INTERNET POLICY REVIEW
Journal on internet regulation

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

Volume 13 **Issue** 3

Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies?

Valerie Hase *LMU Munich* valerie.hase@ifkw.lmu.de
Jef Ausloos *University of Amsterdam* j.ausloos@uva.nl
Laura Boeschoten *Utrecht University* l.boeschoten@uu.nl
Nico Pfiffner *University of Zurich* n.pfiffner@ikmz.uzh.ch
Heleen Janssen *University of Amsterdam* h.l.janssen@uva.nl
Theo Araujo *University of Amsterdam* t.b.araujo@uva.nl
Thijs Carrière *Utrecht University* t.c.carriere@uu.nl
Claes de Vreese *University of Amsterdam*
Jörg Haßler *LMU Munich* joerg.hassler@lmu.de
Felicia Loecherbach *University of Amsterdam* f.loecherbach@uva.nl
Zoltán Kmetty *Centre for Social Sciences* kmetty.zoltan@tk.hu
Judith Möller *University of Hamburg – Leibniz Institute for Media Research (HBI)* j.moeller@leibniz-hbi.de
Jakob Ohme *Weizenbaum Institute for the Networked Society* jakob.ohme@weizenbaum-institut.de
Elisabeth Schmidbauer *LMU Munich* Elisabeth.Schmidbauer@ifkw.lmu.de
Bella Struminskaya *Utrecht University* b.struminskaya@uu.nl
Damian Trilling *Vrije Universiteit Amsterdam* d.c.trilling@vu.nl
Kasper Welbers *Vrije Universiteit Amsterdam* k.welbers@vu.nl
Mario Haim *LMU Munich* haim@ifkw.lmu.de

DOI: https://doi.org/10.14763/2024.3.1793

Published: 16 September 2024
Received: 4 August 2023 **Accepted:** 25 April 2024

Funding: Theo Araujo, Laura Boeschoten, Heleen Jansen & Bella Struminskaya received funding from the Platform Digitale Infrastructuur SSH. Jörg Haßler & Elisabeth Schmidbauer received funding from the Bavarian Research Institute for Digital Transformation. Jef Ausloos received funding from the Dutch Research Council, project: VI.Veni.201R.096. Valerie Hase received funding from the Deutsche Forschungsgemeinschaft, project: 539555013. Zoltan Kmetty received funding by the Bolyai Scholarship, grant number: BO/834/22.
Competing Interests: The author has declared that no competing interests exist that have influenced the text.

Citation: Hase, V., Ausloos, J., Boeschoten, L., Pfiffner, N., Janssen, H., Araujo, T., Carrière, T., de Vreese, C., Haßler, J., Loecherbach, F., Kmetty, Z., Möller, J., Ohme, J., Schmidbauer, E., Struminskaya, B., Trilling, D., Welbers, K., & Haim, M. (2024). Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies? *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1793

Keywords: Social media platforms, Data donation, Digital trace data, Internet policy, Transparency

Abstract: Research into digital platforms has become increasingly difficult. One way to overcome these difficulties is to build on data access rights in EU data protection law, which requires platforms to offer users a copy of their data. In data donation studies, researchers ask study participants to exercise this right and donate their data to science. However, there is increasing evidence that platforms do not comply with designated laws. We first discuss the obligations of data access from a legal perspective (with accessible, transparent, and complete data as key requirements). Next, we compile experiences from social scientists engaging in data donation projects as well as a study on data request/access. We identify 14 key challenges, most of which are a consequence of non-compliance by platforms. They include platforms' insufficient adherence to (a) providing data in a concise and easily accessible form (e.g. the lack of information on when and how subjects can access their data); (b) being transparent about the content of their data (e.g. the lack of information on measures); and (c) providing complete data (e.g. the lack of all available information platforms process related to platform users). Finally, we formulate four central recommendations for improving the right to access.

1. Introduction

Digital platforms, such as Facebook, Instagram, and YouTube, provide information for citizens across the globe (Newman et al., 2023). When people use digital platforms, they leave digital traces that researchers can deploy to study human behaviour (Freelon, 2014; Keusch & Kreuter, 2021). However, researchers often face challenges in accessing such digital trace data (de Vreese & Tromble, 2023). Digital platforms and, relatedly, corporations, such as Meta and Google, store data in proprietary archives, rendering them de facto gatekeepers of research agendas (Ausloos & Veale, 2021).

In light of this limitation, researchers have developed approaches for obtaining digital traces, including negotiating with platforms (Dommett & Tromble, 2022), setting up research collaborations between platforms and researchers (Wagner, 2023) and using tools provided by platforms (e.g. application programming interfaces [APIs][1]). They also employ more adversarial methods that do not rely on the goodwill of platforms to share data (e.g. scraping; for overviews, see Mancosu & Vegetti, 2020; Ohme et al., 2024).

Researchers have also begun capitalising on data access provisions in the law

[1] For a glossary of abbreviations and terms, see Table A1 Supplement.

(Ausloos & Veale, 2021; Bruns, 2019; Freelon, 2018; Halavais, 2019), especially the right of access in the General Data Protection Regulation (GDPR). The GDPR grants *data subjects*, that is, identifiable persons to whom information relates (Art. 4(1) GDPR), the right to obtain a copy of all personal data that platforms process about them. Digital platforms or *data controllers* as actors determining the purposes and means of this processing (Art. 4(7) GDPR) are obliged to enable such access. In particular, Art. 15 of the GDPR requires platforms to give data subjects, i.e. platform users, access to a copy of their personal data and information on how data was processed.

This article discusses how platforms could and potentially should enable research with a focus on data rights based on the GDPR (see similarly European Digital Media Observatory, 2022), although our analysis also applies to more recent frameworks, such as the Data Act (DA), the Digital Markets Act (DMA) or the Digital Services Act (DSA). Individuals can also exercise data access rights as provided by the DMA (Art. 11) and the recently adopted DA (Art. 5(7)). For research purposes, accountability, and pro-competition aims, selective data access rights can similarly be exercised under the DSA (Art. 40) and the DA (see further Leersen, 2024; Veale, 2023). While we focus on the GDPR, our discussion also holds implications and recommendations for the DA, the DMA, and the DSA (see also Ausloos et al., 2023).

In pace with these legal requirements, researchers have begun developing research designs that rely on access provided through GDPR regulation. In so-called data donation studies, researchers ask platform users to request their data from platforms. Ideally, platform users can access and store such data in the form of data download packages (DDPs), that is, files containing their personal data. Individuals can then donate their DDPs to researchers via data donation tools (DDTs)[2], such as Port (Boeschoten et al., 2022) or the Data Donation Module (Pfiffner et al., 2024b). These tools extract and anonymise relevant data from platform users' DDPs on the subjects' devices as a form of privacy-by-design; afterwards, they send the data to researchers (van Driel et al., 2022). Data donation studies have, for example, been used to study how citizens use social media platforms to stay informed or message friends (Hase & Haim, 2024; van Driel et al., 2022), whether they employ search engines to search for political information (Blassnig et al., 2023) or to detect depression-related behaviour by social media users (Kmetty & Bozsonyi, 2022).

[2] We define data donations as *subjects downloading their data from platforms as DDPs and donating it to research via DDTs*. This excludes other approaches, such as tracking or APIs.

As user-centric approaches (Halavais, 2019), data donation studies build on the informed consent of users and make use of rights featured in data protection laws (for an overview of laws beyond the EU, see Greenleaf, 2021). While data donation studies have often been employed in the context of Europe, researchers have also started to rely on this approach elsewhere, for example in China (Wu-Ouyang & Chan, 2023), India (Garimella & Chauchard, 2024) or Pakistan (Ejaz et al., 2023). Under the GDPR, platforms are obliged to comply with data access rights. However, a growing body of case law (Case C-487/21; GDPR Hub, 2020) and empirical research (Ausloos et al., 2020; Syrmoudis et al., 2021) paints a problematic picture of how these rights are implemented. This has also been underlined by concerns in response to the call for evidence related to the DSA (European Commission, 2023; see similarly van Drunen & Noroozian, 2024).

In this paper, we combine perspectives from legal scholars and social scientists engaging in data donation studies. First, we discuss the data access obligations in the GDPR, specifically legal requirements to provide accessible, transparent, and complete data. Second, we draw on the experiences of social scientists who engage in data donation studies to identify challenges resulting from platforms' inadequate compliance with such requirements. To do so, we combine a structured review of challenges researchers encountered across data donation projects with a study on variation in data request/access. Third, by combining a legal and social scientific perspective, we formulate four recommendations for the enforcement of data access rights to improve platform user empowerment and data donation studies.

Our goal is twofold: first, we aim to stimulate discussions among social scientists relying on the right of access to conduct data donation studies. Second, we translate our concerns for a broader group of stakeholders, including companies, data protection authorities, and policymakers. By combining both perspectives, we engage in interdisciplinary efforts to improve data access, as demanded, for instance, by Tromble (2021).

2. Legal background to the right of access in the GDPR: Why and how platforms need to provide users access to their data

2.1 Normative underpinnings of the right to access

The exponential growth of data brings about social, legal, and ethical concerns related to the asymmetries of information and power (Beer, 2017; Kitchin, 2017). Information asymmetries result from the complexity of data infrastructures and engineered opaqueness by those controlling them (Ausloos & Veale, 2021; for an extended discussion, see Nieborg et al., 2024). Power asymmetries stem from the ability to exploit infrastructures in light of commercial or political imperatives at the expense of individuals, communities, and society at large (Giannopoulou et al., 2022). One of the key objectives of data protection law is to challenge these asymmetries by providing effective and complete protection of the fundamental rights of natural persons with respect to the processing of personal data (Case C-131/12; Case C-73/16). The right of access constitutes a cornerstone in this regard (Case C-553/07). As a first objective, this right functions as an emancipatory legal tool. It empowers platform users by making data infrastructures visible, allowing them to govern the use of their data and enabling them to exercise other rights (e.g. to have data erased or ported to other platforms; see also C-434/16 as well as joint cases C-141/12 and C-372/12). A second objective is to enable platform users to monitor platforms' compliance with the GDPR (Case C-434/16; Recital 63 GDPR), including lawfulness, purpose limitation, data minimisation, accuracy, and storage limitation. The right of access is intent-agnostic, meaning that it does not require a motivation vis-à-vis platforms (Case C-307/22; Mahieu, 2023). It can be invoked to safeguard interests, rights, or freedoms, such as non-discrimination.

