CENTER FOR DEMOCRACY & TECHNOLOGY

Researcher Access to Data Held by Online Hosts of User-Generated Content

From CDT's *Making Transparency Meaningful: A Framework for Policymakers*

Independent researchers, public policy advocates, and journalists seek access to data from hosts of user-generated content in order to investigate scientific or other academic questions, publish news or analysis, and inform advocacy and policy making. Improving researcher access to this data requires a common framework for understanding the current methods of access and the key questions – and the tradeoffs involved in their answers – that will shape policy decisions about regulating researcher access to this data.

////

Current Methods of Independent Researcher Access to Data

In general, independent researchers have three methods of obtaining access to data from hosts of user-generated content: (1) access to public data; (2) company-sanctioned access to public or nonpublic data; and (3) independent access to nonpublic data or data that is public but restricted.

Some data is available on the public internet.[1] Researchers collect this data manually or using automated methods such as scraping. For example, the website Pushshift[2] scrapes comments and posts from the social media website Reddit to create an archive of Reddit content that researchers have used to study issues such as social media echo chambers[3] or the effects of social networking deplatforming.[4]

1 Whether online data is "public" may not always be immediately clear, and the definition of "public" may vary based on circumstances or statutory definitions.

2 Pushshift.io; Jason Baumgartner et al., *The Pushshift Reddit Dataset*, Assoc. for the Advancement of Artificial Intelligence (2020).

3 Matteo Cinelli et al., *The echo chamber effect on social media*, Proceedings of the Nat'l Academy of Sciences of the United States of America (Feb. 23, 2021).

4 Shiza Ali et al., *Understanding the Effect of Deplatforming on Social Networks*, Assoc. for Computing Machinery (2021).

As discussed below, the scope of permissible scraping of public data is subject to ongoing policy and legal debate.

Some companies voluntarily make certain data available to researchers, often through Application Programming Interfaces (APIs).[5] APIs may be for general use or for use specifically by researchers. Companies may also voluntarily make data available through other datasets provided directly by the company or in partnership with a third party. Social Science One,[6] CrowdTangle[7] and the Twitter API for Academic Researchers[8] are all examples of company-sanctioned methods of researcher access to data. Company-sanctioned access may require researchers to apply to the company for access, satisfy criteria for access set by the company (such as affiliation with an academic institution), and obtain company approval of their research plans.

Finally, researchers use independent measures to gain access to hosts' data without company sanction, particularly from social networking companies.[9] The "data donation" method allows internet users to give their data directly to researchers, often using a custom web browser or browser extension installed by volunteers or paid participants.[10] The browser or extension collects and provides to researchers certain data from all of the internet sites that users visit or from particular social networks.[11] Researchers use the collected data, often paired with demographic data from the participants, to examine how users encounter or interact with content and how social networks sites target content to users. For example, the MarkUp's Citizen Browser Project,[12] NYU Ad Observer,[13] and Mozilla Rally[14] all rely on data donation to gather social networking data.

Another method of independent access asks internet users to send data that may not be otherwise publicly available to a central platform or repository, which can then be

5 APIs are "tools that allow programmers from outside the company to retrieve a set of data from company servers." Elizabeth Hansen Shapiro et al., *New Approaches to Platform Data Research*, NetGain Partnership at 13 (Feb. 2021).

6 *Social Science One* (last visited Nov. 29, 2021).

7 Will Bleakley, *About Us*, CrowdTangle (last visited Nov. 29, 2021). In April 2021, Facebook integrated CrowdTangle into its "integrity team," a move which some have criticized as intended to weaken the transparency provided by the tool in the face of negative information about Facebook reported as a result of CrowdTangle data.

8 *Twitter API: Academic Research Access*, Twitter (last visited Nov. 29, 2021).

9 This method is sometimes referred to as an "adversarial approach."

10 Giving users the ability to export their data, such as through interoperability services like Google Takeout, may also enable them to share historical data with researchers. *See* Ross James, *'What is Google Takeout?': How to use Google's simple tool for downloading all of your account data at once*, Insider (Jan. 23, 2020).

11 A browser extension is software that enhances the capabilities of a web browser, such as by allowing users to store passwords or block advertisements. Browser extensions used for data donation to researchers often copy specific content from the websites a user visits or a specific subset of those websites and transmits the data to the researcher. For example, the NYU Ad Observer browser extension copies the ads a user sees on Facebook or YouTube. *Ad Observer*, NYU Cybersecurity for Democracy (last visited Nov. 29, 2021).

12 *The Citizen Browser Project—Auditing the Algorithms of Disinformation*, Markup (Oct. 16, 2020).

13 Ad Observer, supra n.11.

14 *It's your data. Use it for a change.*, Mozilla Rally (last visited Nov. 29, 2021).

accessed by researchers. For example, Junkipedia uses user submissions to create an annotated archive of mis- and disinformation from a range of platforms.[15] In a third method of independent access, researchers pose as users or advertisers to gather data. For example, researchers might pose as users by creating accounts with different demographic profiles or indicia to investigate patterns of bias[16] or as advertisers by placing ads on social media sites to investigate ad targeting.[17] Social media companies have resisted or shut down independent methods of data access in the past, such as when Facebook deactivated the accounts of two researchers from the NYU Ad Observatory, effectively blocking their research.

