transparency and well functioning of the process.

On the means of access to data, its formats and documentation

The usability of the data provided to researchers by VLOPs and VLOSEs through the delegated act processes is critical to ensure studies can effectively be led. This entails that the modalities of access to data, as well as the form under which it is provided, in terms of formats and corresponding documentation, must be precisely described within the regulation.

In that regard, **proposed recitals 16, 18, 26 and 27 are very important; they could still be enriched and should be better reflected within Articles 9 and 15** of the regulation's main text.

- Ensure data is usable without any constraint thanks to open formats

While Article 15 is entitled *"Data format and data documentation", it* does not actually provide any requirement regarding formats.

Europe proposed over the past decades different versions of an European Interoperability Framework[1] which defines open specifications for data formats along the following lines:

- *all stakeholders have the opportunity to contribute to the development of the specification and a public review is part of the decision-making process;*
- *the specification is available for everyone to study;*
- *intellectual property rights to the specification are licensed on fair, reasonable and non-discriminatory terms, in a way that allows implementation in both proprietary and open source software.*

As recommended in Recital 27 (*« data providers should not impose any restrictions on the analytical tools employed »*), in order to ensure researchers will be in capacity to work with the data released by data providers without any legal, technical or financial constraints (for instance having to buy specific proprietary software to be able to process the data, hence making the work hardly reproducible), data should always be provided along those same open principles.

Article 15 should therefore include an extra paragraph explicitly mentioning the requirement for data to be provided under such open formats:

> ***« 1 bis. Data providers shall provide vetted researchers the data requested in machine-readable, non-proprietary formats respecting the openness principle of the New European Interoperability Framework. »***

[1] https://ec.europa.eu/isa2/sites/default/files/eif_brochure_final.pdf

- **Ensure technical documentation and metadata is provided with datasets**

  + **Documentation is not an option**

Article 15.2 requires data providers to provide some documentation along with the data which is a critical point to ensure data is useful for the research purposes.

But the second part of the article also leaves data providers the liberty to consider that providing such documentation would result in a *"significant vulnerability",* in which case they shall only have to notify the DSCE and propose an alternative documentation *"where possible"* only.

Such significant vulnerability should only be judged as such by the DSCE and should never lead to a complete absence of documentation, which would result in rendering the data hardly usable to attain any research objective.

Therefore we propose to rewrite **Article 15.2** as such:

> *« 2. Data providers shall provide vetted researchers with relevant documentation related to the data requested. In cases where* ***the data provider considers*** *the provision of such documentation* ***would*** *result~~s~~ in a significant vulnerability, the data provider shall notify the Digital Services Coordinator of establishment and,* ***~~where possible~~*** ***upon its approval,*** *propose alternative documentation. »*

  + **Provide technical documentation as metadata for single data exports**

Whenever data is provided to researchers as a single downloadable snapshot export rather than an ongoing access to live data (through an API or a secure processing environment for instance), it is essential that such data is properly documented in terms of date and means of export, so that the researchers can know exactly what the data they work with corresponds to.

For instance, considering the potential delays between the initial data access application and the final data access delivery, knowing the exact date and time of the export can be critical.

Similarly, if the export corresponds to some form of filtering or querying of the provider's full datasets, it is essential for the researchers to understand how such filtering was realized. For instance, if requesting metadata on all users of a platform describing themselves as medias within their bios, one needs to know whether the provider only searched for the word "media" in lowercase, or also for any kind of case, as well as different spellings such as "média", "medias", etc. The same principles would apply also for instance if the provided data corresponds to an aggregation of data.

Such examples of metadata could be included within **Recital 26:**

*« (26) In order to enable the navigation and usability of the accessed data for the purposes of the research and to put them in the proper context, data providers should provide vetted researchers with the relevant metadata and documentation describing the data made available, such as codebooks, changelogs**, date and time of export, methods of querying or filtering** and architectural documentation. »*

In addition, metadata should be explicitly mentioned along with documentation within the first sentence of **Article 15.2:**

*« 2. Data providers shall provide vetted researchers with relevant documentation **and metadata** related to the data requested. »*

- **Ensure data access modalities enable attainment of research objectives**

Whereas both recitals 16 and 18 take into consideration the critical need to ensure attainment of the research objectives when specifying modalities of access to data by DSCs, Article 9 on that regard does not, and that requirement only appears in Article 12 for the handling of potential amendment requests from data providers.

