Improving Researcher Access to Digital Data

A Workshop Report

August 2022

CENTER FOR DEMOCRACY & TECHNOLOGY

The **Center for Democracy & Technology** (CDT) is a 27-year-old 501(c)3 nonpartisan nonprofit organization that fights to put democracy and human rights at the center of the digital revolution. It works to promote democratic values by shaping technology policy and architecture, with a focus on equity and justice. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

Improving Researcher Access to Digital Data

A Workshop Report

Author

Caitlin Vogus

With Contributions by

Samir Jain, Emma Llansó, Dhanaraj Thakur, Greg Nojeim, Eric Null, Aliya Bhatia, Gabriel Nicholas, Ridhi Shetty, Will Adler, Elizabeth Seeger, and Tim Hoagland.

Acknowledgements

We thank Daphne Keller, Jason Pielemeier, and Alicia Wanless for their help facilitating the workshop, Hugh Grant-Chapman, Michal Luria, Jamal Magby, Hannah Quay-de la Vallee and Elizabeth Remy for their help conducting the workshop, as well as all the workshop participants for their participation in the workshop and feedback on the report.

August 2022

Table of Contents

I. Introduction

Does content moderation on social media affect how users behave? Do social media recommendation systems lead people down rabbit holes and exacerbate filter bubbles? How many bots are on Twitter talking about a particular political issue before an election? Are fact checks of online COVID-19 misinformation effective?

These are just a few of the questions that researchers at a recent workshop hosted by the Center for Democracy & Technology (CDT) said they used data from hosts of user-generated content to explore.[1] Access by independent researchers — *i.e.*, those not affiliated with a platform — to data held by content hosts is an important part of technology company transparency, and one that has garnered increased attention from the policy community in the last year.

Following reports that platforms have stymied, or devoted insufficient attention and resources to, independent researchers' access to data, policymakers are considering whether and how to mandate or regulate independent researchers' ability to obtain platform data. In the United States, lawmakers have proposed at least four bills that would require certain tech companies to provide data to independent researchers, the public, or both: the Platform Accountability and Transparency Act (PATA), Digital Services Oversight and Safety Act (DSOSA), Social Media DATA Act, and Kids Online Safety Act (KOSA). In Europe, Article 31 of the Digital Services Act will become the first major legislation requiring some online services to make certain data available

1 We use the term "hosts of user-generated content" to refer mainly to social media companies and messaging services, which hold user-generated content or metadata about content. Most of the researchers in the workshop were focused on these services and content or metadata; however, researchers may seek access to data from a wide variety of technology companies, such as internet search engines.

to researchers. For a description of these bills, see the Appendix to this report and CDT's chart, *Independent Researcher Access to Social Media Data: Comparing Legislative Proposals*.

CDT believes strongly in the need to improve independent researchers' access to data held by content hosts while still protecting user privacy and security. In March 2022, CDT convened 29 researchers from academia, civil society, and journalism in a workshop designed to explore specific questions around access to data and better inform policy conversations. This workshop built on our December 2021 report describing current methods of access and outlining the tradeoffs that policymakers must consider when establishing policy to improve access. Workshop participants discussed three key questions concerning access to data held by content hosts:

1. What data held by content hosts do researchers believe is valuable and useful to research in the public interest?

2. Who should be given access to this data, and how should researchers be vetted to gain access?

3. How should researchers be given access to this data, i.e., what methods should hosts use to provide access to researchers?

Answers to these three questions will form the basis for any policy that requires certain content hosts to provide data to independent researchers. While other questions are equally important, such as what specific methods and standards researchers and hosts should follow for transferring, storing, and analyzing data to protect user privacy, we focused on these three basic issues that could be informed by researchers' firsthand experiences. We also selected these questions because their answers may be interrelated. For example, the type of data that researchers should be able to access may depend on how strictly they are vetted; the methods used to provide access may depend on the type of data provided.

This report first details the information shared by workshop participants on each of these questions, highlighting feedback that was common across participants. Next, it provides CDT's recommendations for policymakers and hosts looking to enhance independent researchers' access to data held by content hosts. Our recommendations are informed by what we learned from researchers at the workshop, our own research, and assessments of other human rights interests such as privacy.

II. What types of data held by content hosts is valuable and useful to research in the public interest?

A. Types of data currently available to researchers

CDT asked researchers at the workshop to identify what kind of data they already have access to and the purposes for which they use it,[2] as well as what data they wish they had access to but do not, and what research they would conduct with this data.

////

Many researchers identified **advertising data** — including both the content of ads, and other data such as basic information about who was targeted or who viewed an ad — as a rich source of information currently available to them, at least in part. Researchers reported using advertising data to investigate questions specifically related to online ads as well as to explore questions unrelated to advertising, such as using advertising audience estimates to monitor international migration and digital gender gaps. While some researchers identified platform ad libraries as important sources of data, they also raised concerns about the completeness and accuracy of ad libraries, in addition to other concerns and limitations in advertising data, described below. *See* Section II.B.

2 Past research has explored available social media platform data and attempted to categorize it. Accordingly, we will not attempt to present an exhaustive account of all possible data that is currently available to researchers, but rather report the categories of data that researchers most commonly cited in this workshop when discussing data that is currently available to them.

Researchers also reported accessing **public content posted by users** and **content posted by users in semi-public[3] groups** or other online environments that researchers were able to join. Researchers reported using content posted publicly or semi-publicly by users on Facebook, Gettr, Reddit, Telegram, WhatsApp, and, most commonly, Twitter.[4] While some of this data — such as tweets or Reddit posts — is available to anyone on the public internet without restriction, researchers also used content posted by users in restricted Facebook or WhatsApp groups that they joined.[5]

Researchers often use public or semi-public content posted by users to study issues directly related to social media platform use. Some of the examples researchers provided in the workshop were research about how mis- and disinformation spread and how to inoculate against it, the existence of coordinated inauthentic activity on a social networking service, the prevalence of bots on a social network and the type of content they spread, and the posting of racist content by law enforcement officers. Some researchers also reported using content posted by users to investigate other questions not directly related to social media, such as why people stay in abusive romantic relationships and what factors made it difficult to leave; the nature of news reporting on the #MeToo movement from different political perspectives; and how members of Congress communicate with the public.

Researchers also said they rely on **social networks or social graph data**, *i.e.*, data that shows how users of a social network are connected to each other. Researchers most commonly gathered this data by using data from public Facebook groups or publicly available data from Twitter, but they said they also sometimes used data from private Facebook groups to which they gained access. Researchers reported, for example, using social networks data

3 We use the term "semi-public" here to mean content that is not public, in the sense that it is made available to any user of a service generally available to the public, but that also is not sent directly to a single other individual or very small number of people. As discussed below, drawing the line between "public" and "non-public" content can be difficult. *See* Section V.4.

