Commentary

A Multi-Stakeholder Approach for Leveraging Data Portability to Support Research on the Digital Information Environment

Zeve Sanderson and Lama Mohammed

Abstract. In this paper, we aim to situate data portability within the evolving discussions of how to support data access for researchers studying the digital information environment. We explore how data donations, enabled by existing data access rights and data portability requirements, provide promising opportunities for supporting research on critical trust and safety topics. Evaluating other data access mechanisms that are more central to policy debates about platform transparency, we argue that data donations are a powerful additional mechanism that offer key legal, ethical, and scientific benefits. We then assess current challenges with using data donations for research and offer recommendations for various stakeholders to better align portability mechanisms with the needs of research. Taken together, we argue that although portability is often considered within a context of competition and user agency, regulators, industry actors, and researchers should understand and leverage portability's potential impact to empower critical research on the societal impacts of digital platforms and services.

1 Introduction

A key concern for policymakers, journalists, civil society organizations, and academics alike is understanding the myriad impacts of digital platforms and services, which have come to play a central role in social interactions, economic activities, and the dissemination of information.[1] However, a recurring challenge has been that the digital trace data[2] necessary to produce rigorous evidence on platform effects are stored in

1. This paper is an expanded version of a chapter included in the compendium for a policy workshop, hosted by the Data Transfer Initative and held in Washington, DC, in February 2024. The compendium can be found here: https://dtinit.org/assets/DTI-Data-Portability-Compendium.pdf.

2. Howison, Wiggins, and Crowston (2011) define digital trace data as "records of activity (trace data) undertaken through an online information system (thus, digital)."

doi:10.54501/jots.v2i4.215

proprietary databases, often accessible only to the companies themselves and used for commercial purposes (Persily and Tucker 2020; Lazer et al. 2020). This dynamic enables platforms to act as gatekeepers for both academic research agendas and evidence-based policy evaluations, leaving key questions of societal import unanswered and unanswerable given a lack of accessible data (Ausloos and Veale 2020; Vreese and Tromble 2023). Alarmingly, several platforms—such as Meta (Nguyen 2024), Twitter (now X) (Kharpal 2023), and Reddit (Gallagher 2023)—have shut down public application programming interfaces (APIs) in recent years, erecting further barriers for independent researchers to collect requisite data.

Policymakers have made data access a central concern for efforts to increase platform transparency, oversight, and accountability. In the European context, the Digital Services Act (DSA), which is primarily concerned with platform transparency and user protection, includes provisions to grant access to data from very large online platforms (VLOPs) and very large search engines (VLOSEs)[3] to vetted researchers (European Commission 2023). In the United States, the Platform Accountability and Transparency Act (PATA), which was re-introduced in the Senate last year (Coons 2023), includes similar mechanisms for requiring independent data access. While promising, these approaches to data access have key limitations, most notably their narrow application to VLOPs and VLOSEs. This limitation is especially important given recent developments in the digital information environment, such as the rise of generative artificial intelligence (Gen AI) applications (Nicholas 2024) and smaller platforms (Ortiz-Ospina 2019) that do not reach DSA or PATA usage thresholds but have potential social, political, or economic significance.

Researchers have developed a range of direct mechanisms for collecting data, such as webscraping and web tracking (Ohme et al. 2023). A key challenge for collecting data without user or platform consent is that it introduces potential legal risks for researchers and ethical risks for users (Fiesler, Beard, and Keegan 2020). Given these dynamics, one promising approach is data donations, in which users consent to donate digital trace data for research (Meyer et al. 2023). In addition to establishing user consent, data donations generally fall within legal data access rights, such as those established in the European Union General Data Protection Regulation (GDPR) (Wolford 2018), and thus provide greater legal protections for researchers. However, the right to data portability, or the right for users to transfer their data from one digital service to themselves and/or to other digital services, has largely been considered through the lens of enhancing user agency and promoting competition in the digital marketplace (Castro 2021; Gulati-Gilbert and Seamans 2023). For user agency, the ability to port one's data theoretically empowers users to make decisions with and about their data (Zweifel-Keegan 2024). For competition, data portability theoretically lowers the friction for users to switch between

3. VLOPs and VLOSEs are defined as online platforms and search engines that reach an average of more than 45 million monthly users.

platforms (Sharma 2024).[4] However, the ability for portability to support critical research has been nearly entirely left out of the policy conversations around either data portability or platform transparency and accountability.

This dynamic has led to a mismatch between data portability as a mechanism to promote competition and user agency in the digital marketplace and a mechanism to collect user data to facilitate research on the digital information environment. On the one hand, policymakers and platforms have approached the design, implementation, and evaluation of data portability through the lens of competition and user agency. On the other, researchers have leveraged data portability provisions for research, but often with challenges due to the misalignment between these various goals.

In this paper, we aim to situate data portability within the evolving discussions of how to support data access for researchers engaging in critical work on the digital information environment. This research is especially urgent in the context of dwindling platform trust and safety teams who would have carried out internal research on platform effects (Duffy 2023), albeit not always made available in public fora. More specifically, we explore how, given changes in the digital information environment, data donations enabled by portability requirements provide promising opportunities for facilitating research that is aligned with ethical and legal frameworks, illustrating how data donations support areas of inquiry central to trust and safety teams and regulators. We then discuss current challenges for using data donations for research, using TikTok as a case study, and provide recommendations for policymakers, companies, and researchers to align portability mechanisms with researcher needs. Taken together, we argue that, although portability is often considered within a context of competition and user agency, various stakeholders should work together to understand and leverage portability's potential impact to empower critical research on the societal consequences of digital platforms and services.

2 The Current State of Action on Expanding Mechanisms for Data Access

Researchers interested in studying the online information ecosystem can access social media data through mechanisms established by legislative or regulatory actions (e.g., the DSA), through mechanisms established by companies (e.g., platform APIs), or through independent methods (e.g., webscraping) (Persily and Tucker 2020). In this section, we provide an overview of the current landscape of how regulators, companies, and researchers have been working to increase data access. The goal is not to provide an exhaustive summary of each area—a topic that could easily be its own paper—but rather

4. The extent to which data portability can realize these goals has been constrained by various factors (Reimsbach-Kounatze and Molnar 2024), and significant work is still necessary to ensure that users are able to fully realize the theoretical goals of data portability *in practice* (Turner and Tanczer 2024).

to highlight shortcomings in each of the primary approaches for making multi-platform data accessible to researchers in a manner that is robust, resilient, and protected. In this context, we argue that data donations serve as an additional compelling approach to the data access landscape. To be clear, there are trade-offs to every data access mechanism, and progress should be made in each of the areas in parallel. However, given the speed of technological developments in the short-term and the lack of focus on data donations in the current policy discussions around data access, we highlight the benefits of data donations as an area for greater investment.

2.1 Data Access through Policy

Legislative and regulatory actions are a powerful mechanism for mandating researchers' access to platform data, as evidenced by the DSA. Through Article 40 of the DSA (Joint Research Centre 2023), researchers can access data from VLOPs and VLOSEs. Although the law demonstrates progress in legislation's ability to mandate researcher's access to digital data, researchers still experience challenges accessing social media data through the mechanisms established under Article 40 (Carvalho 2024), with many reporting rejected requests for data.[5] Additionally, researchers can only access data if they execute projects on detecting, identifying, and understanding systemic risks in the EU (Engler 2021). Therefore, access to these data is restricted to vetted researchers who meet stringent criteria (Albert 2022). These restrictions, coupled with the law's mandate that researchers focus on large platforms and Europe-centric studies, reduces the diversity of research projects the law can facilitate and thus limits our understanding of the complex and evolving digital information environment. PATA, if passed, would likely run into similar issues: The processes for vetting researchers and research projects would be resource intensive, the jurisdiction over whose data is made available would be limited, and the platforms and services covered by the law would be constrained to the largest companies.

2.2 Data Access through Platform APIs

Documented APIs made available by platforms have been a key data access pathway for supporting research on the digital information environment. This mechanism is effective at facilitating research because it provides access to real-time, structured platform data. However, APIs can change at any moment, at the whims of the platforms, and often with little notice—leaving research projects vulnerable to corporate policies. For example, while X once provided free or heavily discounted API access to researchers with university affiliations, the company introduced substantial price increases after Elon Musk's acquisition, effectively blocking researcher access to platform data through its documented APIs (Stokel-Walker 2023; Calma 2023) and impacting hundreds of projects on critical topics.[6] Similarly, Meta's CrowdTangle, which allowed researchers

5. For more information, see https://www.soscisurvey.de/DSA40applications/.
6. See results from a survey capturing the impacts of Twitter's API change: https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/.

to track data from public pages and groups, shut down in August 2024 (Data for Good 2021; Nguyen 2024), limiting the ability of researchers in academic and other contexts to track information across Meta's platforms in the months leading up to the 2024 US elections (CITR 2024).