Article 9.2 should therefore be completed as such:

*« 2. When determining the access modalities, the Digital Services Coordinator of establishment shall **ensure it enables the attainment of the research objectives of the research project and at the same time** take into account the sensitivity of the data requested, the rights and interests of the data provider, including the protection of confidential information, in particular trade secrets, and the security of its service. »*

- **Ensure researchers' freedom of methods and tools choice**

    + **Data providers should not impose researchers further actions**

Recital 27 and Article 15.3 explicitly forbid data providers to "*impose archiving, storage, refresh and deletion requirements"* to researchers.

In our experience, such constraints are sometimes considered by VLOPs and result in making research work very complex, or even impossible. For instance Twitter's former free API required users to remove data on deleted tweets previously collected without providing the means to be informed of such deletion.

Therefore, we would like to strongly express support to **keep that mention within Recital 27 and Article 15.3**.

+ **Researchers should be able to run the algorithms or codes they need**

Recital 27 also explicitly states that *"data providers should not impose any restrictions on the analytical tools employed by vetted researchers, including relevant software libraries"*, which is also very important to ensure the freedom of choosing whichever method, algorithm, tool, library or software they need.

Although, Article 15.4 actually provides for the opposite in cases where the reasoned request from the DSC would explicitly specify it.

We do not see any reason why the DSC would enforce such an extra constraint to a data access application submitted by researchers, especially since it would result in seriously limiting the analysis capabilities.

Therefore we believe **Article 15.4 should be removed**.

> *« ~~4. Data providers shall be allowed to limit vetted researchers' use of standard analytical tools, including relevant software libraries, for the analysis of the data requested, only if it is specified in the reasoned request.~~ »*

+ **Secure processing environments shall not limit researchers capacities**

Recital 16 and 18, as well as Article 9.4 define the conditions and the modalities under which data providers should provide data through *"secure processing environments"* at the request of the DSCE in cases where data security, confidentiality and protection of personal data might require it.

As former members of Facebook's Social Science One programme as well as beta testers of the Meta Content Library, we already have some experience working within such secure processing environments and believe it is quite important to ensure the secure environment constraints do not limit the capacities of researchers to work with the data.

First, this includes ensuring the environment is made accessible to the researchers without any discrimination, which means no proprietary tool shall be required to access it. Many researchers work under free open source software environments which would not be able to install specific softwares built for Windows or MacOS in order to access or connect to the secure environments.

Moreover, to preserve the academic freedom of running any kind of data analysis, it is critical that those secure processing environments allow researchers to use any external resources they could need, such as other data, or the programming languages and libraries of their own choice, as well as their own algorithms or pieces of code.

Therefore, we suggest to complete **Article 9.4** with the two following extra points:

« **(a bis)** ***ensures that vetted researchers can access the secure processing environments without requiring the installation of any proprietary software;*** ***»***

« (c bis) ***ensures that vetted researchers can upload onto the secure processing environment other existing data, software, library, algorithm or code;*** *»*

On the nature and extent of requestable data

- **Ensure data providers' inventories are easy to find**

Recital 6 and Article 6.4 propose that data providers shall provide "*easily accessible online"* public *"data inventories of their services [...] and suggested modalities to access them"*. This was part of our suggestions in the early stages of the drafting process and we are very enthusiastic about that recommendation which we believe will be a very important tool for researchers to know before requesting anything what is already available, where and how, and therefore what should indeed separately be requested through the data access application process.

Although, in our experience, data providers' online websites are logically not architectured first for researchers but for their users, and these websites are often prone to quick changes, potentially making it quite time consuming to search for that information on each data provider's website. Having all of these data inventories centralized in a single place would be very helpful to easily find and compare them. Since the DSCE and the EU Commission will already be the recipient of diverse informations from the data providers such as their contact details referred to in Article 6.3, we believe it would make things more coherent and considerably easier for researchers if the data inventories would be published, or at least linked, directly on the DSA data access portal. This would also be coherent with the mission of the portal defined within Article 3.1.(b) since it is supposed to *"serve as the single digital point of exchange of information on the data access process"*.

