Original Research Article

MTG

Media, Technology and Governance
1–20


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/29777984261486601
journals.sagepub.com/home/mag


Regulating a Floating Signifier: Interpretations of "Data" in the Digital Services Act Researcher Access Provision

Taylor Annabell[1]

Abstract

The European Union's Digital Services Act promises to disrupt the power held by platform companies to determine how data needed to understand risks is accessed by researchers. Article 40 brings researcher data access into a governance *of* platforms framework, but how data are conceptualized in this regime is overlooked. This article uses discourse analysis to examine how data are understood in the legal texts, public submissions of feedback during the drafting process and published data catalogs by the platforms that fall within the regulation. I argue that the meaning of data "floats" across and within regulatory and commercial discourses, aligning and disconnecting with how data are understood in research disciplines. Platform companies narrowly interpret and operationalize data as aggregated, quantified presentations structured through predefined datasets, which is in tension with the multiplicity of understandings of data in academic research. The attempt to narrow the definitional boundaries of data seeks to impose limitations on how researchers might envisage data for their academic work, which is overlooked in current discussions of this data access provision.

Keywords

platform governance, data, discourse, Digital Services Act, data access

Introduction

As the private governors of our platform society, platform companies not only provide critical infrastructure for interaction, entertainment, work, health, education, and government (van Dijck et al., 2018) but also shape and impact the research infrastructure for understanding our platform society (Flensburg et al., 2025). To understand how platforms mediate our social, political, economic, and

[1]School of Journalism, Media & Culture, Cardiff University, Cardiff, UK

Corresponding Author:
Taylor Annabell, School of Journalism, Media & Culture, Cardiff University, 2 Central Square, Cardiff, CF10 3AT, UK.
Email: annabellt@cardiff.ac.uk

personal interactions, researchers rely on access to data, which is produced, inferred, and circulated within socio-technical environments. But the availability of data from platforms, beyond what we can see (and scrape) at the level of the interface or request from users, such as how content is moderated, reported, or recommended, has been significantly limited. In other words, how we do research about platforms, including how they govern and are governed (Gillespie, 2017), is tied to the governance of data access.

The European Union's (EU's) Digital Services Act (DSA) promises to disrupt the power of platform companies to determine researcher access to data by bringing the issue into the governance of platforms framework. Under the legal procedure established in Article 40(4) and the subsequent Delegated Act, Very Large Online Platforms and Search Engines (VLOPSEs) are obligated to provide vetted researchers with data, previously inaccessible, to investigate systemic risks stemming from their platform design in the EU. While researchers have critically reflected on how transparency and observability are operationalized and drawn attention to issues of data quality and non-compliance with provisions related to publicly available data (see Botero Arcila et al., 2026; Griffin, 2025; Goanta et al., 2026; Leerssen, 2023; Peters & Weller, 2026; Selling et al., 2025), this article focuses on how data are treated as a self-evident, technical concept, which has significant governance implications. What falls within the meaning of "data" and thus can be requested is undefined in the legal texts. I argue that data becomes a site of contestation in which platforms attempt to narrow the definitional boundaries of the term, imposing limitations on how researchers might envisage data for their academic work.

Article 40 of the DSA, entitled "Data access and scrutiny," addresses how platform data can be obtained from different actors: the Digital Services Coordinator of establishment and the Commission (para. 1–3), as well as vetted researchers (para. 4–12). There are two modes of data access: publicly available data and data available through reasoned request. While Article 40 outlines key conditions and timeframes in the reasoned request process, it also clarifies that delegated acts will be adopted to supplement the regulation by harmonizing and clarifying technical conditions, procedures, and rules. In preparation for the delegated act, the Commission received 133 submissions as part of a "call for evidence" in April and May 2023 and then 109 submissions during the feedback period on the draft delegated act between October and December 2024. Some platform companies used these avenues to shape the legal text, which, as the analysis will indicate, included attempts to determine the meaning of data. The Delegated Act was adopted on 2 July 2025 and entered into force on 29 October 2025, with VLOPSEs beginning to publish their data catalogs in November 2025. These catalogs not only are part of their regulatory obligations but are a critical site through which VLOPSEs make visible their understanding of data. While researchers have begun to submit applications to the DSA Data access portal, with those applying earliest learning about outcomes by February 2026 (see Goanta & Iamnitchi, 2026 for reflection on their rejection), at the time of writing (April 2026), the portal is yet to display any research projects that have been approved and thus indicate the operationalization of the process.

This article proposes that what constitutes data in Article (40)4 and the Delegated Act is a taken-for-granted understanding, which problematically enables platform companies to narrow the definitional scope and exert power. Through a discourse analysis on the conceptualization of data in the context of data access for researchers through the DSA, I highlight how data becomes a site of contestation, demonstrating how platform companies interpret data as aggregated, anonymized, and predefined datasets, which is in tension with the multiplicity of understandings of data across disciplines in academic research. Turning to discourse theory, I argue that data functions as a floating signifier (Laclau, 2005). While resonant with interpretative flexibility of boundary objects (Star, 2010; Star & Griesemer, 1989), the discursive analytical concept is oriented toward surfacing the ideological struggle in the domain of language; "data" is articulated differently across

regulatory and commercial discourses, engendering an ambiguity to the term, which is consequential for how data access provision may be actualized and aligned to different disciplinary perspectives.

Conceptualizing Data in Platform Society and Research Communities

To situate the contestation of data within the data access provision of the DSA, it is necessary to return to long-standing debates about the conceptualization of data. This section teases out how data understood in the context of platforms is entangled with and distinct from its operationalization in research practices, which differ across disciplinary boundaries.

First, data can be understood as a resource and economic asset at the foundation of business models in platform society. As van Dijck et al. (2018) explain, the digital architecture governing interaction between users, known as the platform, is "geared towards the systemic collection, algorithmic processing, circulation, and monetization of user data" (p. 4). One of the ways this is exploited for economic gain is through offering third parties controlled access to this data and facilitating targeted advertising. As such, creating and using data are critical to how platforms operate their services—controlling the visibility and access to user-generated content and transactional data—as well as the power of Big Tech as governors of data.

The commercial operation of platforms is premised on this techno-commercial strategy of datafication. This term refers to the transformation of social and material characteristics, behavior, and events into digitally processable data (Mayer-Schönberger & Cukier, 2013). Such quantification through data has engendered a "new paradigm for understanding sociality and social behavior" (van Dijck, 2013, p. 198). Critically, while data are often assumed to be objective and treated as raw material for analysis or processing, it is always "precooked." Information, behavior, and phenomena are transformed into a data format, a representation, and thus cannot be understood as pre-analytical. As Kitchin and Laurialt (2014) explain, when data are understood as representative and reflective of the world, they can be taken at face value, which means the inherent politics and processes that have informed their production are ignored. Critical data studies prioritize questioning how data are "produced and deployed to create knowledge, manage society, derive business value, and perform numerous other task" (Kitchin, 2025, p. 4). Applying this to the data access via platforms context means ensuring that the data are not treated as neutral inputs and naturally occurring but produced by techniques. In addition, what data are collected and stored by platforms are guided by their operations. As Olteanu et al. (2019) point out, platforms "may not capture all relevant data, (p. 6)" echoed by Lazer et al. (2021), who identify that data are often captured for purposes other than research.