////

Enabling researcher access to data: Considering tradeoffs

Who should have access to data from hosts of user-generated content?

Because certain data can include highly sensitive and private information, restricting access to data to only particular entities and individuals is often desirable. Access could be restricted to certain categories such as "researchers" or "journalists." But defining these categories can be difficult and overly exclusive. For example, if "researchers" are defined as those with an academic affiliation, then journalists, civil society, independent analysts, government researchers, and 82% of all scientists and engineers[18] would be excluded from access. "Academic affiliation" would also have to be defined to determine whether, for example, affiliation with for-profit or foreign colleges and universities qualified.

Another approach would restrict access based on the intended use of the data. For example, access could be granted only to researchers whose research is in the public interest or meets other criteria intended to establish the research's importance or rigor, or only to researchers with a non-commercial purpose. Intended-use restrictions would require vetting the merits of proposed research or its non-commercial purpose and giving an entity or person (such as the company who holds the data, a government agency, or some other third party) the power to decide which researchers should be permitted to access data.

15 *About Junkipedia*, Junkipedia (last visited Nov. 29, 2021).

16 *See, e.g.,* Benjamin G. Edelman et al., *Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment* (September 16, 2016). American Economic Journal: Applied Economics 9, no. 2 (April 2017) 1-22, Harvard Business School NOM Unit Working Paper No. 16-069; Sam Levin, *Airbnb blocked discrimination researcher over multiple accounts*, Guardian (Nov. 17, 2016); Kalhan Rosenblatt, *Senator's office posed as a girl on fake Instagram account to study app's effect*, NBC News (Sept. 30, 2021).

17 *See, e.g.*, Piotr Sapiezynski et al., *Algorithms That "Don't See Color": Comparing Biases in Lookalike and Special Ad Audiences*, arXiv (Dec. 16, 2019).

18 *S&E Workers in the Economy*, Nat'l Ctr. for Science and Eng'g Statistics (last visited Nov. 29, 2021).

Vetting research to establish compliance with intended-use restrictions raises the risk of vesting too much power in the vetter to decide what research is in the public interest and what research is not; to lessen that risk, the vetter should be prohibited from discriminating based on viewpoint or the vetter's self interest. Even then, intended-use restrictions may still prohibit some worthy research; a non-commercial purpose restriction, for example, could inadvertently bar researchers who intend to sell books or news articles based on their research. However, given the privacy and other risks of granting researchers access to certain data held by hosts of user-generated content, screening research to determine whether it is in the public interest or meets other criteria may be appropriate.

Finally, access could be restricted based on an entity's or individual's ability to meet certain content-neutral criteria, such as the ability to conduct scientifically valid research (the meaning of which would have to be defined) and meet data security and privacy standards. Academic institutions that receive federal funding for research will typically have an Institutional Review Board (IRB) that could serve some of these functions, but the capacity of IRBs to conduct such assessments and enforce such standards is far from guaranteed.[19]

What types of data do researchers seek access to, and why?

Different researchers seek access to different kinds of data to answer questions in fields such as the social sciences and computer science. Data from hosts of user-generated content can be broken down into a variety of categories.[20] One analysis has divided such data into three categories: (1) content data, such as posts or comments made by social media users or advertisements; (2) moderation data, or data about hosts' content policies and their decisions about enforcement of those policies; and (3) distribution data, or data about how and why users see particular content, including content recommendation algorithms.[21] Researchers may also seek access to other data, such as demographic information about users (which can provide important context to other categories of data), social networks or social graphs data, *i.e.*, data that shows how users of a social network are connected to each other, and other metadata. The data that researchers seek access to may be historical data or real-time data.

19 *See* Simon N. Whitney, *Institutional review boards: A flawed system of risk management*, 12(4) Research Ethics 182 (2016); Prosperi, M., Bian, J. *Is it time to rethink institutional review boards for the era of big data?*, Nat. Mach. Intell. 1, 260 (2019).

20 Access to data unrelated to user speech or access to information, such as data about the finances or employees of hosts of user-generated content, customer data stored by cloud services, or government data held by companies with government contracts are outside the scope of this overview.

21 *See* Shapiro et al., *supra* n.5 at 17-24.

Different kinds of data raise greater or lesser privacy concerns, even within categories.[22] For example, content data about public social media posts may raise few privacy concerns, while content data about direct messages between users of a messaging service may be highly sensitive and protected from disclosure by law. Real-time content data about elections advertising may present different research opportunities, and raise different speech and privacy concerns, from historical data about ad targeting during a past election.

What online services should make data available to researchers?

While many hosts of user-generated content may have data that would inform research, most focus has been on access to data from consumer-facing online companies such as social media platforms. Defining what entities qualify as a "social media platform," however, is not always straightforward, since they may include social networking sites and applications, messaging services, content aggregation services, or even comment sections on news websites. Some of these services may have data that is more or less useful to research in the public interest and more or less sensitive than others. In addition, it may be necessary to draw distinctions in and between what data or how much data should be shared with researchers based on the size of the host to ensure that smaller hosts are not burdened by costs and obligations that may drive them from the market. These distinctions can be based on factors such as the age of the company, number of employees, revenues, or consumer usage, with upsides and downsides to each metric.[23]

How do we safeguard individual privacy while enabling broader access to data by researchers?

