EDMO
European Digital Media Observatory

EUI SCHOOL OF TRANSNATIONAL GOVERNANCE

European Digital Media Observatory

Report on EDMO Workshop on Platform Data Access for Researchers

Lisa Ginsborg

Contributors:
Kalina Bontcheva
Valentin Châtelet
Philipp Darius
Matt Motyl
Andreas Neumeier

September 2024

Report on EDMO Workshop on Platform Data Access for Researchers

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

About the Workshop

On 15 May EDMO organised a workshop on the topic of **Platform Data Access for Researchers**, at the Brussels School of Governance. The workshop (led EUI STG in collaboration with Globsec) took place in hybrid format and was attended by over 50 participants in person and 25 online, including representatives from the EDMO Network, civil society organisations, as well as EC representatives and some VLOP representatives.

The workshop aimed to bring together the research community to discuss and share their experience and views on platform data access for research, including newly developed platform APIs, the type of data available and its accessibility, and whether the new research data access provisions under the DSA meet the needs of the research community. Speakers from the research community (including representatives from the EDMO.eu and Vera.ai projects, the Integrity Institute, the Digital Forensic Research Lab, and the Center for Digital Governance, Hertie School) provided key insights and shared concrete experiences with regard to new and old data access tools provided by VLOP and VLOSEs as well as potential opportunities and challenges for research.

Given the key role of independent research in enabling transparency and independent oversight and a deeper understanding of the disinformation phenomenon, especially in the context of new technological developments and upcoming elections worldwide, wider data access for research organisations conducting independent research in the public interest remains a key priority for the EDMO community.

Background: Data access provisions in EU regulation and self-regulation on disinformation

In the 2022 Strengthened Code of Practice on Online Disinformation many online platforms explicitly committed 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.” Such voluntary commitment has now become a legal obligation for VLOPs/VLOSEs under DSA article 40(12) which explicitly requires platforms to provide access to publicly accessible data, and when technically possible to real-time data. The Delegated Act on data access which was scheduled to be adopted in spring 2024 has not yet been released.

Data access provisions for researchers present a rapidly changing area, with a number of new tools being developed by platforms especially through researcher specific APIs to allow access to researchers to real time data. Yet a number of limitations appear to be currently hindering their use by the research community, including lack of awareness of the new tools, limited access for civil society researchers, complicated or lengthy application procedures and potential legal risks deriving from their use. As a result, the uptake of APIs appears slow and piecemeal, with very few European researchers actively using such APIs at present. Although greater uptake may be expected as the tools continue to be rolled out and refined, as most tools are still being tested it is essential to better understand their utility and how they compare to previous instruments.

Workshop programme

Moderator: Claes de Vreese I EDMO Executive Board, University of Amsterdam

Framing the issues surrounding data access for research

Speakers:

- Lisa Ginsborg I EDMO, School of Transnational Governance, EUI
- Rebekah Tromble I George Washington University, EDMO
- Matt Motyl I Integrity Institute

Exchange of experience using current platform APIs

Speakers:

- Kalina Bontcheva I University of Sheffield, BROD, vera.ai
- Valentin Châtelet I DFRLab, Atlantic Council
- Philipp Darius I Center for Digital Governance, Hertie School
- Andreas Neumeier I Bundeswehr University

Key Takeaways from the EDMO Workshop

Despite the DSA having entered into force, and its key provisions with regard to data access for research purposes as contained in Art. 40, in practice data access for researchers has not yet seen a significant overall improvement to date. In fact, in certain areas or in relation to specific tools data access for researchers appears to present significant limitations and in certain cases may have even deteriorated over the last year. While this may be temporary as the new instruments become operational, and before the delegated act on data access is adopted, a number of positive developments may also be noted in this area. These include public statements by the EC that the interpretation of Art. 40(12) appears to enable non-permissioned scraping,[1] and the fact that researcher data access provisions under Art. 40(12) now exist for the large majority of VLOPs at least on paper, although in many cases the details of the concrete programs and data available are still missing.