This brings us to how data are understood in the research context, which carries significant semantic baggage prior to its use as a concept with digital cultures. As reviews on definitions and debates about what data are attest, the term is contested and connected to other concepts such as information, facts, and knowledge, equally as elusive (see Borgman, 2015; Boyd, 2022; D'ignazio & Klein, 2020; Drucker, 2011; Leonelli, 2022; Markham, 2013; Rosenberg, 2013). For example, Drucker (2011) differentiates between capta and data, situated in constructivist and realist approaches, respectively, and refers to the constructed interpretation or assumed "given" representation of the world. Adding to this, what is made to constitute data depends on discipline, as does how the term itself is used. While social sciences and sciences used "data" to categorize the use of facts, numbers, letters, and symbols, in the humanities, the term emerged in the turn toward digital objects and cultures in research.

For the purposes of this article, research data as "entities used as evidence of phenomena for the purposes of research or scholarship" (Borgman, 2015, p. 28) are adopted. This captures how data in

itself are not treated as possessing a truth value or representation of phenomena but as what Leonelli (2022) refers to as a "relational category." As Markham (2013) puts it, data do more than describe units of information but discourses about knowledge and, "through its ambiguity, can foster a self-perpetuating sensibility that it is incontrovertible, something to question the meaning of, or the veracity of, but not the existence of." What this means for approaching access to platform data is problematizing the presentation of data as neutral and a recognition of potential tension between commercial and platform-anchored understandings of data as framed under datafication, with its conceptualization in research contexts that may be broader and inclusive of nonnumerical and quantitative forms. Whilst the same data may be used by platforms and researchers, other representations or evidence of phenomena might be understood as data by researchers that fall outside of data for platforms.

Access to Data as an Issue of Governance

Due to data being treated and protected as a private economic good, gaining access to platform data has traditionally been part of the private governance of platform companies. This means platforms establish what rules and processes will determine access and use of their data for research purposes, giving rise to the distinction between authorized "platform-governing instruments, providing 'controlled' access to data" such as Application Programming Interface (API) (van Dijck et al., 2018, p. 35), and unauthorized methods, in which researchers scrape digital traces left behind (Ohme et al., 2024).

Academics were not intended users of APIs; APIs were designed for platform profit, wherein developers and "corporations to monitor their customer bases, brand identity, and advertising" (Tromble, 2021, p. 3). Yet, research APIs have been developed to facilitate data access, which has been dependent on the "goodwill" of platforms (Selling et al., 2025) and part of the performance of transparency, given their significant position in society. Access to platform data via APIs, however, has been subject to policy changes with restrictions on data availability following Cambridge Analytica being described as an "APIcalypse" (Bruns, 2021), a "data abyss" (de Vreese & Tromble, 2023), and a "post-API age" (Freelon, 2018). As Tromble (2021) reminds us, even during what is considered the "golden era" of data access through APIs, issues of data quality and privacy were prevalent, and researchers may have given too much weight to gaining access to data at scale and allowed it to dictate the research agenda in computational social media research.

Beyond research APIs, access to platform data has been obtained through what Bruns (2021) refers to as corporate data philanthropy initiatives. For example, Social Science One facilitated access to Facebook data for selected research projects on democracy and elections through the established partnership between Meta and the organization hosted by Harvard's Institute for Quantitative Social Science. More recently, Meta has partnered with the Secure Data Access Centre, which reviews research requests to access data from the Meta Content Library. Whilst these collaborations support access for eligible researchers, as with APIs, terms of use and availability of types of data remain the purview of the platform.

Data donation studies provide an alternative access point for digital trace data, which leverages rights of users under data protection laws such as the General Data Protection Regulation (GDPR). Platforms must provide access to users in concise, transparent ways to all information processed by the platform about them. Through informed consent, users may "donate" their data packages requested from platforms to researchers, demonstrating how legal frameworks are leveraged for data access purposes. This "user-centric approach" to data collection (Ohme et al., 2024) is faced by challenges in how platforms fulfill their obligations, and issues of noncompliance have implications for data donation studies (Hase et al., 2024).

As this section has sought to demonstrate so far, data access is a governance issue, compounding the issue that our understanding of platform governance is also shaped by our access to data. As van Dijck et al. (2018) propose, "public access to data could also help to make platforms more accountable" (p. 142). The relationship between data access for research and governance necessitates that we reflect on how the political economy of Big Tech is not only the object of study but shapes conditions of research (Flensburg et al., 2025). The volatility of data sources and methods is due to dependency on proprietary research infrastructure and means "our epistemic practices are shaped by the commercial interests that dominate the digital environment" (Flensburg et al., 2025, p. 192). Furthermore, while scholars have critiqued data access as providing limited access to data and dictating the availability of metadata, the paradigm of data as digital trace data and quantified, measurable representations overlooks other forms of data that researchers might like to obtain.

The EU's approach to governance of platforms in the DSA relies on a framework of risk management, which includes a novel obligation that brings researchers in as another actor who can meaningfully contribute to platform governance. Leerssen (2024) conceptualizes the DSA as a shift from seeking platform transparency through explanations about specific algorithms to platform observability, drawing on the concept put forward by Rieder and Hofman (2020). In this vein, the DSA aims to tackle conditions and processes in the governance of socio-technical systems using levers to facilitate access to data rather than disclosure of knowledge. Researchers take on an active participatory role in the risk management system through the "regulatory novelty" of data access provisions (Liesenfeld, 2025) for two modes of data access.

First, in terms of access to publicly available data (DSA Article 40.12), platforms have introduced or extended data access options to comply with their obligation (Selling et al., 2025), although preliminary findings from the Commission found TikTok and Meta (European Commission, Directorate-General for Communication, 2025), AliExpress (European Commission, Directorate-General for Communications Networks, Content and Technology, 2025), and X (European Commission, Directorate-General for Communication, 2023) in breach of their obligations. Second, and the focus of this article, is what is referred to as "privileged access to data," where researchers must outline what data they are requesting, amongst other details, in reasoned requests.