Company-held data can expose individuals' personally identifiable information, patterns of their online behavior, and the inferences that companies make about them. Certain data may be so sensitive that researchers should not be granted access to it at all, or should be granted access to it only for certain research projects. As a threshold matter, companies, lawmakers, and others considering the issue of researcher access to data should consider what data, if any, is so sensitive that it cannot be provided to researchers in some or all instances.

To the extent that researchers are granted access to personal or other sensitive data, companies, policymakers, and others must consider what privacy and data security protections to put in place. Privacy protections may be applied to the entirety of a research projector or in a multistage process. For example, a researcher could be

22 In addition, companies may be legally prohibited from sharing certain data, *see, e.g.*, 18 U.S.C. § 2702(a) (prohibiting a person or entity providing an electronic communication service to the public from knowingly divulging to any person or entity the contents of a communication while in electronic storage by that service, with limited exceptions) or may lose certain legal protections for data, such as those for trade secrets, if they disclose it publicly.

23 Eric Goldman & Jess Miers, *Regulating Internet Services by Size*, CPI Antitrust Chronicle, Santa Clara Univ. Legal Studies Research Paper (May 2021).

granted access to an anonymized dataset for their research project, or they could be granted access to an anonymized dataset for their initial research and then later granted access to more sensitive data if they can demonstrate that their research is fruitful and access to additional data is necessary.

Privacy and data security can be protected through technical measures, access controls, legal liability, or a combination of methods. Common technical means of enforcing privacy include data aggregation, by which raw data is combined in a summary form, and differential privacy, which uses mathematical techniques to allow analysis of data while protecting its identifiable characteristics.[24] These methods may require significant expertise and expense to implement and may limit the type of research that can be done. Access controls help protect user privacy by allowing researchers to access data only within environments where hosts can limit the analyses that researchers can perform, prohibit the copying or removal of data, and have in place data security measures such as encryption. This method may significantly constrain the type of research and the type of researchers who are able to conduct research, and it may prevent the sharing of data with research partners at other institutions, or other researchers who may seek to replicate a particular study. Finally, privacy can be protected through imposing legal liability for misuse of data in ways that violate privacy or security requirements, whether through generally applicable law that extends to certain data use, a statute written specifically to govern researcher access to data, or terms of service. Such methods, however, are only as effective as the enforcement mechanism and resources that accompany them.

How can companies and lawmakers eliminate unnecessary legal barriers to researchers' independent access to data?

Researchers that use independent methods to access data in the United States may face civil or criminal barriers to their work that lawmakers could eliminate or ameliorate. For example, changes or updates to the Computer Fraud and Abuse Act (CFAA) or Digital Millennium Copyright Act (DMCA) may remove or lessen the risk of liability for researchers.[25] In addition, voluntary carve-outs in companies' terms of service to permit research would remove the risk of civil liability for researchers who break terms of service by, for example, offering browser extensions that facilitate data donation. Congress could also require such carve-outs or immunize from civil liability researchers who break a companies' terms of service.

However, the CFAA, DMCA, and company terms of service can be important tools for limiting misuse of company data. As a result, companies and lawmakers should consider limiting any such carve-outs to apply only to research in the public interest. One challenge in this approach is how to write provisions that precisely distinguish between "white hat" or research in the public interest that should not be prohibited and other

24 Bennett Cyphers, *Understanding differential privacy and why it matters for digital rights*, Access Now (Oct. 25, 2017).

25 Joseph Lorenzo Hall & Stan Adams, *Taking the Pulse of Hacking: A Risk Basis for Security Research* (Mar. 2018).

activity that in the guise of "research" involves invasions of privacy, infringement of intellectual property, or other misuses that should be prohibited. In addition, the tradeoffs involved in intended-use restrictions on researcher access to data discussed above, such as the potential for abuse in vesting the power to decide what research is in the public interest in companies or government, apply here as well.[26]

Finally, in some instances, companies have used legal provisions or government consent decrees as a pretext for blocking researchers' access to data they hold on privacy grounds.[27] New federal privacy legislation or future government settlements with companies that violate existing privacy laws could state explicitly that research in the public interest or research that complies with particular criteria intended to protect user privacy are not forbidden on privacy grounds, to prevent companies' use of privacy laws or consent decrees as a basis for blocking independent methods of researcher access to data. Again, however, defining research in the public interest presents challenges.

Should researchers' access to data directly from companies continue to be at companies' discretion or be mandated in certain circumstances?

Current company-sanctioned methods of researcher access to data are voluntary. Voluntary provision of data to researchers allows a company and researchers to develop and experiment with different processes for providing access, which may lead to the development of new and innovative data-sharing methods. It also allows a company to decide what and how much data to share based on information that only the company may possess, such as the specific privacy needs of its users and the company's financial and other capacity to provide researchers with access.