In practice current data access for researchers by platforms remains limited to date, with some platform APIs performing better than others while a number of significant shortcomings remain and are presented in the current report. Among such limitations researchers report limited accessibility, complex application procedures which include significant risks with regard to liability and fines, and the requirement for applications to

[1] See for instance Press Release of 12 July 2024 ‘Commission sends preliminary findings to X for breach of the Digital Services Act’, available at https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761

be linked to specific projects rather than approved at organizational level. The legal requirements for researchers imposed in the contracts by the platforms should therefore be clear and not prohibitive for smaller research organisations, in light of the risk of large fines and the lack of existing mechanisms to protect research organizations from such risks. Equality in terms of data access is also key: online platforms and search engines should be discouraged from providing data and/or funding to only a small group of researchers (on specific issues or from selected countries), selected by the companies themselves in a non-transparent manner.

Leaving aside the specific problems of each platform API, which are presented in detail in the current report, greater standardisation in data access API provisions by similar kinds of platforms and search engines is fundamental for researchers to be able to conduct cross-platform research, which is central to studying disinformation in a holistic way. In particular it is crucial for similar types of VLOPs and VLOSEs to provide similar types of data via their APIs in order to allow for comparability across platforms (e.g., TikTok and Instagram; Bing Search and Google Search). While researchers made it clear that platforms providing researchers with data through their APIs should be seen as a very welcome development as few researchers appear to be gaining access to some of the platform APIs at present, data quality and reliability also appears to be a concern among researchers when using these APIs. This underlines the need for an independent entity that could be testing and ensuring data quality under DSA Article 40(12), and potentially under Articles 40(4); 40(8), as it is clear that this cannot be done by individual researchers.

With regard to the way data is accessed through specific APIs, streaming/real-time APIs (e.g., previous Twitter API) allowing to get continuously new sets of data may be preferable to APIs in which periodic downloading is required (YouTube and TikTok) as this may also limit the kind of research that is possible. Finally the clean room approach of the Meta Content Library and API is highly restrictive with respect to repeatability and open science, also in light of the fact that the clean room is cleaned after a certain period of time and the requirement of access through VPN is also problematic for data transparency. Overall it is essential that while new tools are rolled out and tested, data access for researchers should improve urgently and continue to do so over time for all VLOPs and VLOSEs while ensuring the information provided to the research community is clear and user friendly. Imposing fees for data access for researchers does not allow for such incremental access.

With regard to DSA Art. 40(4), the system will only become fully operational once the Delegated Act is released. The work done (and presented during the workshop) by the EDMO Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms (led by Dr Rebekah Tromble), as well as the draft Code of Conduct on how platforms can share data with independent researchers while protecting users' rights will play a key role going forward. Collaboration will be key for researchers to ensure research priorities are prominent and the processes and institutions, including relevant Digital Services Coordinators (DSCs) are not overwhelmed. In this respect, the importance of researchers being aware of the variety of data collected by platforms, which may be requested for research purposes, becomes even more pressing. Greater information sharing and relevant tools are essential in this respect, including the data dictionary currently being developed by the Integrity Institute, including keywords, phrases in metrics, and variables researchers might request for research purposes.

Findings from surveying the EDMO Network on researcher data access

Lisa Ginsborg presented the results from a recent survey conducted by EDMO.eu, *inter alia*, with research representatives in the EDMO Hubs surrounding VLOPSEs implementation of their commitments on the empowerment of researchers under the Strengthened Code of Practice (CoP). The report from the survey describes a number of limitations faced by the research community with regard to access to platform data as a pre-condition for transparency and accountability that is long overdue.

In particular the Research Survey conducted by EDMO shows that data transparency and data access through APIs (Commitment 26) continues to be a key priority for the research community. Despite encouraging reports by all platforms on recent launches of new tools for researchers, the uptake of APIs appears slow and piecemeal, with very few European researchers among EDMO Hubs seemingly using such APIs at present. Several reasons may be ascribed to the lack of use of new APIs by EDMO Hubs, starting most obviously from their recent set up.

Other reasons which emerged from asking researchers in EDMO hubs include lack of awareness of the new tools, complicated or lengthy application procedures and potential legal risks deriving from their use. While greater uptake may be expected as the tools continue to be rolled out and refined it is clear that simply launching an API for researchers may not be enough to meet the requirements of Commitment 26 and ultimately the DSA. Given most tools are still being tested, it remains difficult to understand their utility and how they compare to previous instruments.