Scholars have critically reflected on the power imbalance between researchers and platforms regarding what data are gathered, inferred, and stored by platforms, which then impacts and potentially limits how the Article 40(4) mechanism might be utilized. As Selling et al. (2025) put "its value for knowledge creation depends largely on platforms, as they determine what data to make available and under what conditions" (p. 24). Similarly, Griffin (2025) identifies how researchers, along with regulators and auditors, are reliant on reports, databases, and APIs of Very Large Online Platforms (VLOPs), which brings with it "preferred metrics and problem framings" (p. 246). While Griffin focuses on how risks are defined and managed, which means that "platform governance is dominated by the commercial interests of 'big tech,'" (p. 253) I build on her work through my interest in understanding of data. Indeed, the requirement for VLOPSEs to provide data access catalogs in the Delegated Act seeks to offset the information asymmetry of the "great data standoff" (Goanta et al., 2026), but also reinforces their role as gatekeepers and, as I will aim to demonstrate, is a valuable way to understand how platforms operationalize their definitions of data.

What data become available through reasoned requests have been categorized by many as nonpublic data. Selling et al. (2025) propose that this encompasses publicly available but inaccessible data such as content labels, visibility restrictions, search recommendations, or nonfollower engagement, as well as nonpublic, inaccessible data such as personalized recommendations, engagement

histories, and "even internal documents detailing governance decisions" (p. 13). While Botero Arcila et al. (2026) use the term "system-level data," which refers to "platform information that describes the overall functioning, performance, and state of the platform during a given timeframe, by including aggregated, structural and other process-level data" (p. 3) that researchers need "broad access" to. In Leerssen's (2023) analysis of feedback submitted to the Commission prior to the drafting of the Delegated Act, he identifies data categories as "data related to" the following five themes: users, accounts, and pages; content; content recommendation; ad targeting and profiling; and content moderation and governance. Collectively, these different perspectives usefully surface the use of examples and categories to describe what data might be requested, which might be useful for researchers in particular in the data access process. Even Botero Arcila et al. (2026), who offer a conceptualization of data, see it operationalized through specific types of data. Building on this, my analysis seeks to reveal these underlying understandings of how data are referred to.

Additionally, I extend the recent work of Peters and Weller (2026), who have interrogated the issue of data quality in debates about the implementation of data access provisions. They highlight how the focus on data *access* in the DSA and Delegated Act overlooks key indicators such as accuracy, completeness, consistency, and representativeness of data itself. Through an analysis of the two feedback periods related to the Delegated Act, they also highlight the divergence between data quality as a research standard amongst research communities, and data quality as an economic good and means to undermine critical research according to platforms. They propose that platforms "have an interest in strategically preventing the establishment of data quality standards for their data" (p. 4) because research on critical topics might challenge the legitimacy and dominance of platforms in society. Taking a step back to think through the meaning of data, which implicates issues of quality, similarly seeks to tease out the multiplicity of discourses.

Research Design

This study uses discourse analysis to critically examine the conceptualization of data in the context of the DSA's data access provision for researchers. In keeping with this methodological approach, discourse is understood as occasioned and dialectically connected with social practice (Fairclough, 2001), necessitating the close examination of language as situated across specific contexts (Taylor, 2001). I operationalize this by paying attention to the social construction of data across different types of texts and actors over key stages of the regulation process.

I focus on meaning-making in the regulatory discourse by examining relevant articles and recitals in the DSA, the draft Delegated Act, and the Delegated Act that address "data" or "information." The analysis centers on surfacing when and how data as a term is configured in the DSA and how the use types, categories, and examples of data reveal a conceptualization of data. Drawing on the guidance offered by the Irish Digital Service Coordinator for applicants, I reflect on what clarification is offered from a regulatory standpoint on data.

I also explore how platform companies negotiate the meaning of data through publicly submitted responses in the second period of feedback: responses to the draft delegated act. Here, the analytical focus is on attempts to impose limitations on what constitutes data. Out of the 25 designated as VLOPs or VLOPSEs, 15 are represented through 10 submissions, as shown in Table 1.[1] In these 10 submissions, providers raise several issues with the Draft Delegated Act, which I detail in Appendix 1 by articles and recitals. When providers did not reference articles and recitals, I identify the ones that most closely align with the issues raised. The colors visualize the concentration of topics across the submissions. Issues related to having access only to the application request summary, the burden of preparing and maintaining a data inventory, and inadequate evidence provided

Table 1. VLOPSE Categorized by Provider and Type; Word Count of Feedback Submissions.

Provider	VLOPSE	Type	Wordcount of Feedback Submission
AliExpress	AliExpress	Marketplace	–
Amazon	Amazon	Marketplace	7,150
Apple	Apple App Store	App store	–
Aylo Freesite	Pornhub	Content host	–
Booking.com	Booking.com	Marketplace	3,680
Google	Google Maps	Content host	1,179
Google	Google Play	App store	
Google	Google Search	Search engine	
Google	Google Shopping	Marketplace	
Google	YouTube	Social media	
Infinite Styles Services	Shein	Marketplace	–
LinkedIn	LinkedIn	Social media	–
Meta	Facebook	Social media	2,359
Meta	Instagram	Social media	
Microsoft	Bing	Search engine	–
NKL Associates	XNXX	Content host	–
Pinterest	Pinterest	Social media	1,177
Snap	Snapchat	Social media	1,108
Technius	Stripchat	Content host	–
TikTok	TikTok	Social media	2,657
Twitter	X	Social media	576
Webgroup	XVideos	Content host	–
Whaleco Technology	Temu	Marketplace	–
Wikimedia	Wikipedia	Content host	1,229

Note. VLOPSE = Very Large Online Platforms and Search Engine.

to Digital Services Coordinators regarding researcher independence and data protection were the most frequently mentioned. I also turn to their publicly available data catalogs to understand how they interpret data in practice by steering researchers toward data types and categories, which in turn is indicative of definitions of data.

Data in Regulatory Discourse

The term "data" is never defined in the DSA; it is not part of the 24 definitions provided in DSA Article 3, nor the eight definitions provided in Article 2 of the Delegated Act. While personal data as defined through the GDPR is cross-referenced, my starting point is that "data" is presented as a neutral, unproblematic term with assumed common ground.

In Article 40, the purpose of data for different actors is clearly asserted. While my focus is on researchers, Digital Services Coordinators and the Commission addressed in para 1 can request "access to data *necessary* to monitor and assess compliance with this Regulation." Recital 96 offers examples, such as VLOPSEs to give access to or report on "data on the processes and outputs of content moderation or internal complaint-handling systems"; but what form this might take or where the boundaries lie in what constitutes data to reveal such insights is not exemplified.

When it comes to vetted researchers, in comparison, the purpose of access to "previously undisclosed or under-disclosed data" narrows to systemic risk and anchors the understanding of data availability in a particular present where what is available from platforms might be known.