2.2 Requirements of the right to access

We now briefly introduce the legal framework with which data platforms must comply when responding to users' access requests (for an overview, see Mahieu, 2023). This framework is dynamic, as its interpretation is the topic of a growing body of case law by national courts, the Court of Justice of the European Union (CJEU), and data protection authorities (DPAs).

Art. 15 of the GDPR contains *content* requirements that specify what data platforms need to provide on request, for example, the purpose of the data processing, the type of data processed, or the recipients of the disclosed data. Art. 12 mainly defines *formal* requirements (including obligations) that have been further clarified by the European Data Protection Board (EDPB, 2023). In short, platforms must inform platform users in a concise and easily accessible way, information must be presented in a transparent way, and platforms must provide complete information.

2.2.1 Concise and easily accessible form

GDPR Art. 12(1) requires platforms to provide access in a concise, transparent, intelligible, and easily accessible form, using clear and plain language. As explained by the EDPB (2023), "the controller should provide appropriate and user-friendly communication channels that can easily be used by the data subject” (p. 3).

2.2.2 Transparency

It is critical to consider data rights in light of the GDPR principle of transparency. Platforms are obliged to support users in their efforts to benefit from the protection of their fundamental rights; they have to actively implement data rights (Art. 25(1) GDPR) and facilitate the exercise of such rights (Art. 12(2) GDPR). These requirements are closely linked to the right of access according to the transparency obligations required by Art. 5(1)(a) and Art. 12(1) (see also Art. 29 Working Party, 2018). Platform users are entitled to be informed, at an individualised level, about data processing. This means that platforms are required to tailor responses to access requests to specific users. This is especially relevant for the data categories listed in GDPR Art. 15(1), such as the processing purposes of personal data, the identity of who the data have been shared with, what has been shared, etc. Mere referrals to platforms’ privacy policy, often phrased in generic terms, do not meet this requirement.

2.2.3 Completeness

Platform users are entitled to full disclosure on all information processed by platforms, including (i) information relating to *content* requirements laid down in Art. 15(1)–(3) and (ii) information held throughout platforms’ IT and non–IT filing systems. Platforms might enable users to download data by providing self-service tools; however, such tools should never limit the scope of users’ entitlements. Platforms could provide users with vast data files that may be complete but do not meet the conditions of information being offered in a transparent and easily accessible manner (see Section 2.2.1). Instead, platforms might consider providing layered information while bearing in mind that providing only a summary will not meet the requirement of completeness. Depending on the reasonable expectations of users, platforms must select a strategy that meets this requirement and demonstrate that it adds value to users (Ausloos et al., 2020; EDPB, 2023). In any case, when asked to provide *all* personal data, platforms cannot limit responses to parts of the respective data. Even more, platforms may be required to provide additional information necessary to understand respective data sets (see also Case C-487/21).

3. Methodology

These three requirements will now serve as points of reference to discuss challenges in data donation studies that emerge from platforms’ non-compliance with legal frameworks. To identify challenges, we rely on a structured review of challenges researchers encountered in data donation projects and a systematic study on data request/access across platforms.

3.1 Structured review

We systematically reviewed challenges encountered in previous or on-going data donation projects. Researchers engaged in such projects were identified through related conferences, such as yearly data donation symposia. Second, researchers who had already published related projects were contacted as well. Third, and as a form of snowball sampling, identified researchers were asked to think of additional scholars who conduct similar projects (but may not yet have published related work). By doing so, a total of twelve data donation projects were identified. These projects have been running across different years (2018–2024) and collecting DDPs from various digital platforms (Facebook, Google, Instagram, Netflix, TikTok, WhatsApp, X/Twitter, YouTube; for an overview, see Table 1). However, and as should be noted critically, most projects were conducted by researchers in Western Europe, thereby limiting the generalisability of our results, since data donation studies are also employed elsewhere (Ejaz et al., 2023; Garimella & Chauchard, 2024; Wu-Ouyang & Chan, 2023).

Researchers identified challenges they had encountered in their data donation projects through virtual brainstorming. Here, researchers described relevant challenges and noted examples for how these became prevalent. Lastly, a core team of authors discussed and sorted the fourteen challenges into the legal requirements of concise and easily accessible form (see Challenge I–VII), transparency (see Challenge VIII–XI), and completeness (see Challenge XII–XIV). Table 2 depicts challenges encountered across projects (for details, see Tables A2.1, A2.2, A3, and A4 Supplement).

3.2. Study on data request/access across platforms

Related to the requirement of concise and easily accessible form, we further analysed Challenges I–VII to understand how data requests and data access differ across platforms and time (for similar approaches see Pins et al., 2022; Syrmoudis et al., 2021, 2024; Wong & Henderson, 2019). We requested and downloaded DDPs for those six platforms most frequently studied across identified projects (Facebook, Instagram, TikTok, Twitter/X, WhatsApp, and YouTube; see Table 1). As part of the D3I volatility project (Carrière et al., 2024), initial data collection was conducted between January and May 2023 by researchers in the Netherlands. To account for variation over time as well as within-country variation, two researchers at

Utrecht University and two researchers at LMU Munich repeated the process in May 2024. Table 3 depicts the results of this updated data collection.

4. Challenges in data donation studies due to platform non-compliance

Based on our findings, platforms seemingly fail to comply with requirements related to the right of access. This affects not only platform users, who may have trouble accessing their data, but also researchers, who face increasing drop-out rates and biases in data donation studies (see for example Hase & Haim, 2024; Pfiffner & Friemel, 2023). Most importantly, by impeding access rights and, consequently, data donation studies, platforms hamper studies on questions of societal relevance, such as platform users' exposure to misinformation, how platforms affect individual well-being, or the role of algorithmic recommendation (see further Ohme et al., 2024).

TABLE 1: Overview of data donation projects by authors. Note: ID describes the number of the project (listed in chronological order)

ID	PROJECT DESCRIPTION	DATA COLLECTION	PLATFORM(S)	REFERENCE
1	Identifying depression-related behaviour online	2018	Facebook	Kmetty & Bozsonyi (2022)
2	Information behaviour related to political referendums	2021	Google, YouTube	Blassnig et al. (2023)
3	Detecting algorithmic bias & fringe bubbles	2022	YouTube	Möller et al. (2023)
4	Evaluating a data donation application in a survey & field study	2022	Google, YouTube	Welbers et al. (2024)
5	Willingness & nonparticipation in data donation studies	2022	Google	Struminskaya (2022)
6	News engagement on social media	2022–2023	Facebook, Instagram, Twitter/X, YouTube	Haim et al. (2023), Haim & Hase (2024)
7	Digital political footprints	2023–2024	Facebook, Google, Instagram, TikTok, YouTube	Centre for Social Sciences (n.d.)
8	Assessing WhatsApp networks with donated data	2023	WhatsApp	– (analysis ongoing)
9	Information exposure on social media	2023	TikTok	Wedel et al. (forthcoming)
10	Social influence, health behaviour, & social media	2023–2024	TikTok, YouTube	Pfiffner et al. (2024a)
11	Exposure to & engagement with news/politics during Dutch elections	2024	TikTok, YouTube	– (analysis ongoing)

ID	PROJECT DESCRIPTION	DATA COLLECTION	PLATFORM(S)	REFERENCE
12	Behind the screens – exploring data donations via Netflix	2024	Netflix	– (analysis ongoing)

TABLE 2: Challenges in data donation studies

ID	CHALLENGES						
	I. INSTANCE RESTRICTIONS	II. INTERFACES	III. LIMITING REQUESTED DATA	IV. DIVERSE DDP FORMATS	V. MULTIPLE ACCESS REQUESTS	VI. NO NOTIFICATION ABOUT DDPS	VII. LIMITED ACCESS OVER TIME
	Challenges I–VII related to requirement: Concise and easily accessible form						
1				X			
2	X		X	X			
3							
4		X		X		X	X
5			X				X
6	X	X	X	X		X	X
7	X	X	X	X		X	
8	X	X	X				X
9	X	X	X	X		X	X
10	X		X	X		X	X
11	X	X	X	X		X	X
12	X	X	X				

	VIII. NO INFORMATION ON COMPLETENESS OF DDPS	IX. NO INFORMATION ON DDP STRUCTURE	X. NO INFORMATION ON DDP MEASUREMENTS	XI. NO INFORMATION ON DDP CHANGES	XII. MISSING INFORMATION ON ACTIVITIES	XIII. MISSING INFORMATION ON CONTENT	XIV. MISSING INFORMATION ON CONTEXT
	Challenges VIII–XI related to requirement: Transparency				*Challenges XII–XIV related to requirement: Completeness*		
1	X	X	X	X	X		X
2	X	X	X	X			X
3	X						X
4	X	X		X			X
5	X	X	X				
6	X	X	X	X	X	X	X
7	X	X	X	X	X	X	X

	VIII. NO INFORMATION ON COMPLETENESS OF DDPS	IX. NO INFORMATION ON DDP STRUCTURE	X. NO INFORMATION ON DDP MEASUREMENTS	XI. NO INFORMATION ON DDP CHANGES	XII. MISSING INFORMATION ON ACTIVITIES	XIII. MISSING INFORMATION ON CONTENT	XIV. MISSING INFORMATION ON CONTEXT
8	X		X	X			X
9	X		X	X	X	X	
10	X	X	X		X	X	X
11	X		X	X	X	X	X
12	X	X	X	X			

TABLE 3: Variation in data request/access across platforms[3]

FACEBOOK	INSTAGRAM	TIKTOK	TWITTER/X	WHATSAPP	YOUTUBE
Challenge I: *How can users access the platform (reg. use) & how can they request/access their DDPs?*					
Reg. use: App, browser DDP request: App, browser DDP access: App, browser	Reg. use: App, browser DDP request: App, browser DDP access: App, browser	Reg. use: App, browser DDP request: App, browser DDP access: App, browser	Reg. use: App, browser DDPs request: Browser DDP access: Browser	Reg. use: App, browser DDP request: App (account, channel reports, chat), browser (account, channel reports) DDP access: App (account, channel reports, chat), browser (account, channel reports)	Reg. use: App, browser DDP request: Browser DDP access: Browser
Challenge II: *Which verification procedures can platform users encounter after signing in when accessing their DDPs?*					
Request: None Download: Password	Request: None Download: Password	Request: None Download: Password	Request: Password & 2FA, copy of ID Download: Password, 2FA	Request: None Download: None	Request: None Download: Password
Challenge III: *Can platform users limit DDPs to specific data points?*					
Type of data, time	Type of data, time	Type of data	No	Type of data	Type of data
Challenge IV: *Can platform users specify file formats for DDPs?*					
HTML, JSON	HTML, JSON	JSON, TXT	No	No	HTML, JSON (depending on type of data)
Challenge V: *Do platform users have to make multiple access requests for DDPs?*					
No	No	No	No	Yes: account, channel report, chats	No
Challenge VI: *How are platform users notified when DDPs are accessible?*					

[3] For WhatsApp, requesting access to account information or channel activities differs from exporting chats, with the latter being related to data portability.

