GLOBSEC
IDEAS SHAPING THE WORLD

Access to data for researchers: A state of play

Centre for Democracy & Resilience

Introduction

The work of independent researchers and analysts who study content that has a potential to cause harm to individuals, groups of people or the whole society is indispensable for democracy. Since the spread of mis- and disinformation became one of the key societal threats, a strong community of experts has emerged with a capacity to study all ABCDEs (actors, behaviours, content, degree and effect)[1] linked to malign content. However capable, the community is still strongly dependent on one factor – data availability (open APIs, web-based repositories) of the platforms where the content, including disinformation, spreads. The degree of openness of data – and the choice of which data would be available for research and in which form (API, dashboard, etc.) - has been dependent upon each platform's decision. The introduction of the Digital Services Act (DSA) has created new opportunities in this respect, with the Article 40 outlining procedures under which researchers should be able to get access to data for activities that contribute "to the detection, identification and understanding of systemic risks in the Union". The regulation distinguishes between two types of data access for vetted researchers – publicly accessible data in the platforms' online interfaces (according to Art 40.12) and non-public data accessible through applications submitted to the national Digital Services Coordinators.[2]

Through the "publicly available" provision the DSA should, in theory, address a core need of immediate access to publicly available (scrapped) data, which is key for monitoring spread and potential virality of malign content in real time and for prompt addressing. The Strengthened Code of Practice on Disinformation also addresses the need in the Commitment 26, which states "Relevant Signatories commit to provide access, wherever safe and practicable, to continuous, real-time or near real-time, searchable stable access to non-personal data and anonymised, aggregated, or manifestly-made public data for research purposes on Disinformation through automated means such as APIs or other open and accessible technical solutions allowing the analysis of said data."[3] The still ambiguous co-regulatory nature of the Code, however, provides little guarantee.

As the Commission's overview of data accesses published in April 2024[4] and our survey results from September-October 2024 confirm, the access as outlined above is not a given and differs greatly by platform, with researchers either relying on lengthy procedures or expensive monitoring tools. This situation demands urgent improvement, as data access for experts committed to studying

1 https://www.jstor.org/stable/pdf/resrep26180.6.pdf
2 https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en
3 https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation
4 https://digital-strategy.ec.europa.eu/en/library/status-report-mechanisms-researcher-access-online-platform-data

and countering the spread of malign content that undermines democratic values, processes, and institutions is an essential prerequisite for strengthening EU resilience. The absence of such access diminishes overall situational awareness, hampers swift response mechanisms and should, therefore, be recognised as a security threat.

About this report

The report outlines the results of an online survey that ran from September 19 to October 25, 2024, developed by GLOBSEC as a part of the Central European Digital Media Observatory (CEDMO) and EDMO projects[5] for experts conducting research in the area of countering foreign information manipulation and interference and disinformation.

The subsequent analysis is based on 54 responses collected from experts from think tanks, academia and research institutions and organisations primarily from across Europe to provide the current state of data access for research. It is primarily intended for EU and national policy makers to understand the extent of the limits and burden the lack of regulation that would enforce real-time data sharing puts on the research in disinformation and related phenomena. All the questions included in the questionnaire can be found in Annex no.1 at the end of the report.

This report has been funded by the Central European Digital Media Observatory (CEDMO) Project, which has received funding from the European Union under the call: DIGITAL-2023-DEPLOY-04, project 101158609. This report reflects the views only of the authors, and the Commission cannot be held responsible for any use that may be made of the information contained herein.