2.3 Data Access through Independent Methods

Given the lack of reliable methods for data collection directly from platforms, researchers have turned to independent data collection methods, such as webscraping and undocumented APIs, to circumvent platform control over researchers' access to social media data.

Webscraping—the automated collection of data rendered on a webpage or application—has been a key focus of data access advocates, as this method enables researchers to collect large datasets without government regulation or platform involvement. However, this technique may pose legal and ethical challenges for researchers (University of Pennsylvania Social Behavioral Research 2023). For example, X filed a lawsuit against the Center for Countering Digital Hate for scraping data for research purposes (CCDH 2024). In the lawsuit, X cited privacy concerns and negative impacts on its advertising business. The X example also highlights that when platforms implement legal measures to control researcher data access (Windwehr and Selinger 2024), researchers risk violating a platform's Terms of Service or the Computer Fraud and Abuse Act. Although recent court rulings have largely supported the rights of researchers to scrape data—especially data in the public domain—the specter of litigation may nonetheless chill critical research projects.

Additionally, undocumented APIs are hidden objects on a platform that researchers can leverage for data collection. For example, researchers can use these APIs to explore whether a platform has blocked a specific term from appearing in search results (Yin and Sankin 2021). While undocumented APIs can provide access to platform data, they are not supported by platforms for public use and thus lack official documentation for external users, making the APIs challenging to access and the data difficult to understand. Like documented APIs, they are also subject to unannounced changes.

2.4 Data Access through Data Donations

In many countries, including those covered by the GDPR, online users have rights to data portability, in which they are able to receive personal data from a digital platform or service in a "structured, commonly used, and machine readable format."[7] Researchers have leveraged rights to data portability for research purposes through data donations: data subjects are able to download their data from platforms and consent to donate their digital trace data to researchers.

7. See Article 20 of the GDPR: https://gdpr-info.eu/art-20-gdpr/.

While used in a number of research studies, this mechanism for accessing platform data exists in a liminal policy space: policy conversations around data portability focus on user agency and competition (Castro 2021; Gulati-Gilbert and Seamans 2023), while policy conversations around data access rarely include portability as a key area of inquiry and investment. However, data donations offer critical benefits relative to the three mechanisms described above: regulations that establish data access rights and mandate data portability have been already passed and implemented; the data collection mechanism is less vulnerable to platform policy changes, as the right to data portability (unlike API access) is mandated by legal requirements; and data donations leverage user data access rights, providing greater legal protection for researchers and privacy protections for users.

The above demonstrates that while each data access pathway has trade-offs, data donations offer key policy benefits given the current accessibility of data portability, resilience to corporate policy, and protection against legal action. Interestingly, the supply of portability by policymakers and companies has exceeded user demand for it (Riley 2024); in other words, while portability requirements have been been enshrined into law and portability systems have been developed by platforms, user demand to port their data has not increased to match this supply. Data donations for research could be one avenue for increasing aggregate demand for portability.

Building on data donation's advantages as a mechanism for data access, in the next section we discuss the scientific benefits of leveraging data donations to study the digital information environment.

3 Beyond the Streetlight: Data Donations in a Multi-Platform Digital Information Environment

A challenge for researchers studying the digital information environment is that research agendas have been, to a certain extent, shaped by the data made available to them (Matamoros-Fernández and Farkas 2021). The clearest impact of data availability is the amount of research undertaken on Twitter (before it was rebranded as X and API access was removed): Twitter is overrepresented in research not because it is seen by scholars as the most important platform for political or social outcomes, but because its once easily accessible API enabled the collection of granular, dynamic, and networked datasets that could support a wide range of research projects (Persily and Tucker 2020). For example, a stark illustration of the agenda-setting power of Twitter's API is that the number of studies on Twitter in communications journals surpasses studies on YouTube (Lukito et al. 2023), even though YouTube has remained the most popular social media platform among US adults for multiple years (Auxier and Anderson 2021). The dynamic of data availability shaping research agendas—colloquially referred to as the streetlight effect—has led to significant blind spots in our understanding of the digital information

environment (Moritz 2016). Given differences across user bases, platform affordances, algorithmic structures, and content moderation policies, insights on one platform do not necessarily generalize to others (Chen and Peng 2023; Shahbaznezhad, Dolan, and Rashidirad 2021). As a result, key stakeholders—ranging from policymakers to civil society groups—lack comprehensive evidence with which to evaluate critical topics in the multi-platform digital environment (Persily and Tucker 2020), ranging from electoral integrity (Faust and Arnaudo 2024) to mental health (Hendrix 2023).

In recent years, scholars have engaged in several data collection strategies to facilitate a broader research agenda on digital platforms. Borrowing from Ohme et al. (2023), there are two approaches to collecting platform data. In a *platform-centric approach*, data is collected directly from platforms without the involvement of users. Examples of this approach include the use of APIs, both documented and undocumented (Yin 2023), and webscraping. Within a platform-centric approach, there are a number of specific data collection strategies, each of which comes with its own trade-offs. APIs, while often providing access to large structured data collections, are subject to deprecation by platforms (Freelon 2018; Vreese and Tromble 2023; Bruns 2019) and have potential biases (Ruths and Pfeffer 2014; Allen et al. 2021). Webscraping can be a powerful tool for collecting large-scale data, but introduces significant legal and ethical risks (Fiesler, Beard, and Keegan 2020; Krotov, Johnson, and Silva 2020). Collaborations with platforms, though able to support ambitious projects for select researchers (Kupferschmidt 2023), have introduced issues of researcher independence (Wagner 2023) and accessibility (Walker, Mercea, and Bastos 2019). Notably, a platform-centric approach has largely dominated policy discussions around data access (Persily 2021), with legal mandates through the DSA structured around researchers' ability to request data directly from VLOPs and VLOSEs (Husovec 2023). But are there other mechanisms for policymakers to support independent researcher data access?

In a *user-centric approach*, researchers directly involve the user in data collection; two primary strategies in this approach are browser plug-ins (Haim and Nienierza 2019) and data donations (Prainsack 2019). While browser plug-ins (custom software that can capture data from a person's browser) can be a powerful tool for data collection, they are technically challenging to build and often tailored for the specific research project (Breuer et al. 2022). For example, two recent papers on Google Search,[8] both published in *Nature*, developed and used different browser plug-ins to collect search results (Robertson et al. 2023; Aslett et al. 2023). The use of browser extensions may also introduce legal risks for researchers, as was the case with the NYU Ad Observatory (Bobrowsky 2021), and potential privacy risks for users. However, a key reason that browser plug-ins or other tracking tools are not the focus of this analysis is that, as a policy area, they would require new regulations (and, arguably, new regulatory frameworks) if they were to be mandated by governments, rather than leveraging regulatory regimes that are already in existence (as in the European Union and South Korea) or under consideration (as

8. Sanderson is a co-author on Aslett et al. (2023)

in the United States and Saudi Arabia) for data portability (Derakhshani 2023).[9] While plug-ins collect data directly from a user's browser, data donations require that users can download their data from platforms. Data access rights through the GDPR grant users the ability to download data from the digital services and platforms they use, as well as mandate that platforms provide the ability to do so (Mondschein and Monda 2018; De Hert et al. 2018). In addition to transferring personal data to another online platform or service that someone might use, these data can be donated to researchers for secondary use. Indeed, data donations enabled by the GDPR's data access rights have already been used in several studies (Halavais 2019; Driel et al. 2022; Boeschoten et al. 2020).

To be clear, significant trade-offs are present with any approach to data collection based on the particular research question (Ohme et al. 2023; Pfiffner and Friemel 2023), and data donations are far from a panacea. However, given the platform-centric orientation of policy interventions that aim to increase data access, which were detailed in Section 2, it is important to note that data donations have a number of characteristics that make this strategy promising for both researchers studying the digital information environment and policymakers working on transparency efforts.