4 Researchers also identified limits to the use of public and semi-public data, such as the time and expense of gathering and analyzing it and how reliance on public and semi-public data may shape or limit the research questions they pose. These limitations are discussed in greater detail below. *See* Section IV.

5 Researchers acknowledged that access to content data is not always possible. One researcher suggested that access to metadata, especially from hosts that offer end-to-end encrypted services, would enable important research in the absence of access to content data, such as research about how content spreads across groups on encrypted messaging services.

to identify trends in Facebook group membership, tracking how Facebook users used groups in general, and, more specifically, how they used them to spread misinformation or racist content within a Facebook group and across different Facebook groups with common members. Researchers also said they use social networks data to study issues concerning online polarization and echo chambers.

Researchers also said that they currently use **engagement data** — information about users' engagement with posts through their reactions or comments — for research. The most common source of engagement data cited by researchers was from CrowdTangle, a tool owned by Facebook that provides data about posts shared on public pages and groups. Researchers explained that they have used engagement data from CrowdTangle to examine which Facebook pages consistently receive high engagement, how Facebook pages spread narratives online, and how changes to ranking and recommendation algorithms affect engagement with Facebook pages. However, many researchers were critical of engagement data for being a "black box" metric that does not allow them to examine content's reach, *i.e.*, the content that users actually see and consume, even if they do not actively "engage" with it.

Finally, researchers said they had limited access to data on **content moderation**, which, when available, allows them to investigate both platforms' policies and the accuracy, fairness, transparency, and efficacy of a service's content moderation decisions. Researchers reported obtaining information about policies by looking at publicly available sources, such as content hosts' terms of service or content guidelines. Other researchers said they gathered data about platforms' specific enforcement decisions based on content moderation decisions that are visible to the public (or can be made visible), such as some moderation actions on Twitch. They also said they used research accounts or data donated from real users to conduct experiments about whether the same content posted by different types of users would be subject to the same moderation decisions.

////

B. Data that researchers currently lack and want access to

Several researchers said that they need to know more about **what data content hosts have** before they can determine the specific types of data they would like to be able to access. In other words, if researchers do not know what data a host collects and maintains, they do not know what data to ask the host for. This lack of knowledge, some researchers said, limits the research questions that they ask, because they do not know whether certain platforms may have data that would allow them to answer different kinds of questions.

Researchers also said that to effectively use sampling techniques and have confidence in their results, they need to know **basic information about the overall volume of data on a service**. Some past research has described this as the "denominator problem": without information about the total volume of content on a platform, researchers cannot "compare content frequencies between platforms, or compare frequencies on the same platform over time."

Greater access to **advertising data** is a high priority for many researchers. While researchers reported using political ad libraries for their work, they also raised concerns that political ad libraries are often incomplete and do not accurately capture all of the political ads on a service. Researchers also said they wanted access to data about more types of advertisements, beyond political ads. In addition, many said that the types and granularity of data that hosts currently provide about ads is insufficient, and that they needed more information at greater levels of detail. In particular, researchers said they wanted more data about ad purchasers and full ad targeting data, to study issues such as the use of ads to target political messages and to spread mis- and disinformation.[6] Some researchers said they also wanted more data on expenditures on ads, especially political ads, across platforms and companies.[7]

6 Some social media companies may be considering or planning to provide additional data on advertisements to researchers. For example, in May 2022, Meta announced that it plans to provide "detailed targeting information for social issue, electoral or political ads" to certain vetted academic researchers through the Facebook Open Research and Transparency (FORT) environment.

7 Outside the workshop, researcher Laura Edelson has also noted that, to access certain ad libraries, researchers may have to "sign an agreement that limits how they use and share the data, which significantly hampers meaningful publication of any research findings."

Many researchers also seek data about hosts' use of **ranking and recommendation algorithms**. Several researchers said they needed more information about ranking and recommendation algorithms to investigate what kinds of content they favor or disfavor and the factors or data that cause algorithms to prioritize or promote content — or, on the other hand, deprioritize or downrank it. Researchers were interested in using this data to examine the consequences of algorithmic ranking and recommendation of content, and especially whether services are biased against particular types of content or users and the impact of that bias.

Researchers also sought access to additional data about the **content moderation algorithms** hosts use to detect and take action against content that violates their policies (other than downranking content), to understand how a host's content moderation is or is not working. In general, researchers expressed a desire for more data about content moderation, including information about specific content that hosts moderate, in order to evaluate content moderation efforts more completely.

Researchers also identified **historical content data and deleted content data** as types of data to which they desire more access. Researchers expressed frustration with a lack of standards around hosts' retention of data and with data access mechanisms that allow them to obtain only recent data or do not provide them access to content data that services or users have deleted. Access to historical data, researchers said, would allow them to compare current and historical events. For example, one researcher said they would use such data to compare false online narratives about the 2014 Ebola virus outbreak in West Africa with false online narratives about COVID-19. Researchers identified access to content data deleted by hosts as particularly important to understanding services' content moderation practices and investigating their fairness, accuracy, and efficacy. In addition, some research has relied on deleted content data that researchers were able to obtain to study how bad actors use deletions to manipulate users and violate platforms' terms of service.

However, researchers raised several ethical issues concerning access to historical or deleted content data, including whether allowing such access undermines services' efforts to prevent or ameliorate harm to users (such as invasions of privacy) by deleting certain content. In addition to these concerns raised by researchers, there may also be legal barriers to retaining and

sharing deleted content, including prohibitions on distributing illegal content and limits imposed by data protection law.

Some researchers, especially journalists, emphasized that they need access to **real-time data**, *i.e.*, data that is available as it is generated, or that is at least relatively recent. Real-time data is important for research on current events and other newsworthy topics, like mis- and disinformation concerning elections or public health issues, wars and military conflicts, disaster relief, and humanitarian aid. Researchers, for example, studying the impact of coordinated inauthentic activity campaigns, or other disinformation campaigns related to the Russian invasion of Ukraine, need access to recent information to publish their findings in a timely and impactful manner. Another researcher reported that access to real-time data has been crucial for research that informs the responses of public health officials to mis- and disinformation. Some academic researchers, however, suggested that real-time data access was less important for some of their research, which typically occurs on longer timelines than research conducted by journalists or civil society. The ethical and privacy law concerns raised by disclosure of historical or deleted content may also arise when researchers collect data in real time.

Researchers in general currently lack access to these types of data because the data are controlled by private companies that have not granted researchers access. However, researchers also identified **cost** as a barrier to accessing and using certain types of data. Researchers who reported using commercial social media monitoring or marketing firms to obtain data also said that this access is costly. In addition, even if data can be obtained for free or low cost, using or analyzing it can be expensive. Researchers said this is particularly true for video and audio content, which are increasingly posted by both advertisers and users, because researchers usually must transcribe video or audio files to use them for research. Transcription is expensive and often relies on automated tools that transcribe inaccurately. Because of the high costs of transcription, some researchers identified audio and video as types of data that are, in practice, currently inaccessible to them.