However, company-sanctioned methods also allow companies to control which researchers can access their data, which may allow them to select researchers they perceive as sympathetic to their interests or with whom they have previous relationships, potentially excluding researchers from less well-known or well-connected institutions. Some critics also argue that company-sanctioned methods give companies too much control over what data they will make available, for what purposes, and for how long. In addition, purely voluntary company-sanctioned access raises the possibility that a company will intentionally manipulate data[28] or release erroneous datasets.[29]

Accordingly, some researchers, advocates, and lawmakers have proposed creating

26 *See supra* Researcher Access to Data at 3 ("Who should have access to data from hosts of user-generated content?")

27 *See, e.g.*, Issie Lapowsky, *The FTC hits back at Facebook after it shut down NYU research*, Protocol (Aug. 5, 2021).

28 Hubert Horan, *Uber's "Academic Research" Program: How to Use Famous Economists to Spread Corporate Narratives*, Promarket (Dec. 5, 2019).

29 Craig Timberg, *Facebook made big mistake in data it provided to researchers, undermining academic work*, Wash. Post (Sept. 10, 2021).

legal incentives[30] or even requiring companies to provide data to researchers. In choosing between incentives and mandates, lawmakers should consider that the First Amendment may prohibit the government from requiring hosts to provide certain moderation data and distribution data to researchers because doing so could violate their right to exercise editorial discretion over the user-generated content they host.[31] Incentivizing or mandating researcher access to data will also require policymakers to resolve all of the prior questions raised in this section: Who should have access to the data? What data should be provided? From what companies? And what privacy protections should be in place?

What is the best mechanism for providing researchers access to data from companies?

Company-sanctioned access to data – whether voluntary or in response to mandates or incentives – can occur through several possible methods, including:

- Making data directly available to researchers;
- Contributing data to a repository administered by a government entity; and
- Contributing data to a repository administered by a third party, such as an academic institution, existing non-profit, or new entity established for this purpose.

There are pros and cons to each of these methods. Directly sharing data with researchers allows use of existing mechanisms and infrastructure for access, such as APIs. However, this approach may be more burdensome for researchers and limit cross-company comparisons. Also, if the data is put in the hands of researchers, it may present privacy and security risks, such as researchers abusing their access by sharing data or inadequately protecting against leaks or other exposure of the data.

Creating a repository administered by either a government entity or third-party would potentially allow for standardization in data formats, methods of access, and privacy controls (while creating additional burdens and costs on companies to standardize data); however, it could create concerns about data security since the repository would be an attractive target for malicious actors seeking to gain unauthorized access to the data. A third-party repository could remove some of the self-interest involved if companies themselves are vetting researcher access, though it would need to be carefully designed to ensure that the third-party administrator was independent from companies that contribute data. In determining whether a repository administered by the government or a third-party is preferable, companies, policymakers, and others should consider whether it is preferable to have the government or a third-party in charge of vetting researchers. A repository administered by the government will also raise concerns about government surveillance of users, particularly if government access to the repository is not strictly limited.

30 Incentives could include offering companies protection from liability for privacy violations that result from the sharing of data with researchers.

31 *Herbert v. Lando*, 441 U.S. 153 (1979); *Miami Herald v. Tornillo*, 418 U.S. 241 (1974).

This brief is a part of the December 2021 CDT report, *Making Transparency Meaningful: A Framework for Policymakers*.

Additional CDT work on this topic: https://cdt.org/insights/report-making-transparency-meaningful-a-framework-for-policymakers

For more info, please contact **Emma Llansó**, Director of the CDT Free Expression Project or **Caitlin Vogus**, Deputy Director of the CDT Free Expression project.

ellanso@cdt.org

@ellanso

cvogus@cdt.org

@CaitlinVogus

The **Center for Democracy & Technology** (CDT) is a 25-year-old 501(c)3 nonpartisan nonprofit organization working to promote democratic values by shaping technology policy and architecture. The organisation is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.

@CenDemTech

CENTER FOR DEMOCRACY & TECHNOLOGY

CENTER FOR DEMOCRACY & TECHNOLOGY

Independent Researcher Access to Social Media Data: Comparing Legislative Proposals

Researchers use data from social media companies and other hosts of user-generated content to study important topics of public concern, such as the efficacy of different content moderation efforts and ideas to improve them, the spread of dis- and mis-information online, ranking and recommendation algorithms, and online advertising. But some researchers have been stymied by the type and amount of data available, the level of control that social media companies exert over researchers' access, and other barriers.

Lawmakers in both the United States and Europe are increasingly focused on how to meet the needs of independent researchers who want better access to data from social media companies to conduct research in the public interest, while at the same time balancing user privacy and other concerns.

In the last year, members of the US Congress have introduced or published at least four bills or discussion drafts with provisions about researcher access to data held by online services: The Platform Accountability and Transparency Act, Social Media Data Act, Digital Services Oversight and Safety Act, and Kids Online Safety Act.