First, data donations allow participants to donate data from multiple platforms in the same study, enabling a richer and more comprehensive view of their online information diets. This capability is especially important given that people increasingly use multiple platforms, (Auxier and Anderson 2021; Krishnan 2023), particularly young people (Anderson, Faverio, and Gottfried 2023). It also allows donations from platforms that do not surpass the size threshold to be classified as VLOPs or VLOSEs under the DSA, but are nonetheless important for understanding social and political outcomes. These include alt platforms (e.g., Gab or Parler), local platforms (e.g., Nextdoor), video game platforms (e.g., Twitch), and private messaging apps (e.g., Telegram).[10] Relatedly, tracking tools are difficult, if not impossible, to deploy for all mobile contexts, leading extension-based studies to often be limited to collecting data from desktop usage. For example, Screenomics, a popular and sophisticated tool for recording data on a mobile screen every five seconds, is only available on Android, missing large portions of the US smartphone market that use iOS (Reeves et al. 2021). Data donations may be able to better capture usage across device types, as is the case with TikTok's data takeout process, in which usage across both mobile and desktop is included in a user's data download (covered in greater detail in Section 4).

9. One area for policy intervention could be protecting researchers who develop browser extensions. While protecting researchers is a critical area, there are no policy proposals, to our knowledge, that would actively promote or require the development of browser plug-ins, and it seems unlikely that this would become a focus for policymakers or regulators. One related area where government involvement could be useful is funding shared infrastructure and tooling, such as the recent NSF-funded National Internet Observatory (see https://nationalinternetobservatory.org/), but this falls outside of the scope of this paper.

10. Somewhat ironically, one of the reasons that data sharing mandates in the DSA and PATA are only applied to the largest online platforms is the potential anti-competitive effects of enacting onerous requirements on smaller platforms that may not have the resources for compliance (Keller 2022). However, portability, which is seen as competition-promoting, has the potential to enable research on these smaller platforms.

Second, while some research questions only require digital trace data *per se*, others require researchers to be able to collect both digital trace data and survey data to connect the online and offline—the relationship between online activity and offline behavioral or attitudinal measures (Salganik 2019). For example, a key area of interest for both scholars and policymakers is the impact of social media on mental health. To study this phenomenon, it is likely that researchers would need to both directly observe a user's social media behavior on multiple platforms and collect survey responses to measure shifts in mental health outcomes; it is also likely that researchers would need to use both of these methods longitudinally. Similar questions of societal import, such as how online (mis)information impacts support for democratic institutions, would also require the pairing of survey and digital trace data. Focusing on key topics of import to policymakers and trust and safety teams, Table 1 provides a comparison of the type of research projects that can be facilitated by platform-centric data (often collected via platform APIs) as compared to user-centric data (often donated directly from study participants) that can be paired with survey data. While not the only mechanism for user-centric data collection, data donations carry myriad benefits enumerated in this article and thus could serve as a critical approach for collecting digital trace data directly from study participants.

Third, there are a number of online harms that are not common and are not randomly distributed across the population, but instead occur unevenly in subpopulations. Ronald E. Roberston refers to this dynamic as "uncommon yet consequential online harms" (Robertson 2022). For content production and diffusion, a minority of users account for large shares of spreading so-called fake news (Guess, Nagler, and Tucker 2019) and producing hate speech (Zannettou et al. 2020); for consumption, misinformation (Grinberg et al. 2019) and radical content (Hosseinmardi et al. 2021) exposure is concentrated in small groups. Similarly, certain subpopulations may be targeted more by online harms, such as Spanish-language communities in the US (Sanchez and Bennett 2022). These patterns mean that large data collections through platforms may not capture the so-called "long tails" of distributions where specific harms are concentrated. Welles (2014) reminds us that "Big Data researchers must choose to examine very small subsets of otherwise large datasets." One way of doing so is recruiting study participants who are in the subpopulations of interest and collecting data donations directly from them, such as a recent bilingual panel of Latinos in the US that pairs survey data with digital data donations (Abrajano et al. 2022).

Finally, data donations include the explicit consent of users who donate data (Halavais 2019; Boeschoten et al. 2020; Driel et al. 2022). Many users see their own digital trace data as potentially sensitive (Hemphill, Schöpke-Gonzalez, and Panda 2022), are unaware of its use in research (Fiesler and Proferes 2018), and have different levels of comfort based on the goal of the study (Gilbert, Vitak, and Shilton 2021). Whereas the data made available through platform APIs and through the DSA may not involve the explicit and informed consent of users whose data are included, data donations directly involve the user and require informed consent (Crutzen, Ygram Peters, and Mondschein

2019).[11] Data donations also fall within the legal regimes that establish user data access rights (Boeschoten et al. 2020; De Hert et al. 2018), thus avoiding a number of legal risks for researchers that have accompanied methods like webscraping.

Table 1: Illustrative examples of the differences in the research questions that platform-centric data vs. user-centric data can support.

Topic	Data from Platforms	Data from Users
Mental health	Measure levels of self-harm content across a platform and changes in levels over time	Measure associations between exposure to self-harm content and mental health outcomes
Foreign influence campaigns	Identify accounts in coordinated foreign influence campaigns	Measure who is exposed to foreign influence campaigns and what association exposure has with beliefs and behaviors
Hate speech	Measure levels of hate speech and changes over time	Measure the individual-level characteristics of users targeted by / exposed to hate speech
Algorithmic recommendations	Describe the most frequently recommended content on average	Measure the association between recommendations and user demographics

4 Current Limitations of Portability for Data Donations

In Sections 2 and 3, we aim to establish the policy and scientific benefits of using data donations for the study of the digital information environment. However, key challenges limit researchers' ability to use data donations. There are three stages to a data donation study. The first is a consideration stage, in which potential participants are provided with information about the study—such as the research topic and details about participation—and decide whether they will participate. The second is the donation stage, in which consenting participants donate their data. And finally, the third is the analysis stage, in which researchers use donated data.

The first stage requires users to consent to participate, and previous work has measured the individual-level characteristics associated with willingness to participate in data donation studies (Pfiffner and Friemel 2023). While the ability for data donation is dependent on the right of access that regulations like the GDPR have established, the consideration stage is determined by an individual's willingness to donate data, and it is not clear how policymakers could (or should) influence an individual's willingness to participate in research. As a result, this stage does not directly involve new policy

11. To be clear, data donations may contain information from other users who did not provide consent, and so privacy and ethical considerations are still present. However, this data collection approach at least involves the informed consent of the person donating data, which is not involved in many other approaches.

questions; thus, we focus on the challenges that impact the next two stages. We both describe the challenges in general and use the process of requesting user data from TikTok as an illustrative example of these challenges (TikTok 2021). In doing so, we aim to illustrate how policymakers, regulators, and companies could better align data portability with the needs of researchers.

The donation stage requires a study participant to request a data download package (DDP) from a digital platform or service. This process requires navigating multiple screens, determining the type of data in the DDP, requesting the DDP, and downloading it to a local device. This multistep process involves a high level of digital literacy, potentially impacting the representativeness of the study sample. In addition, complex tasks in a research project may lead to attrition among those who expressed willingness to participate, and so require clear instructions and ongoing participant support (which still might not be enough to mitigate attrition) (Ohme et al. 2021; Breuer, Bishop, and Kinder-Kurlanda 2020). Another challenge is that users generally need to download DDPs directly to their devices before donating them to researchers. Depending on the size of the files, participants may need to have access to a desktop and high-speed internet. In turn, this may limit users who do not have access to a desktop or high-speed internet to donate their data, leading to within- and between-country variations in the ability to download DDPs. Taken together, these challenges contribute to sampling biases in the group of participants who are willing to donate data for research purposes (Keusch et al. 2024). While sampling biases in data donation studies may be impossible to fully overcome (Hase and Haim 2024), decreasing the technical burdens placed on participants may help increase sample representativeness and the quality of the data collected.[12]

As an example, take a hypothetical project in which researchers want to measure the political content algorithmically served to different demographic groups on TikTok. If a study participant consents to donate their TikTok data to research, the person can do so by sharing the data file from the mobile application or browser. On the mobile app (at the time of writing this article), users must proceed through five screens in their settings on the app. After requesting a data takeout, the DDP may be available immediately or it may take up to several days for the file to be made available in the data takeout portion of the app; in our testing, in cases where the process took several days, the file was made available without a push notification to alert the user of the availability of the DDP. Furthermore, once a takeout file is available, users have only four days to download their file before the link expires and the process must be restarted. In practice, this means

12. To be clear, all data collection strategies suffer from sampling biases. For example, the previously accessible Twitter API, which was the foundation of significant academic literature over the last decade, was limited to users whose accounts were public. In addition, the Twitter API provided data associated with published tweets without providing data about who was exposed to those tweets, limiting our understanding of the platform experience for roughly half of the platform users who were infrequent posters but frequent consumers (Odabas 2022). Sampling biases are also present in the Meta Content Library, which limits data on Facebook to public groups and pages, while providing no data from ordinary users. While far from a comprehensive accounting of biases in various data collection mechanisms, the purpose here is to illustrate that no data collection process is free from sampling biases and so researchers must consider them throughout the research process, from question formulation to analysis decisions.

that, in cases where the file takes multiple days to be made available, the user must remember to navigate through the five screens to determine its availability, and must do so in the four-day window before it expires. In the app, users also have no control over which types of data they want to download—the takeout file includes all data, including direct messages (DMs) and watch history. The lack of options at the request stage is problematic for users who may be concerned with privacy and for researchers who want to collect the minimum amount of data required for their project. Finally, the takeout file is delivered in a JSON or TXT format, and due to the complex file structure of a mobile operating system, users may find it challenging to locate the takeout file to share with researchers.

