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).

December 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