Unlike Recital 96, the contextualization in Recital 97 centers on potential data points such as "number of views or, where relevant, other types of access to content by recipients of the service prior to its removal by the providers of very large online platforms or of very large online search engines." Here, it appears that the operation of the provider's content moderation system is being analyzed through the aggregated interactional data for content that the platform has removed access to. While this concretizes a data access request and facilitates future-proofing, ambiguity remains present in the inclusion of "other types of access to content," which is seemingly offered as an illuminating example of how to interpret what data could be requested but might not fulfill requirements for a researcher to be specific in their reasoned request.

Research access under para. 12 differs from para. 4 (reasoned request) due to the public nature of the data, emphasized through the recital as content from *public* pages, *public* groups, or from *public* figures. When we should understand data as public, however, is unclear. Through the examples in Recital 98, "aggregated interactions with content from public pages, public groups, or public figures, including impression and engagement data such as the number of reactions, shares, comments from recipients of the service," we see a slightly more expansive understanding of data. Not only does interactional data in aggregated form constitute data, but user content itself does as well.

In the Delegated Act, "data" similarly is not defined, but a greater range of examples are provided in Recital 11:

> Current examples of such data include data related to users of the services, such as profile information, relationship networks, individual-level content exposure and engagement histories; interaction data such as comments or other engagements; data related to content recommendations, including data used to personalise recommendations; data related to the targeting of advertisements and profiling, including cost per click data and other measures of advertising prices; data related to the testing of new features prior to their deployment, including the results of A/B tests; data related to content moderation and governance, such as data on algorithmic or other content moderation systems and processes, including changelogs, archives or repositories documenting moderated content, including accounts as well as data related to prices, quantities and characteristics of goods or services provided or intermediated by the data provide

In stark comparison to the singular example of aggregated interactional data offered in the recital connected with Article 40, the Delegated Act offers a nonexhaustive list which presents a wider vision of data. As such, here it is proposed that data might be represented through numerical values or textual forms such as documentation on "content moderation and governance." In other words, these examples point to an understanding of data for researchers beyond metrics about user content.

The Delegated Act also necessitates that VLOPSEs provide information on data that could be requested. As Article 6(c) states:

> a DSA data catalogue, which describes the data assets, that may be accessed for the purposes set out in Article 40(4) of Regulation EU 2022/2065, as well as their data structure and metadata.

This article, thus, introduces the term "data assets"; assets offer a semantic reminder that data are understood as resources owned by platform companies, which generate value. In Recitals 7 and 8, it is clarified that the catalog "should describe the available data assets, their data structure and metadata," which "may rely on existing resources used for other purposes and audiences, such as advertising, content creation, or third-party app development." While useful in addressing issues of information asymmetry and the catalogs "should not bind or limit applicant researcher," catalogs reinforce the power that VLOPSEs have in suggesting how data should be conceptualized. In

repurposing data as structured and accessed by other actors such as advertisers, developers, and content creators, what might be considered relevant data for researchers is overlooked.

The ambiguity of "data" as a concept is in tension with the need for specificity from both VLOPSEs and researchers. Article 8 of the Delegated Act establishes what needs to be provided by the researcher, including:

> a description of the data requested, including format, scope and, where possible, the specific attributes, relevant metadata and data documentation, also considering the information made available pursuant to Article 6(4) of this Regulation.

This has been refined from the draft in which "granularity and type of data" (Article 8(c)) and "where possible, specific data points" (Article 8 (9c)) were also included, emphasizing the operationalization of data as a concept in terms of type and even data points. Even with the terminological changes from the draft, a need for granularity persists. The clarification for the researcher to take into account the data catalog also gestures to the primacy afforded to VLOPSEs' understanding of data.

DSA Article 40(5) and the Delegated Act Article 12 address the process for VLOPSEs to submit amendment requests. There are only two reasons that VLOPSEs can provide for not giving access:

(a) they do not have access to the data; and
(b) giving access to the data will lead to significant vulnerabilities in the security of their service or the protection of confidential information, in particular, trade secrets.

The conceptualization of data is fundamental to how the first ground may be used, which infers a naturally occurring status to data, obfuscating the processed and constructed nature. As van Drunen and Noroozian (2024) reflect, the flexibility in what data can be requested may be constrained as platforms may propose they "do not have access to the data (or more specifically, that they do not have the precise form of access requested by researchers)" (p. 5).

I propose that the Delegated Act indicates an expansiveness to what "data" can be requested, which is anchored in the systemic risk but recognizes the multiplicity of processes that underpin platform services: not only the circulation of user-generated content but advertising content, moderation and reporting, and testing. This resonates with data in a research context, which extends beyond, yet encompasses, platform data in which digital traces and activities are collected, processed, and inferred within the operations of the services. The examples provided regarding researcher access in DSA Article 40 primarily illustrate the latter compared to the Delegated Act.

Contesting a Broad Understanding of Data

Feedback to the Regulator. Moving to how platforms conceptualize data, I now examine feedback and data catalogs provided by VLOPSEs. In terms of the conceptualization of data through VLOPSE feedback on the Delegated Draft Act (DDA), the most explicit engagement with how data should be defined comes from Booking.com and Amazon. Both critique the broad interpretation in the DDA that they claim goes beyond the intention of the DSA; a strategic move to, in the absence of definitions, assert a drift from the purpose of data access.

According to these VLOPSEs, Article 40 is about providing access only to quantitative raw datasets (Booking.com) or quantitative data and raw datasets (Amazon), which is how data should be defined. Interestingly, X objects to access to raw data, expressing concern over research

misinterpretation, and furthering the collective narrative from platforms about the need for data to be processed and represented in structured, platform-determined ways. Amazon elaborates on its position by outlining how the DDA introduces a "more expansive definition" of data. Researchers should not "be empowered to audit data providers' internal processes," unlike supervisory authorities. They point to the use of "data" in Article 40 rather than the terms "documents" or "information," which are used in other parts of the DSA to clarify the distinction they see between data as purely quantitative compared with textual representations or other forms that might offer access into processes not considered to be about their use of data. This submission views the references, exclusively, to examples of quantitative data in the DSA's Recitals 97 and 98 as reflective of a narrow understanding of data. As such, they propose that the examples provided in Recital 12 of the DDA (which is unchanged in Recital 7 of the Delegated Act) deviate too far from data as conceived in the DSA, becoming too broad.

Concern with how the examples offered in the recital go beyond the scope is also made in submissions by Meta, TikTok, and Zalando. It "*actually* encompasses almost all the data we or any other VLOP would hold" (Zalando, emphasis added); TikTok specifically refers to data and documentation concerning the identification and moderation of content, while Meta refers to testing new features before deployment, or content moderation and governance. Here, we see pushback from providers not only regarding the narrowing of what constitutes data based on (re)defining the term data, but also about what types of data researchers should be able to request through Article 40.