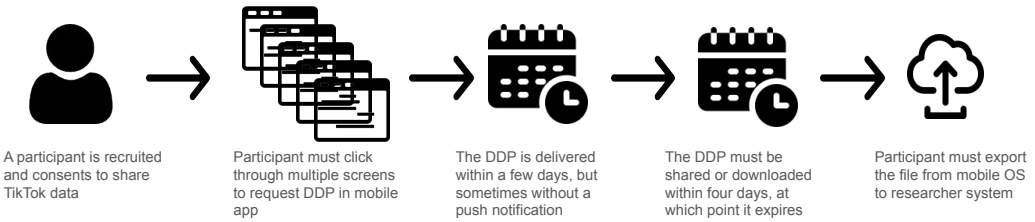

Figure 1: The process for consenting study participants to request, download, and upload their DDPs through the mobile application is challenging, potentially limiting its applicability to research.

The method is different if a user wishes to request their TikTok data file from the browser interface rather than the mobile application. This process is more direct, since users can go to their settings on the interface and click the "request data" option. Unlike in the mobile app, users can also select specific data types to download, such as just browsing history or direct messages. The timing challenges, such as the lack of push notifications and the limited download window, may still be present in the takeout process through the desktop. Once a user does receive a link and downloads the DDP, it goes directly to a user's desktop, allowing them to more easily share it with researchers' desktops. While this process is more streamlined and well-suited for research needs, users generally use TikTok on a mobile device and likely stay signed in on their device. Since TikTok is not typically accessed via desktop, users may not be signed in, requiring them to log in first before accessing their data. Additionally, the richer a users' TikTok digest is, the larger the file size, which may require users to have sufficient network bandwidth or storage capacity on their machine to open and share their data with researchers; this process is entirely inaccessible to users without access to a desktop, forcing these study participants to use the more onerous process through the mobile application.

If study participants are able and willing to undertake this process, researchers then need to implement a technically secure donation process. While some projects aim to support data donation, such as Port (Boeschoten et al. 2023) or the Data Donation Module,[13] researchers often need to create their own implementation of the donation process for the particular study (Ausloos and Veale 2020), limiting such study designs to scholars

13. See the Data Donation Module GitHub repository: https://github.com/uzh/ddm.

with the technical expertise and resources to do so.

The analysis stage requires that researchers have access to documented, structured data in machine-readable formats (Ohme et al. 2023). However, in the context of the TikTok takeout file, internal research teams have discovered that data from the payload is missing, without clarity as to when or why this issue may be occurring. Previous research using DDPs has also shown that data structures were unclear (e.g., posts showing up multiple times) and metadata categories were not well documented in DDPs, leading to confusion about how to transform data for analysis and measure key concepts (Driel et al. 2022). At best, these challenges require significant work from researchers to clean and transform data for analysis; at worst, these challenges make some data impossible to use for research given the lack of clarity.

5 Charting a Path Forward

While data donations, enabled by data access rights and supported by legal data portability requirements, offer benefits described in Sections 2 and 3, the current systems for donating data through DDPs is insufficient for the needs of researchers, limiting its broad utility for supporting critical research. In this section, we identify key opportunities for stakeholders across government, industry, and academia to improve data portability systems for research on the digital information environment.

5.1 The Role of Government

5.1.1 Engaging with Researchers

Regulators should engage directly with researchers to ensure that the pursuit of portability for competition enhances opportunities for using portability in research. By using competition metrics, regulators could incentivize platforms to design robust data donation portals if there are potential regulatory benefits or reliefs for supporting academic research. For example, the US government has the Research & Experimentation (R&E) Tax Credit to encourage companies to invest in research and development (R&D) activities (Tax Analysis 2016), helping to deduct the cost of research expenses from their profits before paying federal taxes (Americans for Tax Fairness 2023). To qualify for this credit, an online platform could establish an effective data donations portal as part of its R&D efforts, which could be modeled after initiatives like the European Digital Media Observatory—an independent intermediary body with experts across academia, industry, and civil society to support research on digital platforms.[14] Moreover, the tax credit could be especially beneficial for smaller platforms, which may be burdened by the resource requirements of building robust portability systems that are suitable for research.

14. For more information, see https://edmo.eu/about-us/edmoeu/.

5.1.2 Designing Portability for Research

While some regulations with portability requirements have already come into effect (e.g., the GDPR and the Digital Markets Act (European Commission 2022)), others are still being considered. During the process of designing policy or regulation with data portability provisions, policymakers and regulators could consider how to design portability *for* research, such as standardizing file formats and requiring clear documentation to ensure data donations are clean, accessible, and secure for research (see Table 2). Metrics for evaluating data portability could also include the ease of use for research purposes (Riley 2023).

Table 2: Policymakers and regulators can mandate clear documentation, standardized structures, and machine readable formats for DDPs.

	Current Challenges	Potential Solutions
Documentation	Documentation lacks clear explanations of variables	Mandate clear documentation of variables included in DDPs
Structure	Platforms do not provide DDPs structured for research	Require standardization of data structures, such as file and variables names
Machine readability	Platforms do not always provide files that are machine readable (e.g., HTML)	Ensure DDPs can be downloaded in machine readable formats

5.1.3 Building on Lessons Learned from the GDPR

Regulators could improve researchers' use of data donations by implementing lessons learned from the GDPR's Article 20, which provides for the right to data portability (RtDP) (Services 2016). This provision empowers EU citizens to take control of their data by having the right to transmit their data from one service or platform to another (Turner et al. 2021) .

Given the lack of a comprehensive federal privacy law in the United States, existing competition laws must ensure that data donations via portability protocols involve informed user consent. Current laws can adopt the GDPR's language on transparency and user control over personal data to mandate that platforms' data donation tools include explicit consent mechanisms. Drawing on the GDPR's transparency language is essential for users to understand what data is collected, control what data to donate, and understand the research purposes and benefits of contributing their data. Additionally, informed and empowered users may be more likely to participate in data donation efforts, supporting future trust and safety research.

Furthermore, some studies have shown that legal uncertainties and a lack of standardized portability mechanisms have hindered GDPR's RtDP (Lazarotto 2024). The United States could address these challenges by leveraging competition law—rather than privacy law—to enhance data portability protocols (Gill and Kerber 2020). Leveraging competition law

helps address legal uncertainties while promoting innovation and empowering users. For example, Australia implemented a hybrid consumer and sector-specific data law that offers a more flexible design of data portability rights (OECD Secretariat 2020). This flexibility allows users to port personal and non-personal data, increasing consumer empowerment and diversity of data for researchers. Additionally, a hybrid measure makes it easier for users to exercise their portability rights, thereby increasing the use of data donations and improving researchers' access to social media data.

5.2 The Role of Industry

5.2.1 Integrating Data Donations into Infrastructure

Platforms could integrate data donation tools directly into their infrastructure. In doing so, platforms should implement privacy-by-design principles (Cavoukian 2011), which embed privacy mechanisms into their data donations portal at the development stage, addressing privacy concerns from the start and as a foundation for data donations. Enabling a privacy-by-design approach to data donation portals protects users when donating their data, enhancing trust in the platform and research projects. Privacy-protected data donation portals also remove the burden on researchers to implement complex privacy safeguards when handling user data, especially if they lack the technical expertise or resources.