CEDMO
Central European Digital Media Observatory

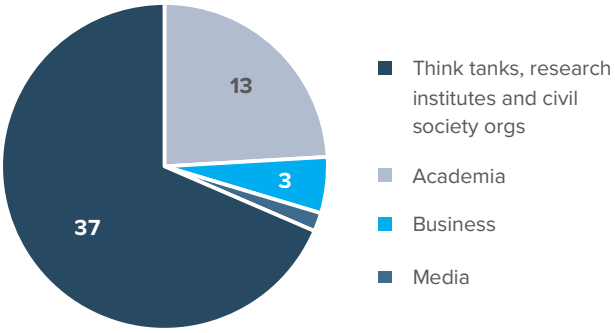
EDMO
European Digital Media Observatory

Survey results

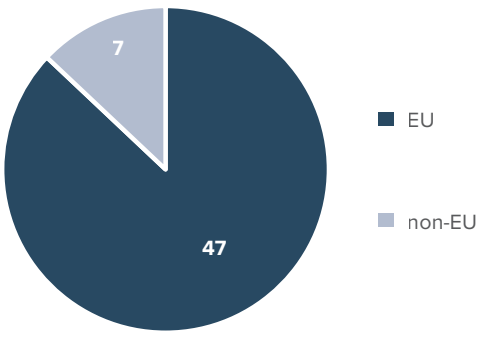
54 researchers filling out the questionnaire are working from 21 different countries, with some covering multiple countries or regions. The greatest representation was secured by Slovakia with 7 respondents, Poland with 6 respondents and Czechia and Romania with 4 respondents. Most - 46 - respondents were from the European Union, 3 had a workplace in the United Kingdom, 1 from Ukraine, 2 from the US, with having offices and researchers based in the EU, and the remaining 2 from other parts of the world.

Most of the respondents are working in non-governmental research institutes, think tanks or smaller civil society organisations, seven came from academia, three from SMEs and one from the media sector. The structure of respondents highlights that a significant portion of research is conducted by non-academics, which demonstrates the importance of ensuring that representatives of NGOs and think tanks have equal rights to access data as universities do.

Type of work of surveyed researchers



Place of work



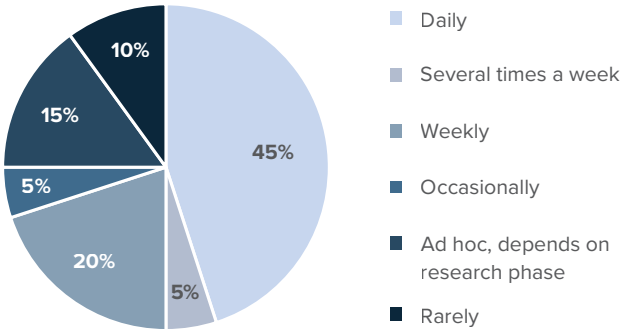
5 https://cedmohub.eu/; https://edmo.eu/

Tools used for research

From the 54 respondents, 40 are currently using some tools for research, while 2 used them in the past. 30 different tools were mentioned by researchers, ranging from commercial social media listening providers, self-built data scrappers, to official platforms' API accesses. Meta Content Library was mentioned the most often, together with a commercial tool Gerulata Juno, followed by Meltwater and Sentione. 70% of researchers use the tools from daily to weekly basis.

Tool used for research	No. of mentions
Meta Content Library Gerulata Juno	8
Meltwater	6
Sentione, Own platform	5
Newswhip Youtube API access Facebook Ads library Google Ads library	4
Brandwatch Tiktok Research API Junkipedia	3
Bright data, Pulsar	2
Apify, Brand24 Buzzsumo, Exolyt Infegy Atlas Letsdata, Newsvibe Newton Media, Pyrra Quid Monitor, Sensika Sherlock, Telethon Twint, X Pro Zeeschuimer	1

Frequency of using the tools



Case of CrowdTangle

CrowdTangle, first independent, later Meta-acquired tool with functionalities that allowed researchers monitor publicly available data on Facebook and Instagram for free, ceased to function in August 2024 without a proper replacement, despite open calls from NGOs and academia to wait until proper replacement.[6]

Already during the tool mapping, CrowdTangle was mentioned 10 times as one of the main tools respondents used to use for their work, 29 respondents claimed they had access to CrowdTangle when specifically asked about it. Out of those who had access, 76% used it on a daily or max. weekly basis for their work. The cancellation of CrowdTangle, was, by the majority, seen as a loss for the research community which is difficult to fully replace. Only 1 respondent claimed they had managed to fully replace the tool from the perspective of functionalities, while **15 claimed they had partial replacement, and 13 said they had no replacement. Researchers with no replacement** for CrowdTangle varied in terms of the work they used it for – they represented academia, think tanks, as well as smaller NGOs and came from all parts of Europe.

Below are some quotes from the surveyed researchers regarding how CrowdTangle's cancellation impact their ability to monitor and analyse data.