There was widespread concern among platforms regarding the data catalog (referred to as an inventory in the draft). The recommendations and solutions for what an inventory should include also illustrate how platforms envision data as information about type of internal data (Snap), high-level description of categories of data (Pinterest), overview of examples of available datasets (TikTok), existing Transparency Center, assumed to fulfill data inventory requirements (Meta), data related to risk categories from DSA Transparency reports, risk reports, audits reports and audit compliance reports (Zalando), types of data related to risk in risk assessment (Booking.com). What this indicates is negotiation over data in terms of content (e.g., some providers limit data to risk categories; internal data), form (e.g., existing datasets; high-level categories and types), and availability. At stake with the latter is how providing access to data might necessitate production of data. Meta's submission in particular pushes for an understanding of data in the context of datasets. They ask researchers to use data available through existing datasets rather than "requesting bespoke datasets," alluding to data as requiring curation that involves labor and resources.

Meta requests clarification in the Delegated Act so providers are "only required to share data that exists; in other words, providers should not be required to create new data by, for example, implementing new loggers, creating new pipelines for data processing, or conducting additional analysis or research." They provide an example of requesting data about the quantity of content removed under a specific policy as "reasonable." However, if the request asked for "hyper-specific data segments," such as information about specific groups of users regarding content removal under a policy that is not ordinarily logged, this would be "unreasonable." This indicates a tension between data as made and accessible through datasets and data as requiring assemblage. Meta points to the issue of providing "data of low accuracy," insisting that data should be complete and accurate. They make a distinction between "existence" and "accuracy." Here, they express concern about how researchers will analyze data and what types of conclusions might be drawn based on requested data, alluding to perceived incompetence of researchers.

Finally, the debate surrounding how personal data fit into data access is also reflected in submissions. Some providers, such as Google and Amazon, argue that data should be anonymized and

pseudonymized, while TikTok seeks clarification on whether nonpublic user data fall within the scope.

In sum, the submissions reveal issues underlying Article 40 regarding the understanding of data and what researchers can request. As I argue, an unresolved tension in the delegated act is the dual nature of data as both an integral resource and an asset extracted, aggregated, and utilized by providers as part of services in platform capitalism, contrasted with data as evidence employed by researchers to understand, examine, and analyze phenomena. Critically, platforms used the feedback process to counteract the broader interpretation of the latter, in which other documentation and types of information would become requestable. Platforms positioned themselves as authorities over data, claiming to know what exists, what does not exist, and what constitutes good quality. Additionally, they advocated for varying limitations on the scope of data, ranging from granularity (e.g., only quantitative and aggregated metrics) to availability (e.g., existing in datasets). The different adjectives applied to data—raw data, personal data, internal data, and available data—hint at the discursive battle within the data terrain.

Operationalizing Data in Catalogs. In the data catalogs made publicly available by platforms after the adoption of the Delegated Act, platforms further demonstrate how they understand data. Table 2 provides an overview of the different approaches to operationalization and how data are being positioned.

First, some platforms rely on making visible existing datasets, data archives and access points. This is most narrowly conceived by Meta, in keeping with their remarks in the feedback period, who briefly describe and offer hyperlinks to four projects (Instagram Data Access Pilot for Well-being Research, URL Shares, US 2020 Facebook and Instagram Election Study, Data for Good), but only the URL Shares description outlines what the dataset entailed. The subheading for this webpage, "Datasets that empower researchers to study the political, economic and social impact of our platforms," further demonstrates Meta's framing of data through structured datasets, and while the hyperlinks show what data could be obtained, they remain constrained to the specificity of data-sharing arrangements between researchers and Meta in the previous or ongoing research projects.

Google also uses links to other webpages to direct researchers to "learn more about our data" through YouTube's Data API and Google Search's Trends and Search Results Result API. Unlike Meta, however, they do offer a brief description of what data are available within a single paragraph, which includes user-generated content such as description or comment text and aggregated user data such as like or view content. However, because of the reliance on existing documentation addressed at developers or researchers, other type of data not accessible to these groups, such as moderation or monetization processes, are not addressed.

However, hyperlinking practices to existing data availability is also used to demonstrate transparency in the case of Wikipedia. Their webpage emphasizes the "very high degree of open data and tooling already available to data," offering hyperlinks to how datasets connected with different parts of their service (e.g., MediaWiki content history or Zendesk Support ticket) are available and in what forms. In a similar but to a lesser extent, X points to their open-source code on algorithmic recommendations available through GitHub repositories, using the data catalog as a way to showcase transparency. For Wikipedia, "sensitive data," however, is considered to be not requestable, which includes the content of support tickets or names of who submitted or handled the ticket. Although this is the most explicit assertion of what data fall outside of the boundaries of what can be requested (according to the platform), despite its availability, other platforms similarly exercise their authority over data by relegating nonanonymized or aggregated data through their articulation of data through catalogs.

Table 2. How Data Are Organized and Represented in VLOPSE Data Catalogs.

						Data Type		Data Category							
VLOPSE	Word Count	Data Catalog Format	Existing Dataset URL	Data Description	Term for Data Points	Metrics	Content	User content & interaction	Reporting	Moderation	Illegal content	Transactions	Advertising	Algorithms	A/B Testing
AliExpress	409	Table		Per category	Fields	✓		✓	✓		✓	✓	✓		
Amazon	831	Table		Per metric	Metric	✓			✓	✓	✓	✓			
Apple App Store	262	List				✓	✓	✓	✓	✓		✓			
Pornhub	250	List	✓		Data asset	✓	✓	✓		✓			✓		
Booking.com	392	Table		Per data asset	Data asset	✓			✓	✓	✓				
Google Maps	76	Paragraph					✓	✓							
Google Play	45	Paragraph				✓	✓	✓				✓			
Google Search	141	Paragraph	✓			✓		✓							
Google Shopping	74	Paragraph	✓				✓	✓				✓			
YouTube	401	Table		Per category	Data asset	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Shein	475	Table				✓	✓	✓	✓	✓	✓	✓			
LinkedIn	1057	List		Per category	Data fields	✓	✓	✓					✓		
Facebook	477	Paragraph	✓			✓	✓	✓							
Instagram	477	Paragraph	✓			✓	✓	✓							
Bing	609	Paragraph		Approach		✓		✓							
XNXX	621	List		Per category	Data fields	✓	✓	✓	✓	✓	✓		✓		
Pinterest	565	List				✓	✓	✓	✓	✓	✓	✓	✓		
Snapchat	96	List				✓	✓	✓							
Stripchat	419	Table		Per category	Fields	✓		✓	✓	✓	✓	✓			
TikTok	677	Paragraph	✓	Per category		✓	✓	✓	✓	✓					
X	621	List		Per category	Data fields	✓	✓	✓	✓	✓	✓		✓		
XVideos	129	Paragraph	✓			✓	✓	✓							
Temu	1085	Table		Per category	Variables	✓		✓	✓	✓	✓				
Wikipedia	1044	Table	✓	Per category	Data structure	✓	✓	✓	✓	✓			✓	✓	

Note. VLOPSE = Very Large Online Platforms and Search Engine.