5.2.2 Enabling Direct Data Transfers

Since technical requirements for implementing data portability may be high (Engels 2016), companies should talk to researchers to ensure that portability systems can transfer data to researchers as well. One particularly effective mechanism, depicted in Figure 2 on the following page, would be direct data donations via transfers from a data host straight to a researcher data store, which avoids the logistical complexities and technical burdens of asking users to download and upload. The ease of use would improve sample quality and decrease attrition; the direct transfer would remove the need for participants to have the device storage or bandwidth necessary for large data downloads; and the common infrastructure would increase accessibility and researchers' ability to engage in data donation-based research. While this system could be managed by academics, it could also be built and maintained through third-party organizations, such as the Data Transfer Initiative,[15] with funding coming through regulations or companies.

5.2.3 Enhancing Users' Understanding of Privacy Policies

Platforms could write their privacy policies more clearly for users, especially concerning data portability rights. While data portability protocols exist legally and on many platforms already, a significant amount of their effectiveness lies in consumers' awareness of their

15. For more information, see https://dtinit.org/.

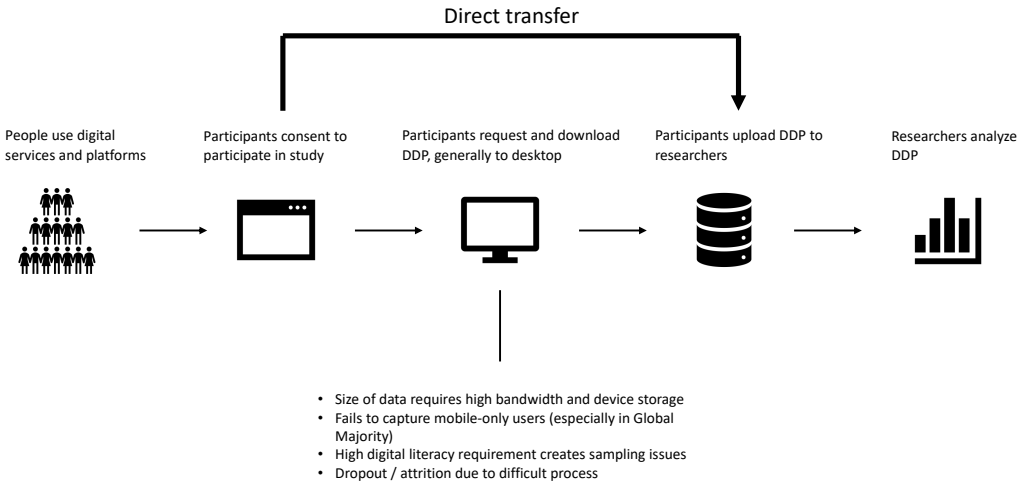

Figure 2: Direct, privacy-preserving transfer of DDPs from consenting participants to researchers would lower the barriers for using data donations in research.

rights and how to use them. For example, (Turner et al. 2021) found that users had difficulty understanding the purpose and meaning of the GDPR's RtDP, likely resulting in fewer users taking advantage of data portability mechanisms.

Additionally, while users can learn about data portability rights from regulators and the media, platforms bear more responsibility in making those rights accessible (Woodall 2024). Therefore, the design of a platform's data download and portability interfaces will impact whether a user engages with such a feature. Without sufficient portability mechanisms on platforms, users may have to navigate a poorly designed interface with confusing steps to download their data, lack clarity around which specific data are being downloaded, and face challenges when sharing their data files with researchers, as detailed in Section 4. This issue highlights the importance of having accessible standard or automated tools for users. Platforms that address this with a user-friendly interface will support research initiatives and gain a better competitive advantage.

5.3 The Role of Researchers

5.3.1 Engaging across Academic and Non-Academic Stakeholders

A primary role that researchers can play is engaging policymakers, regulators, and companies through the process of designing and implementing portability systems that align with research needs. This process requires engaging diverse stakeholders: policy experts in data rights and portability, policy experts in platform transparency, and company employees designing these systems, among others.

One possible mechanism for aligning and communicating research needs is to invest in intermediary structures that could be effective bridges and thus reduce the burden of translating researcher interests, especially given that researchers will be involved from

across disciplines and geographies. For example, this could take the shape of a research consortium that would set up mechanisms for transfers from major platforms, and researchers could interact with that consortium to support their particular projects. There are already models for this type of consortium approach for negotiating and provisioning data access between companies and researchers, such as the Social Media Archive at the Inter-university Consortium for Political and Social Research (ICPSR) [16]. A similar model could be developed here, though it would need platform buy-in.

5.3.2 Investing in Shared Infrastructure

While other suggestions require time, resources, and multistakeholder collaboration, researchers can facilitate data donations through building and maintaining shared infrastructure for data donations. Projects such as Port and the the Data Donation Module are ongoing and could benefit from greater resources from funders and coordination across research groups.

6 Conclusion

We aim to situate data portability, which is often considered through the lens of user agency and competition, within ongoing work on expanding data access for independent researchers. By leveraging data access rights established by regulations with portability provisions, data donations offer a powerful mechanism for researchers to collect multi-platform digital trace data from consenting users. While laws like the DSA have platform-centric data access provisions, data donations offer another compelling avenue for facilitating research on core trust and safety topics. However, significant challenges currently limit researchers' ability to utilize data donations, including the complex process of obtaining a DDP, non-standardized data structures, and insufficient documentation.

Our intended audience is threefold: policy experts across the areas of data portability and platform transparency who rarely, if ever, engage with one another; platform employees developing privacy-preserving data-sharing mechanisms to support independent research; and researchers studying the digital information environment. There are opportunities for each group to address the current challenges associated with leveraging data donations for research. Meaningful actions include stakeholders engaging with one another to standardize file formats and simplify data management, fostering cross-stakeholder engagement to design data portability systems for the needs of researchers, and integrating data donation tools directly into platform infrastructure with privacy-by-design principles. A multistakeholder approach to aligning data portability systems with research offers opportunities to support research on critical topics with academic, policy, and public importance.

16. See https://socialmediaarchive.org/pages/?page=About&ln=en

References

Abrajano, Marisa, Marianna Garcia, Aaron Pope, Robert Vidigal, Edwin Kamau, Joshua A. Tucker, and Jonathan Nagler. 2022. "Social Media, Information, and Politics: Insights on Latinos in the U.S." *OSF* (November 3, 2022). https://csmapnyu.org/research/academic-research/social-media-information-and-politics-insights-on-latinos-in-the-u-s.

Albert, John. 2022. "A Guide to the EU's New Rules for Researcher Access to Platform Data." AlgorithmWatch, December 7, 2022. https://algorithmwatch.org/en/dsa-data-access-explained/.

Allen, Jennifer, Markus Mobius, David M. Rothschild, and Duncan J. Watts. 2021. "Research Note: Examining Potential Bias in Large-scale Censored Data." *Harvard Kennedy School Misinformation Review* (July 26, 2021). https://doi.org/10.37016/mr-2020-74.

Americans for Tax Fairness. 2023. *The Research Expensing Tax Deduction vs. The Research Tax Credit.* https://americansfortaxfairness.org/research-expensing-tax-deduction-vs-research-tax-credit/.

Anderson, Monica, Michelle Faverio, and Jeffrey Gottfried. 2023. "Teens, Social Media and Technology 2023." Pew Research Center: Internet, Science and Tech. https://www.pewresearch.org/internet/2023/12/11/teens-social-media-and-technology-2023/.

Aslett, Kevin, Zeve Sanderson, William Godel, Nathaniel Persily, Jonathan Nagler, and Joshua A. Tucker. 2023. "Online Searches to Evaluate Misinformation Can Increase Its Perceived Veracity." *Nature* 625 (December 20, 2023): 1–9. https://doi.org/10.1038/s41586-023-06883-y.

Ausloos, Jef, and Michael Veale. 2020. "Researching with Data Rights." *Amsterdam Law School Research Paper,* no. 2020-30 (December 31, 2020): 136–57. https://doi.org/10.2139/ssrn.3465680.

Auxier, Brooke, and Monica Anderson. 2021. "Social Media Use in 2021." Pew Research Center, April 7, 2021. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/.

Bobrowsky, Meghan. 2021. "Facebook Disables Access for NYU Research into Political-Ad Targeting." *Wall Street Journal* (August 4, 2021). https://www.wsj.com/articles/facebook-cuts-off-access-for-nyu-research-into-political-ad-targeting-11628052204.

Boeschoten, Laura, Jef Ausloos, Judith Moeller, Theo Araujo, and Daniel L. Oberski. 2020. *Digital Trace Data Collection through Data Donation,* November 13, 2020. https://doi.org/10.48550/arXiv.2011.09851. arXiv: 2011.09851 [cs.CY].