FACEBOOK	INSTAGRAM	TIKTOK	TWITTER/X	WHATSAPP	YOUTUBE
Email, push notification	Email	Push notification	Email, push notification	Push notification (account, channel report); chat can be exported right away	Email
Challenge VII: *How long can platform users access DDPs?*					
Four days	Four days	Four to five days	Seven days	Thirty days	Seven days

4.1 Concise and easily accessible form

According to GDPR Art. 12(1), platforms should present DDPs so that platform users can easily understand their content. However, we identified seven challenges impeding the fulfilment of this requirement (for variation in request/access, see Table 3; for Challenges I–VII, see Table 4).

TABLE 4: Challenges related to concise and easily accessible form

REQUIREMENT: CONCISE AND EASILY ACCESSIBLE FORM	
DEFINITION	Platforms must present DDPs so that users can easily understand their content.
LEGAL BASIS	GDPR Art. 12(1)
CHALLENGES	I. *Instance restrictions*: Platform users can only request/access DDPs via certain instances. II. *Interfaces*: Platform users may encounter different interfaces/verification. III. *Limiting requested data*: Platform users cannot limit requests to specific data points. IV. *Diverse DDP formats*: Platform users are presented with diverse file formats. V. *Multiple access requests*: Platform users have to submit multiple requests to obtain all their data. VI. *No notification about DDPs*: Platform users are not informed when DDPs are accessible. VII. *Limited access over time*: Platform users can only access DDPs for a limited time.

Challenge I. Instance restrictions

Social media users often consume content from platforms via smartphone apps. However, our review indicates that if users want to request or download their data from these platforms, they can often only do this via web applications accessible through browsers, for example via desktop computers. In two-thirds of the data donation projects, researchers encountered such restrictions (see Table 2). For example, social media users often consume YouTube content via smartphone apps. However, if they want to access their data, such requests can only be filed through the web application in the browser where they have to login again and with which they may not be familiar. While this was more often the case up until 2023 (see projects 6–7, Table A.2.1), our study on data request/access indicates that most platforms – except for Twitter/X and YouTube – have lifted instance restrictions by May 2024 (see Table 3).

For users, such additional steps pose an unnecessary burden. Importantly, platforms cannot require that users submit requests in a certain form (EDPB, 2023). Ease of accessibility implies that users should be able to submit access requests through interfaces based on which they use platforms. For researchers, such restrictions are difficult in that they lower response rates: respondents often report that technical difficulties, for example, switching application instances between platform use and data requests, lead them to drop-out in studies related to digital trace data (Gil-López et al., 2023). Attrition may not only increase costs for data donation studies. Systematic drop-out may lead to biased samples, for example if technically less savvy users cannot participate (Hase & Haim, 2024). Overall, instance restrictions could thus limit the generalisability of findings stemming from data donation studies.

Recommendation: Platforms should enable data request/access via all instances through which platform users can use their infrastructure (i.e. via apps and browsers).

Challenge II. Interfaces

Platform users may encounter different interfaces when requesting or accessing DDPs: when submitting access requests through an app, for example WhatsApp, interfaces can depend on operating systems (see project 8, Table A.2.1). Moreover, platforms sometimes perform design experiments, leading to further variation (see project 7, Table A.2.1). Additionally, verification procedures for accessing DDPs often vary: some platforms ask for verification via two-factor authentication (2FA) or even hardcopy identification via official IDs (see projects 4, 6, 9, or 11, Table A.2.1; see similarly Pins et al., 2022; Syrmoudis et al., 2021; Wong & Henderson, 2019).

While, from a user perspective, data should be appropriately secured, varying security measures make it challenging for researchers to guide users through access requests. Respondents who are concerned about their privacy are often less likely to participate in data donation studies. As such, heightened security measures, such as sending in an ID, may further bias studies in that privacy-concerned participants drop-out (Hase & Haim, 2024), which would limit the generalisability of findings.

Recommendation: Platforms should standardise interfaces and verification procedures across instances (see Challenge I), operating systems, user profiles, and platforms. Platforms could facilitate access by providing a consistent link where subjects can request data. Moreover, platforms should refrain from interface design experiments or they should announce such as part of public data documentation (see also Section 4.2).

Challenge III. Limiting requested data

Often, platform users cannot indicate which type of data or for which time data should be included. In more than two-thirds of the data donation projects, researchers encountered related challenges (see Table 2), many of which are still prevalent according to our study on data request/access in May 2024 (see Table 3).

From a data protection-by-design perspective, platform users should be able to download only parts of their data (Syrmoudis et al., 2024). This allows users to feel empowered, for example, because they can exclude sensitive data. Moreover, downloading all data increases the risks of data breaches on users' devices. For researchers, limiting access requests is relevant as this could reduce the size of data, which makes it easier to process DDPs. According to our review, large DDPs often led to participants failing to upload their DDPs, which may introduce bias (see project 6, Table A.2.1).

Recommendation: Platforms should allow access requests to be restricted to specific data points while retaining the ability to download all data (see Section 4.3.). Filtering options should be tied to a consistent link for data access requests.

Challenge IV. Diverse DDP formats

Only some platforms provide options to choose the format in which DDPs can be accessed. This includes JSON formats (JavaScript Object Notation formats) or CSV formats (Comma-Separated Values formats), which are more easily readable for machines. It extends to HTML formats (Hypertext Markup Language formats), which are more easily readable for humans (see similarly Pins et al., 2022; Syrmoudis et al., 2021; Wong & Henderson, 2019). According to our review, researchers considered this a challenge in two-thirds of the data donation projects (see Table 2). To date, most platforms still provide diverse formats, as indicated by our study on data request/access (see Table 3).

Although there is no explicit requirement for machine readability in the GDPR, its intention to enable platform users to port data certainly prompts a machine-readable option vis-à-vis the explicitly required readability for humans. For users, this means that both human- and machine-readable formats should be provided to understand but also port data. For researchers, diverse file formats are, however, problematic in that users may wrongly request DDPs in human- instead of machine-readable data, which requires increased resources for data processing via

DDTs.

Recommendation: Platforms should offer at least one machine-readable format. They should turn this file-format selection into an opt-out prompt so that a machine-readable option is included by default, while also providing a human-readable option.

Challenge V. Multiple access requests

Some platforms do not enable users to request their data via single access requests. For example, if users want to access WhatsApp data, they have to export each chat separately (see project 8, Table A.2.2).

For users, multiple access requests mean more effort. In turn, researchers may have to grapple with missing data precisely because users do not want to engage in multiple access requests. Oftentimes, users have trouble finding respective buttons on digital platforms (Pins et al., 2022). For example, if participants have to export each WhatsApp chat separately, this complicates the analysis of social interactions (Kohne & Montag, 2023).

Recommendation: Platforms should centralise access requests via a single link. Here, platform users should be able to request all their data via a single request and download it in a single DDP.

Challenge VI. No notification about DDPs

According to the GDPR, platforms must comply with access requests within 30 days. In practice, large platforms often comply within a couple of days, as indicated by our review (see projects 4 and 10, Table A.2.2; see also Wong & Henderson, 2019). However, time periods between platform users requesting their data and DDPs being accessible differ across platforms and users. Neither platform users nor researchers can estimate when DDPs will be ready. This is cumbersome, especially since DDPs are often deleted after a few days (see Challenge VII).

Moreover, platforms often use different means – e.g. emails or push notifications – to inform platform users that DDPs are available, as indicated by our study on data request/access (see Table 3). If platform users rarely use platforms, such notifications may go unnoticed and DDPs may be deleted before they can be accessed. Some platforms, such as TikTok, did not provide notifications in previous years (see projects 10 and 11, Table A.2.2), although this seems to have changed by May 2024 (see Table 3).

Recommendation: Platforms should inform users at the moment of their access requests of when their DDPs will be available and send notifications once DDPs are accessible. Preferably, this should be done via email rather than notifications on platforms, or in combination, as users who rarely use platforms may miss the latter.

Challenge VII. Limited access over time

Access to DDPs often expires a couple of days – according to the review (see Table 2) and the study on data request/access between four days and thirty days after the data became available (see Table 3).

As such, platform users may miss the opportunity to download and store their data, especially if platforms do not notify them (see Challenge VI). For researchers, this complicates the process of data collection: they have to remind participants to download and donate their DDPs within this time frame, which may lead to low response rates in data donation studies.

Recommendation: Platforms should standardise and extend the time during which DDPs can be downloaded. This should be a more reasonable amount of time, at least thirty days, especially since platforms themselves need to respond to access requests within thirty days.

4.2 Transparency

According to GDPR Art. 5(1)(a) and Art. 12(1), platforms must transparently provide platform users with information on data processing. However, we identified four challenges impeding the fulfilment of this requirement (for Challenges VII–XI, see Table 5).