In Europe, Article 31 of the Digital Services Act is poised to become the first major legislation requiring some online services to make certain data available to researchers. The European Council, Commission, and Parliament have each adopted positions on the DSA and are now engaged in the "trilogues," through which they will negotiate a joint position, including on researcher access to data. *(*Because the trilogue process and draft joint positions of the European Council, Commission, and Parliament are not open to the public, the chart below summarizes the European Parliament's position on Article 31. Many parts of Article 31—such as the specific criteria for the vetting of researchers—are being discussed during the trilogues and may differ in the final version of the DSA.*)*

CDT has compiled a chart (last updated on April 21, 2022) comparing how these different researcher access to data proposals answer seven key questions.

CENTER FOR DEMOCRACY & TECHNOLOGY

Table of Contents

I. Who would have access to data?

Platform Accountability and Transparency Act	"Qualified researchers" = "a **university-affiliated researcher** specifically identified in a research proposal that is approved by the NSF to conduct research as a qualified research project" (Sec. 2) **Public access to some data:** Gives the FTC rulemaking authority to require covered platforms to report certain other data or information to **the public,** qualified researchers, or some combination of the two (Sec. 12(a)) and requires the FTC to issue rules requiring platforms to make **public reports** about **content that has been highly disseminated** (Sec. 12(b)), **advertising** (Sec. 12(c)), **algorithms** (Sec. 12(d)), and **content moderation** (Sec. 12(e)).
Social Media Data Act	**Academic researchers** and the **FTC**. (Sec. 2(a)(1)) Academic researcher = an individual that conducts research in collaboration with an **institution of higher education** (as defined in section 6 101(a) of the Higher Education Act of 1965) and research is not for commercial purposes. (FTC may update definition as needed) (Sec. 2(d)(1))
Digital Services Oversight and Safety Act	**Researchers affiliated with an institution of higher education** or **nonprofit** whose mission includes developing a deeper understanding of the impacts of platforms on society. Both organizations and researchers must be certified by the Office of Independent Research Facilitation to be established at the FTC. "**Host**

	organizations" must meet requirements TBD by the FTC and commit to training researchers, reviewing research projects, and other commitments. "**Certified researchers**" must meet requirements established by the FTC and make commitments, such as compliance with information or security requirements established by the FTC, agreeing not to attempt to reidentify data, agreeing to publish their research, and more. (Sec. 10(b)) **Public access to some data**: Requires FTC to issue regulations requiring a provider of a hosting service to issue **publicly available transparency reports** relating to content moderation. (Sec. 6(b)) Requires FTC to issue regulations requiring providers of a large covered platform to maintain a public version of an **advertising library** (Sec. 10(f), 10(f)(3)) and a public version of a **high-reach public content stream** (Sec. 10(g), 10(g)(4)).
Kids Online Safety Act	"Qualified researchers" = (1) Affiliated with an **institution of higher education** or a **nonprofit organization**, including any 501(c); *and* (2) **Approved** by Assistant Secretary of Commerce for Communications and Information (NTIA) To gain approval, a researcher must: Conduct the research for **noncommercial purposes**; Demonstrate a **proven record of expertise** on the research topic and related research methodologies; and Commit to fulfill, and demonstrate a capacity to fulfill, specific **data security and confidentiality requirements** corresponding to the application. (Sec. 7(a)(2), 7(a)(5), (b))
DSA Art. 31	Vetted **researchers who are affiliated with academic institutions and vetted not-for-profit bodies, organisations or associations** representing the public interest. Vetted researchers and not-for-profits must: Be **independent from commercial interests** **Disclose the funding** financing the research Be **independent from government/state bodies** (except for public academic institutions) Have **proven records of expertise** in the fields related to the risks investigated or related research methodologies Preserve **data security and confidentiality** requirements. (Art. 31 para. 2 & 4)

CENTER FOR DEMOCRACY & TECHNOLOGY

II. What types of data would be accessible to "researchers," specifically?

Platform Accountability and Transparency Act	"Qualified data and information" = "data and information from a platform that the NSF determines is necessary to allow a qualified researcher to carry out the research contemplated under a **qualified research project**." (Sec. 2). The criteria for "qualified data and information" is **TBD** by the NSF, but it must at least be (1) **feasible** for the platform to provide; (2) **proportionate** to the needs of the qualified researchers to complete the qualified research project; and (3) not cause the platform **undue burden**. (Sec. 4). Qualified data and information **could include non-public content data and personally identifiable information**.
Social Media Data Act	**Ad library** with certain **specified information about any advertiser** that purchases $500 or more of advertising in a calendar year: name and unique identification number of the advertiser, digital copy of the ad, targeting method & description of target audience, optimization objective chosen by advertiser, description of the actual audience, number of views, ad conversion, date and time of ad display, amount advertiser budgeted and paid, ad category (such as politics, employment opportunity, housing opportunity, or apparel), ad language, and platform's advertising policy. (Sec. 2(a)(1)) Would also establish a **Working Group for Social Media Research Access at the FTC** to study making other data and information available to academic researchers (Sec. 2(c))
Digital Services Oversight and Safety Act	The FTC must issue regs identifying the precise types of information that will be available. The FTC regs can specify any relevant information, but it **must consider** particular information: info about **internal platform studies**; info about **content moderation decisions and policies,** the people setting the policies and making decisions, & the training of moderators; **third party requests** to act on a user, account, or content; **engagement and exposure data**; classification of **information sources**; **archives of removed content and accounts**; **Advertisements** and influencer marketing content; detailed information about a platform's **algorithms**. (Sec. 10(c)). The information required to be disclosed by FTC regulations **could**