Boeschoten, Laura, Niek C. de Schipper, Adriënne M. Mendrik, Emiel van der Veen, Bella Struminskaya, Heleen Janssen, and Theo Araujo. 2023. "Port: A Software Tool for Digital Data Donation." *Journal of Open Source Software* 8, no. 90 (October 3, 2023): 5596. https://doi.org/10.21105/joss.05596.

Breuer, Johannes, Libby Bishop, and Katharina Kinder-Kurlanda. 2020. "The Practical and Ethical Challenges in Acquiring and Sharing Digital Trace Data: Negotiating Public-Private Partnerships." *New Media & Society* 22, no. 11 (October 4, 2020): 2058–80. https://doi.org/10.1177/1461444820924622.

Breuer, Johannes, Zoltán Kmetty, Mario Haim, and Sebastian Stier. 2022. "User-centric Approaches for Collecting Facebook Data in the 'Post-API Age': Experiences from Two Studies and Recommendations for Future Research." *Information, Communication & Society* 26, no. 14 (July 8, 2022): 2649–68. https://doi.org/10.1080/1369118X.2022.2097015.

Bruns, Axel. 2019. "After the 'APIcalypse': Social Media Platforms and Their Fight against Critical Scholarly Research." *Information, Communication & Society* 22, no. 11 (July 11, 2019): 1544–66. https://doi.org/10.1080/1369118X.2019.1637447.

Calma, Justine. 2023. "Twitter Just Closed the Book on Academic Research." The Verge, May 31, 2023. https://www.theverge.com/2023/5/31/23739084/twitter-elon-musk-api-policy-chilling-academic-research.

Carvalho, Mateus Correia de. 2024. "Researcher Access to Platform Data and the DSA: One Step Forward, Three Steps Back." Tech Policy Press, May 31, 2024. https://www.techpolicy.press/researcher-access-to-platform-data-and-the-dsa-one-step-forward-three-steps-back/.

Castro, Daniel. 2021. *Improving Consumer Welfare with Data Portability.* Technical report. Information Technology and Innovation Foundation.

Cavoukian, Anna. 2011. *Privacy by Design: The 7 Foundational Principles - Implementation and Mapping of Fair Information Practices.* https://privacy.ucsc.edu/resources/privacy-by-design---foundational-principles.pdf.

Center for Countering Digital Hate. 2024. "Elon Musk vs. Center for Countering Digital Hate: Nonprofit Wins Dismissal of 'Baseless and Intimidatory' Lawsuit Brought by World's Richest Man," March 25, 2024. https://counterhate.com/blog/elon-musk-vs-ccdh-nonprofit-wins-dismissal-of-baseless-and-intimidatory-lawsuit/.

Chen, Meng, and Altman Yuzhu Peng. 2023. "Why Do People Choose Different Social Media Platforms? Linking Use Motives with Social Media Affordances and Personalities." *Social Science Computer Review* 41, no. 2 (April 26, 2023): 330–52. https://doi.org/10.1177/08944393211049120.

Coalition for Independent Technology Research. 2024. *Blocking Our Right to Know: Surveying the Impact of Meta's CrowdTangle Shutdown.* https://independenttechresearch.org/wp-content/uploads/2024/07/CrowdTangle-Survey-Report-Final.pdf.

Coons, Chris. 2023. "Senator Coons, Colleagues Introduce Legislation to Increase Transparency around Social Media Platforms." Office of Sentator Chris Coons, June 8, 2023. https://www.coons.senate.gov/news/press-releases/senator-coons-colleagues-introduce-legislation-to-increase-transparency-around-social-media-platforms.

Crutzen, Rik, Gjalt-Jorn Ygram Peters, and Christopher Mondschein. 2019. "Why and How We Should Care about the General Data Protection Regulation." *Psychology & Health* 34, no. 11 (May 21, 2019): 1347–57. https://doi.org/10.1080/08870446.2019.1606222.

Data for Good. 2021. "Data for Good at Meta CrowdTangle." Meta. https://dataforgood.facebook.com/dfg/tools/crowd-tangle.

De Hert, Paul, Vagelis Papakonstantinou, Gianclaudio Malgieri, Laurent Beslay, and Ignacio Sanchez. 2018. "The Right to Data Portability in the GDPR: Towards User-centric Interoperability of Digital Services." *Computer Law & Security Review* 34, no. 2 (March 28, 2018): 193–203. https://doi.org/10.1016/j.clsr.2017.10.003.

Derakhshani, Delara. 2023. "Global Developments in Data Portability Law." Data Transfer Initative, October 24, 2023. https://dtinit.org/blog/2023/10/24/global-developments.

Driel, Irene I. van, Anastasia Giachanou, J. Loes Pouwels, Laura Boeschoten, Ine Beyens, and Patti M. Valkenburg. 2022. "Promises and Pitfalls of Social Media Data Donations." *Communication Methods and Measures* 16, no. 4 (September 12, 2022): 266–82. https://doi.org/10.1080/19312458.2022.2109608.

Duffy, Clare. 2023. "'It's an Especially Bad Time': Tech Layoffs Are Hitting Ethics and Safety Teams." CNN, April 6, 2023. https://edition.cnn.com/2023/04/06/tech/tech-layoffs-platform-safety/index.html.

Engels, Barbara. 2016. "Data Portability among Online Platforms." *Internet Policy Review* 5, no. 2 (June 11, 2016): 1–17. https://doi.org/10.14763/2016.2.408.

Engler, Alex. 2021. "Platform Data Access is a Lynchpin of the EU's Digital Services Act." Brookings Institution, January 15, 2021. https://www.brookings.edu/articles/platform-data-access-is-a-lynchpin-of-the-eus-digital-services-act/.

European Commission. 2022. *The Digital Markets Act: Ensuring Fair and Open Digital Markets,* October 12, 2022. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-markets-act-ensuring-fair-and-open-digital-markets_en.

———. 2023. *Commission Designates First Very Large Online Platforms and Search Engines under the Digital Services Act.* https://ec.europa.eu/commission/presscorner/detail/en/IP_23_2413.

Faust, Rachel, and Daniel Arnaudo. 2024. "The Urgency of Social Media Data Access for Electoral Integrity." Tech Policy Press, March 4, 2024. https://www.techpolicy.press/the-urgency-of-social-media-data-access-for-electoral-integrity/.

Fiesler, Casey, Nathan Beard, and Brian C. Keegan. 2020. "No Robots, Spiders, or Scrapers: Legal and Ethical Regulation of Data Collection Methods in Social Media Terms of Service." In *Proceedings of the International AAAI Conference on Web and Social Media,* 14:187–96. May 26, 2020. https://doi.org/10.1609/icwsm.v14i1.7290.

Fiesler, Casey, and Nicholas Proferes. 2018. "'Participant' Perceptions of Twitter Research Ethics." *Social Media + Society* 4, no. 1 (March 10, 2018): 2056305118763366. https://doi.org/10.1177/2056305118763366.

Freelon, Deen. 2018. "Computational Research in the Post-API Age." *Political Communication* 35, no. 4 (October 25, 2018): 665–68. https://doi.org/10.1080/10584609.2018.1477506.

Gallagher, Josh. 2023. "Reddit Will Begin Charging for Access to Its API." TechCrunch. https://techcrunch.com/2023/04/18/reddit-will-begin-charging-for-access-to-its-api/.

Gilbert, Sarah, Jessica Vitak, and Katie Shilton. 2021. "Measuring Americans' Comfort with Research Uses of Their Social Media Data." *Social Media + Society* 7, no. 3 (July 24, 2021): 20563051211033824. https://doi.org/10.1177/20563051211033824.

Gill, Daniel, and Wolfgang Kerber. 2020. "Data Portability Rights: Limits, Opportunities, and the Need for Going Beyond the Portability of Personal Data." *SSRN* (November 10, 2020): 1–11. https://doi.org/10.2139/ssrn.3715357.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake News on Twitter During the 2016 US Presidential Election." *Science* 363, no. 6425 (January 25, 2019): 374–78. https://doi.org/10.1126/science.aau2706.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5, no. 1 (January 9, 2019): eaau4586. https://doi.org/10.1126/sciadv.aau4586.

Gulati-Gilbert, Sukhi, and Robert Seamans. 2023. "Data Portability and Interoperability: A Primer on Two Policy Tools for Regulation of Digitized Industries." Brookings Institution, May 9, 2023. https://www.brookings.edu/articles/data-portability-and-interoperability-a-primer-on-two-policy-tools-for-regulation-of-digitized-industries-2/.