Second, other platforms offer tables or lists of categories of data where data and metadata are often presented. In these instances, platforms tend to offer greater levels of specificity and granularity about data from across their platform services. For example, TikTok refers to "example variables" across categories that are classified by systemic risk (illegal content, fundamental rights, civic discourse and electoral processes, public health, minors, and well-being). Amazon adopts the term "metric," reinforcing its position on quantifiable data, referring to "metric split" and "metric definition," as well as data structure (integer, datatime, text, and float) that clearly align with a computational approach to data.

Across the platforms who take this approach to displaying data across categories, data are expressed as aggregate, average or total numbers (Pinterest and TikTok), these metrics are allocated to more specific defined classified types (e.g., product category or report category, AliExpress; detection method or country of listing, Booking.com; report category or decision result, Temu; product type and notice type, Amazon) and in some instances, quantitative data and content itself are recognized as data (e.g., App description or review text, Apple; bio description, X; thumbnail and model names, XNXX and X Videos; project or patent descriptions, LinkedIn; comment string or public link, Snap).

These structures and fields demonstrate a range of interpretations of data by VLOPSEs and patterns in terms of platform types. While social media and content sharing sites differ in whether data pertaining to individual content items, users, and interactions are mentioned, online marketplaces, for example, are largely more aligned in what they classify as data. However, Shein, who similarly produces a table to display their catalog, is notably vague in their articulation of "data." In the description of the nonpublicly available types of data, classified as either "content moderation data" or "risk assessment and mitigation measures data," the opening structure "data on…" is always used. For example, the data related to mental health is described as "data on approval of new features and functionalities." Whilst not offering clarity in terms of data points, fields, or metadata that might be available, the absence of what constitutes data in this context may open up wider possibilities for requests and enable multiple understandings of data to be leveraged. Zalando, by comparison, offers the most extensive description of data and the widest range. Alongside categories that align with online marketplaces, they are the only VLOPSEs in their data catalog to include A/B testing described as "Experiment-level metadata (descriptions, hypotheses) and anonymized input metrics used in experiments" or content production and moderation policies, described as "Documentation on policies, community guidelines, and content steering data that relate to well-being risk." This illustrates how examples provided in the Delegated Act recital may translate into the formulation of the data catalog and how a broad interpretation of data is evident.

Bing also emerges as another outlier in their data catalog. They assert "how it works," outlining how datasets of anonymized search queries, frequencies, and metadata on selected topics will be provided, which, although presented differently, resonates with Google Search's data catalog. In addition, Bing includes an example of what a structured dataset would look like. This reveals the operationalization of data in the context of aggregated user interaction, solely focused on one form of engagement through the search engine rather than engagement with search results, search result recommendations, autocomplete search queries, integration of Co-pilot or sponsored content, for example.

Discussion

While the legal texts governing researcher access to data from VLOPSEs in the EU do not formally define data, they offer examples that hint at what the term could encompass through recitals that

accompany articles. Critically, examples that showcase a broader interpretation of data beyond aggregated metrics about interaction on platforms are found in the Delegated Act, which some VLOPSEs contended during the drafting process represented an illegitimate creep beyond the understanding put forward in Article 40 itself. Turning to the feedback illustrates how definitional boundaries were one of the central points that VLOPSEs sought amendment about. Although this was largely unchanged in the Act that came into effect, the operationalization of "data" through the data catalog evidences a tendency to narrowly approach data through structured datasets and clearly defined data types and structures, privileging quantified representations.

Attending to the discursive construction of data reveals how the term is a floating signifier. Following a Laclauian perspective, the floating signifier entails an "expression of the ambiguity" where there is instability in the construction of the signifier's meaning (Laclau, 2005, p. 110). As Torfing (1999, p. 301) puts it, the signifier is "overflowed with meaning because it is articulated differently within different discourses." The reappropriation and reinterpretation of "data" within the data access provision, I propose, can be understood in this vein, akin to arguments about the deployment of the term Artificial Intelligence or Big Data (see Barassi, 2016; Suchman, 2023). Based on my analysis of legal and platform texts, the meaning of data as narrowly tied to quantified and aggregated numerical representations of interaction on platforms (Lazer et al., 2021; Mayer-Schönberger & Cukier, 2013; Olteanu et al., 2019; van Dijck, 2013) or more broadly understood as representations and symbols that can be made meaningful to a researcher (Borgman, 2015; Drucker, 2011; Leonelli, 2022; Markham, 2013) "floats" across regulatory and commercial discourses.

In regulatory discourse, I propose that the meaning of "data" is unsettled. This is evident in recitals across both legal texts that offer examples of what data could be requested in different forms. Recitals connected to Article 40 gesture toward numerical forms, while those in the Delegated Act are more wide-ranging and include content and textual representation as data. Alongside this uncertainty around what could constitute data, the regulatory discourse assigns the ability to conceptualize data to both researchers (who must account for their data requirements in their request) and platform companies (who must provide a data catalog). On the one hand, this acknowledges that expertise and knowledge about data are distributed across industry and the research community, with the latter potentially afforded flexibility, and the disciplinary distinctions are recognized by not limiting data (at least in the Delegated Act) to a platform-centric view. Yet, on the other hand, while not bound to the nonexhaustive information, data catalogs offer categorical assertions of what data are and, in the face of information asymmetry, may shape the horizon of expectation of data for other actors. In this case, data as a floating signifier has "performative effects that serve the interests of those with stakes in the field" (Suchman, 2023, p. 4). The issue of how the scope will be governed is unaccounted for in legal texts; the Digital Services Coordinator of establishment is positioned as a mediating authority over the process that seems to preclude what counts as data.

In commercial discourse, the meaning of data is anchored in platform capitalism, with few VLOPSEs recognizing data outside of their circulation and moderation of user-generated content. Critically, data are assumed to be quantitative and aggregated metrics, and processed by the platform, ideally existing in datasets and databases. As some VLOPSEs responded to the draft version, they drew attention to what they perceived as a shift in the regulatory discourse from how Article 40 was aligned with their revision. This feedback advocated for the narrow operationalization of data and reinforced their position as the sole authority in the domain. The data catalogs from VLOPSEs largely reinforce the collective position put forward in the feedback period. As the analysis demonstrated, companies differ in how they operationalize data as well as in how they present to researchers the granularity and specificity of metadata and documentation. However, for researchers in computer and data science, the articulation of data is in keeping with disciplinary conventions in

comparison with qualitative traditions in the social sciences and humanities, in which platform-related phenomena are examined in data outside of quantitative representations. At risk is that not only do platforms shape the direction of data requests, but alternative understandings of data are marginalized or excluded from the commercially driven, datafication paradigm that comes to dominate what counts as data.