TABLE 5: Challenges related to transparency

REQUIREMENT: TRANSPARENCY	
DEFINITION	Platforms must, at an individualised level, provide information about data processing.
LEGAL BASIS	GDPR Art. 5(1)(a) & Art. 12(1)
CHALLENGES	VIII. *No information on completeness of DDPs*: Platform users lack information on the completeness of DDPs. IX. *No information on DDP structure*: Platform users lack information on the structure of DDPs. X. *No information on DDP measurements*: Platform users lack information on how measurements were created. XI. *No information on DDP changes*: Platform users lack information on changes concerning DDPs.

Challenge VIII. No information on completeness of DDPs

To assess whether platforms provide users with complete data, it is necessary to determine what information platforms collect and store (Rau, 2023). Without this information, users cannot know whether platforms fulfilled their obligation. According to our review, researchers in all twelve projects critically discussed that it was unclear if any data is missing, mostly because platforms did not provide information on what data was collected in the first place (see Table 2). Some projects indicated that it was clear that data was missing, for example, because participants indicated that data was absent from their DDPs (see project 4, Table A3), similar to existing studies (Syrmoudis et al., 2024).

For users, this is problematic as they cannot know whether they can access all data platforms collected on them. For researchers, missingness could introduce bias, for example, if data is systematically missing for some participants or platforms but not others.

Recommendation: Platforms should publish a complete and legally binding list of the information (i.e. "variables"/"features") they collect and store about users as part of public data documentation.

Challenge IX. No information on DDP structure

Almost equally as often, researchers encountered the challenge of platforms lacking transparent information on what files in DDPs signify (see Table 2). While some file names are seemingly self-explanatory, others are not. Furthermore, DDPs sometimes combine different types of information in the same list, making it difficult for users to understand where they can access which information. For example, watch and search histories stored in DDPs from YouTube contain a mix of forced-to-view content, such as advertisements, and self-selected content (see projects 6 and 10, Table A3). For users, this makes it hard to understand their DDPs. For researchers, this may introduce measurement error, for example, when studying how users select and consume information on platforms. If they do not know which content users choose to watch (as opposed to being shown as part of advertisement), this complicates analysis, for example, on the effects of political campaigns on platforms during elections.

Second, platforms use different file or variable names in DDPs, depending on users' devices or account settings. For example, Google relies on localised file names, while Instagram uses localised variable names according to individual language settings (see projects 2 and 4, Table A3). This complicates data portability for users. Similarly, it may introduce measurement error in data donation studies.

Thus, platforms should be transparent about how account settings affect DDPs. Ideally, data provided in a machine-readable format, such as JSON or CSV, should always look the same, regardless of individual settings.

Recommendation: Platforms should provide a description of all the variables collected as part of public data documentation, including a description of files in DDPs (van Driel et al., 2022). This extends to using the names of standardised files, variables, and values in machine-readable DDPs, for example, via consistent date formatting according to ISO 8601. These recommendations have also been supported by recent CJEU case law, which explained that additional information may have to be provided “where the contextualization of the data processed is necessary in order to ensure the data are intelligible” (Case C-487/21, para. 41).

Challenge X. No information on DDP measurements

Platforms do not provide information on how they measure variables (Nonnecke & Carlton, 2022; Rau, 2023), an issue mentioned throughout our review. One example is information on exposure, such as “seen videos” (example file from YouTube DDPs). Here, it is unclear whether the content is marked as “seen” once it appears on a user’s display or only after the content was displayed for a certain amount of time. Moreover, DDPs can also include information that is algorithmically derived. Such information is not a direct measurement of actions by users (e.g. clicking a link) but results from algorithmic classifications based on, for example, users’ past actions. An example would be “inferred interests” for advertisements (example file from Instagram DDPs). Without contextual information on classification algorithms, it is difficult for users to understand such data (Cotter et al., 2021; Rieder & Hofmann, 2020).

For users, missing information on measurements makes it impossible to understand what information platforms collect on them. For researchers, it complicates addressing measurement error and, thus, may bias results. For example, when spotting that algorithmically inferred travel modes in DDPs from Google are likely invalid (e.g. measures indicating that users completed a trip of 200 km by foot in 30 minutes, see project 5 in Table A3), researchers cannot correct such errors because they do not know how measurements were created.

Recommendation: As part of public data documentation, platforms should provide a description of how measures were created. This includes indicating when measurements were algorithmically inferred, in which case platforms should provide information on the algorithms used for inferences, the variables used as inputs for classifications, and what the potential outputs are.

Challenge XI. No information on DDP changes

Platforms often change the content and structure of DDPs. This includes altering variable names, file names, or the format in which the information is provided, adding new variables or removing variables. Such changes can make it difficult to compare DDPs across time (Rieder & Hofmann, 2020; van Driel et al., 2022). Additionally, unexpected changes add obstacles for researchers engaged in data donation studies. For example, Google started to provide different file formats throughout a data donation study from our review without any information prior to this change. When researchers contacted the platform, they did not receive any response on how to fix this (see project 2, Table A3).

For users, shifts in how and what data is provided make it harder to access data. For researchers it could – as a worst-case-scenario – lead to a sudden halt of ongoing data donation studies.

Recommendation: Platforms should announce and document changes made to the structure of DDPs and included measurements before they take effect – at least 30 days in advance – as part of public data documentation.

4.3 Completeness

According to Art. 15(1)–(3) of the GDPR, users are entitled to full disclosure of all available information that platforms process on them. This includes metadata (e.g. which data is collected and how it is used) according to Art. 15(1) and information on data transfers (e.g. whether this data is transferred to other organisations) according to Art. 15(2). It extends to a copy of the data platforms collect according to Art. 15(3), which should include all this information. However, we identified three challenges impeding the fulfilment of this requirement (for Challenges XII–XIV, see Table 6).

TABLE 6: Challenges related to completeness

REQUIREMENT: COMPLETENESS	
DEFINITION	Platforms must grant platform users access to all the information they collect, process, and store on platform users.
LEGAL BASIS	GDPR Art. 15(1)–(3)
CHALLENGES	XII. *Missing data on activities:* Platform users cannot access complete information on activities. XIII. *Missing data on content:* Platform users cannot access complete information on content. XIV. *Missing data on contexts:* Platform users cannot access complete information on the contexts in

REQUIREMENT: COMPLETENESS	
	which they engaged in activities or with content.

Challenge XII. Missing data on activities

Oftentimes, platforms provide incomplete information on user activities (van Driel et al., 2022), such as the overall time spent on platforms. This extends to more passive activities, such as the content that users watched or were exposed to. In previous years, such exposure information was, for example, not included in the DDPs from Facebook or Twitter/X (see projects 1 or 9, Table A4) although newer DDPs now include this information, which indicates that such data is, in fact, collected by platforms.

For users, this means that they cannot see what activity-related data platforms process on them. For researchers, it undermines studies on questions of societal relevance. For example, data donation studies cannot study exposure to misinformation on digital platforms, as exposure data is often missing from DDPs (Ohme et al., 2024). To understand the role of misinformation on social media during elections, researchers currently rely on other means of data access, like research collaborations with platforms (Wagner et al., 2023).

Recommendation: Platforms should provide complete information on all activities related to platform users, including the time they spent on a platform as well as passive (e.g. watching content) and active (e.g. searches) activities.

Challenge XIII. Missing data on content

Platforms often provide incomplete information regarding the content that users were exposed to or actively interacted with. DDPs from YouTube, for example, often contain links to or IDs of accounts that platform users follow rather than account names (see projects 10–11, Table A3), although such information is available as meta data.

For users, this means that they would have to look up such content manually, which requires an unrealistic degree of effort to understand, for example, how they were targeted by advertisers on digital platforms. Even more troublesome, researchers have to look up such information via APIs, given the amount of data included in DDPs. Seeing that most platforms shut down or drastically restrict their APIs (Bruns, 2019; Freelon, 2018), this renders data donation studies unfeasible.

Recommendation: Platforms should provide complete information on content, both related to what the content is about and the names of the accounts publishing it. Information that can only be provided by APIs should be included in DDPs.

Challenge XIV. Missing data on contexts

Lastly, platforms rarely provide access to necessary context information for users to fully understand their behaviour on digital platforms, contrary to the GDPR requirements specified by the CJEU (Case C-487/21). Most DDPs include a list of user activities or the content users engaged with, but it remains unclear how such engagement came about, as indicated by our review (see Table 2): DDPs do not indicate whether the content users engaged with was recommended to them or whether they encountered it elsewhere, for example, through messages by friends. In turn, it is impossible for users to understand non-exposure or non-engagement: since DDPs do not contain information on which videos were visible or recommended to them, they cannot determine which content they could have watched or liked but did not.

This lack of information is especially troublesome in light of the DSA where Art. 34(1)(c) underlines the importance of assessing the systemic risks created by platforms, including “actual or foreseeable negative effects on civic discourse and electoral processes and public security”. For example, to understand the spread of political ads or misinformation during electoral processes, researchers require data on how such content was promoted to users. Here, the DSA offers a pathway for data access that may strengthen data donation studies. To understand how civic discourses emerge and may be harmed, both platform users and researchers need access to contextual information, including feeds as streams of information on platforms.

Recommendation: Platforms should provide contextual information on activities and content, for instance, as metadata (i.e. whether or not content was recommended to platform users by platforms; timestamps) and by providing access to feeds on platforms.

5. Improving the right of access: Four recommendations for the road ahead

Our empirical investigation underscores that platforms’ non-compliance with the GDPR data access rights has a significant impact on users, but also researchers, engaging in data donation studies. While platforms allow platform users to exercise these rights *de facto* somehow, platforms' compliance does not necessarily hold up with a stricter *de-jure* interpretation (Ausloos et al., 2020; de Vreese & Tromble, 2023). By complicating the exercise of data access rights (in a concise and easily accessible form), by not communicating what various data points mean or where they originate from (transparency), and by providing only some of the data they collect on users (completeness), platforms constrain the scope of the right of access.