Haim, Mario, and Angela Nienierza. 2019. "Computational Observation: Challenges and Opportunities of Automated Observation within Algorithmically Curated Media Environments Using a Browser Plug-in." *Computational Communication Research* 1, no. 1 (December 2, 2019): 79–102.

Halavais, Alexander. 2019. "Overcoming Terms of Service: A Proposal for Ethical Distributed Research." *Information, Communication & Society* 22, no. 11 (June 20, 2019): 1567–81. https://doi.org/10.1080/1369118X.2019.1627386.

Hase, Valerie, and Mario Haim. 2024. "Can We Get Rid of Bias? Mitigating Systematic Error in Data Donation Studies through Survey Design Strategies." *Computational Communication Research* 6, no. 2 (July 24, 2024). https://doi.org/10.5117/CCR2024.2.2.HASE.

Hemphill, Libby, Angela Schöpke-Gonzalez, and Anmol Panda. 2022. "Comparative Sensitivity of Social Media Data and Their Acceptable Use in Research." *Scientific Data* 9, no. 1 (October 22, 2022): 643. https://doi.org/10.1038/s41597-022-01773-w.

Hendrix, Justin. 2023. "Twitter API Changes Set to Disrupt Public Interest Research." Tech Policy Press, February 6, 2023. https://www.techpolicy.press/twitter-api-changes-set-to-disrupt-public-interest-research/.

Hosseinmardi, Homa, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M. Rothschild, and Duncan J. Watts. 2021. "Examining the Consumption of Radical Content on YouTube." *Proceedings of the National Academy of Sciences* 118, no. 32 (June 7, 2021): e2101967118. https://doi.org/10.1073/pnas.2101967118.

Howison, James, Andrea Wiggins, and Kevin Crowston. 2011. "Validity Issues in the Use of Social Network Analysis with Digital Trace Data." *Journal of the Association for Information Systems* 12 (12): 2. https://doi.org/10.17705/1jais.00282.

Husovec, Martin. 2023. "How to Facilitate Data Access under the Digital Services Act." *SSRN* (May 19, 2023). https://ssrn.com/abstract=4452940.

Joint Research Centre. 2023. "FAQs: DSA Data Access for Researchers." European Commission, December 13, 2023. https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en.

Keller, Daphne. 2022. *Statement Before the United States Senate Committee on the Judiciary, Subcommittee on Privacy, Technology and the Law, Hearing on Platform Transparency: Understanding the Impact of Social Media.* https://www.judiciary.senate.gov/imo/media/doc/Keller%20Testimony1.pdf.

Keusch, Florian, Paulina K. Pankowska, Alexandru Cernat, and Ruben L. Bach. 2024. "Do You Have Two Minutes to Talk about Your Data? Willingness to Participate and Nonparticipation Bias in Facebook Data Donation." *Field Methods* (January 11, 2024). https://doi.org/10.1177/1525822X231225907.

Kharpal, Natasha. 2023. "Twitter Announces New API with Only Free, Basic and Enterprise Levels." TechCrunch. https://techcrunch.com/2023/03/29/twitter-announces-new-api-with-only-free-basic-and-enterprise-levels/.

Krishnan, Sriram. 2023. "Opinion | Threads, Twitter, and the Future of Social Media." *New York Times* (July 15, 2023). https://www.nytimes.com/2023/07/15/opinion/social-media-threads-twitter-reddit.html.

Krotov, Vlad, Leigh Johnson, and Leiser Silva. 2020. "Tutorial: Legality and Ethics of Web Scraping." *Communications of the Association for Information Systems,* no. 47 (December 10, 2020). https://doi.org/10.17705/1CAIS.04724.

Kupferschmidt, Kai. 2023. "Does Social Media Polarize Voters? Unprecedented Experiments on Facebook Users Reveal Surprises." *Science* 381, no. 6656 (July 27, 2023). https://doi.org/10.1126/science.adj9982.

Lazarotto, Bárbara da Rosa. 2024. "The Right to Data Portability: A Holistic Analysis of GDPR, DMA and the Data Act." *European Journal of Law and Technology* 15, no. 1 (May 2, 2024): 1–15. https://ejlt.org/index.php/ejlt/article/view/988.

Lazer, David M.J., Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. "Computational Social Science: Obstacles and Opportunities." *Science* 369, no. 6507 (August 28, 2020): 1060–62. https://doi.org/10.1126/science.aaz8170.

Lukito, Josephine, Megan A. Brown, Ross Dahlke, Jiyoun Suk, Yunkang Yang, Yini Zhang, Bin Chen, Sang Jung Kim, and Kaiya Soorholtz. 2023. *The State of Digital Media Data Research, 2023.* Research report. Media & Democracy Data Cooperative. https://doi.org/10.26153/tsw/46177.

Matamoros-Fernández, Ariadna, and Johan Farkas. 2021. "Racism, Hate Speech, and Social Media: A Systematic Review and Critique." *Television & New Media* 22, no. 2 (January 22, 2021): 205–24. https://doi.org/10.1177/1527476420982230.

Meyer, Michelle N., John Basl, David Choffnes, Christo Wilson, and David M.J. Lazer. 2023. "Enhancing the Ethics of User-sourced Online Data Collection and Sharing." *Nature Computational Science* 3, no. 8 (July 23, 2023): 660–64. https://doi.org/10.1038/s43588-023-00490-7.

Mondschein, Christopher F., and Cosimo Monda. 2018. "The EU's General Data Protection Regulation (GDPR) in a Research Context." In *Fundamentals of Clinical Data Science,* edited by P. Kubben, M. Dumontier, and A. Dekker, 55–71. Springer International Publishing, December 22, 2018. https://doi.org/10.1007/978-3-319-99713-1_5.

Moritz, Mark. 2016. "Big Data's 'Streetlight Effect': Where and How We Look Affects What We See." The Conversation, May 17, 2016. https://theconversation.com/big-datas-streetlight-effect-where-and-how-we-look-affects-what-we-see-58122.

Nguyen, Britney. 2024. "Facebook is Killing a Data Tool that Helps Researchers Monitor Facebook." Quartz, March 14, 2024. https://qz.com/meta-replacing-crowdtangle-monitor-content-us-election-1851334958.

Nicholas, Gabriel. 2024. "Red Teaming Isn't Enough." Foreign Policy, July 8, 2024. https://foreignpolicy.com/2024/07/08/artificial-intelligence-ai-election-misinformation-technology-risks/?tpcc=recirc_latest062921.

Odabas, Meltem. 2022. "5 Facts about Twitter 'Lurkers'." Pew Research Center, March 16, 2022. https://www.pewresearch.org/short-reads/2022/03/16/5-facts-about-twitter-lurkers/.

OECD Secretariat. 2020. "Consumer Data Rights and Competition - Background Note." Organisation for Economic Co-operation and Development. https://one.oecd.org/document/DAF/COMP(2020)1/en/pdf.

Ohme, Jakob, Theo Araujo, Laura Boeschoten, Deen Freelon, Nilam Ram, Byron B. Reeves, and Thomas N. Robinson. 2023. "Digital Trace Data Collection for Social Media Effects Research: APIs, Data Donation, and (Screen) Tracking." *Communication Methods and Measures* (February 27, 2023): 1–18. https://doi.org/10.1080/19312458.2023.2181319.

Ohme, Jakob, Theo Araujo, Claes H. de Vreese, and Jessica Taylor Piotrowski. 2021. "Mobile Data Donations: Assessing Self-report Accuracy and Sample Biases with the iOS Screen Time Function." *Mobile Media & Communication* 9, no. 2 (September 30, 2021): 293–313. https://doi.org/10.1177/2050157920959106.

Ortiz-Ospina, Esteban. 2019. "The Rise of Social Media." Our World in Data. https://ourworldindata.org/rise-of-social-media.

Persily, Nathaniel. 2021. "A Proposal for Researcher Access to Platform Data: The Platform Transparency and Accountability Act." *Journal of Online Trust and Safety* 1, no. 1 (October 28, 2021). https://doi.org/10.54501/jots.v1i1.22.

Persily, Nathaniel, and Joshua A. Tucker. 2020. "Conclusion: The Challenges and Opportunities for Social Media Research." In *Social Media and Democracy,* edited by Nathaniel Persily and Joshua A. Tucker, 313–31. SSRC Anxieties of Democracy. Cambridge University Press. https://doi.org/10.1017/9781108890960.

Pfiffner, Nico, and Thomas N. Friemel. 2023. "Leveraging Data Donations for Communication Research: Exploring Drivers behind the Willingness to Donate." *Communication Methods and Measures* (March 1, 2023): 1–23. https://doi.org/10.1080/19312458.2023.2176474.