	include non-public content data and personally identifiable information, but the FTC must require platforms to **deidentify certain data** (non-public data, personal health information, biometric information, and information related to a person under 13 years old), before it may be disclosed and also restricts sharing of precise location information. (Sec. 10(c)(6)) In addition, the FTC must issue regulations requiring covered platforms to submit a **data dictionary** describing the information that can be provided to certified researchers. (Sec. 10(d)) The FTC must also issue regulations requiring large covered platforms to give researchers and the FTC access to an **ad library** (Sec. 10(f)) and a **"high-reach public content stream."** (Sec. 10(g)).
Kids Online Safety Act	Data assets that can be used to conduct **public interest research regarding harms to the safety and well being of minors**, including the following types of matters: (1) promotion of self-harm, suicide, eating disorders, substance abuse, and other matters that pose a risk to physical and mental health of a minor; (2) patterns of use that indicate or encourage addiction-like behaviors; (3) physical harm, online bullying, and harassment of a minor; (4) sexual exploitation, including enticement, grooming, sex trafficking, and sexual abuse of minors and trafficking of online child sexual abuse material; (5) promotion and marketing of products or services that are unlawful for minors, such as illegal drugs, tobacco, gambling, or alcohol; and (6) predatory, unfair, or deceptive marketing practices. (Sec. 3(b), 7(b)(1)) The term "data assets" is not defined in the statute, and could include **non-public content data and personally identifiable information.**
DSA Art. 31	Any data that serves the **permissible purposes of research** specified in Art. 31 para. 2 [See Section III], **unless** the 'very large online platform' **(VLOP) does not have access to the data** or giving access would lead to **significant security vulnerabilities** or reveal **confidential information**. (Art. 31 para. 6) Also grants access to "**aggregate numbers** for the total views and view rate of content **prior to a removal** on the basis of" orders for removal of illegal content under Art. 8 or content moderation under a provider's own TOS. (Art. 31, para. 2a)

CENTER FOR DEMOCRACY & TECHNOLOGY

III. Are there restrictions on the purpose of the research or research project?

Platform Accountability and Transparency Act	Only "qualified research projects" approved by NSF. A qualified research project must (1) have **IRB approval** or be exempt or excluded from IRB approval; (2) "**aim to study activity on a platform**"; (3) meet other criteria **TBD** by the NSF. (Sec. 4)
Social Media Data Act	**None**.
Digital Services Oversight and Safety Act	Researchers may be certified to gain access to information only for the purposes specified in the Act: "to gain understanding and measure the **impacts of the content moderation, product design decisions, and algorithms** of covered platforms **on society, politics, the spread of hate, harassment, and extremism, security, privacy, and physical and mental health**." (Sec. 10(b)(1) & (2))
Kids Online Safety Act	Researchers may access data only to conduct **public interest research** pertaining to **harm to the safety and wellbeing of minors.** Public interest research = scientific or historical analysis of information that is performed for the primary purpose of advancing a broadly recognized public interest. (Sec. 7(a)(4), (b)(1))
DSA Art. 31	Data may be used only for research that contributes to the **identification, mitigation and understanding of specified systemic risks** set out in Art. 26(1) and Art. 27(1). In addition, the **Commission must adopt delegated acts** laying down, among other things, the purposes for which the data may be used. (Art. 31 para. 5)

CENTER FOR DEMOCRACY & TECHNOLOGY

IV. Which online services must make data available?

Platform Accountability and Transparency Act	"Platforms" = Subject to FTC jurisdiction under section 5(a)(2) of FTC Act; and Is a website, desktop application, or mobile application that **allows users to establish accounts to share user-generated content** and whose primary purpose is for users to interact with user-generated content and for the **platform to deliver ads to users**; and has at least **25 million unique monthly users** in the United States for a majority of the months in the most recent 12-month period. (Sec. 2)
Social Media Data Act	"Covered platform" = any website, desktop application, or mobile application that is **consumer-facing**; and **sells digital advertising space**; and has more than **100 million monthly active users** for a majority of months during the preceding 12 months. **FTC can update definition** as needed. (Sec. 2(d)(3))
Digital Services Oversight and Safety Act	"Covered platform" = A hosting service that **stores information provided by, and at the request of, users** and which, **at the request of users, stores and disseminates information to the public**; and has at least **10 million monthly active users**. The methodology for determining MAU will be determined through rulemaking. (Sec. 2(11); Sec. 10(c)) In issuing the regulations about the types of information that must be disclosed, the manner of disclosure, and whether disclosure is mandatory or optional, the FTC must "**vary the specifications based on the size and scope of a covered platform**, including by having different specifications for different services.