6 https://foundation.mozilla.org/en/campaigns/open-letter-to-meta-support-crowdtangle-through-2024-and-maintain-crowdtangle-approach/

How did cancellation of CrowdTangle impact your work?

"Significantly - CrowdTangle gave us option to work globally on research related to elections & minorities"

"Loss of the API was the critical issue for us, but the Meta Content Library is a good replacement for Crowdtangle and we had access early."

"It means we have to find workarounds anytime we want to get large-scale content from Meta platforms, and those workarounds are not stable and/or cost money and/or break often and/or don't scale as well."

"Our ability to research Facebook/Instagram has been significantly disrupted. Projects tracking extremism across platforms have seen their volumes on those platforms drop to 0."

"The monitoring requires more time and effort without CrowdTangle. Also, it affects the efficiency of our work – much more content flies under the radar."

"Ongoing projects had to be stopped."

"It was a nice tool, because it was quite easy to use, but it lacks of all the data about personalization, making it useless for a lot of investigation I am carrying out.."

"It has made a big negative impact as it was the only monitoring tool we had continuous access to. We are unsure about how to go forward because as an NGO, we cannot afford to pay for expensive monitoring tools."

"We have access to Meta Content Library (MLC) API. However, there are some changes that makes this tool less valuable than CrowdTangle."

"MCL API is not very intuitive. We had to go around a couple of times to understand how we can download the data."

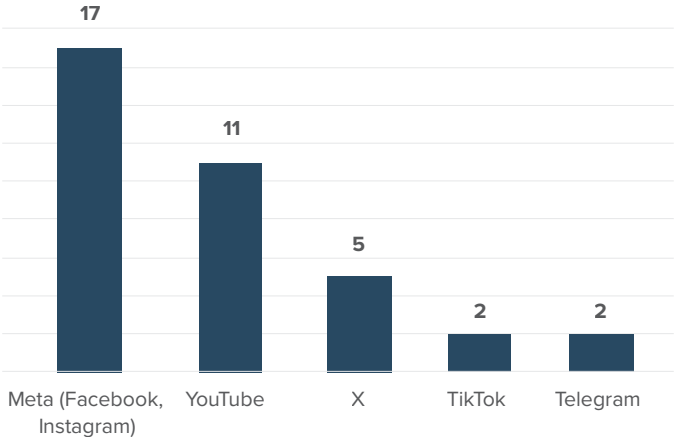
"It forced us to completely change our data pipelines for our tools and to purchase an external service to access at least some FB and Instagram data. In short, it cost us significant time and money."

Access to social media platforms' API and content libraries

23 respondents stated they had some access to platforms' APIs, out of 26 who applied. Most have access to Meta's Facebook and Instagram and YouTube, with only some to TikTok and X, with some using Telegram's open API. The experience with getting access to data was labelled more negative than positive, receiving an overall score of 4.6/10 from 23 different organisations.

The quality of data provided also varies, with Meta and Google receiving more average and positive scores than the other platforms. The differences in experiences and evaluations also point to the broad variety of types of research that is being done and needs to be conducted in the field of disinformation and other types of malign or illegal content.

No. of researchers using access to the following platforms' API or content libraries



Main issues or deficiencies faced by researchers include a **difficulty in accessing data**, characterised by **slow, complex, and often bureaucratic application processes**. This issue is particularly pronounced for Meta Content Library (MCL) and X, where academic and NGO researchers face repeated access barriers, extensive Terms & Conditions, and long waiting times for approval. In addition, several respondents mentioned that contracts offered by the platforms often include liability rules that the organisations and universities cannot sign with conditions requiring potential costly audits, which results in not submitting the application.

The current state of play which requires researchers to re-apply for access, for example at Meta Content Library, with every specific project is another major barrier, that disables quality real-time data gathering for general situational awareness and crisis situations.

"Slow, cumbersome data access applications, which often feel suited for academic institutions rather than nonprofits. Or simply no workable solution available."