The training sessions, workshops, and collaboration environment for fact-checkers and researchers offered by CrowdTangle in the past (which allowed researchers to exchange experience, tools and methodology and provided a clear point of contact for researchers) constitute good practices in this regard. Researchers reported them to be especially useful around election periods.

A number of problems were also reported by EDMO Hubs, including: length of the application process by platforms, making it difficult for researchers to realise the intended research projects within the funding period; the challenges of reconciling the contractual requirements with the conditions of independent research and the protection of employees by the university; as well as complicated authorisation procedures and concerns in respect of the current conditions imposed in contracts by the platforms, including the risk of large fines for research organizations and the lack of existing mechanisms to protect research organizations from such risks.

A number of **concrete recommendations** also emerged from EDMO Hubs representatives from the survey including:

- The need for platform tools and interfaces for data access to be rolled out fast;
- The need to raise awareness about the newly developed APIs, including making the information clear and user friendly and providing training and support to the research community on these tools;

- The need for application procedures to be clear and for applications by researchers to be approved swiftly;
- The suggestion for applications to be approved at organizational level, as opposed to being linked to specific projects;
- The data provided should meet the needs of the research community by enabling incremental and near real-time access to data, which should also include providing researchers with information on data that is flagged as misinformative;
- Allowing access to the APIs also for researchers from civil society;
- Providing support to researchers entering legal contracts, e.g., establishment of a legal protection/insurance system for public research institutions, NGOs and independent researchers;
- Greater standardization of APIs will allow the research community to use them more widely, rather than having to gain specific skills for each platform;
- Going forward the question may be raised of whether specific tools could be developed which could enable the use of APIs by researchers from different disciplinary backgrounds beyond data scientists

Access to data for researchers under DSA Art. 40(4), 40(12) and EDMO's work on data access

Rebekah Tromble provided a number of updates and insights on the state of state of data access under the DSA, in particular with regard to article 40(12) and article 40(4) and what may be expected going forward as well as EDMO's work on data access in preparation for Art. 40(4) becoming fully operational. While some elements are still missing including relevant delegated acts and guidance from DCSs, there are a number of actions researchers may already start working on now.

Dr. Tromble started by emphasising some of the positive developments in light of DSA Art. 40(12), including the public statement from a number of key members of the EC that their interpretation of Art. 40(12) is that it enables non-permissioned scraping. Programs exist with regard to data access under Art. 40(12) for the large majority of VLOPSEs at least on paper, although in many cases the details of the concrete programs and data are still missing. Both the platforms and the researchers need more guidance from the EC about the requirements for such programs, but this is unlikely to come before the Delegated Act for Art. 40(4) is released, given the need for harmonisation between the different provisions. EC interest in this area is demonstrated by its inquiries to all of the VLOPs and VLOSEs for more information about their Art. 40(12) programs, as well as by the formal investigations into X and Meta, including for their compliance with Art. 40(12). Other organizations are also stepping in, including EDMO, the Institute for Data, Democracy & Politics with its Tracker aiming to describe existing research access tools, the Coalition for Independent Technology Research with its DSA Data Access Audit and Survey, which aims to gather information about the experience of researchers with the data access programs and in the future aims to work with researchers to put together applications in order to test the systems directly.

With regard at Art. 40(4), researchers who are affiliated with research organizations as defined by the EU Copyright Directive will have the ability to apply for specific non-public data under Article 40(4). While the exact details of this still need to be clarified, it is clear that both civil society and academic organizations that have a substantial research mission will qualify under Article 40(4). It remains to be clarified whether non-European research institutions will qualify, raising the importance of collaboration going forward.

The data researchers can apply for must be ‘in scope’, applying to European systemic risks and potential mitigations for those risks. The request must be proportionate, the data must ‘exist’, must not jeopardise trade secrets, and must comply with GDPR. The EDMO Draft Code of Conduct aims to help researchers and regulators in this respect and is providing a partial blueprint for the delegated act, and aims to provide further guidance also to the DSCs. With regard to the application process, this will be further clarified in the Delegated Act, but is likely to see researchers submitting their applications to local DSCs, the DSC in most instances would send the application to an independent intermediary body to provide an advisory opinion, which the EDMO Working Group for the Creation of an Independent Intermediary Body to Support Research on Digital Platforms has been working on. The local DSC would then send the recommendation to the DSC of establishment (in most cases the Irish DSC), which would make the final decision. If positive, designated researchers would be involved as “vetted researchers” and make “reasoned request” for the data of platform(s). Platforms will then have 15 days to respond, and may also object either because they don’t have the data or on trade secrets grounds.

