CENTER FOR
DEMOCRACY
& TECHNOLOGY

Executive Summary

From CDT's *Improving Researcher Access to Digital Data: A Workshop Report*

In March 2022, the Center for Democracy & Technology convened 29 researchers from academia, civil society, and journalism in a workshop. The workshop was designed to explore specific questions around researchers' access to data held by social media companies and other hosts of user-generated content. 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?

The report describes certain key findings from the workshop:

- Researchers say they currently lack access to a variety of data that could be used for public interest research, including **more extensive advertising data**, data about hosts' use of **ranking and recommendation algorithms and content moderation algorithms, historical content data and deleted content data**, and **real-time data**. Researchers also 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.

- Most of the researchers at the workshop **opposed requiring a researcher to have an academic affiliation** to access 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 (such as academics, members of civil society, or journalists), **vetting should be done on a project-by-project basis**. Researchers also emphasized the **importance of improving access to public data**, which may require less vetting than access to non-public data.

- Researchers often rely on **Application Programming Interfaces** (APIs), through which certain hosts voluntarily provide data access for public interest research purposes, but they identified concerns about the accuracy and completeness of this data, the "streetlight effect" that use of APIs can cause, and the limits that hosts may place on API access. Researchers also rely on **independent methods** to gather data, such as scraping or data donation, but they are concerned about legal risks that can arise from the use of these methods and other actions hosts can take to thwart them. In some instances, they also use **commercial social media monitoring or marketing firms** to obtain data.

In this report, CDT makes seven recommendations to policymakers and hosts looking to enhance independent researchers' access to data held by content hosts:

1. **Policymakers and hosts should help researchers understand what potential data is available.** Publicly available "codebooks" that clearly define what data hosts possess, and may make available to researchers, will help overcome researchers' lack of knowledge about what data they may be able to access.

2. **Hosts should have established processes through which researchers can request access to non-public data or tools to make public data more accessible.** An established and transparent process is more equitable and efficient than ad hoc methods for determining researchers' data access requests.

3. **Hosts and policymakers should make accessing and using data for research in the public interest less expensive.** Hosts should make paid tools that they offer for accessing, using, processing, or analyzing data available for free or lower cost to researchers conducting research in the public interest. Lawmakers should allocate additional funding or other resources for non-government research using social media data and other data held by content hosts.

4. **Policymakers should prioritize improving researchers' access to public data through legislation.** Doing so would enable socially important research while avoiding some of the thornier questions around access to non-public data. However, policymakers should understand that enhancing access to public data is not entirely without risks or challenges. Access to public data could be enhanced by requiring some hosts to make tools for bulk disclosures of public data available, like APIs, and through safe harbors that protect independent methods of gathering data for noncommercial, public interest research.

5. **Policymakers should prioritize access to advertising data.** Policymakers 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 balancing risks to user privacy and businesses from requiring such disclosures.

6. **When vetting is necessary, hosts or legislation should evaluate specific research projects and plans based on established, transparent criteria. Hosts that voluntarily provide data or legislation that requires hosts to provide data should not rely simply on whether a researcher falls within a particular category,** such as academia. Requests for access should be evaluated on several criteria described in the report, such as whether the research plan demonstrates a valid research methodology, whether the researcher has the necessary qualifications to conduct the proposed research, and whether the research plan includes adequate privacy and security safeguards for using, transferring, and storing data and publishing results.

7. **Policymakers should strictly limit or eliminate the ability of commercial social media monitoring and marketing firms to collect and sell users' data,** even though some researchers rely on commercial sources to access data. Lawmakers should prohibit or significantly limit data broker collection and processing of information, and they should ensure that consumers are given sufficient information to understand what information data brokers have collected about them and their options for controlling that data.

This executive summary is of the August 2022 CDT report, *Improving Researcher Access to Digital Data: A Workshop Report*.

Additional CDT work on this topic: https://cdt.org/insights/improving-researcher-access-to-digital-data-a-workshop-report/

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

✉ cvogus@cdt.org

@CaitlinVogus

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.

@CenDemTech

CENTER FOR
DEMOCRACY
& TECHNOLOGY