On a scale of 1 to 10, researchers rated their experience with data access at an average of 4.6

(1 = terrible, 10 = amazing)

Reasons for not applying

The main reasons cited for not applying for data access revolve around lack of necessity, complexity of the processes, and limited resources, which force them to delay or deprioritise applications for platform APIs. From 27 responses, approximately:

1. **30%** of respondents stated that their **current research needs were being adequately met by existing tools or method**s. They felt there was no pressing requirement for additional data access.
2. **20%** mentioned being **unaware of the availability** of data access options or unclear about the eligibility and application processes.
3. **15%** cited the **complexity and administrative burden** of applying as a deterrent, emphasizing that the process seemed daunting or inefficient.
4. **15%** noted they lacked the **resources, capacity, or organizational readiness** to pursue applications, often needing to establish their priorities or finalise organizational structures first.
5. **10%** of respondents explained that they were in the **early stages of organizational growth** and not ready to engage with platform APIs.
6. **5%** preferred to rely on **existing data access arrangements** or partner programs.
7. **5%** expressed worries about meeting the platforms' **strict criteria or the efficiency of evaluation processes**, some wrongly assumed that access is limited to universities.

Indicative waiting time

One of the key barriers mentioned by the researchers was the time the researchers had to wait to have their access approved /or rejected by each platform. Below are the responses to an open-ended question "How long have you been waiting for a response? (please provide information per platform)", with a reference to previous questions on application for direct access to social media researchers' APIs or content libraries granted by social media platforms. The table below summarises the responses that depend on which platforms the respondents chose to assess, divided depending on whether the application was approved, rejected or the respondents are still waiting for a decision.

The table below showcases that a current system is not at all reflecting the needs of researchers to conduct effective research and have access to real-time data to be able to respond to immediate crises or information influence operations. The regulators should ensure that the access to libraries according to Art 40.12 will be granted based on organisations or individual researchers, not projects. In practice, if an organisation applies for an access, its licence to use the platforms' tools should be granted for min. 2 years without a need for re-application. Also, time limits of max. 1 month should be imposed to platforms with regards to decision-making on data access, including a solid reasoning and a possibility to appeal against a decision.

Several researchers mentioned their applications and consequent denial of data access were not properly reasoned by the platforms, which hinders the possibility of appealing and making necessary adjustments to succeed.

	Access approved after the waiting period of approximately:	Currently awaiting a decision for approximately:	Access rejected (time and reasoning)
Meta Content Library	3 months 3 months 2 months 2 months 1-2 months 1 month	2 months 1,5 months	No timing specified, with a reasoning that the organisation did not qualify as a research institution (it was a think tank in Bulgaria).
Meta API	5 months 2 months 2 months 1 month	1,5 months	
X API	7 months 4 months Less than 1 month	7 months 5 months	After less than 1 month, with a reasoning that the research project would not contribute to the study of DSA systemic risks (despite the fact that the DSA office disagreed). Another respondent mentioned it was impossible to get access to academic APIs, having all applications rejected.
YouTube	1 month		
TikTok	7 months 1 month 1 month	5 months after an appeal following a rejection after 4 months (see on the right) 6 months 3 months	4 months, with a reasoning that the applicant was from "an institution that is not a US/EEA university". Eligibility criteria, however, did not mention university at all, instead the requirement is to be "a non-profit academic institution in the U.S. or Europe". Another applicant was rejected on the grounds of not meeting requirements.
Bing Qualified Researcher			1 week, with a reasoning that datasets requested fell outside the scope of "publicly accessible data", as referenced in Article 40, paragraph 12 of the DSA.

Quality of data

Another significant problem is the **lack of data quality and granularity**. Data provided is often limited, with constraints like user thresholds that strongly restrict conducting effective analysis on a representative sample. Many noted that Meta's content library lacks essential features, such as fact-checking labels and other content moderation actions taken by the platforms,, and lacks many features that had previously been available with CrowdTangle, including access to full archive of content.

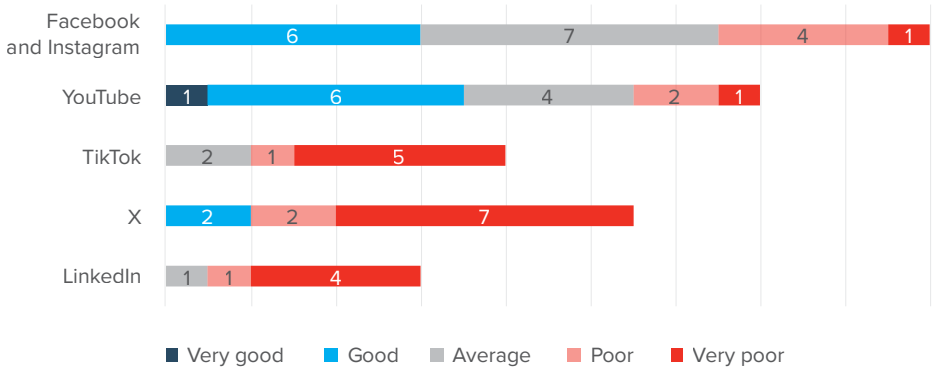
Additionally, there are **limitations in API functionality**, including restrictions on volume, and reduced query allowances, which hinder efficient research. The necessity of sharing sensitive information, such as ID photos, to gain access raises privacy concerns. Overall, while tools like MCL are functional once accessed, the overarching sentiment is that the **processes and data limitations significantly impede research efforts**.

"Data quality, barriers for access that are set up for academic researchers, limited ability to do real-time analysis, access often limited to either specific project questions or individuals within a team."

"Data is blurred by default and therefore unsuitable for many analysis methods (URL Shares)."

The responses to the desired features for new social media data platforms and APIs reveal a consistent need for **greater data accessibility, granularity, and functionality**. A common theme is the desire for **downloadable and customizable data sets**, preferably in CSV format, to streamline research processes. Many users emphasised the importance of **access to detailed and downloadable engagement and content data**, including specific

What is the quality of data provided by platforms' API and content libraries per platform?



metrics (total and unique views, reactions, shares and comments), types of content included in the post – video, picture, source URLs, as well as interactions and content of comments under the post.

Researchers also want **more transparent and precise advertising data**, such as exact amounts spent on ads and targeting criteria (instead of ranges), to better analyse the influence of paid content. There is a strong call for **Boolean search capabilities and aggregated filters** with a user-friendly dashboard, akin to those previously offered by CrowdTangle, to enhance the efficiency of data retrieval. Importantly, no researchers advocated for access to personal data which would breach GDPR and allow for individual tracking.

Accessibility is another major issue: respondents highlighted the need for **machine-readable formats,** without a need for a complex pre-processing with manual steps, as well as non-technical interfaces, tutorials to assist those without programming skills, and ability to **share access across teams**.

Overall assessment

The survey results showcase that researchers studying disinformation, foreign information manipulation and interference and other forms of malign actors, behaviours and contents currently do not have enough data to conduct their work effectively. The lack of user friendly and cross-platform tools remains a key challenge that is seriously impeding any effort to defend and build democratic resilience in Europe.

At the moment, too many actors, especially smaller organisations, are left to rely on expensive social media monitoring tools, which creates an equity issue not only between the organisations but also member states. With larger teams and institutional backing, bigger organisations and universities are more likely to overcome the challenge of data access either through persistent application processes to receive access to platform content libraries or through purchasing monitoring tools, which smaller organisations cannot afford due to a lack of financial and human resources. Likewise, those in member states with a higher purchasing power generally, may have more ability to do research.

92% of respondents "definitely agreed" that social media platforms should provide more free data to researchers, both from academia as well as civil society organisations.

There are limited opportunities to conduct research with real-time data and on large-scale in-depth datasets with all required metrics to assess the type, virality and impact of the content, conduct network analyses and identify potential harmful actors.

On a scale of 1 to 10, researchers evaluated their access to data and information needed to conduct their current or upcoming research as 4.7 on average.

The Commission should thus push the platforms to speed up the processes of granting data access to researchers, especially real-time, publicly accessible data under Article 40.12 of the DSA. Based on the data above and the suggestions from the respondents, the following should be ensured:

1. **API keys to satisfy Art. 40.12 of the DSA should be shared widely with generous quotas with no clean room requirement for each platform, similar to CrowdTangle or current YouTube data API standards. These should be shared via a user-friendly interface available also for organisations with a lack of resources for big data analysts.**
2. **Researchers should have the ability to download the data in a user-friendly format and work with it beyond the platforms' libraries. The data for download should include all engagement and impression metrics from the content and its comments.**
3. **Eligibility criteria for data access should be defined clearly and given on an organisational basis, not project basis to prevent recognised organisations being evaluated as non-eligible, to limit confusion among the research community, and alleviate the burden of constant access requests. NGOs with proven research and ethical standards and non-partisan nature should be included into all data access schemes equally as large universities and research institutions. The CrowdTangle cancellation is a proof of the misuse of the lack of principles and rules on access to data.**
4. **Criteria should be clearly defined on what kind of research should be allowed under requests within Art 40.8 to prevent vague formulation in reasonings for the rejections in access provision.**
5. **Deadlines to evaluate the proposals to prevent months-long waiting times should be set.**

In addition, surveyed researchers suggested to:

- **explore the possibility of creating legislation that protects researchers from prosecution for scrapping;**
- **create a support mechanism within the research community to increase others' awareness of what data is available, how to make the requests properly, and how to use the data;**
- **and put together a code of conduct for researchers to facilitate the selection process, which is currently under process within EDMO.**

On a scale of 1 to 10, researchers rated their access to data for conducting research prior to the European Parliament elections at an average of 5.09.

Annex no.1

Full list of questions

Note: The branching was applied where logically appropriate. For example, those who answered "No" to question no.8 were automatically redirected to question no.12.

1. Name and surname (optional)

2. Position

3. Organisation

4. Country where you work

5. Do you use any social media monitoring tools for your work?

a) Yes
b) No

6. (Only if 5a) Which social media monitoring tool(s) do you use? (open question)

7. (Only if 5a) How often do you use social media monitoring tool(s)? (open question)

8. Did you have access to CrowdTangle?

a) Yes
b) No

9. (Only if 8a) How often did you use CrowdTangle?

a) Daily
b) 2-3 times a week
c) Once a week
d) Once a month
e) Less than once a month
f) Other (open)

10. (Only if 8a) Do you have a replacement for CrowdTangle with tools with similar functionality?

a) Yes, full
b) Yes, partial
c) No
d) Other (open)

11. (Only if 8a) How did the cancellation of CrowdTangle impact your work? (for example, financial implications, personnel, research) (open)

12. Do you or your organisation have direct access to social media researchers' APIs or content libraries granted by social media platforms?

a) Yes
b) No
c) Other (open)

13. (Only if 12a) What social media platforms do you have direct data access to? (multiple)

a) Meta (Facebook, Instagram)
b) YouTube
c) TikTok
d) LinkedIn
e) X
f) Other (open)

14. (Only if 12a) What is your experience with getting researcher access/ API access to data?
(1 = terrible, 10 = amazing)

15. (Only if 12a) What is the quality of data provided by platforms APIs and content libraries?

a) Very good
b) Good
c) Average
d) Poor
e) Very poor

Per platform:

i) Facebook and Instagram
ii) YouTube
iii) TikTok
iv) X
v) LinkedIn

16. (Only if 12a) A space for any comments to the question above. (open)

17. (Only if 12a) What are the main issues or deficiencies that you observed and experienced? (open)

18. (Only if 12a) What kind of data or features would you want the new platform libraries and APIs to encompass for your research? (open)

19. Have you or your organisation applied for direct access to social media researchers' APIs or content libraries granted by social media platforms?

a) Yes
b) No

20. (Only if 19b) Why have you not applied? (open)

21. (Only if 19a) How long have you been waiting for a response? (please provide information per platform) (open)

22. (Only if 19a) What requirements/conditions of social media platforms do you find the most demanding or unreasonable to fulfil? (please provide information per platform) (open)

23. (Only if 19a) Have you been denied access to researcher data by a social media platform?

a) Yes
b) No

24. (Only if 23a) Which platform(s) have denied you the access? (open)

25. (Only if 23a) What was the reasoning? (open)

26. Do you have enough data/information to conduct your (running or upcoming) research effectively? (1 = I have no data, 10 = I have all the data I need)

27. Did you have enough data to conduct research prior to the elections to the European Parliament? (1 = I had no data, 10 = I had all the data I needed)

28. Should social media platforms provide more free data to researchers, both from academia as well as civil society organisations?

a) Definitely yes
b) Rather yes
c) Rather not
d) Definitely not
e) I am not sure

29. Would you be interested in getting data from other researchers who are vetted and have access to APIs?

30. Anything we did not think of?