This becomes even more problematic as other roads to data access, for example, collaborations between researchers and platforms, cannot – and, arguably, should not – become the standard (Wagner, 2023). Such collaborations depend on the goodwill of platforms and limit the degree of control researchers have over samples or measurements. Moreover, researchers involved in such collaborations often stem from more resourceful countries from the Global North, which may narrow research foci and the generalisability of results, as Parry (2024) critically points out.

Instead of being able to conduct research with data that should be available pursuant to the GDPR, researchers are, to date, forced to argue and deal with illegitimate behaviour from platforms. Platform users, in turn, cannot use DDPs to understand their behaviour on digital platforms. We think that these challenges *could* – and in light of our legal understanding of the matter *should* – be overcome. This would be for the greater good of platform users, research and, finally, society, where solutions to problems such as the spread of misinformation require independent research. To illustrate the road ahead, we propose four central recommendations (see Figure 1) for how platforms could (and should) facilitate data donation studies.

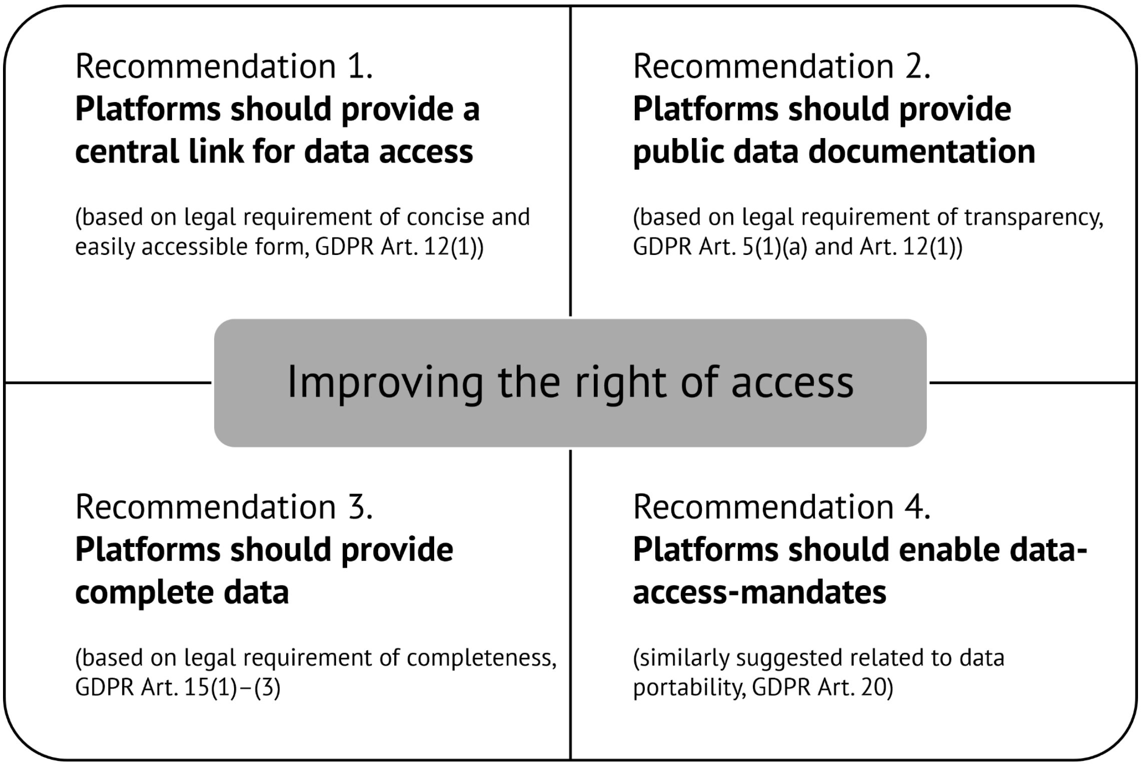
Recommendation 1.
Platforms should provide a central link for data access
(based on legal requirement of concise and easily accessible form, GDPR Art. 12(1))
Recommendation 2.
Platforms should provide public data documentation
(based on legal requirement of transparency, GDPR Art. 5(1)(a) and Art. 12(1))
Improving the right of access
Recommendation 3.
Platforms should provide complete data
(based on legal requirement of completeness, GDPR Art. 15(1)–(3)
Recommendation 4.
Platforms should enable data-access-mandates
(similarly suggested related to data portability, GDPR Art. 20)

Figure 1: Key recommendations.

Recommendation 1. Platforms should provide a central link for data access

Related to Challenges I–VII and the requirement to provide data in a concise and easily accessible form, platforms must facilitate data access. This includes the possibility of exercising data access rights independent of instances, individual settings, or usage. Across and between platforms, this also requires interoperable formats, such as machine-readable DDP's consisting of JSON or CSV files in ZIP packages via an opt-out function. DDPs should be available for at least 30 days. Lastly, platforms should standardise security measures, including how users are notified about their data being available. A persistent link where platform users can request their data, including the possibility to filter for data points, would benefit users and researchers who could more easily direct study participants to the right location in data donation studies. In line with common "forgotten password" security measures, such a persistent link could request that platform users provide their login credentials (e.g. their email address or phone number) and verification via 2FA (e.g. via smartphones for app-based platforms or via email for browser-based platforms).

Recommendation 2. Platforms should provide public data documentation

Related to Challenges VIII–XI and the requirement of transparency, platforms should not be the de-facto gatekeepers of research agendas. While it is not the primary goal of the GDPR to facilitate research, recent frameworks such as the DSA have further strengthened data access rights, especially for researchers (Leersen, 2024) and for questions of high societal relevance. As such, platforms should facilitate independent research by providing public data documentation detailing data origins, levels of retrievable data, operationalisations, measures of completeness, and transparency about algorithmic inference. This includes standardising the names of files, variables, and values in DDPs. Any changes to data structures and measures should be communicated in advance.

Recommendation 3. Platforms should provide complete data

Related to Challenges XII–XIV and the requirement of completeness, platforms must provide complete data. This includes data on activities, content, and contextual information, for example as-of-yet often missing data on what content users are exposed to on digital platforms or how they are exposed to such content.

Recommendation 4. Platforms should enable data-access mandates

Lastly, we consider data-subject-centric designs the most beneficial route forward for platforms, users, and researchers. While our first three recommendations are more closely tied to existing legal requirements, a mandate for data portability has been suggested by Art. 20 of the GDPR. Deployed through platforms, users could be equipped with the ability to issue data-access mandates, signalling platforms to provide access to data on their behalf (see similarly Nonnecke & Carlton, 2022). Portability could allow users to issue mandates to researchers to access, for example, exposure data within a given time frame or advertisement users were targeted with. This would enable users to reign over their data and who shall have access to it. For platforms, it would provide a consistent pattern of data provision and specify clear boundaries alongside which they can operate sustainably. Finally, it would allow researchers access to machine-readable data without the burden of supporting a wide variety of access-right exercises while ensuring users' approval. In turn, this would benefit researchers' autonomy when studying digital public spheres, civic discourse, and democratic processes and how platforms influence these dynamics.

6. Conclusion

Our study illustrates that platforms do not comply with legal requirements related to data access rights: they do not provide accessible, transparent, and complete data to platform users. Not only could this undermine users' trust in and use of platforms (Syrmoudis et al., 2024). It also hampers research aiming to address questions of societal relevance, such as how misinformation on digital platforms may interfere with elections or the reduced digital well-being of users as an outcome of platform usage. As Parry notes, research and, as such, society at large, currently risks "being left behind without any robust means of studying the potential systemic risks at play" (2024, p. 2). Independent research could address these issues – and platforms could (and should) support this through compliance with data access rights.

Our four central recommendations could be implemented in several ways. First, researchers and policymakers should lobby for their inclusion in legal frameworks, for example related to the upcoming delegated act on data access in the DSA (Windwehr & Selinger, 2024). This includes clear guidelines on what data and documentation platforms should provide (Jaursch et al., 2024). Second, researchers themselves could and should develop guidelines on data access rights, for example protocols on data handling (Tromble, 2021; see for example European Digital Media Observatory, 2022) or research ethics (Lukito, 2024). Third, researchers and especially policymakers should monitor (Leersen, 2024) and, where applicable, sanction non-compliance with data access rights. Recent examples include observing users during data access requests (Pinks et al., 2022; Syrmoudis et al., 2024) or providing status reports for researchers on data access across platforms (European Commission, 2024a). In terms of sanctioning, the investigation of the European Commission related to Meta's non-compliance with the DSA offers a warning for platforms going forward (European Commission, 2024b). Lastly, researchers, policymakers, and platforms should collaborate to extend existing and propose new infrastructures for improving data access (Jaursch et al., 2024; van Drunen & Noroozian, 2024). Here, research institutions and politics need to provide funding for research infrastructures to ensure that a diversity of researchers can partake in data access (Nonnecke & Carlton, 2022).

ACKNOWLEDGEMENTS

We thank further contributors, namely Johannes Breuer, Erik Knudsen, Daniel Oberski, Amy Orben, Maas van Steenbergen, and Wouter van Atteveldt, for adding further valuable insights.