III. Who should be given access to data held by hosts, and how should individuals or entities seeking access to data and their research projects be vetted?

A. Current methods used by hosts to vet researchers

CDT asked researchers at the workshop to describe how they have been vetted by content hosts or other institutions before obtaining access to data and to discuss the pros and cons of different types of vetting procedures.

////

Some researchers noted that vetting is not a prerequisite for access to all data held by hosts. When using tools to access publicly available data, such as APIs for Reddit or YouTube[8] content, researchers said they sometimes underwent **no vetting**; in other words, these tools are available for anyone to use. Similarly, some researchers noted that they did not undergo vetting when using independent methods of gathering data, such as scraping.

Several researchers described filling out an **online application** to apply to access tools or datasets made available by hosts. For example, access to the Twitter API for academic researchers requires researchers to complete an online application that describes their academic credentials, affiliation, and research project, including what questions it seeks to answer; how it will use, analyze, and present

8 Researchers were referring to the YouTube Data API for Developers; after the Workshop, YouTube announced the YouTube Researcher Program, which will provide additional API access to YouTube data to certain vetted academic researchers.

Twitter data; and how the researcher will publish or share their work. Researchers said that hosts used the information provided in applications to verify the legitimacy of the researchers and ensure their research projects did not violate terms of service.

Researchers also described entering into **individual contractual agreements** with hosts for access to specific data, on a one-off or ongoing basis. In some instances, hosts also funded research, in addition to providing specific data. Some researchers had their contractual agreements for access vetted by lawyers, and others did not.

Especially for access to data at smaller or newer companies that are content hosts, researchers said **informal and ad hoc** vetting methods are common. At smaller or newer companies, there are typically no formal policies or procedures around researcher access to data, and decisions about whether to grant access may be made by a single employee. Researchers interested in studying these types of services described relying on networks and connections to "ask around" at a company about whether they could be given access to specific data. Researchers reported mixed success with actually gaining access to data, and they said they often did not receive any explanation about why a host had denied specific requests for data access.

////

B. Pros and cons of different vetting methods

Researchers identified several problems and tradeoffs with different vetting methods currently in use or proposed by lawmakers.

Researchers said that **slow vetting processes** are a problem. Certain types of researchers, like journalists and some civil society organizations in particular, rely on timely access to data to investigate and report information quickly, at the time that it is most newsworthy. Some academic researchers also said that their work can be time sensitive, and that they are also negatively impacted by slow or "kludgy" vetting processes. Particularly when discussing individual contractual agreements with hosts, researchers raised concerns about delays in finalizing the agreements or in obtaining the data once the agreements were reached. They also raised concerns that such agreements are not sustainable and will be impossible to use to give access to large numbers of researchers.

In addition, researchers emphasized that having hosts vet researchers and research projects creates **conflicts of interest** and

gives hosts too much power. Researchers suggested that hosts may be biased in favor of their own commercial interests and that they will make access decisions based on those interests, rather than the best interests of researchers or society. Researchers also said that certain vetting methods — like individual contractual agreements — give hosts too much power to decide what data researchers can access and what research they can conduct. For example, one researcher described using an agreement with a host to obtain a list of accounts, profiles, or content that the host had identified as fake. The host provided the researcher with this list before it removed the fake accounts or content from its service, to allow the researcher to study the activity of the accounts and corroborate the host's findings that they were fake. However, the researcher noted that the host controlled which accounts, profiles, or content appeared on the list it provided, inhibiting the researcher from having insight into the process by which the host identified the items to include on the list.

Researchers also identified a **lack of expertise in research** as a problem with models that rely on hosts to vet researchers. Researchers said that data access is not a priority for hosts and that employees often lack the expertise to evaluate the importance or legitimacy of requests for data access for research purposes. This problem is magnified with smaller or newer hosts that use informal and ad hoc vetting, researchers said.

In response to concerns about hosts' conflicts of interest and lack of expertise, some researchers supported using **government entities**, like the National Science Foundation or Federal Trade Commission, to vet researchers or research projects. However, other researchers raised concerns about empowering government actors — especially in less democratic countries — to vet researchers, noting that governments may have their own biases and conflicts of interest. For example, one researcher expressed concern about allowing the government of her country — which has attempted to restrict online speech by pressuring, investigating, or bringing legal action against hosts who do not remove speech critical of government officials, among other things — to vet researchers and determine who should be granted access to data held by hosts.

Others suggested that a **third-party body that is neither a host or a government entity** should be in charge of vetting researchers or research projects, based on a set of objective criteria. This idea is consistent with a report published in May 2022 by the European

Digital Media Observatory's (EDMO) Working Group on Platform-to-Researcher Data addressing how researchers can access platform data in compliance with the General Data Protection Regulation (GDPR). The EDMO Working Group recommended that an "independent intermediary body" be established to oversee certain aspects of the process for platform-to-researcher data access in a manner that complies with the GDPR, as set forth in the report and accompanying Code of Conduct. Among other functions, this independent intermediary body would vet researchers and research proposals to ensure they meet criteria detailed in the Code of Conduct.

Researchers also discussed whether access should be limited to researchers with an academic affiliation. Researchers recognized that limiting access to researchers with an academic affiliation can serve a valuable gatekeeping function, mainly because their research projects may be required to undergo review by an Institutional Review Board (IRB).[9] As a result, limiting access to some or all data, especially highly sensitive data, to academics may help ensure that the data is not publicly disclosed or used for disallowed purposes (such as commercial use), that the researchers are capable of conducting scientifically valid research, and that they are meeting obligations concerning data ethics, security, and privacy.

Nevertheless, most of the researchers at the workshop **opposed requiring an academic affiliation as part of the vetting process**. Researchers said that requiring an academic affiliation is limiting because it prevents access by journalists and researchers at civil society organizations. Indeed, some journalists and civil society researchers said they had been prevented from accessing certain data because they do not have an academic affiliation, which is often one of the questions hosts ask during the vetting process. Journalists and civil society researchers noted they can sometimes

9 At the workshop, Researchers noted both pros and cons to IRB review of research projects. Some research has suggested that IRBs were not designed to address online data collection and that IRBs, as currently constituted, may be an inadequate safeguard against misuse of online data for research purposes for a variety of reasons, such a lack of technical expertise or a lack of guidelines on data security standards and privacy protection that IRBs should apply to research projects that rely on online data collection. *See, e.g.*, Vitak et al., Ethics Regulation in Social Computing Research: Examining the Role of Institutional Review Boards, Journal of Empirical Research on Human Research Ethics (Aug. 23, 2017); Nebeker et al., Ethical and regulatory challenges of research using pervasive sensing and other emerging technologies: IRB perspectives, AJOB Empirical Bioethics (Dec. 8, 2017). In addition, researchers at the workshop acknowledge that researchers at civil society organizations or journalists would be excluded from data access in many instances if IRB review is required, because they do not have access to an IRB or their research is not appropriate for IRB review.

arrange collaborations with academics as a way to meet the requirement of academic affiliation, but they would prefer to access data without having to find an academic collaborator.