CENTER FOR DEMOCRACY & TECHNOLOGY

Kids Online Safety Act	"Covered platforms" = a commercial software application or electronic service that connects to the internet and that **is used, or is reasonably likely to be used, by a minor**. (Sec. 2(2), Sec. 7(b)(3))
DSA Art. 31	Very Large Online Platforms (VLOP) = **average monthly active recipients of the service** in the EU equal to or higher than **45 million** for at least **four consecutive months**. Number of average monthly active recipients **can be adjusted** based on changes to the EU population. (Art. 25) "Online platforms" = a provider of a hosting service which, at the request of a recipient of the service, **stores and disseminates to the public information**, unless that activity is a minor or a purely ancillary feature of another service or functionality of the principal and cannot be used without that other service, and the integration of the feature or functionality into the other service is not a means to circumvent the applicability of the DSA. (Art. 2)

CENTER FOR DEMOCRACY & TECHNOLOGY

V. What privacy and security safeguards would there be for data made available to researchers?

Platform Accountability and Transparency Act	Newly-established **FTC Platform Accountability and Transparency Office** (Sec. 3) would establish criteria for privacy and cybersecurity safeguards required for qualified data and information related to a qualified research project, and **can require** reasonable privacy and cybersecurity safeguards for particular data sharing, such as **encryption of data**; delivery of **deidentified data**; use and monitoring of a **secure environment** to facilitate delivery of data. (Sec. 4(j))
Social Media Data Act	**None**. The **Working Group for Social Media Research Access** would study privacy preserving techniques for **other data** that could be made accessible to academic researchers. (Sec. 2(c))
Digital Services Oversight and Safety Act	**Tiered access:** More sensitive info has more safeguards and is accessed by fewer researchers than less sensitive info. (Sec. 10(c)(2)) The FTC must issue regulations specifying the manner in which information is to be accessed, including **when privacy protecting techniques** "such as differential privacy and statistical noise" should be used, **what information security standards should be in place**, and other privacy and security measures. (Sec. 10(c)(4)) The FTC must issue regulations specifying when the Commission should **review research before it is published to protect user privacy** or trade secrets. (Sec. 10(c)(5)) FTC regulations must ensure that provision of access to information does not infringe upon reasonable expectations of personal privacy and must **require platforms to deidentify certain information before it can be provided:** nonpublic information, personal health information, biometric information, information about a child under 13 years old. Also restricts sharing of precise location information. (Sec. 10(c)(6)).

	Users who do not post public content must be given the ability to **opt-out** of having their information shared with researchers. (Sec. 10(c)(6)(C))
Kids Online Safety Act	The NTIA must **establish standards for privacy, security, and confidentiality** required to participate in the program for a researcher to receive, and a covered platform to provide, data assets. (Sec. 7(b)(4)(C)) Imposes a **duty of confidentiality** on a qualified researcher with respect to data assets provided by a covered platform. The duty of confidentiality may be defined further by the NTIA. (Sec. 7(b)(5))
DSA Art. 31	**TBD**: The Commission must adopt delegated acts laying down, among other things, "the specific conditions under which such sharing of data with vetted researchers or not-for-profit bodies, organisations or associations can take place **in compliance with [the GDPR]** taking into account the rights and interests of the very large online platforms and the recipients of the service concerned, including the **protection of confidential information**, in particular **trade secrets**, and maintaining the **security of their service**." (Art. 31 para. 5.)

CENTER FOR DEMOCRACY & TECHNOLOGY

VI. What would be the method or mechanism for vetting researchers and providing access?

Platform Accountability and Transparency Act	The NSF **vets the researcher and research project** and **determines what data** a platform must make available; the Platform Accountability and Transparency Office informs the platform and **establishes the privacy and cybersecurity safeguards** for the particular data at issue. Step 1: A researcher submits a research application to the NSF Step 2: The NSF determines if it is a "qualified research project" by a "qualified researcher." Step 3: The NSF identifies the "qualified data and information" that platforms will be required to make available to the researcher, and in what form. Step 4: The NSF refers the qualified research project to the FTC Platform Accountability and Transparency Office Step 5: The Office notifies the platform that it will be required to provide data and establishes reasonable privacy and cybersecurity safeguards for the data. Step 6: The platform can comment on the privacy and cybersecurity safeguards; following the platform's comments, the Office makes a final determination re: the safeguards. (Sec. 4).
Social Media Data Act	A covered platform must maintain, and **grant academic researchers and the Commission access** to, an **ad library** that contains in a **searchable, machine readable** format. (Sec. 2(a)(1))
Digital Services Oversight and Safety Act	The FTC establishes a "**research certification process**" under which an organization can apply and be qualified as a host organization and an individual associated with a host organization can apply and be certified as a certified researcher. (Sec. 10(b)) The FTC issues **regulations specifying the manner in which researchers will access information** from covered platforms. (Sec. 10(c)(1) & 10(c)(4)) The FTC must consider, among other things, **size and sampling techniques** used to create data sets, under what circumstances **APIs** are required, and designate "**secure facilities and computers** to analyze information through a Federally Funded Research and Development Center" established by the Act. (Sec. 10(c)(4).