We therefore propose to complete **Article 6.4** as such:

« 4. Data providers shall make available and easily accessible on their online interfaces the details of the point of contact, a link to the DSA data access portal as well as an overview of the data inventory of their services, including examples of available datasets and suggested modalities to access them. ***Data providers should also communicate to the Commission the link to that overview of the data inventory, so that it can be publicly listed for all data providers on a single information page on the DSA data access portal.*** *»*

- **Provide for as many kinds of requestable data as possible**

Recital 12 provides an important list of examples of data which should be made available for the study of systemic risks, including already many good elements such as user profiles and

relationship networks as well as data on personalised content recommendations, ad targeting or content moderation.

Some more examples we can think of are not included yet though and we believe it can only be fruitful to add these within Recital 12 as such:

> *« Current examples of such data include data related to users such as profile information, relationship networks, individual-level content exposure and engagement histories; interaction data such as* ***views, likes, clicks, answers,*** *comments or other engagements****, both as aggregated counts as well as detailed individual timed actions; logs of active and passive individual users activities****; data related to content recommendations, including data used to personalise recommendations* ***and content feeds; data related to internal quantitative or qualitative qualification or ranking of users or elements used in any form of sorting algorithm*** *; data related to ad targeting and profiling, including cost per click data and other measures of advertising prices; data related to the testing of new features prior to their deployment, including the results of A/B tests; data related to* ***any form of modification, reporting, addenda or deletion of individual elements by their author, the other users or the data provider itself, including*** *content moderation and governance, such as data on algorithmic or other content moderation systems and processes, archives or repositories documenting moderated content, including accounts as well as data related to prices, quantities and characteristics of goods or services provided by the data provider.* ***Data related to any individual user or content accessible online should always be provided with means of verification such as direct URLs towards corresponding elements online.*** *»*

On the submission and processing of data access applications

- **Ensure more transparency on the data access applications**
    - **+ Also publish information on the refused data access applications**

The proposed draft provides within Article 11 that the DSA data access portal will publicly host overviews of each reasoned request formulated by the DSCE to the data providers. Overviews shall contain an abstract of the research topic and systemic risks involved, the name of the data provider, a description of the expected data and the access modalities determined by the DSCE.

We strongly advised for such transparency of the data access applications within earlier consultation steps of the regulation drafting process. We are therefore very supportive of this article.

Although, we believe that the process could be even more transparent and that all applications should be publicly documented, whether approved or not by the DSCE. Indeed, it will be critical for researchers to know in advance whether the data they are interested in has already been requested, approved or not and why, before investing time on formulating

an actual data access application. This should also help reduce the potential flow of applications to be handled by the DSCs.

We propose therefore that Article 11 should be rewritten to not only apply to reasoned requests, but also to the reasons why a reasoned request could not be formulated as defined in Article 7.2.(b).

This could for instance be the case with the following reformulation of **Article 11**:

> *« Publication of the overview of the ~~reasoned request~~**data access application** in the DSA data access portal*
>
> 1. *Upon **action from**~~formulation of the reasoned request,~~ the Digital Services Coordinator of establishment **as provided in Article 7 paragraph 2, it** shall publish an overview of the ~~reasoned request~~**data access application** in the DSA data access portal. The overview shall contain:*
>
> *(a) the summary of the data access application referred to in Article 8, point (i);*
>
> *(b) **if a reasoned request was formulated,** the access modalities for the sharing of the data to the vetted researchers**;***
>
> *(c) **if a reasoned request could not be formulated, the reasons why**.*
>
> 2. *Where necessary as a result of an amendment request or a dispute settlement procedure, the Digital Services Coordinator of establishment shall update the overview referred to in paragraph 1 accordingly. »*

+ Include the timeline of a whole application within the public information

Evaluating the procedure, delays and impacts of the proposed regulation will with no doubt be a matter of interest to both researchers, journalists, data providers and institutions themselves.