At the same time, researchers recognized the difficulty of defining "journalists" or "civil society researchers" in a way that is not over- or under-inclusive, or could allow for manipulation that results in anyone, including bad actors, gaining access to sensitive data held by hosts. Some suggested that, rather than vetting applicants for data access based on whether they met definitions for particular *kinds of researchers*, **vetting should be done on a project-by-project basis**. Under this method, regardless of what type of researcher sought access to data, their request would be vetted by examining the proposed research question, the data sought, the proposed research methods, their plans for data protection, and their plans for addressing other ethical concerns.[10] In particular, researchers suggested that vetting could examine whether the research would be conducted using methods that meet best practices in methodology, privacy, ethics, and technical security measures. Some researchers noted, however, that such a process would be resource intensive and difficult to scale, potentially resulting in less data access by researchers.

Some researchers also suggested that **access to public data** does not and should not require any vetting, because the data is, by definition, already publicly available. These researchers said that improving access to public data should be a policy priority because it would significantly enhance research while avoiding difficult questions about whether and how to grant special access to non-public data to only certain researchers, including questions about how to vet researchers. For example, laws could protect independent methods of gathering public data, such as scraping publicly available data, or prohibit platforms from cutting researchers off from accessing public data or tools used to access public data.

10 As explained above, researchers were skeptical of empowering platforms to conduct this vetting and suggested that a government entity or independent third party, such as that proposed by the EDMO Working Group on Platform-to-Researcher Data, might be better suited to vet research projects.

IV. What methods should hosts use to provide data to researchers?

Researchers described a number of methods they currently use to access data held by hosts. Roughly speaking, these methods fall into three categories: (1) methods or tools made available by hosts; (2) methods or tools developed independently by researchers; and (3) methods or tools developed by commercial firms. Researchers also discussed pros and cons of each of these methods.

Among the methods of access made available by hosts, researchers mostly discussed using **Application Programming Interfaces** of companies like Facebook, Reddit, or Twitter, to access publicly available data in a bulk format. Some said that they found APIs to be a robust, comprehensive, and easy way of accessing data on at least some services. However, they also highlighted the "streetlight effect" of API access: Researchers do a lot of research on platforms that make APIs available, because they make data accessible. Because these results might not be generalizable to other social media, research under existing conditions may not accurately or completely represent how social media works or its impacts. One researcher also said that the streetlight effect limits research on the impacts of social media outside the United States, where many people may use social media services that are not Facebook, Reddit, or Twitter and which do not have APIs.

Researchers noted that the data shared through APIs is entirely within the control of the host. Hosts are often unwilling to make full or particular data available,

do not know what data researchers would like to access or use, or have made mistakes in providing data in the past, according to researchers. Researchers expressed skepticism that current APIs provide accurate and complete information, and there have been high profile examples of hosts disclosing incomplete or inaccurate datasets. Some also raised concerns that hosts may eliminate certain tools entirely, such as CrowdTangle.[11] Other researchers said that limits on the amount of data that can be accessed from APIs — such as the Twitter Academic API's "Tweet cap" of 10 million Tweets per month or the YouTube Developer API's default quota allocation of 10,000 units per day — stifle or prevent research. Researchers also said that time limits on data included in APIs, such as data from only the past seven days, limit APIs' usefulness for historical research.

Researchers described several independent methods they use to gather data, including **surveying users**, deploying **research accounts** to investigate services' features and dynamics from the perspective of users with different characteristics, creating or using **data donation** tools that allow users to voluntarily give them data directly, and **scraping data**. Some researchers also said that they use **free tools or repositories** of data created by third parties, such as Junkipedia, a database of online misinformation across a range of platforms created by submissions and annotations from organizations and the public.

Researchers noted that some of these methods, particularly data donation and scraping, may be prohibited by platforms' terms of service, making them legally risky. Freelance researchers and researchers from less well-resourced institutions may be especially deterred by the legal risks from using these methods because they lack access to legal counsel and support. In addition, researchers said that they face technical barriers to scraping data, with some services deploying anti-scraping measures that technically limit or entirely prevent them from scraping. As a result, researchers said that scraping is most effective for data collection from smaller websites or non-social media services that do not use anti-scraping techniques. Finally, researchers said obtaining data through data donation may result in biased sampling and incomplete data. For example, one researcher mentioned that it is difficult to get data from elderly populations using data donation. As a result, research results based on data donation may be less reliable than those based on statistically valid sampling methods.

11 Following the workshop, news outlets reported that Meta is expected to shut down CrowdTangle.

Researchers also said they sometimes turn to **commercial social media monitoring or marketing firms** to obtain data — tools like Brandwatch, BuzzSumo, Infegy, Meltwater Explore, and Synthesio. While these tools can be useful for obtaining data, researchers said they are expensive, costing tens of thousands of dollars for the necessary access. This makes them less available for independent and less well-resourced researchers.

V. Recommendations for Policymakers and Hosts

Based on the information learned in the workshop, CDT makes seven recommendations to policymakers and content hosts that are seeking to improve independent researchers' access to data held by hosts. While our recommendations are informed by what researchers at the workshop said, they do not necessarily reflect the views of any researchers who participated in the workshop.

////

1. Policymakers and hosts should help researchers understand the potential data that is available.

Some researchers at the workshop said they do not always know what data to request from a host because they are unaware of what data the host has. *See* Section II.B. Hosts who voluntarily make data available to researchers upon request could address this barrier to access by providing comprehensive data dictionaries and data maps — or "codebooks"[12] — that clearly define what data they possess and may make available to researchers. In addition, hosts should publish a description of the data that they will *not* disclose to researchers and the reasons for this decision. For example, a host may be legally prohibited from voluntarily disclosing some data. Providing this information will ensure that researchers do not waste their time and resources pursuing data that the host will not or cannot voluntarily disclose. Hosts should also publicly explain the types of data that they make available to advertisers, other businesses, and law enforcement. This will help

12 The EDMO Working Group on Platform-to-Researcher Data has also recommended that platforms make available codebooks with respect to data or datasets that may be available for research and that contain certain specified information.

researchers and the public understand what data platforms make available to different actors. Finally, hosts should create transparent and open processes for researchers to share feedback on the available data and information about what data they would like access to for research purposes.

Codebooks should be public and centrally located, so anyone can understand what data is available and researchers do not have to piece together information about the available data from multiple locations.