References

Article 29 Data Protection Working Party. (2018). *Guidelines on transparency under Regulation 2016/679* (No. WP260 rev.01). https://ec.europa.eu/newsroom/article29/items/622227/en

Ausloos, J., Mahieu, R., & Veale, M. (2020). Getting data subject rights right. A submission to the European Data Protection Board from international data rights academics, to inform regulatory guidance. *JIPITEC – Journal of Intellectual Property, Information Technology and E-Commerce Law*, *10*(3), 283–309. https://www.jipitec.eu/archive/issues/jipitec-10-3-2019/5031

Ausloos, J., Meiring, A., Buijs, D., van Echeoud, M., Boss, S., & Strycharz, J. (2023). *Information law and the digital transformation of the university. Part II. Access to data for research* [Report]. Institute for Information Law. https://www.uva.nl/binaries/content/assets/uva/nl/over-de-uva/over-de-uva/beleid-en-financien/digitale-agenda/part-ii-access-to-data-for-research.pdf

Ausloos, J., & Veale, M. (2021). Researching with data rights. *Technology and Regulation*, *2*, 136-157 Pages. https://doi.org/10.26116/TECHREG.2020.010

Beer, D. (2017). The social power of algorithms. *Information, Communication & Society*, *20*(1), 1–13. https://doi.org/10.1080/1369118X.2016.1216147

Blassnig, S., Mitova, E., Pfiffner, N., & Reiss, M. V. (2023). Googling referendum campaigns: Analyzing online search patterns regarding Swiss direct-democratic votes. *Media and Communication*, *11*(1), 19–30. https://doi.org/10.17645/mac.v11i1.6030

Boeschoten, L., Mendrik, A., van der Veen, E., Vloothuis, J., Hu, H., Voorvaart, R., & Oberski, D. L. (2022). Privacy-preserving local analysis of digital trace data: A proof-of-concept. *Patterns*, *3*(3), 1–10. https://doi.org/10.1016/j.patter.2022.100444

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566. https://doi.org/10.1080/1369118X.2019.1637447

Carrière, T. C., Schipper, N. C., Boeschoten, L., & Araujo, T. (2024). *A systematic collection of Data Download Packages*. Open Science Framework Preprints. https://osf.io/be2q6

Case C-73/16. *Judgment of the Court (Second Chamber) of 27 September 2017. Peter Puškár v Finančné riaditeľstvo Slovenskej republiky and Kriminálny úrad finančnej správy*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62016CJ0073

Case C-131/12. *Judgment of the Court (Grand Chamber) of 13 May 2014. Google Spain SL, Google Inc. V Agencia Española de Protección de Datos (AEPD), Mario Costeja González*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0131

Case C-141/12 and Case C-372/12. *Judgment of the Court (Third Chamber), 17 July 2014. YS v Minister voor Immigratie, Integratie en Asiel and Minister voor Immigratie, Integratie en Asiel v M and S*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX:62012CJ0141

Case C-307/22. *Opinion of Advocate General Emiliou delivered on 20 April 2023. FT v DW*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62022CC0307

Case C-434/16. *Judgment of the Court (Second Chamber) of 20 December 2017. Peter Nowak v Data Protection Commissioner*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62016CJ0434

Case C-487/21. *Judgment of the Court (First Chamber) of 4 May 2023. F.F. v Österreichische Datenschutzbehörde and CRIF GmbH*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62021CJ0487

Case C-553/07. *Judgment of the Court (Third Chamber) of 7 May 2009 (reference for a preliminary ruling from the Raad van State (Netherlands)) – College van burgemeester en wethouders van Rotterdam v M.E.E. Rijkeboer*. The Court of Justice of the European Union. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62007CA0553

Centre for Social Sciences. (n.d.). *Digital political footprints*. https://recens.tk.hu/en/digital-political-footprints

Cotter, K., Medeiros, M., Pak, C., & Thorson, K. (2021). “Reach the right people”: The politics of “interests” in Facebook’s classification system for ad targeting. *Big Data & Society*, *8*(1), 1–16. https://doi.org/10.1177/2053951721996046

de Vreese, C., & Tromble, R. (2023). The data abyss: How lack of data access leaves research and society in the dark. *Political Communication*, *40*(3), 356–360. https://doi.org/10.1080/10584609.2023.2207488

Dommett, K., & Tromble, R. (2022). Advocating for platform data access: Challenges and opportunities for academics seeking policy change. *Politics and Governance*, *10*(1), 220–229. https://doi.org/10.17645/pag.v10i1.4713

Ejaz, W., Altay, S., & Naeem, G. (2023). Smartphone use and well-being in Pakistan: Comparing the effect of self-reported and actual smartphone use. *Digital Health*, *9*. https://doi.org/10.1177/20552076231186075

European Commission. (2023). *Delegated Regulation on data access provided for in the Digital Services Act* [Announcement]. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13817-Delegated-Regulation-on-data-access-provided-for-in-the-Digital-Services-Act_en

European Commission. (2024a). *Commission opens formal proceedings against Facebook and Instagram under the Digital Services Act* [Press release]. https://ec.europa.eu/commission/presscorner/detail/en/ip_24_2373

European Commission. (2024b). *Status report: Mechanisms for researcher access to online platform data* [Report]. https://digital-strategy.ec.europa.eu/en/library/status-report-mechanisms-researcher-access-online-platform-data

European Data Protection Board (EDPB). (2023). *Guidelines 01/2022 on data subject rights – Right of access* (Guidelines No. Version 2.1). https://edpb.europa.eu/system/files/2023-04/edpb_guidelines_202201_data_subject_rights_access_v2_en.pdf

European Digital Media Observatory. (2022). *Report of the European Digital Media Observatory’s Working Group on platform-to-researcher data access* [Report]. https://edmoprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf

Freelon, D. (2014). On the interpretation of digital trace data in communication and social computing research. *Journal of Broadcasting & Electronic Media*, *58*(1), 59–75. https://doi.org/10.1080/08838151.2013.875018

Freelon, D. (2018). Computational research in the post-API age. *Political Communication*, *35*(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Garimella, K., & Chauchard, S. (2024). How prevalent is AI misinformation? What our studies in India show so far. *Nature*, *630*(8015), 32–34. https://doi.org/10.1038/d41586-024-01588-2

GDPR Hub. (2020, January 12). *Category: Article 15 GDPR*. https://gdprhub.eu/index.php?title=Category:Article_15_GDPR

Giannopoulou, A., Ausloos, J., Delacroix, S., & Janssen, H. (2022). Intermediating data rights exercises: The role of legal mandates. *International Data Privacy Law*, *12*(4), 316–331. https://doi.org/10.1093/idpl/ipac017

Gil-López, T., Christner, C., de León, E., Makhortykh, M., Urman, A., Maier, M., & Adam, S. (2023). Do (not!) track me: Relationship between willingness to participate and sample composition in online information behavior tracking research. *Social Science Computer Review*, *41*(6), 2274–2292. https://doi.org/10.1177/08944393231156634

Greenleaf, G. (2021). *Global tables of data privacy laws and bills* (Dataset No. 7th Ed.). Privacy Laws & Business International Report. https://doi.org/10.2139/ssrn.3836261

Haim, M., Leiner, D., & Hase, V. (2023). Integrating data donations in online surveys. *Medien & Kommunikationswissenschaft*, *71*(1–2), 130–137. https://doi.org/10.5771/1615-634X-2023-1-2-130

Halavais, A. (2019). Overcoming terms of service: A proposal for ethical distributed research. *Information, Communication & Society*, *22*(11), 1567–1581. https://doi.org/10.1080/1369118X.2019.1627386

Hase, V., & Haim, M. (2024). Can we get rid of the bias? Mitigating systematic error in data donation studies through survey design strategies. *Computational Communication Research*, *6*(2), 1–29. https://journal.computationalcommunication.org/article/view/8596

Jaursch, J., Ohme, J., & Klinger, U. (2024). *Enabling research with publicly accessible platform data: Early DSA compliance issues and suggestions for improvement* (Position Paper No. 9; Weizenbaum Policy Papers). Weizenbaum Institute. https://doi.org/10.34669/WI.WPP/9

Keusch, F., & Kreuter, F. (2021). Digital trace data. In U. Engel, A. Quan-Haase, S. Liu, & L. Lyberg (Eds.), *Handbook of computational social science: Theory, Case Studies and Ethics* (Vol. 1, pp. 100–118). https://doi.org/10.4324/9781003024583

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society*, *20*(1), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

Kmetty, Z., & Bozsonyi, K. (2022). Identifying depression-related behavior on Facebook – an experimental study. *Social Sciences*, *11*(3), 1–19. https://doi.org/10.3390/socsci11030135

Kohne, J., & Montag, C. (2023). ChatDashboard: A framework to collect, link, and process donated WhatsApp chat log data. *Behavior Research Methods*, *56*(4), 3658–3684. https://doi.org/10.3758/s13428-023-02276-1

Leerssen, P. (2024). Outside the black box: From algorithmic transparency to platform observability in the Digital Services Act. *Weizenbaum Journal of the Digital Society*, *4*(2), 1–29. https://doi.org/10.34669/WI.WJDS/4.2.3

Lukito, J. (2024). *Platform research ethics for academic research* [Report]. Center for Media Engagement. https://mediaengagement.org/research/platform-research-ethics

Mahieu, R. L. P. (2023). *The right of access to personal data in the EU: A legal and empirical analysis* [Doctoral dissertation, Vrije Universiteit Brussel]. https://researchportal.vub.be/en/publications/the-right-of-access-to-personal-data-in-the-eu-a-legal-and-empiri

Mancosu, M., & Vegetti, F. (2020). What you can scrape and what is right to scrape: A proposal for a tool to collect public Facebook data. *Social Media + Society*, *6*(3). https://doi.org/10.1177/2056305120940703

Möller, J., Linnert, E., & Araujo, T. (2023, August). *Detecting algorithmic bias and fringe bubbles in social media* [Conference presentation]. ECREA Political Communication Section, Berlin, Germany.

Newman, N., Fletcher, R., Eddy, K., Robertson, C. T., & Nielsen, R. K. (2023). *Digital news report 2023* [Report]. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-06/Digital_News_Report_2023.pdf

Nieborg, D., Poell, T., Caplan, R., & van Dijck, J. (2024). Introduction to the special issue on Locating and theorising platform power. *Internet Policy Review*, *13*(2). https://doi.org/10.14763/2024.2.1781

Nonnecke, B., & Carlton, C. (2022). EU and US legislation seek to open up digital platform data. *Science*, *375*(6581), 610–612. https://doi.org/10.1126/science.abl8537

Ohme, J., Araujo, T., Boeschoten, L., Freelon, D., Ram, N., Reeves, B. B., & Robinson, T. N. (2024). Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures*, *18*(2), 124–141. https://doi.org/10.1080/19312458.2023.2181319

Parry, D. A. (2024). Without access to social media platform data, we risk being left in the dark. *South African Journal of Science*, *120*(3/4). https://doi.org/10.17159/sajs.2024/17008

Pfiffner, N., & Friemel, Thomas. N. (2023). Leveraging data donations for communication research: Exploring drivers behind the willingness to donate. *Communication Methods and Measures*, *17*(3), 227–249. https://doi.org/10.1080/19312458.2023.2176474

Pfiffner, N., Tribelhorn, T., & Friemel, T. N. (May 2024a). *Investigating the influence of friendships on YouTube usage history similarity using data donations* [Conference presentation]. Data Donation Symposium 2024, Amsterdam, Netherlands.