Kids Online Safety Act	The NTIA must establish a program under which a researcher can apply for access to data and the NTIA can approve their application. (Sec. 7(b)(1)-(4)) For applications that are approved, a covered platform must provide to a qualified researcher access to data assets **through online databases, application programming interfaces, and data files** as appropriate for the qualified researcher to undertake public interest research. (Sec. 7(b)(3)(A)(ii))
DSA Art. 31	Researchers would be vetted by the Digital Services Coordinator of establishment or the Commission. (Art. 31, para. 4) Access to data would be provided through **online databases** or **application programming interfaces**, as appropriate, **and with an easily accessible and user-friendly mechanism to search for multiple criteria.** (Art. 31 para. 3) **More details TBD**: The Commission must adopt delegated acts laying down, among other things, "the technical conditions under which [very large online platforms] are to share data . . . ." (Art. 31 para. 5)

CENTER FOR DEMOCRACY & TECHNOLOGY

VII. Is there a safe harbor for independent methods of data access?

Platform Accountability and Transparency Act	**Yes**. No civil or criminal liability for **any person** for **collecting covered information** as part of a **newsgathering or research project** on a platform. (Sec. 11). Conditions: Only applies to "**covered methods of digital investigation**"; purpose must be to **inform the general public about matters of public concern**, and the information in fact is used only that way; the person takes **reasonable measures to protect the privacy** of the platform's users; w/r/t research accounts, the person takes reasonable measures to **avoid misleading users**; and the project **does not materially burden the technical operation of the platform.** (Sec. 11). **"Covered method of digital investigation"** = **TBD** by FTC regulations, but must include collection of data through automated means, through data donation, or through research accounts. (Sec. 11). **"Covered information"** = **publicly available** information, information about **ads**, **other information TBD** by FTC that **does not unduly burden user privacy.** (Sec. 11).
Social Media Data Act	**No**.
Digital Services Oversight and Safety Act	**Yes. Certified researchers** granted immunity for liability under **state, federal, and local law** for **violating platform's TOS** for two specified research activities: creating a research account (if researcher takes reasonable means to avoid misleading users and does not burden technical operation of platform) and data donation with informed consent of users. (Sec. 10(c)(10)) Also **prohibits a covered platform from discriminating against a certified researcher** in the provision of services because of those two research activities. (Sec. 10(c)(10)).
Kids Online Safety Act	**Yes.** No cause of action for **violating platform's TOS** may be brought based on actions a researcher takes while collecting data assets as part of public interest research regarding harms to minors. (Sec. 7(c))
DSA Art. 31	**No**.

CENTER FOR DEMOCRACY & TECHNOLOGY

European Policy

CDT Europe Joins Coalition Letter Calling for the DSA to Be Used to Shed Meaningful Light on Platforms' Impact on Our Public Sphere

December 1, 2021

Asha Allen, Ophélie Stockhem

CDT Europe has co-signed an open letter initiated by AlgorithmWatch and Local Witness, along with civil society organisations, international academics, researchers, and independent think tanks. The statement calls on all members of the Internal Market and Consumer Protection Committee (IMCO) of the European Parliament to ensure the Digital Services Act empowers a broad base of vetted public interest researchers' access to data; researchers whose independent scrutiny can be vital to holding large tech platforms accountable, and whose analysis can prove useful in assessing trends and improving transparency.

The statement highlights shortcomings in scrutiny measures directed at big online platforms, which would severely undermine researchers' ability to assess the risks that platforms may pose to our public sphere, in particular the tabled amendments made:

- to restrict data access and scrutiny solely to those researchers affiliated with academic institutions in Art. 31(4) of the draft DSA; and
- to allow for broad exemptions which would allow platforms to deny data access based on protection of "trade secrets" in Art. 31(6)b.

The letter calls on lawmakers to widen data access in the DSA to vetted public interest civil society organisations and to remove the trade secrets exemption on the basis of which Very Large Online Platforms (VLOPs) could deny requests for data access. If both are fulfilled, these demands would considerably increase the EU's ability to hold VLOPs to account. While it is appropriate, in CDT's view, for the Commission to consider the importance of protecting genuine trade secrets in the delegated acts and guidance it would develop to implement Article 31, it is vital that research not be stymied by mere assertion of trade secret concerns by individual companies.

An extract of the letter can be found below. For the full letter + list of signatories, read more here.

***

Data access and scrutiny by third-party vetted researchers, via Article 31, goes to the heart of the DSA's oversight structure. That is why we strongly support IMCO's amendment to Art. 31(4), which extends data access to

civil society organisations with proven expertise, representing the public interest and following strict privacy guidelines. Preserving this amendment is vital to expanding the network of experts and watchdogs working to help ensure systemic risks are identified, understood, and acted on, even as the risks are constantly evolving.

At the same time, the proposal for platforms to be able to deny access to their data for independent scrutiny based on protection of "trade secrets" risks making Article 31 entirely meaningless. As there is no definition of what constitutes a trade secret, it would give huge discretionary power to platforms to block any public interest research – particularly where it raises issues that are uncomfortable for the platform – with little opportunity for recourse or challenge.

As the European Parliament nears agreement on its position for the DSA, we strongly urge you to support widening data access and scrutiny of Very Large Online Platforms to vetted public interest civil society organisations and journalists in Art. 31(4) and removing the trade secrets exemption in Art. 31(6)b. Both these demands would considerably increase the EU's ability to hold Very Large Online Platforms to account, and they must not be traded against each other.

Read the full letter + the list of signatories here.