Prainsack, Barbara. 2019. "Data Donation: How to Resist the iLeviathan." In *The Ethics of Medical Data Donation,* edited by J. Krutzinna and L. Floridi, 137:9–22. Philosophical Studies Series. Springer International Publishing, January 16, 2019. ISBN: 978-3-030-04363-6. https://doi.org/10.1007/978-3-030-04363-6_2.

Reeves, Byron, Nilam Ram, Thomas N. Robinson, James J. Cummings, C. Lee Giles, Jennifer Pan, Agnese Chiatti, M.J. Cho, Katie Roehrick, Xiao Yang, et al. 2021. "Screenomics: A Framework to Capture and Analyze Personal Life Experiences and the Ways That Technology Shapes Them." *Human–Computer Interaction* 36, no. 2 (March 13, 2021): 150–201. https://doi.org/10.1080/07370024.2019.1578652.

Reimsbach-Kounatze, Christian, and Andras Molnar. 2024. "The Impact of Data Portability on User Empowerment, Innovation, and Competition." *OECD Going Digital Toolkit Notes, No. 25* (June 29, 2024). https://doi.org/10.1787/319f420f-en.

Riley, Chris. 2023. "Metrics for Success in Data Portability Work." Data Transfer Initative, October 10, 2023. https://dtinit.org/blog/2023/10/10/metrics-for-success.

———. 2024. "DTI's Portability Predictions for 2024." Data Transfer Initiative, January 2, 2024. https://dtinit.org/blog/2024/01/02/portability-predictions.

Robertson, Ronald E. 2022. "Uncommon Yet Consequential Online Harms." *Journal of Online Trust and Safety* 1, no. 3 (August 31, 2022). https://doi.org/10.54501/jots.v1i3.87.

Robertson, Ronald E., Jon Green, Damian J. Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. 2023. "Users Choose to Engage with More Partisan News Than They Are Exposed to on Google Search." *Nature* 618 (May 24, 2023): 1–7. https://doi.org/10.1038/s41586-023-06078-5.

Ruths, Derek, and Jürgen Pfeffer. 2014. "Social Media for Large Studies of Behavior." *Science* 346, no. 6213 (November 28, 2014): 1063–64. https://doi.org/10.1126/science.346.6213.106.

Salganik, Matthew J. 2019. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press, August 6, 2019. ISBN: 9780691196107.

Sanchez, Gabriel R., and Carly Bennett. 2022. "Why Spanish-language Mis-and Disinformation is a Huge Issue in 2022." Brookings Institution, November 4, 2022. https://www.brookings.edu/articles/why-spanish-language-mis-and-disinformation-is-a-huge-issue-in-2022/.

Services, Intersoft Consulting. 2016. *Art. 20 GDPR Right to Data Portability.* https://gdpr-info.eu/art-20-gdpr/.

Shahbaznezhad, Hamidreza, Rebecca Dolan, and Mona Rashidirad. 2021. "The Role of Social Media Content Format and Platform in Users' Engagement Behavior." *Journal of Interactive Marketing* 53, no. 1 (January 31, 2021): 47–65. https://doi.org/10.1016/j.intmar.2020.05.001.

Sharma, Chinmayi. 2024. "A Marketplace for Data Portability." In *The Present and Future of Data Portability,* edited by Chrile Riley, 32–50. Data Transfer Initiative, March 28, 2024. https://doi.org/10.2139/ssrn.4741065.

Stokel-Walker, Chris. 2023. "Twitter's $42,000-per-Month API Prices Out Nearly Everyone." WIRED, March 18, 2023. https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/.

Tax Analysis, Office of. 2016. "Research and Experimentation (R&E) Credit." U.S. Department of the Treasury. https://home.treasury.gov/system/files/131/RE-Credit.pdf.

TikTok. 2021. *Requesting Your Data.* https://support.tiktok.com/en/account-and-privacy/personalized-ads-and-data/requesting-your-data.

Turner, Sarah, July Galindo Quintero, Simon Turner, Jessica Lis, and Leonie Maria Tanczer. 2021. "The Exercisability of the Right to Data Portability in the Emerging Internet of Things (IoT) Environment." *New Media & Society* 23, no. 10 (July 10, 2021): 2861–81. https://doi.org/10.1177/1461444820934033.

Turner, Sarah, and Leonie Maria Tanczer. 2024. "In Principle vs in Practice: User, Expert and Policymaker Attitudes Towards the Right to Data Portability in the Internet of Things." *Computer Law & Security Review* 52. https://doi.org/10.1016/j.clsr.2023.105912.

University of Pennsylvania Social Behavioral Research. 2023. "Use of Social Media as a Research Activity." Human Research Protections Program Office of the Institutional Review Board, University of Pennsylvania. https://irb.upenn.edu/homepage/social-behavioral-homepage/guidance/types-of-social-behavioral-research/use-of-social-media-as-a-research-activity/.

Vreese, Claes de, and Rebekah Tromble. 2023. "The Data Abyss: How Lack of Data Access Leaves Research and Society in the Dark." *Political Communication* (May 20, 2023): 1–5. https://doi.org/10.1080/10584609.2023.2207488.

Wagner, Michael W. 2023. "Independence by Permission." *Science* 381, no. 6656 (July 27, 2023): 388–91. https://doi.org/10.1126/science.adi2430.

Walker, Shawn, Dan Mercea, and Marco Bastos. 2019. "The Disinformation Landscape and the Lockdown of Social Platforms." *Information, Communication & Society* 22, no. 11 (August 29, 2019): 1531–43. https://doi.org/10.1080/1369118X.2019.1648536.

Welles, Brooke Foucault. 2014. "On Minorities and Outliers: The Case for Making Big Data Small." *Big Data & Society* 1, no. 1 (April 1, 2014): 2053951714540613. https://doi.org/10.1177/2053951714540613.

Windwehr, Svea, and Joschka Selinger. 2024. "Can We Fix Access to Platform Data? Europe's Digital Services Act and the Long Quest for Platform Accountability and Transparency." Internet Policy Review, March 27, 2024. https://policyreview.info/articles/news/can-we-fix-access-to-platform-data.

Wolford, Ben. 2018. "What is GDPR, the EU's New Data Protection Law?" GDPR.EU. https://gdpr.eu/what-is-gdpr/.

Woodall, Angela. 2024. "Data Portability IRL: A Stakeholder Assessment of Data Portability Methods." In *The Present and Future of Data Portability,* 17–31. Data Transfer Initiative, May 26, 2024. https://doi.org/10.2139/ssrn.4738496.

Yin, Leon. 2023. "Journalists Should Be Looking for Undocumented APIs. Here's How to Start." Nieman Lab, March 15, 2023. https://www.niemanlab.org/2023/03/journalists-should-be-looking-for-undocumented-apis-heres-how-to-start/.

Yin, Leon, and Aaron Sankin. 2021. "How We Discovered Google's Hate Blocklist for Ad Placements on YouTube." The Markup, April 8, 2021. https://themarkup.org/google-the-giant/2021/04/08/how-we-discovered-googles-hate-blocklist-for-ad-placements-on-youtube.

Zannettou, Savvas, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. "Measuring and Characterizing Hate Speech on News Websites." In *Proceedings of the 12th ACM Conference on Web Science,* 125–34. July 6, 2020. https://doi.org/10.1145/3394231.3397902.

Zweifel-Keegan, Cobun. 2024. "Portable Trust: Fostering Both Autonomy and Privacy in Data Portability." In *The Present and Future of Data Portability,* edited by Chrile Riley, 4–16. Data Transfer Initiative, February 27, 2024. https://dtinit.org/assets/PortableTrustCZK.pdf.

Authors

Zeve Sanderson is the Executive Director at NYU's Center for Social Media and Politics and a PhD student at Vrije Universiteit Amsterdam.

Lama Mohammed is the Tech Policy Fellow at NYU's Center for Social Media and Politics.

Acknowledgements

We thank Chris Riley and the Data Transfer Initiative for feedback on and support of this paper.

Funding statement

We gratefully acknowledge that the Center for Social Media and Politics at New York University is supported by funding from the John S. and James L. Knight Foundation, the Charles Koch Foundation, Craig Newmark Philanthropies, the William and Flora Hewlett Foundation, the Siegel Family Endowment, and the Bill and Melinda Gates Foundation.

Keywords

Data Portability; data donations; data access; internet policy; competition; platform transparency.