Similarly, if lawmakers require that certain online hosts[13] make data available to researchers through a process in which researchers must request specific data, they should also require those hosts to disclose what data they possess in codebooks, so researchers know what data they can request. For example, Section 10(d) of DSOSA would require the Federal Trade Commission to issue regulations requiring certain "covered platforms"[14] to submit a data dictionary with specific information annually, to help researchers formulate requests for access. A law requiring publication of codebooks should specify what information should be contained in them based on what researchers say will be useful, though it will likely be difficult for hosts to publish a comprehensive codebook that meaningfully informs researchers of all of the data they can request.[15]

---

13 Any law mandating access to data held by hosts must carefully define which hosts are covered by the mandate. These definitions should be based both on the type of service the host provides and its size, which should be defined based on multiple criteria measured over time. *See* Eric Goldman & Jess Miers, Regulating Internet Services by Size, Santa Clara Univ. Legal Studies Research Paper (June 10, 2021).

14 DSOSA defines a "covered platform" as a "hosting service" that disseminates information to the public and has been designated as a covered platform by the FTC, based on a calculation that the number of average monthly active users in the United States is equal to or greater than 10,000,000. H.R. 6796 at Sec. 2(11). This number can be adjusted based on an increase or decrease in the population of the United States. *Id.* Sec. 2(11)(E). DSOSA defines a "hosting service" as an interactive computer service that stores information provided by, and at the request of, a user; and at any point in the preceding 2 calendar years, was owned or controlled by an entity with net annual sales or a market capitalization greater than $2,500,000, adjusted annually for inflation. *Id.* Sec. 2(14).

15 Lawmakers may wish to consider the information that the EDMO Working Group on Platform-to-Researcher Data recommended must be included in codebooks, such as "[a] description of the categories of data contained within the dataset"; "[a] description of the categories and approximate number of data subjects represented within the dataset"; "[a] description of what the dataset represents and its fitness for research"; and "[a] description of any relevant privacy or other settings that apply to the data."

In addition to requiring hosts to provide codebooks or other information about the data they have, any laws on this topic should include hosts early in the process of determining what data they will be required to make available in response to a specific request. Once a researcher makes a request for specific data from a host, the entity adjudicating that request should provide the request to the host, so the host can respond with information about its ability to provide the data to the researcher. If possible, the law should encourage direct communication between the researcher and the host to avoid inefficiency, delay, or miscommunication. A host is likely the only actor that knows what data exists and in what formats,[16] that can provide an estimate of the time and expense of producing specific data for researchers and that can give information about the risks to its users of providing specific data. Getting this information from the host early in the process of determining a researcher's request for access will save researchers' and the adjudicator's time and resources. In some cases it may allow researchers to reformulate requests to pursue research using additional or alternate data; in others it may ensure that hosts are not mistakenly required to turn over data that they do not already collect or create.

////

2. Hosts should have established processes through which researchers can request access to non-public data or tools to make public data more accessible.

A host that voluntarily makes data available to researchers should have an established and transparent process through which researchers can request access to non-public data or a host's existing tools that make public data more accessible, like APIs, that includes clear and known criteria against which requests are evaluated.

Without such a process, researchers are often left to rely on personal and professional connections at companies to make data requests. *See* Section III.A. As with other forms of networking, this type of informal system perpetuates inequality, because it is most likely to benefit researchers at elite institutions, researchers with backgrounds in industry, and researchers whose networks reflect the majority white and male make-up of many technology companies. Informal systems of requesting access also give hosts more power to decide access requests arbitrarily or ignore requests entirely. Establishing a set process by which researchers can

16 While publication of data dictionaries or data maps should help reduce the number of requests from researchers for data that a host does not have, it is still possible that researchers will request nonexistent data if, for example, they misinterpret a data dictionary or data map, or a host publishes an overly-general, incomplete, inaccurate or confusing data dictionary or data map.

request access and have their request evaluated on known criteria would help counteract this network gap and make hosts more accountable to researchers, even when they are providing data voluntarily.

While a process that allows researchers to request access to data may take the form of a sophisticated online application, it could be as simple as a website with brief instructions and an email address to which to submit requests. Hosts should provide at least basic information about to whom the request should be made, what criteria will be used to assess the request, and how a decision on the request will be communicated. Hosts should also be transparent about whether and why a request for access is granted or denied and the basis for that denial. These decisions should be made in a timely manner. A generalized application process is preferable to individual bespoke contractual agreements, which researchers at the workshop described as slow and cumbersome.

Hosts should also share information publicly about the number of research requests they receive, general information about the research proposal and data sought, and for which proposals the host accepted or rejected a request for data. This information could be shared through a periodic transparency report on researchers' requests for access to data held by a host.

////

3. Hosts and policymakers should make accessing and using data for research in the public interest less expensive.

Both hosts and policymakers should consider ways to reduce costs for accessing and using data for research in the public interest. Reducing costs would not only ensure that more public interest research could be done, but it would also create greater opportunities for researchers who lack institutional affiliations to support and fund their work and researchers who cannot obtain research grants. This is especially critical to improving research opportunities for Black, Latinx, Indigenous and multiracial researchers, who are underrepresented among the faculty at U.S. colleges and universities, particularly for Black researchers, who have also been found to be less likely to receive federal funding for their research compared to other researchers, in some cases.

To the extent that hosts offer paid tools to advertisers or others for accessing their data, they should make those tools available for free or lower cost to vetted researchers. If hosts have paid tools or services that would make it easier for researchers to use, process, and analyze the data they collect, they should also consider

donating these resources to researchers conducting research in the public interest.

Lawmakers should allocate additional funding or other resources[17] for non-government research using social media data and other data held by content hosts. Funding decisions for particular projects should be based on objective criteria and insulated from political or other pressures. In addition, in light of studies that have shown that Black researchers are less likely to receive some federal grants for research because of biases against the research questions they propose, lawmakers should require the collection and study of the demographics of the researchers to whom federal funding for platform research is awarded, and the research questions those researchers proposed. If this study reveals racial or other biases in the allocation of funding, funds should be reallocated to correct for them.

////

4. Policymakers should prioritize improving researchers' access to public data through legislation.

Researchers at the workshop identified many important research questions that can be answered using publicly available data, especially public content data. *See* Section II.A. Because providing access to that data does not present many of the complications that can arise from mandating access to non-public data, policymakers should prioritize facilitating access to public data.

To be sure, some researchers also expressed a desire to access non-public data, which can include data provided by users (such as content data sent directly to another individual or restricted to only particular people) and data created and held by hosts (such as information about how hosts' ranking and recommendation algorithms work). Researchers identified important questions that can be answered only by accessing non-public data, like how extremist content spreads through private social media groups or messaging services and how hosts are engaging in content moderation. But access to non-public data raises greater risks of invading users' privacy and revealing trade secrets or security measures used by hosts. Depending on the legal regime, it may

17 For example, some researchers have proposed that the federal government establish programs that would provide computational power to researchers for certain kinds of research using online data. In another example, the Carnegie Endowment for International Peace and Princeton University have announced a project to create the Institute for Research on the Information Environment (IRIE) "an international resource to study information ecosystems that can spur evidence-based policy solutions." Modeled on CERN, IRIE will, among other things, support "large-scale shared infrastructure" and maintain technical resources to support research on information ecosystems.

also be unlawful for hosts to disclose non-public data.[18] As a result, requiring them to provide researchers with access to non-public data is more complicated than requiring them to facilitate greater or easier access to public data. It requires policymakers to determine (or delegate to others to determine) when the benefits of access outweigh the risks and resolve difficult questions about how to vet researchers and research projects, provide data, and review researchers' publications to mitigate those risks.