The ongoing debates about data access with the DSA (Botero Arcila et al., 2026; Leerssen, 2023; Peters & Weller, 2026; Selling et al., 2025) necessitate a wider discussion of what we mean by data. As a floating signifier, data are invested with different approaches to collecting, analyzing, and understanding platform-related phenomena, which traverse academic disciplines, the boundaries between industry, regulation, and academia, and companies that comprise VLOPSEs. These distinctions have implications for the roles that actors are expected to play in platform governance, which are obfuscated and collapsed when data are treated as a self-evident concept. While discourse theory is analytically attuned to this discursive struggle, as data requests are processed and approved, there may be a point to stabilization of meaning and as such turning to other frameworks will ensure material and infrastructural arrangements are examined. In keeping with the adoption of boundary objects in platform governance literature, future work could trace how the floating signifier of data also comes to function as a boundary object as data access materializes.

Conclusion

This article has focused on what data means in the context of the novel provision in the DSA to regulate how researchers gain access to data held by platforms to analyze systemic risks in the EU. The absence of a definition in the DSA and Delegated Act indicates a taken-for-granted assumption of data as a concept, despite long-standing academic debates surrounding its conceptualization that demonstrate it is not a neutral term. Employing a discourse analysis, I have examined how data are understood in Article 40 and the Delegated Act, paying close attention to the examples provided in the recital to guide legal interpretation of the provision. I have brought this into dialog with how VLOPSEs asserted what constitutes data through publicly submitted feedback on the draft Delegated Act, in which I highlighted that this is one of the core issues that platform companies push back on. Turning to the data catalogs published by VLOPSEs, I propose this is the site of operationalization. These lists, tables, and descriptions reveal a platform view of data as aggregated and quantified presentations, which are structured through predefined datasets and types. As such, I identify tensions between data in platform capitalism and data science with alternate academic disciplines, which are present both in legal texts and between the regulatory and commercial discourses.

As researchers are granted access to platform data through Article 40(4), future studies can trace how these tensions play out in practice, testing the boundaries of how data are understood. This might take the form of analyzing successful applicants as well as engaging with the research community about this specific issue of the definition of data. Consequently, there is an opportunity for clarification about the form that data can and should take in discussions on the viability of Article 40(4) between the research community and the Commission.

More broadly, this study points to the need for reflection on the role of terminology and our discursive construction of concepts in the research context. In ongoing debates about data access regimes in the EU (see Botero Arcila et al., 2026; Griffin, 2025; Goanta et al., 2026; Leerssen, 2023; Peters & Weller, 2026; Selling et al., 2025) and beyond (Bruns, 2021; Freelon, 2021; de Vreese & Tromble, 2023; Tromble, 2021), what we mean by data are assumed to be commonly held, overlooking critical distinctions in conceptualization across our disciplinary homes. This risks commercial platform-centered definitions or data science frameworks being positioned as

representative and the default, restricting how other paradigms, including humanities and social sciences in which qualitative methods are employed, may think about what counts as data. As the field of platform governance continues to advocate for access to data to understand, analyze, and recommend regulatory arrangements in digital society, embracing the multiplicity of data and clarifying our deployment of the term when we discuss data requirements is vital.

ORCID iD

Taylor Annabell https://orcid.org/0000-0002-6384-6754

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Note

1. As of 27 May 2025, the Commission terminated the designation of Stripchat as a VLOP and as such, fourth months later DSA obligations for VLOPs did not apply.

References

Barassi, V. (2016). Contested visions: Digital discourses as empty signifiers from the 'network'to 'big data'. *Communication and the Public*, *1*(4), 423–435. https://doi.org/10.1177/2057047316680220

Borgman, C. L. (2015). *Big data, little data, no data*. MIT Press.

Botero Arcila, B., Ramaciotti, P., & Cabale, E. (2026). Seeing in the dark: Towards a broad construction of the access to data provisions of the DSA. *Internet Policy Review*, *15*(1). https://doi.org/10.14763/2026.1.2085

Boyd, C. (2022). Data as assemblage. *Journal of Documentation*, *78*(6), 1338–1352. https://doi.org/10.1108/JD-08-2021-0159

Bruns, A. (2021). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. In *Disinformation and data lockdown on social platforms* (pp. 14–36). Routledge. https://doi.org/10.4324/9781003206972-2

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

D'ignazio, C., & Klein, L. F. (2023). *Data feminism*. MIT Press.

Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly*, *5*(1), 1–21. https://doi.org/10.63744/r4ysrh7ae534

European Commission, Directorate-General for Communication. (2023). Commission opens formal proceedings against X under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_23_6709

European Commission, Directorate-General for Communication. (2025). Commission preliminarily finds TikTok and Meta in breach of their transparency obligations under the Digital Services Act. https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503

European Commission, Directorate-General for Communications Networks, Content and Technology. (2025). Commission makes AliExpress' commitments under the Digital Services Act binding. https://digital-strategy.ec.europa.eu/en/news/commission-makes-aliexpress-commitments-under-digital-services-act-binding

Fairclough, N. (2001). Critical discourse analysis as a method in social scientific research. In R. Wodak & M. Meyer (Eds.), *Methods of critical discourse analysis* (pp. 122–138). SAGE Publications.

Flensburg, S., Lai, S. S., & Ørmen, J. (2025). At the mercy of the objects, we study: Epistemic consequences of proprietary digital research infrastructures. *New Media & Society*, *27*(4), 1928–1944. https://doi.org/10.1177/14614448251314397

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

Gillespie, T. (2017). Regulation of and by platforms. In J. Burgess, T. Poell, & A. Marwick (Eds.), *The SAGE handbook of social media* (pp. 254–278). Sage.