Pfiffner, N., Witlox, P., & Friemel, T. N. (2024b). Data donation module: A web application for collecting and enriching data donations. *Computational Communication Research*, *6*(2). https://journal.computationalcommunication.org/article/view/8597

Pins, D., Jakobi, T., Stevens, G., Alizadeh, F., & Krüger, J. (2022). Finding, getting and understanding: The user journey for the GDPR’S right to access. *Behaviour & Information Technology*, *41*(10), 2174–2200. https://doi.org/10.1080/0144929X.2022.2074894

Rau, J. (2023, May 4). Clearing the data fog: German far-right research’s requirements for the DSA research data access. *Media Research Blog*. https://web.archive.org/web/20231210002301/https://leibniz-hbi.de/de/blog/clearing-the-data-fog-german-far-right-research-s-requirements-for-the-dsa-research-data-access

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4). https://doi.org/10.14763/2020.4.1535

Struminskaya, B. (2022, September). *Willingness and nonparticipation bias in data donation* [Conference presentation]. ODISSEI Conference for Social Science in the Netherlands, Utrecht, Netherlands.

Syrmoudis, E., Luzsa, R., Ehrlich, Y., Agidigbi, D., Kirsch, K., Rudolf, D., Schlaeger, D., Weber, J., & Grossklags, J. (2024). Unlocking personal data from online services: User studies on data export experiences and data transfer scenarios. *Human–Computer Interaction*, 1–25. https://doi.org/10.1080/07370024.2024.2325347

Syrmoudis, E., Mager, S., Kuebler-Wachendorff, S., Pizzinini, P., Grossklags, J., & Kranz, J. (2021). Data Portability between Online Services: An Empirical Analysis on the Effectiveness of GDPR Art. 20. *Proceedings on Privacy Enhancing Technologies*, 351–372. https://doi.org/10.2478/popets-2021-0051

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988929

van Driel, I. I., Giachanou, A., Loes Pouwels, J., Boeschoten, L., Beyens, I., & Valkenburg, P. M. (2022). Promises and Pitfalls of Social Media Data Donations. *Communication Methods and Measures*, *16*(4), 266–282. https://doi.org/10.1080/19312458.2022.2109608

van Drunen, M. Z., & Noroozian, A. (2024). How to design data access for researchers: A legal and software development perspective. *Computer Law & Security Review*, *52*. https://doi.org/10.1016/j.clsr.2024.105946

Veale, M. (2023). *Denied by design? Data access rights in encrypted infrastructures*. SocArXiv Papers. https://doi.org/10.31235/osf.io/94y6r

Wagner, M. W. (2023). Independence by permission. *Science*, *381*(6656), 388–391. https://doi.org/10.1126/science.adi2430

Wedel, L., Ohme, J., & Araujo, T. (In press). Augmenting data donations: Integrating TikTok DDPs, video metadata, and the multimodal nature of audiovisual content. *Method, Data, Analysis*.

Welbers, K., Loecherbach, F., Lin, Z., & Trilling, D. (2024). Anything you would like to share: Evaluating a data donation application in a survey and field study. *Computational Communication Research*, *6*(2). https://journal.computationalcommunication.org/article/view/8598

Windwehr, S., & Selinger, J. (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*. https://policyreview.info/articles/news/can-we-fix-access-to-platform-data

Wong, J., & Henderson, T. (2019). The right to data portability in practice: Exploring the implications of the technologically neutral GDPR. *International Data Privacy Law*, *9*(3), 173–191. https://doi.org/10.1093/idpl/ipz008

Wu-Ouyang, B., & Chan, M. (2023). Overestimating or underestimating communication findings? Comparing self-reported with log mobile data by data donation method. *Mobile Media & Communication*, *11*(3), 415–434. https://doi.org/10.1177/20501579221137162

Supplement

TABLE A1: Glossary of abbreviations and terms

TERM	DEFINITION
Application programming interface (API)	Software interface enabling data access to digital platforms, for instance, via programming scripts

TERM	DEFINITION
Court of Justice of the European Union (CJEU)	The Court of Justice of the European Union ensures that EU law is interpreted in the same way across countries.
Comma-separated values (CSV) file	Text file that uses commas to separate values, here abbreviated using the file format ending .csv
Data Act (DA)	Regulation in EU law concerned with fair use of data generated by internet of things devices
Data controller	Actor determining the purposes and means for the processing of personal data based on Art. 4(7) of the GDPR, a term we use here interchangeably with platform
Data Donation Tool (DDT)	Tool to extract, anonymise, and send data from data subjects' DDPs for data donation studies
Data Download Package (DDP)	File(s) containing personal data collected by platforms
Digital Markets Act (DMA)	Regulation in EU law concerned with the regulation of large technology companies in the digital sector
Data processing	Any operation performed on personal data, including collection, structuring, storage, anonymisation, deletion, or sharing based on Art. 4(2) of the GDPR
Data protection authority (DPA)	An independent national supervisory authority tasked with monitoring and enforcement of the GDPR based on Art. 51(1) which has investigative powers to impose administrative fines whenever controllers do not comply
Digital Services Act (DSA)	Regulation in EU law concerned with updating existing regulations for content moderation on digital platforms
Data subject	Person who can be directly or indirectly identified by reference to an identifier based on GDPR Art. 4(1), a term we use here interchangeably with platform user
European Data Protection Board (EDPB)	Independent advisory EU body ensuring consistent application of the GDPR by data controllers and national supervisory authorities across the EU based on GDPR Arts. 69 and 70
Feed	Streams of information on digital platforms, often sorted either chronologically or via algorithms
General Data Protection Regulation (GDPR)	Regulation in EU law concerned with data protection and privacy
HyperText Markup Language (HTML)	Form of human-readable formatting of content that should be displayed in a web browser, here abbreviated using the file format ending .html
JavaScript Object Notation (JSON)	Form of machine-readable formatting of content, here abbreviated using the file format ending .json
Personal data	Any information relating to an identified or identifiable person (data subject) based on GDPR Art. 4(1)
Text file (TXT)	A file containing unformatted text, usually ending with .txt

TABLE A2.1: Challenges I–IV (Concise and easily accessible form)

STUDY	PLATFORMS	I. INSTANCES RESTRICTIONS	II. INTERFACES	III. LIMITING REQUESTED DATA	IV. DIVERSE DDP FORMATS
1	Facebook	–	–	–	Participants could choose between file formats, such as

STUDY	PLATFORMS	I. INSTANCES RESTRICTIONS	II. INTERFACES	III. LIMITING REQUESTED DATA	IV. DIVERSE DDP FORMATS
					JSON- and HTML-files.
2	Google, YouTube	Respondents could only request and download data via browsers, not apps.	–	Participants could not select for which time they wanted to request data.	Participants could choose between file formats, such as JSON- and HTML-files.
3	YouTube	–	–	–	–
4	Google, YouTube	–	Interfaces for data requests differed across language settings by participants. Also, Google sometimes requested two-factor authentication.	–	Participants could choose between file formats, such as JSON- and HTML-files.
5	Google	–	–	Participants could not select for which time they wanted to request data.	–
6	Facebook, Instagram, Twitter/X, YouTube	Respondents could only request and download data via browsers, not apps.	Twitter/X asked some participants to send in a copy of their national ID for data requests to be continued.	For YouTube and Twitter/X, participants could not select which data to request (e.g. type of data, time). Especially for YouTube, participants often had trouble uploading large DDPs.	Participants could choose between file formats, such as JSON- and HTML-files.
7	Facebook, Google, Instagram, TikTok, YouTube	For some platforms, respondents could only request and download data via browsers, not apps.	For Facebook, interfaces for data requests differed across respondents potentially due to design tests, rendering download instructions incorrect.	For TikTok, participants could not select which data to request (e.g. type of data, time).	Participants could choose between file formats, such as JSON- and HTML-files. However, some data (e.g. playlists on YouTube) was only available as CSV-files.
8	WhatsApp	Participants could only request and download chat histories via apps, not browsers.	Interfaces for data requests differed across smartphone operating systems.	Participants could not select for which time they wanted to request data.	–
9	TikTok	Respondents could request data via apps but had to download data via browsers.	TikTok sometimes requested users to re-enter passwords.	Participants could not select which data to request (e.g. type of data, time).	Participants could choose between file formats, such as JSON- and HTML-files.
10	TikTok, YouTube	For YouTube, respondents could only request and download data via browsers. For TikTok,	–	For YouTube, participants could not select for which time they wanted to request data.	Participants could choose between file formats, such as JSON-, HTML- and TXT-files.