While resolving these issues and giving access to at least some non-public data is important, policymakers should prioritize enhancing researchers' access to public data, which would enable socially important research while avoiding some of the thornier questions around access to non-public data. Mandating hosts to facilitate researchers' access to public data would require little vetting of researchers or research projects, because accessing public data is usually less privacy-invasive than non-public data. For the same reason, it likely would not require aggregating data, using data clean rooms, or using differential privacy to protect users' privacy. And it likely would not require restrictions on researchers sharing data or prepublication review of researchers' writing to ensure it does not reveal non-public data.

However, enhancing access to public data is not entirely without risks or challenges. For one thing, lawmakers must define "public data." Some categories of data, such as content posted on the open internet and accessible to anyone, are more obviously public, but the definition quickly becomes murkier after that. For example, content that is available only to users with accounts on a service, but not otherwise restricted, is in practice widely publicly accessible, assuming that anyone with an email address can create an account.[19] But content posted in a closed group on a service, to which users must request access, may or may not be considered "publicly" available, depending on factors such as whether requests are always granted or sometimes denied. Content posted in a closed group that contains millions of members is available to large parts of the public but is not entirely public. Content that a user posts publicly initially but then deletes (but the hosts retains

18 *See* 18 USC § 2702(a) & (b) (prohibiting electronic communications services and remote computing services from voluntarily disclosing the contents of electronic communications except in certain narrow, statutorily-defined circumstances).

19 Some proposed laws on researcher access to data would explicitly define this as "public content." *See, e.g.*, DSOSA (defining "public content" as " information on a covered platform that is available to a potentially unlimited number of third parties" and stating that "such term does not exclude information merely because an individual must log into an account in order to see the information.").

a record of it) may or may not be considered public data. The permutations and hypotheticals are myriad, and how a law defines "public" data will impact the potential privacy risks of making that data widely available.

In addition, even data available to anyone on the open internet can be misused in ways that invade people's privacy. Governments have used publicly-available online data to engage in harmful surveillance. For example, civil society organizations have documented how U.S. law enforcement agencies engage in social media surveillance of protesters and activists, including by using tools (one, two) that allow them to collect large amounts of public content data. Private entities have also used public online data in ways that invade privacy. For instance, facial recognition tools created by private companies that scrape public data online create risks of harassment, stalking, or doxxing, and some have been sold to law enforcement agencies. To help mitigate these risks, any law that seeks to improve researchers' access to public data must also include limits on government access to and use of the data, including by researchers who are hired by or otherwise act as agents of the government. Laws that enhance researchers' access to public data should also prohibit researchers' use of the data for any purpose other than noncommercial research.

One way of improving researchers' access to public data would be to require platforms to make tools for bulk disclosures of public data available, like APIs. While some hosts make some public data available in bulk formats already, requiring this form of access would expand that practice to more hosts. It would also require them to continue to provide such tools even if they determine that it is against their business interests, for example, if the resulting research casts a negative light on the host's service. Legislation could also require hosts to make older data available through APIs, require hosts to make exceptions to caps on data obtained through APIs for some research, or set more generous limits on all researchers' use of APIs. However, laws should not require hosts to retain data they would not otherwise retain so it can be disclosed through APIs.[20] Lawmakers should also explore the reasons why some hosts currently impose caps on data access through APIs, and whether there are legitimate concerns about costs or user privacy that support allowing such limitations.

20 As CDT has explained, minimizing the amount of data that hosts collect and retain is an important privacy protection for users.

Legislative mandates could increase the number of hosts that provide tools aiding in bulk disclosure of data to independent researchers, and help combat the "streetlight effect."[21] However, it may be too expensive for hosts of smaller and newer services to make data available or build tools to allow for bulk disclosure of data, and if they do not have employees with expertise in data access, they may be more likely to mistakenly release data that, for example, violates users' privacy. As a result, lawmakers should distinguish between larger hosts that are required to make public data readily available and smaller hosts that are not, to ensure that public data access mandates do not have an anti-competitive effect of driving smaller competitors out of the market and that covered hosts have the capacity and capability to prevent mistaken releases of data.[22] These size distinctions would ideally be based on multiple criteria, measured over a period of time. For example, a law could apply a data access mandate only to hosts that have a particular number of unique monthly users and also meet a minimum revenue requirement, measured over the most recent 12-month period.

To ensure that public data disclosed through an API or otherwise is accurate and complete, lawmakers should also consider establishing an independent auditing requirement. While some of the errors found in data made available through APIs have been discovered by researchers or journalists, or disclosed by hosts, requiring hosts to undergo an independent audit would help ensure the accuracy and completeness of disclosures.[23]

Finally, another way of improving researchers' access to public data is through safe harbors. These can shield researchers who conduct noncommercial, public interest research from legal liability for using independent methods to access public data, or prohibit hosts from cutting off or otherwise penalizing those researchers. A safe harbor could provide that a researcher who scrapes publicly available data, obtains it through data donation, or uses a research account to gather it cannot face civil liability under a host's terms of service or other civil or criminal liability for doing so for the purpose of public

21 Such a mandate will not eliminate the streetlight effect entirely. Some hosts have little or no public data, and requiring only certain hosts with public data to make data available may drive more research based on those sources, which may not be generalizable to other services with less or no public data.

22 For example, Article 31 of the Digital Services Act requires providers of "Very Large Online Platforms" — those with at least 45 million average monthly active recipients of the service in the EU — to make certain data available to researchers.

23 For a more in-depth discussion of how the independence of such audits could be established and other considerations for third-party audits of public data disclosures, see *Making Transparency Meaningful: A Framework for Policymakers, Analysis, Assessments, and Audits*.

interest research.[24] For maximum protection, a safe harbor should also prohibit a host from barring a researcher's account or using technological measures to block access to researchers who qualify for the safe harbor.

Lawmakers should also consider whether and how a safe harbor could be extended to those who make research tools, such as scraping or data donation tools,[25] but who do not directly engage in research themselves. However, a safe harbor for tool makers is more complicated than one for public interest researchers, since a tool that can be used for public interest research may also be misused for commercial purposes or even by a host's competitors.