Goanta, C., & Iamnitchi, A. (2026). *If at first you don't succeed: Reflections on a rejected Art. 40 DSA data access request*. DSA Observatory Blog. https://dsa-observatory.eu/2026/03/12/if-at-first-you-dont-succeed-reflections-on-a-rejected-art-40-dsa-data-access-request/

Goanta, C., Zannettou, S., Kaushal, R., van de Kerkhof, J., Bertaglia, T., Annabell, T., Gui, H., Spanakis, G., & Iamnitchi, A. (2026). The great data standoff: Researchers vs. platforms under the Digital Services Act. *Proceedings of the Twentieth International AAAI Conference on Web and Social Media*, *20*, 874–888. https://doi.org/10.1609/icwsm.v20i1.42671

Griffin, R. (2025). Governing platforms through corporate risk management: The politics of systemic risk in the Digital Services Act. *European Law Open*, *4*(2), 223–253. https://doi.org/10.1017/elo.2025.17

Hase, V., Jef, A., Laura, B., Nico, P., Heleen, J., Theo, A., Carrière, T., de Vreese, C., Haßler, J., Loecherbach, F., Kmetty, Z., Möller, J., Ohme, J., Schmidbauer, E., Struminskaya, B., Trilling, D., Welbers, K., & Mario, H. (2024). Fulfilling data access obligations: How could (and should) platforms facilitate data donation studies? *Internet Policy Review: Journal on Internet Regulation*, *13*(3), 1–37. https://doi.org/10.14763/2024.3.1793

Kitchin, R. (2025). *Critical data studies: An A to Z guide to concepts and methods*. Polity Press.

Kitchin, R., & Lauriault, T. P. (2014). Towards critical data studies: Charting and unpacking data assemblages and their work. The Programmable City Working Paper, 2, http://papers.ssrn.com/sol3/papers.cfm? abstract_id¼2474112. https://doi.org/10.2307/j.ctt21h4z6m.6

Laclau, E. (2005b). Populism: What's in a name? In F. Panizza (Ed.), *Populism and the mirror of democracy* (pp. 32–49). Verso.

Lazer, D., Hargittai, E., Freelon, D., Gonzalez-Bailon, S., Munger, K., Ognyanova, K., & Radford, J. (2021). Meaningful measures of human society in the twenty-first century. *Nature*, *595*(7866), 189–196. https://doi.org/10.1038/s41586-021-03660-7

Leerssen, P. (2023). *Call for evidence on the delegated regulation on data access provided for in the Digital Services Act – summary & analysis*. European Commission. https://digitalstrategy.ec.europa.eu/en/library/digital-services-act-summary-report-call-evidence-delegatedregulation-data-access

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

Leonelli, S. (2015). What counts as scientific data? A relational framework. *Philosophy of Science*, *82*(5), 810–821. https://doi.org/10.1086/684083

Liesenfeld, A. (2025). The legal significance of independent research based on Article 40 DSA for the management of systemic risks in the Digital Services Act. *European Journal of Risk Regulation*, *16*(1), 184–196. https://doi.org/10.1017/err.2024.61

Markham, A. N. (2013). *Undermining 'data': A critical examination of a core term in scientific inquiry*. *First Monday*.

Mayer-Schönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. John Murray Publishers.

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

Olteanu, A., Castillo, C., Diaz, F., & Kıcıman, E. (2019). Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, *2*(13), 1–33. https://doi.org/10.3389/fdata.2019.00013

Peters, Y., & Weller, K. (2026). Little mentioning, moderate attention, great relevance: The quality of online platform data in the Digital Services Act. *Platforms & Society*, *3*, 1–18. https://doi.org/10.1177/29768624261438624

Rieder, B., & Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, *9*(4), 1–28. https://doi.org/10.14763/2020.4.1535

Rosenberg, D. (2013). Data before the fact. In *Raw data is an oxymoron* (pp. 15–40). MIT Press. https://doi.org/10.7551/mitpress/9302.003.0003

Selling, L. K., Iglesias Keller, C., Ohme, J., Klinger, U., & Vreese, C. D. (2025). Data *access* for *researchers* under the Digital Services Act: From *policy* to *practice* (No. 14). Weizenbaum Policy Paper.

Star, S. (2010). This is not a boundary object: Reflections on the origin of a concept. *Science, Technology, & Human Values*, *35*(5), 601–617. https://doi.org/10.1177/0162243910377624

Star, S., & Griesemer, J. R. (1989). Institutional ecology, 'translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907–39. *Social Studies of Science*, *19*(3), 387–420. https://doi.org/10.1177/030631289019003001

Suchman, L. (2023). The uncontroversial 'thingness' of AI. *Big Data & Society*, *10*(2), 1–5. https://doi.org/10.1177/20539517231206794

Taylor, S. (2001). Evaluating and applying discourse analytic research. In M. Wetherell, S. Taylor, & S. Yates (Eds.), *Discourse as data: A guide for analysis* (pp. 311–330). SAGE Publications.

Torfing, J. (1999). *New theories of discourse: Laclau, Mouffe, and Žižek*. Blackwell.

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), 1–8. https://doi.org/10.1177/2056305121988929

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. Oxford University Press.

van Dijck, J., Poell, T., & De Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press.

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*, 105946. https://doi.org/10.1016/j.clsr.2024.105946

Appendix 1: Articles Mentioned or Alluded to in Publicly Submitted Feedback From the Providers of VLOPSEs on the Delegated Act.

	Amazon	Booking.com	Google	Meta	Pinterest	Snap	TikTok	Wikimedia	X	Zalando
Article 6(4): Burden of data inventory		X		X	X	X	X	X		X
Article 7: VLOPs lack of input during draft of reasoned requests				X		X	X		X	X
Article 8: Insufficient evidence and requirements concerning researcher independence and protection of data	X	X	X	X		X	X		X	
Article 9(1): Lack of details about or input access modalities		X	X	X		X	X		X	
Article 9(2): Safety of data subjects	X	X	X	X				X	X	
Article 9(3): Transfer of data concerns			X	X			X			
Article 9(4): Secure processing environment as required/ costly				X		X	X	X		
Article 10(1)(b): Timeline for access	X	X	X			X		X		
Article 10: Being only provided a summary of the access request is insufficient for providers	X	X	X	X		X	X	X	X	
Article 12: Limited reasons for amendment requests	X			X					X	X
Article 13(4): Provider absorbs mediation costs						X		X		
Article 13: Provider should be able to object to request on other grounds	X	X		X					X	
Article 14: Consultation with external experts inaccessible to providers	X		X	X					X	
Article 15(1): Proposed timeframe too short for notification	X	X	X			X	X			
Article 15(2): provision of documentation for dataset as beyond scope		X								
Article 15(3): Provider cannot impose necessary requirements on data access	X	X	X				X			
Article 15(4): Provider cannot limit analytic tools		X	X					X		

(continued)

Continued.

	Amazon	Booking.com	Google	Meta	Pinterest	Snap	TikTok	Wikimedia	X	Zalando
Rec. 9: Commitment letter for data protection as lacking		X	X							
Rec. 12: Data inventory and examples							X	X	X	X
Rec. 18: Safety of data subjects		X						X		
Rec. 23: Consultation with experts							X			
Rec. 28: Legal obligations of personal data			X					X		

Note. VLOPSE = Very Large Online Platforms and Search Engine; VLOP = Very Large Online Platforms.