STUDY	PLATFORMS	I. INSTANCES RESTRICTIONS	II. INTERFACES	III. LIMITING REQUESTED DATA	IV. DIVERSE DDP FORMATS
		they could request data via apps but had to download data via browsers.			
11	TikTok, YouTube	For YouTube, respondents could only request and download data via browsers. For TikTok, they could request data via apps but had to download data via browsers.	Both platforms sometimes requested users to re-enter passwords or two-factor authentication.	Participants could not select which data to request (e.g. type of data, time).	Participants could choose between file formats, such as JSON- and HTML-files.
12	Netflix	Respondents could only request and download data via browsers.	Netflix required users to ask the account holder (i.e. the paying account) to accept/verify the data request.	Participants could not select which data to request (e.g. type of data, time).	–

TABLE A2.2: Challenges V–VII (Concise and easily accessible form)

STUDY	PLATFORMS	V. MULTIPLE ACCESS REQUESTS	VI. NO NOTIFICATION ABOUT DDPS	VII. LIMITED ACCESS OVER TIME
1	Facebook	–	–	–
2	Google, YouTube	–	–	–
3	YouTube	–	–	–
4	Google, YouTube	–	It was not transparent when data would be accessible. For example, Google gave a too pessimistic outlook on the waiting time (e.g. “possibly hours or days” although data was accessible earlier), which confused participants.	DDPs were no longer accessible after seven days.
5	Google	–	–	DDPs were no longer accessible after seven days.
6	Facebook, Instagram, Twitter/X, YouTube	–	Notifications about DDPs being available were provided differently (e.g. via email or platform notifications). It was not transparent when the data would be accessible. The amount of time it took to provide participants with DDPs differed across platforms.	For some platforms, DDPs were no longer accessible after a certain time span (e.g. seven days).
7	Facebook, Google, Instagram, TikTok, YouTube	–	Notifications about DDPs being available were provided differently (e.g. via email or platform notifications). TikTok did not provide notifications about DDPs being available. The amount of time it took to provide DDPs differed across platforms and the amount of data participants requested.	–

STUDY	PLATFORMS	V. MULTIPLE ACCESS REQUESTS	VI. NO NOTIFICATION ABOUT DDPS	VII. LIMITED ACCESS OVER TIME
8	WhatsApp	Participants needed to export each chat separately.	–	DDPs were no longer accessible after a couple of weeks.
9	TikTok	–	Once requested, it was not transparent when the data would be accessible.	DDPs were no longer accessible after four days.
10	TikTok, YouTube	–	TikTok did not provide notifications about DDPs being available. Also, the amount of time it took to provide DDPs differed. In 2023, TikTok provided data within two to four days. In April 2024, data was often available within minutes – making it necessary to change reminder emails to participants to donate their DDPs.	For TikTok, DDPs were no longer accessible after four days.
11	TikTok, YouTube	–	TikTok did not provide notifications about DDPs being available. It was not transparent when the data would be accessible.	For TikTok, DDPs were no longer accessible after four days.
12	Netflix	–	–	–

TABLE A3: Challenges VIII–XI (Transparency)

STUDY	PLATFORMS	VIII. NO INFORMATION ON COMPLETENESS OF DDPS	IX. NO INFORMATION ON DDP STRUCTURE	X. NO INFORMATION ON DDP MEASUREMENTS	XI. NO INFORMATION ON DDP CHANGES
1	Facebook	Facebook did not provide documentation on which data was collected, making it unclear whether data was missing.	Facebook did not provide a complete documentation on what folders/files signified, making it impossible to understand data. Files included language-specific or localised names.	Facebook did not provide documentation on how measures were created.	There was no information policy about changes in DDPs.
2	Google, YouTube	Platforms did not provide documentation on which data was collected, making it unclear whether data was missing.	Platforms did not provide complete documentation on what folders/files signified, making it impossible to understand data. Files included language-specific or localised names, which was especially problematic because the study was conducted in Switzerland, a multilingual country.	Platforms did not provide documentation on how measures were created.	There was no information policy about changes in DDPs. During data collection, Google started to only provide HTML files (instead of requested JSON files) without any information prior to this change. Contacting Google did not yield any reaction as to how to solve this issue.
3	YouTube	YouTube did not provide documentation on	–	–	–

STUDY	PLATFORMS	VIII. NO INFORMATION ON COMPLETENESS OF DDPS	IX. NO INFORMATION ON DDP STRUCTURE	X. NO INFORMATION ON DDP MEASUREMENTS	XI. NO INFORMATION ON DDP CHANGES
		which data was collected, making it unclear whether data was missing.			
4	Google, YouTube	Platforms did not provide documentation on which data was collected, making it unclear whether data was missing. Participants indicated missing data when inspecting DDPs.	Platforms did not provide complete documentation on what folders/files signified, making it impossible to understand data. Files included language-specific or localised names.	–	There was no information policy about changes in DDPs.
5	Google	Google did not provide documentation on which data was collected, making it unclear whether data was missing.	Google did not provide complete documentation on what folders/files signified, making it impossible to understand data.	Google did not provide documentation on how measures were created. For example, Google infers travel modes via algorithms – but it is unclear how. Data indicated measurement errors (e.g. by-foot trip of 200 km in 30 min.).	–
6	Facebook, Instagram, Twitter/X, YouTube	Platforms did not provide documentation on which data was collected, making it unclear whether data was missing.	Platforms did not provide documentation on what folders/files signified, making it impossible to understand data. For some platforms, files included language-specific or localised names. For YouTube, watch and search histories contained a mix of forced-to-view content, such as advertisements, and self-selected content, making it hard to distinguish between activities.	Platforms did not provide documentation on how measures were created.	Interfaces for data requests changed throughout the study, rendering download instructions incorrect.
7	Facebook, Google, Instagram, TikTok, YouTube	Platforms did not provide documentation on which data was collected, making it unclear whether data was missing. For YouTube, data was missing (e.g. watch histories limited in time).	Platforms did not provide documentation on what folders/files signified, making it impossible to understand data. For some platforms, files included language-specific names.	Platforms did not provide documentation on how measures were created.	There was no information policy about changes in DDPs.

STUDY	PLATFORMS	VIII. NO INFORMATION ON COMPLETENESS OF DDPS	IX. NO INFORMATION ON DDP STRUCTURE	X. NO INFORMATION ON DDP MEASUREMENTS	XI. NO INFORMATION ON DDP CHANGES
8	WhatsApp	WhatsApp did not provide documentation on which data was collected, making it unclear whether data was missing.	–	WhatsApp did not provide documentation on how measures were created.	There was no information policy about changes in DDPs.
9	TikTok	TikTok did not provide documentation on which data they collected, making it unclear whether data was missing.	–	While some general information was provided, TikTok did not provide documentation on how each measure was created.	There was no information policy about changes in DDPs.
10	TikTok, YouTube	Platforms did not provide documentation on which data was collected, making it unclear whether data was missing. DDPs indicated missingness (e.g. for some participants information from the watch history that was available for others was absent).	Platforms did not provide documentation on what folders/files signified, making it impossible to understand data. For YouTube, files included language-specific names. For YouTube, watch and search histories contained a mix of forced-to-view content, such as advertisements, and self-selected content, making it hard to distinguish between activities.	Platforms did not provide documentation on how measures were created.	–
11	TikTok, YouTube	Platforms did not provide documentation on which data was collected, making it unclear whether data was missing. For TikTok, it became clear that data was missing (e.g. search and watch histories were partly or completely empty).	–	While some general information was provided, platforms did not provide documentation on how each measure was created.	There was no information policy about changes in DDPs. The process of retrieving DDPs from TikTok changed throughout the study without any information prior to this change.
12	Netflix	Netflix did not provide documentation on which data was collected, making it unclear whether data was missing.	Netflix did not provide documentation on what folders/files signified.	Netflix did not provide documentation on how measures were created.	There was no information policy about changes in DDPs.

TABLE A4: Challenges XII–XIV (Completeness)

STUDY	PLATFORMS	XII. MISSING DATA ON ACTIVITIES	XIII. MISSING DATA ON CONTENT	XIV. MISSING DATA ON CONTEXTS
1	Facebook	Facebook provided incomplete data on activities, e.g. content participants were exposed to.	–	Facebook provided incomplete data on the context in which users engaged in activities or were exposed to content (e.g. via feeds, via private messages?).
2	Google, YouTube	–	–	Platforms provided incomplete data on the context in which users engaged in activities or were exposed to content. For example, Google search data only contained information on what participants clicked on after a search query, not the results shown as a response, making it difficult to study content selection.
3	YouTube	–	–	YouTube provided incomplete data on the context in which users engaged in activities or were exposed to content, e.g. whether content was recommended.
4	Google, YouTube	–	–	Platforms provided incomplete data on the context in which users engaged in activities or were exposed to content, e.g. whether content was recommended.
5	Google	–	–	–
6	Facebook, Instagram, Twitter/X, YouTube	Platforms provided incomplete data on activities, such as exposure to content or timestamps of activities.	Platforms provided incomplete data on content, e.g. IDs rather than titles of watched content, making it impossible for users to understand DDPs.	Platforms provided incomplete data on the context in which users engaged in activities or were exposed to content (e.g. via feeds, via private messages?).
7	Facebook, Google, Instagram, TikTok, YouTube	Platforms provided incomplete data on activities, such as exposure to content or timestamps of activities.	Platforms provided incomplete data on content, such as the names of followed/ liked pages, which were not unique.	Platforms provided incomplete data on the context in which users engaged in activities or were exposed to content, e.g. whether content was recommended.
8	WhatsApp	–	–	WhatsApp provided incomplete data on the context in which users engaged in activities or were exposed to content, e.g. when someone was quoted as part of the structure of chats.
9	TikTok	TikTok provided incomplete data on activities, for example by limiting watch histories in time.	TikTok provided incomplete data on content, e.g. links to watched content instead of video titles.	–
10	TikTok, YouTube	Platforms provided incomplete data	Platforms provided incomplete data on content, e.g. links to	Platforms provided incomplete data on the context in which users engaged in activities or were exposed to content, e.g. whether content

STUDY	PLATFORMS	XII. MISSING DATA ON ACTIVITIES	XIII. MISSING DATA ON CONTENT	XIV. MISSING DATA ON CONTEXTS
		on activities, such as liked videos on YouTube (such data was available a few years back).	watched content instead of video titles.	was recommended.
11	TikTok, YouTube	Some platforms provided incomplete data on activities, such as watch and search histories.	Platforms provided incomplete data on content, e.g. links to watched or shared content instead of video titles.	Platforms provided incomplete data on the context in which users engaged in activities or were exposed to content, e.g. whether content was recommended.
12	Netflix	–	–	–

Published by

Alexander von Humboldt
Institute for Internet
and Society

in cooperation with

CREATe

UOC R&I IN3 Internet interdisciplinary Institute
Universitat Oberta de Catalunya

cnrs centre internet et societe

University of Tartu
Johan Skytte Institute of
Political Studies