////

5. Policymakers should prioritize access to advertising data.

Researchers at the workshop identified advertising data as particularly valuable and important. *See* Section III.A & B. Many of those researchers emphasized that ad libraries that included both political and non-political advertisements are important to answer questions like how dis- or misinformation spread through physical and mental health advertising, or whether and how hosts are applying their own content moderation rules or other policies to advertisements.

Lawmakers could improve researchers' access to advertising data by requiring hosts to maintain searchable ad libraries of all of the advertisements that have appeared on their services and to disclose other specific data about ads. While some data about advertisements — such as the content of the ad and, on most services, who sponsored it — is public, and therefore less likely to raise privacy concerns, other data that would be useful is frequently obscured. The Social Media DATA Act would require covered platforms to disclose, for example, information about the method used to target an ad, a description of the targeted audience, and a description of the actual audience for the ad, including demographic information. Non-public information about advertisements may be critical to research in the public interest, such as research about discriminatory ad practices by hosts and advertisers, giving rise to strong arguments as to why it should be disclosed.

24 The Knight First Amendment Institute at Columbia University has proposed a safe harbor to immunize certain research from legal liability based on a services' terms of service, the Computer Fraud and Abuse Act (CFAA), and state-law analogs to the CFAA. Several proposed bills on researcher access to data also include a safe harbor provision. *See* Appendix.

25 One example of such a tool is Pushshift.io, a "big-data storage and analytics project" that, among other things, makes available an API of Reddit data to users, including researchers.

However, when considering requiring disclosure of non-public advertising data, lawmakers should also think carefully about whether its disclosure could violate users' privacy — by, for example, allowing a researcher to infer a user's demographic information based on the ads she is served. Public disclosure of targeting criteria for advertisements may also negatively impact businesses that advertise on hosts' services, since revealing how they choose to target ads to their competitors could reveal information about their marketing or other strategies. Risks to user privacy and businesses that advertise on hosts' services may require imposing conditions on access to this data, such as restricting non-public advertising data to only researchers with vetted research projects or requiring other privacy-protective measures to be applied to the data itself.

////

6. When vetting is necessary, hosts and legislation should evaluate specific research projects and plans based on established, transparent criteria, rather than whether a researcher falls within a particular category.

Unlike with public data, more extensive vetting is necessary and appropriate before a researcher is permitted to access non-public data. Rather than relying on a categorical approach, wherein only academics are granted access to non-public data and journalists and researchers at civil society organizations are excluded, requests for access should be evaluated on several criteria.[26] Those should include the particular research project and plans, and the researchers' qualifications to conduct the proposed research and ability to take appropriate steps to protect the privacy and security of data. Hosts that voluntarily offer data to researchers could follow this approach, as could laws that require hosts to provide non-public data. Under this approach, vetters would examine transparent, established criteria about the research project and research plan, such as:

- Whether the research project is conducted for non-commercial purposes;[27]
- Whether the research plan demonstrates a research methodology considered valid within the researchers' field

26 Vetting may be done by hosts, government entities, or independent third-parties, and there are pros and cons to each approach, as discussed above. *See* Section III.B. This recommendation addresses the criteria upon which vetting should be done, regardless of who conducts the vetting.

27 Journalists and researchers sometimes offer their work for sale to the public. Dissemination of research to the general public for journalistic or informational purposes in the form of a news report, book, research paper, or similar publication by itself should not be considered a "commercial purpose" even if readers or viewers must pay to access the content.

of study or profession, and whether the researcher has the necessary qualifications to conduct the proposed research;

- Whether the requested data is necessary to conduct the research, and whether more privacy-protective means are reasonably available;
- Whether the research plan has been approved by an independent body tasked with examining the ethics of conducting the research; and
- Whether the research plan includes adequate privacy and security safeguards for using, transferring, and storing data and publishing results, and whether the researcher has demonstrated their ability to comply with those safeguards.

While this list is not exhaustive, it demonstrates the type of questions about the purpose, ethics, and methodological validity of the research and about the researchers' ability to protect users' privacy and maintain the confidentiality and security of data that should be the focus of vetting inquiries.

////

7. Policymakers should strictly limit or eliminate the ability of commercial social media monitoring and marketing firms to collect and sell users' data.

Several researchers said that they rely on data brokers, such as commercial social media monitoring or marketing firms, to obtain data for their research. *See* Section IV. While access to data through data brokers may enable research, it also raises privacy concerns. The data available from commercial sources often goes beyond public data.[28] As CDT reported in *Legal Loopholes and Data for Dollars: How Law Enforcement and Intelligence Agencies Are Buying Your Data from Brokers*, some data brokers obtain and sell access to non-public data, including to law enforcement. And, even if commercial social media monitoring or marketing firms collect and sell only public data, users sometimes post sensitive data publicly, and public data can also be misused in ways that harm users, such as through law enforcement monitoring or invasive commercial practices. *See* Section V.4.

The harms these data brokers cause are not outweighed by the limited benefits researchers see from using them to gather

28 "The range of platform data commercially available from platforms and from third parties to assist marketers and advertisers is sometimes richer in detail and insight than publicly-available data. Private third-party data sets often combine social media data with other sources of media data."

data. Accordingly, lawmakers should, as CDT has previously recommended, significantly limit data broker collection and processing of information – and in at least some cases, prohibit it altogether. They should also ensure that consumers are given sufficient information to understand what information data brokers have collected about them, and what meaningful choices they have to shape that data, such as preventing the sale of their data to brokers in the first place.

VI. Conclusion

Independent researchers' access to data held by social media services and other content hosts holds great potential, both to increase public understanding of how online services impact our society and to inform policymaking. Researchers have already made tremendous use of data from voluntary disclosures by hosts or independent research methods. Their impact could be even greater with key improvements to data access.

In Europe, as a result of the Digital Services Act, we are about to see what a first attempt at mandatory researcher access to data can enable in terms of public interest research and societal and policy change, and what unanticipated risks or concerns may arise. Guidance from the new Code of Conduct proposed by the EDMO Working Group on Platform-to-Researcher Data Access may prove useful to platforms and researchers in both Europe and the U.S. Soon, U.S. law may also take its own approach to mandating researcher access to data.

Properly balancing competing interests, such as the risks to user privacy, may require policymakers to take incremental steps to improve researchers' access to data, and to carefully assess whether those steps are serving the public interest. Across the Atlantic, and indeed around the world, the process will necessarily be iterative, and flexibility and thoughtful review of progress and outcomes will be key. As this workshop and report demonstrate, there is much that can be done to improve researchers' access to data, and much we can learn from taking initial steps, even if they may not immediately get researchers access to maximum amounts of data from all hosts. However, even gradual change can have big effects, and will improve our understanding of the online services that increasingly shape our lives.