In that regard, it will be very important that the published overviews include the detailed timelines of the different steps composing a full data access application process. This could easily be achieved simply by including, within the publication of the data access applications overviews, the different dates involved:
- date of the first receival of the data access application by a DSC;
- date of first answer by the DSC to the principal researcher responsible (Article 7.1);
- date of formulation or refusal to formulate a reasoned request by the DSCE (Article 7.2);
- date of initiation of access by the data provider as communicated to the DSCE in Article 15.1.(d);
- date of termination of access by the data provider as communicated to the DSCE in Article 15.1.(e);
- when necessary the dates of initiation and settlement of an amendment request or dispute.

We propose therefore to edit **Article 11** by adding an extra point to paragraph 1 as well as an extra paragraph 3:

*« **(a bis) the dates of processing of the data access application by the Digital Services Coordinator, the Digital Services Coordinator of establishment and the data provider;** »*

*« **3. The Digital Services Coordinator of establishment shall update the dates referred to in paragraph 1 point (a bis) following the initiation or settlement of an amendment request or a dispute settlement procedure and the notification of access and termination of access referred to in Article 15 paragraph 1.** »*

- **Ensure Digital Services Coordinators' capacity to process all requests**

Article 7 proposes strict delays within which DSCs will have to examine, validate and convert into reasoned requests the researchers' data access applications. The proposed delays of 5 and 21 days are very ambitious and are a strong sign of goodwill that should be saluted.

Although, our experience for instance with Freedom of Information processes in France illustrates that such legal delays might not always be respected if the administration is not in human capacity to process important volumes of requests submitted within short periods of time, resulting in a loss of trust into the legal process.

In that regard, extending slightly those delays could seem reasonable, especially considering the huge expectations from the research communities towards the settlement of this new data access process provided by the DSA.

We therefore recommend **adjusting those delays to respectively 10 and 42 days within paragraphs 1 and 2 of Article 7**.

- **Ensure researchers can also amend reasoned requests**

A research project's methodology and means can greatly vary along the work depending on the material and circumstances. Regarding external data especially, whenever interested in manipulating data one has never explored before, it is most of the time very difficult to know in advance exactly the nature or volume of the data that will turn out to be the most interesting and practical to work with, or what field or metadata will reveal useful.

But in the proposed regulation, Article 8 establishes that a proper application shall define precisely a wide range of information which can very easily evolve during the research project. For instance Article 8.3 requires to provide detailed information about funding, which can be granted or refused depending on the capability to access the requested data. Similarly, Article 8.5 requires *"information about the volume, scope, granularity and type of the data requested"* whereas quite often simply knowing the volume of data related to a specific subject or request is actually part of the reasons motivating the request.

For all these reasons, we believe it is very important to provide for researchers the capacity to amend their data access applications later in the process, as well as raise disputes the same way data providers can as defined in Article 13.

First of all, in order to do that, **the definition of "amendment request" within Article 2 paragraph 8 should be enlarged to include principal researchers** as well as data providers.

Moreover, **Article 12 should be completed or mirrored to provide for duly motivated amendment requests of data access applications by the principal researchers, which should be examined by the DSCE and communicated to the data providers upon approval**.

- **Ensure dispute settlements procedures always reach a solution**

Article 13 defines very precisely all the steps for dispute settlement procedures to be handled. Although, it leaves an open window for such procedures to end without reaching an agreement between the data provider, the DSCE and potentially the principal researcher. Since the article does not provide any means of action in the latter situation, this could potentially leave the option to data providers to raise disputes as a way to evade some legitimate data access applications. To prevent such a situation, the regulation should probably provide for some sanctions mechanism in such cases.

Therefore we would like to advocate for **adding to Article 13 an extra paragraph explicitly providing the consequences for all parties involved in refusing to reach an agreement within such procedures**.

Minor legislative drafting notes

Reading the proposed draft revealed a couple legislative drafting mistakes which should probably be fixed:

- **Articles 10 & 11**

Articles 10.1.(d) and 11.1.(a) both refer to *"Article 8, point (i)"* although such point does not exist in Article 8.

Articles 10.1.(d) and 11.1.(a) should refer in that place to *"Article 8, **paragraph 9**"*.

- **Article 15**

Article 15.1 provides a list of points which start with point *"(d)"*.

Article 15.1.(d) and 15.1.(e) should be adjusted as 15.1**.(a)** and 15.1**.(b)**.