A negotiation process may be expected in this context involving the regulators, the platforms and the researchers themselves with the likely involvement of the intermediary body. Potential hurdles and unknowns remain, including in relation to how to tailor the GDPR risk assessment in relation to the data received, assessing who qualifies and what qualifies for research projects, what might the platform fees be if any, but the biggest potential hurdle is overwhelming the regulators with data requests. It is therefore important for researchers to provide feedback to the regulators but most importantly cooperate with one another to not overwhelm the system. Once the intermediary body is established, it aims to bring researchers together for consultations in Fall 2024 to identify the most pressing data needs and priorities to be shared with DSCs.

Online Platform Data Collection and Access

Matt Motyl from the Integrity Institute aimed to support researchers outside of industry to learn about what data is currently collected by the relevant companies. This includes the obvious data provided by users, the less obvious data extracted by platforms, and additional data that is learned about users often through machine learning, AI or by buying data from third parties. While the presentation started from Facebook, the lessons were seen to be applicable to most social media platforms.

Dr. Motyl, presented some of the categories of data users normally provide to Facebook, either directly or through their consumption patterns or posts and additional data platforms may extract from the data provided by users. Further user behaviours on the platforms (e.g. time spent on platforms, clicks including on ads, purchases, reactions, comments, reporting behaviour, etc.) provide signals to platforms on content to be fed to users. Time spent engaging in such behaviours is also logged. Further data is provided on how users interact with each other including in relation to all of the behaviours above.

Less obvious data that platforms extract about users includes device metadata, from which they can extract where people are located, what network they are on, whether they use multiple devices, whether multiple people use the same devices for access to those platforms, GPS location, and all kinds of network information, including phone numbers and names, whether they are also on that platform, and from there establish networks, which can help establish by way of example what people in specific networks may buy, or even instances of coordinated inauthentic behaviour.

More sophisticated information that may be collected includes the probability users are real and are who they say they are, probability users will buy specific things, who users are likely to vote for, what parties they may belong to or how likely they are to show up to a political event, whether the user is working in coordination to interfere with another country's election, whether the user produces *unwanted social interactions*. Most of the above information will exist at the user level but also at the level of content, group, ad, page, etc.

Dr. Motyl also introduced the way data is stored by the platforms, in particular the two main types of categories for data storing: dimension tables and fact tables. Dimension tables are mostly structured, i.e., containing one row per aggregated object (e.g., user, post, page, post, video) with a key id and one column per variable (and there may often be 1000s of variables). Fact tables, on the other hand, are mostly unstructured with often one row per event, which are not stored in columns but often in JSON strings, MAPs, Arrays (theoretically could be 0 to infinite length) with key-value pairs. For large platforms it is probably impossible to have a single table containing all variables, as it would probably involve many exabytes of data if not more. Different categories of variables can be stored in different tables and even different warehouses or cloud services. Joining data is therefore quite complicated.

Further, while thousands of tables may exist per user, page, group, etc., these tables may also have different retention times, until the data is transferred to a different format and becomes much more difficult to obtain. Different platforms will also have different data retention policies and even within a company, different security / privacy settings may apply to different tables and even columns. In light of the above accessing specific data may become quite complicated or take a long time.

Matt Motyl and the Integrity Institute are currently building a data dictionary, including keywords, phrases in metrics, variables researchers might want, how to map users activities to these data table, community requests, with the aim that researchers will know what to ask for when they are requesting platform data, in order to not overburden regulators going forward.

Experiences challenges and opportunities with platform APIs

TikTok Research API

Dr. Philipp Darius from the Center for Digital Governance at the Hertie School and Andreas Neumeier from the Bundeswehr University in Munich presented their experiences, current requirements and problems in using the TikTok Research API. In a collaboration with the SPARTA project the team collected political party communication on Facebook, Instagram, TikTok, Twitter/X and YouTube during the EU elections. For accessing the TikTok Research API it took TikTok approximately 4 weeks to grant the application for the Research API.

The **TikTok Research API**, it was made accessible to European researchers on July 20, 2023. The websites entailed contradicting information on whether the API is available only to European researchers or to researchers globally that focus on European systemic risks. The stated aim is to 'support research and increase transparency 'and a collaboration team with up to 10 researchers in one lab on the developer platform may be formed, which should enable pooling of API keys to increase the quotas of API requests. TikTok has also initiated a commercial content library that includes ads, advertiser metadata and targeting information.

In order to get access to the Researcher API an application form needs to be submitted, and the researchers must adhere to the Community Guidelines and the TikTok Research API Services Terms of Service. The research proposal needs to be approved by the research institution's ethics committee, there should be no conflict of interest and no commercial purposes and the researcher should have demonstrated experience and expertise and be employed by a not-for profit organization in the EU or in the US.

With regard to the data that can be collected with the TikTok Research API, researchers can generally collect content on public accounts by “creators” or users with public profiles and who are aged 18 and over. Information may in theory be collected on videos, comments, users, liked/reposted/pinned videos, follower and following lists (if following list is made public).

Dr. Darius went on to describe their experience of using the TikTok API in practice while working on a cross-platform study collecting communication on TikTok, YouTube, Instagram, Facebook, and X by all European political parties participating in the 2024 EU elections. Especially for the TikTok side of the collection they have been facing serious issues regarding the data retrieved via the TikTok Research API. In particular, the data provided between March and June 2024 cannot be used for research on party campaign communication behaviour as it deviates strongly from metrics shown in the app or on the webpage as TikTok user interfaces. Mr. Neumeier proceeded to provide concrete examples of the problem experienced, in which comparing the JSON retrieved from the API to the relevant TikTok webpages a number of disparities could be noticed in particular with regard to the likes count, comments count (which in the API appears to include the favorites count), the share count, and the view count in which very significant differences can be noted. Overall, the data provided by the API on some of the metrics was not reliable and cannot be used for their current research project.

They faced the following issues:

- The API drastically underreported view and share counts of videos in contrast to metrics visible on TikTok's user interfaces on the website and in the TikTok App.
- Conflation of comment count and favourites count metric.
- The follower collection was interrupted at around 3,000 accounts, and there appears to be some kind of limit on it but the reason for this remains unclear.
- API data is not available in real-time, as obligated in DSA 40(12) but with an approximately 10-days delay.

The researchers have reached out to the TikTok API support team, and after some time they have now received availability to schedule an appointment to discuss the problems with the data and potential underlying issues with the API. The team thereafter also reported the issues to the Bundesnetzagentur, as the German Digital Service Coordinator, and staff members of DG Connect. While biases of other research APIs (e.g. Twitter's Streaming API (see Morstatter et al., 2013) are known, in the case of the TikTok API more than a bias, the data was faulty.

In July 2024 the team found that the TikTok Research API seems to have been repaired and provides only marginally deviating metrics from the user interface (which is acceptable for distributed systems). However, the identified problems underline the need for an independent entity that could be testing and ensuring data quality under DSA Article 40(12), and potentially under Articles 40(4); 40(8), since individual researchers cannot test all the data provided by platforms via APIs or via data access requests.

The data tracker by the Weizenbaum Institute was also mentioned as a good practice that could be combined with other data access trackers to bundle researchers' experiences when working with platform data.

YouTube API, Google Search and other platform APIs

Kalina Bontcheva from the University of Sheffield provided an account and reflections on her experience with data access at the University of Sheffield and as part of the Vera.ai EU-funded project. In particular she presented her experience with the YouTube API as well as a number of other APIs.

In terms of positives the **YouTube API** gives access to the video descriptions, the subtitle data (if present), channel descriptions, etc. Based on the video ID resarchers can retrieve all comments for that video, which is important when analysing responses to a specific video that is spreading disinformation. The video comments also tend to have accurate timestamps.

However, a number of issues were also encountered with regard to the YouTube API, which could be improved over time as the dialogue on data access with platforms progresses. When retrieving comments, all comments always appear, and there is no way of retrieving only new comments posted since a certain date, which would be more efficient with respect to the usage on the researcher API data quota. Similarly, to check for new replies to existing top-level comments researchers must query each top-level comment individually to get all of its children and then work out which ones are new. Researchers would benefit from an API call that gave all top-level and reply comments on a given video that are dated after a given timestamp.

Further, the "search" API (to find videos matching search terms, or all videos posted by a given channel) is extremely expensive in terms of quota. The daily quota is thereby exhausted after just a few searches, which put limits on the speed at which research may be conducted. While there is an explicit provision stating that "On request, and with sufficient justification, you will receive sufficient API quota for use as specified by this Program ToS", it would be important to clarify the protocol for increasing the quota. The biggest problem, which may relate also to the TikTok API, is that when the data is downloaded it is a snapshot at the time it was downloaded. Differently from the previously available Twitter streaming API, it is therefore not a streaming/real-time API allowing to get continuously new sets of data, but periodic pulling is required and data may therefore be a bit out of date, which also limits the kind of research you can do.

Prof. Bontcheva also presented her experience with other platform APIs, starting from the **TikTok Research API**, for which their application was successful and the researchers are getting rich data via their comprehensive research API and documentation, despite some of the issues with data quality discussed earlier. **Bing** provides access to thematic datasets for Search, which from the perspective of a scientist are very important as they allow open science and repeatability and it means that different researchers can compare results also with regard to their AI models. Bing also provides API access for news, web, image, and other search, which is very valuable. Finally from **Google Search** it is possible to get access to Google Trends, the Fact-check API and the Ad library, which have been available for some time. Google Search however does present a number of access limitations.

In particular, researchers would benefit from access to free quotas to the Google Custom Search JSON API, which would seem easy to provide, as in the case of YouTube. Further, Google search thematic collections would also be extremely valuable for researchers comparable to those from Bing. This is important because disinformation does not live on one platform and researchers need to be able to answer cross-platform research questions, e.g., are citizens exposed to disinformation in their search results, irrespective of which search engine they use? What is the difference between search results provided by the different search engines that are signatories to the CoP? At present it is impossible to answer those types of questions, because different search engines are interpreting the idea of research data access in very different ways. Shared research collections across platforms are very important for open science and repeatability and provide insight into the history.

Further, while APIs may of course differ across platforms it is important to have access to the same kinds of data in order to allow for comparability across platforms. The importance of cross-platform research capabilities makes it essential for similar types of VLOPs to ideally provide similar types of data via their APIs that allows researchers to carry out cross-platform research. By way of example the clean room approach of

Instagram is highly restrictive with respect to repeatability and open science, also in light of the fact that the data access provisions as of April 2024, meant that the clean room is cleaned after a certain period of time. Research experiments should be repeatable over time and the methodology should be documentable. At present a number of research questions that EU-funded research projects on disinformation aim to investigate (e.g., Comparative impact analysis of disinformation videos/images on popular video and image sharing platforms) cannot be answered because the data access provisions by Instagram and TikTok are very different. Further, with the Instagram clean room it is not currently straightforward to import other URLs, e.g. debunked disinformation from other platforms, making potential research highly limited.

Prof. Bontcheva also highlighted that in certain European countries the majority of political campaigning continues to take place on **X/Twitter**. Governments use it as a platform to communicate with the electorate, but it is no longer possible for researchers to study it. In particular researchers in the UK have received numerous requests from policy makers, regulators, and government bodies for in-depth, large-scale quantitative research on the prominence and impact of disinformation in public political discourse, which can no longer be satisfied as the X/Twitter API is no longer available for free to UK researchers. Further, data access for researchers from associated countries who are part of Horizon Europe projects such as the UK is currently being restricted by X, which is currently flatly rejecting all UK applications claiming that DSA provisions do not extend to UK researchers, even when working on EU projects funded by the EU and looking into disinformation in the EU.

Finally equality in terms of data access among researchers is essential, and online platforms and search engines should be discouraged from providing data and/or funding to only a small group of researchers (on specific issues or from selected countries), selected by the companies themselves in a non-transparent manner. The importance of collaboration among researchers was further emphasised as well as the need for a shared understanding among all stakeholders on data access and research and its essential role for society and democratic integrity. This should include positive engagement by policy makers and platforms, beyond current narrow concerns about reputational risk. Lastly, in terms of data access provisions, and their monitoring, there is a need for more prominent input from AI/CS experts on the technological and standardisation aspects of data access provisions and their adequacy, which can have huge implications for research.

Meta Content Library and API

Valentin Châtelet from the DFRLab introduced the current data access landscape for Meta. As is well known Meta is decommissioning **CrowdTangle (CT)** starting from 14 August 2024, which is widely used in the research community including the participants to the workshop. In relation to this, the EC has initiated an enquiry on the potential infringement to the DSA of Meta's current data access policy and tools. While none of the participants to the workshop, including Mr. Châtelet, had access to the Meta Content Library and API, he shared insights on the experience of some of his colleagues at the DFRLab. Meta also provides access to researchers to certain datasets, in particular to take-downs and CIB related deplatforming.

The procedures to receive access to the **Meta Content Library and API** were reported by Mr. Châtelet to be very long. Applications are reviewed by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. This includes questions about the researcher's experience of handling confidential data as well as large datasets, and optional documentation from an Ethics Committee or Institutional Review Board (IRB). The Content Library is a Graphic User Interface (GUI) to query content from Meta, while the API allows to query through code, either R or Python, to public accounts with more than 25,000 followers, including public Pages, public groups, public events, public profiles, and even comments according to Meta's documentation.

Once researchers gain access the Content Library interface appears like a search engine, which includes the following sections:

- Influence operations datasets, including CIB and other deplatformed taken-down content that is downloadable as CSVs
- Saved searches: textual search queries that researchers can save for later use and continued monitoring
- Producer list: collection of pages that populate a collection
- Downloads (history of CSV files that have been downloaded by researcher)

The Datasets section resembles the collections of CIB datasets from CT, but in terms of labelling contains significantly less information regarding how the content has been edited (or removed) when compared to CT, and explanations about what are the grounds justifying the existence of the 'dataset' also do not appear to be available. On the positive side, the CSV files of the datasets (including Meta takedowns and CIB) may be downloaded, as well as trend graph, which show the evolution of the amount of posts found when users query the search engine of the Content Library.

In terms of limitations the Meta Content Library does not allow access to cross-platform searches. Each query on the search engine will therefore only be able to search for that specific query on Facebook or on Instagram, not on both. In this respect the Content Library is different from the Meta Ad Library which gives the opportunity to search across both platforms. In terms of the search interface using the textual search, researchers will still have access to several filters which were available on CT. While the results of the search can be saved in the personal account, they cannot be downloaded as a CSV file and it appears to be a less potent search engine, with no cross-reference search with Facebook & Instagram.

Another limitation of the Meta Content Library is that researchers have access to even less search operators. Since the workshop, the use of quotes to perform literal searches was implemented, however, the search engine still does not support Boolean operators. At the time of the workshop, the use of quotes to perform literal searches was not available. In addition, handles and URLs of contents cannot be used to perform searches, which presents a significant setback compared to CT. Finally, there appears to be less search and query results than the numbers indicated for each search. The same appears to be true for the Meta Ad Library, where the number of Ads indicated is expressed in ranges instead of giving researchers an exact figure. A final significant limitation is that to access the Meta Content Library requires the use of a VPN, casting potential doubt on the monitoring of researchers' activity and data on the platform.

DFR Lab observes that the amount of public data that is available via Meta tools increasingly lacks cross-compatibility with existing Meta products. Regarding the Meta Content Library and API, the Unique ID provided when searching for a specific Facebook page is only created within the Meta Content API and cannot be used to search on Meta's technologies, including Facebook or Instagram, or and don't provide access to the assets' page using the URL facebook[.]com/<UniqueID>.

The community also expressed worry regarding the contractual terms for accessing Meta research tools from a legal perspective especially for research institutions with fewer resources. Civil society actors such as the media and journalists are not intended to be granted access to the Meta Content Library and API. In addition, there is a requirement regarding publications to notify Meta of any forthcoming publication, which also raises issues with concern to the independence of research.[2]

[2] See https://transparency.meta.com/researchtools/product-terms-meta-research last accessed 23 July 2024

EDMO

EUI SCHOOL OF TRANSNATIONAL GOVERNANCE

www.edmo.eu

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