Appendix: Background on legislative proposals

This appendix summarizes four bills — or discussion drafts of bills — in the United States that would require particular technology companies to provide data to independent researchers, the public, or both: the Digital Services Oversight and Safety Act (DSOSA), Kids Online Safety Act (KOSA), Platform Accountability and Transparency Act (PATA), and Social Media DATA Act. It also summarizes Article 31 of the Digital Services Act, which will require providers of Very Large Online Platforms to disclose certain data to "vetted researchers."[29]

For a comparison of bills and proposals in the United States and Europe concerning researcher access to data held by hosts, see CDT's chart, *Independent Researcher Access to Social Media Data: Comparing Legislative Proposals*.

Digital Services Oversight and Safety Act (DSOSA) (H.R. 6796): Introduced by Representative Trahan and cosponsored by Representatives Schiff and Casten, DSOSA would establish an Office of Independent Research Facilitation within the Federal Trade Commission (FTC) to certify academic institutions or 501(c)(3) organizations and researchers affiliated with them. It would also administer access to covered platforms' data by certified organizations and researchers for research into "the impacts of the content moderation, product design decisions, and algorithms of covered platforms on society, politics, the spread of hate, harassment, and extremism,

29 The European Parliament adopted the DSA in July 2022. The provisions of Article 31 are expected to go into effect in late 2022 or early 2023.

security, privacy, and physical and mental health." DSOSA would require the FTC to issue regulations establishing the types of information that covered platforms will make available to certified researchers, the manner in which the information will be accessed, and the circumstances under which it will be optional or mandatory for covered platforms to provide certified researchers with access to the information. DSOSA would also establish a safe harbor from liability under law, or for violation of a platform's terms of service, for certified researchers who collect data using research accounts or receive data donations from users. Finally, DSOSA would require the FTC to issue regulations making an advertising library and "high-reach public content stream" available to certified researchers, the FTC, and the public.

The Kids Online Safety Act (KOSA) (S. 3663): A bipartisan bill introduced by Senators Blumenthal and Blackburn, KOSA would establish a process under which the National Telecommunications and Information Administration (NTIA) would vet researchers from academic institutions or 501(c) organizations. It would require covered platforms to make data assets available to qualified researchers to conduct public interest research about harms to children's safety and well-being. KOSA would also create a safe harbor from certain causes of action related to terms of service violations brought against a researcher for collecting data assets to conduct public interest research about harms to children.

The Platform Accountability and Transparency Act (PATA): PATA is a discussion draft of a bipartisan bill by Senators Coons, Portman, and Klobuchar. The bill would establish a process by which the National Science Foundation and a newly established office within the FTC, the Platform Accountability and Transparency Office, would vet academic researchers and their projects aimed at studying activity on a platform. This process would also determine the particular data that a covered platform should be required to make available to an approved researcher, and establish privacy and cybersecurity safeguards for the data. Covered platforms would be required to make qualified data available to qualified researchers or risk losing their immunity under Section 230 for civil claims related to their failure to comply. They could also face an enforcement action by the FTC for unfair and deceptive trade practices if they do not make qualified data available to qualified researchers as ordered. PATA would also establish a safe harbor from civil and criminal liability for collection of certain data for newsgathering or research through automation, data donation from users, or the use of research accounts, as long as certain statutory criteria are met. Finally, PATA would authorize — or in some cases, require

— the FTC to issue regulations that require covered platforms to proactively disclose other data, metrics, or information to the public.

The Social Media Disclosure and Transparency of Advertisements (DATA) Act (H.R. 3451): The Social Media DATA Act, a bill introduced by Representative Trahan and cosponsored by Representative Castor, would require the FTC to issue regulations requiring covered platforms to make an advertising library available to researchers and the FTC with certain specified information. It would also require the FTC to establish a Working Group for Social Media Research Access to study best practices for making data from interactive computer services available to academic researchers and recommend a code of conduct for researchers working with such data and make policy recommendations concerning data access.

Article 31 of the Digital Services Act would require providers of Very Large Online Platforms (VLOPs) to provide data to "vetted researchers" upon request from the Digital Services Coordinator of establishment (DSC). Researchers may use the data only for a limited purpose: to conduct research that contributes to the detection, identification, understanding, and mitigation of systemic risks in the EU that are identified in the DSA.

A provider of a VLOP may ask that a data request from the DSC be amended on the basis that it does not have access to the data or that providing access "will lead to significant vulnerabilities for the security of its service or the protection of confidential information, in particular trade secrets." If a provider requests an amendment, it must propose an alternative means of access to the requested data or suggest other data that could be used to fulfill the request. Article 31 also specifies that when providers of VLOPs are required to provide data to vetted researchers, they must provide access to data through "appropriate interfaces" including online databases or APIs.

Under Article 31, the DSC will declare researchers "vetted researchers" based on an application submitted by researchers. "Vetted researchers" must satisfy certain criteria:

> (a) they are affiliated to a research organisation as defined in Article 2, point (1), of Directive (EU) 2019/790;[30]

30 This includes a university, research institute, or any other entity whose primary goal is to either conduct scientific research or carry out educational activities also involving the conduct of scientific research. This research should occur on a not-for-profit basis, or all of the profits should be reinvested either in the entity's scientific research or pursuant to a public interest mission recognised by a Member State.

> (b) they are independent from commercial interests;
> (ba) the application submitted by the researchers discloses the funding of the research;
> (c) they are in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request and to protect personal data, and they describe in their request the appropriate technical and organisational measures they put in place to this end;
> (d) the application submitted by the researchers justifies the necessity and proportionality for the purpose of their research of the data requested and the timeframes within which they request access to the data, and they demonstrate the contribution of the expected research results to the purposes laid down in [Article 31] paragraph 2;
> (e) the planned research activities will be carried out for the purposes laid down in [Article 31] paragraph 2;
> (f) commit to making their research results publicly available free of charge, within a reasonable period after the completion of the research and taking into account the rights and interests of the recipients of the service concerned in compliance with Regulation (EU) 2016/679.

In addition, Article 31 requires providers of VLOPs to provide some data to other researchers, including researchers at nonprofits, who meet some but not all of the criteria for "vetted researchers." It states that providers of VLOPs must provide data that is "publicly accessible in their online interface" including "real time data where technically possible" to researchers who meet criteria (b), (ba), (c), and (d) and use the data solely to perform research that contributes to the detection, identification, and understanding of systemic risks in the European Union that are identified in the DSA.

Finally, Article 31 empowers the European Commission to adopt delegated acts "laying down the technical conditions under which providers of [VLOPs] are to share data" with vetted researchers "and the purposes for which the data may be used." Article 31 instructs that the delegated acts should address how data can be shared with researchers consistent with the General Data Protection Regulation (GDPR). For more information about how the GDPR applies to researchers who access data held by content hosts, see the EDMO Working Group on Platform-to-Researcher Data Access's report on researcher access to platform data.

cdt.org

cdt.org/contact

Center for Democracy &
Technology
1401 K Street NW, Suite 200
Washington, D.C. 20005

202-637-9800

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY