



DATA ACCESSIBILITY AS A PLATFORM AFFORDANCE

The Shaping of Research Possibilities on Online Platforms with Prospective Legal Methods

By Sophie Stalla-Bourdillon* and Marion Lieutaud*

* Sophie Stalla-Bourdillon, Co-Director of the Brussels Privacy Hub, Researcher at Vrije Universiteit Brussel, Sophie.Stalla-Bourdillon@vub.be.

* Marion Lieutaud, Lecturer in Social Data Science, Edinburgh University, marion.lieutaud@ed.ac.uk.

The Brussels Privacy Hub publications are intended to circulate research in progress for comment and discussion. Available at <https://brusselsprivacyhub.com/>. Reproduction and translation for non-commercial purposes are authorised, provided the source is acknowledged.

DISCLAIMER

The opinions expressed in this paper are those of the author/s.

Contents

Abstract	4
1. Introduction	5
2. Data Accessibility as a De Jure Affordance	7
2.1 Perceived v. Real Affordance	7
2.2 Closed v. Open Affordance	9
2.2.1 Access to Publicly-Accessible Data	10
2.2.2 Access to Enclosed Data	12
3. Data Accessibility for Researchers	14
3.1 Identifying Researchers' Data Needs	15
3.2 Interpreting the Law to Meet Researchers' Data Needs	20
3.2.1 Secure processing environments	20
3.2.2 Web scraping	21
3.2.3 Data donations	28
4. Conclusion	31

Abstract

This working paper examines data accessibility as a key affordance of online platforms for researchers and civil society actors. Building on Gibson's theory of affordances and Treem and Leonardi's taxonomy of social media affordances, it conceptualises platform affordances as possibilities for action shaped jointly by technical design and governance arrangements. The working paper argues that data accessibility is not only a technical affordance but also a legally shaped one, whose scope is profoundly influenced by legal interpretation. Combining digital methods with legal methods, the working paper highlights a persistent gap between perceived and real affordances, as well as between de facto and de jure affordances, driven in part by strategic exploitation of legal ambiguity by platforms. To address this gap, the working paper distinguishes between traditional doctrinal analysis and prospective doctrinal analysis, advancing speculative interpretation as a method for exploring alternative legal meanings and reconfiguring affordances. Introducing the distinction between closed and open affordances, it shows that legally shaped affordances are inherently open-ended and evolve through interpretive practices, practices in which both regulators and research organisations should proactively engage in order to resist interpretative entrepreneurship by very large platforms. Through an analysis of key interpretive battles re-invigorated by the adoption of the DSA in the European Union—particularly around web scraping and trade secrets—the working paper illuminates the indeterminacy of the DSA and warns against reductionist interpretations. Drawing lessons from researchers' actual needs, the working paper calls for preserving diverse contexts of data access, including web scraping, and ultimately argues for stackable data-access regimes across the GDPR, DSA, DMA, and Data Act, alongside institutional support structures, with a view to further strengthen platform research.

1. Introduction

As Venturini et al explain, “[d]igital methods may (...) be defined as techniques for the ongoing research on the affordances of online media.”¹ Gibson initially introduced the theory of affordances in the domain of ecology,² a theory which was later leveraged in the field of product design³ and technology,⁴ and then became popular in the field of social media research,⁵ where Treem and Leonard’s taxonomy of social media affordances (visibility, editability, persistence, association)⁶ has been gaining significant traction.⁷

Following Treem and Leonard and understanding affordances as potential for action,⁸ the affordances of online platforms can be defined as the possibilities for action they provide to a range of stakeholders, including users, developers, organisations, and institutions, shaped by the ways in which these platforms are designed, structured, and governed.

This paper is interested in affordances available to one category of stakeholders: researchers broadly defined to include civil society actors. More specifically, it examines an affordance that is crucial for platform research: data accessibility. In this context, data accessibility refers to the extent to which data, including publicly-available content, generated on/through online platforms can be reached, retrieved, and meaningfully used by researchers. Data accessibility, as a type of online platform affordances, is shaped both by technical design (e.g., APIs, search functionalities, data export tools) and by governance rules (e.g., terms of service).

This paper argues that legal methods, together with digital methods, can illuminate how the affordances of online platforms, and data accessibility in particular, are shaped, in part, by strategic interpretations and exploitations of legal ambiguity. A gap therefore emerges, not only between perceived and real affordances, but also between *de facto* (in fact) and *de jure* (in law) affordances, a gap that stakeholders invested in sustaining long-term research efforts, such as research organisations, regulators, and courts, should work to bridge.

On legal methods more specifically, the paper distinguishes between traditional doctrinal analysis, focused upon what the Law is, and prospective doctrinal analysis, focused upon what the Law could be. It underscores the value of speculative interpretation as a means of critically interrogating both *de facto* and *de jure* affordances. Speculative interpretation is viewed as an interpretive stance that not only seeks to reconstruct the realistic or to-be expected meaning of legal norms but also projects alternative *possible* meanings, opening interpretive space for critique and transformation. Such a stance highlights how affordances, particularly legally shaped affordances, are never fixed or closed

¹ Tommaso Venturini and others, ‘A Reality Check(List) for Digital Methods’ (2018) 20 New media & society 4195.

² James J Gibson, *The Ecological Approach to Visual Perception* (1st Edition, Taylor and Francis Group 2014) <<https://doi.org/10.4324/9781315740218>>.

³ Donald A Norman, *The Psychology of Everyday Things* (Basic Books 1988).

⁴ Ian Hutchby, ‘Technologies, Texts and Affordances’ (2001) 35 Sociology 441.

⁵ See e.g., Manju Ahuja, Patel Pankaj and Suh Ayoung, ‘The Influence of Social Media on Collective Action in the Context of Digital Activism: An Affordance Approach’ *Proceedings of the 51st Hawaii International Conference on System Sciences* (2018) <<http://hdl.handle.net/10125/50163>>.

⁶ Jeffrey W Treem and Paul M Leonardi, ‘Social Media Use in Organizations: Exploring the Affordances of Visibility, Editability, Persistence, and Association’ (2013) 36 Annals of the International Communication Association 143.

⁷ Alexander Ronzhyn, Ana Sofia Cardenal and Albert Batlle Rubio, ‘Defining Affordances in Social Media Research: A Literature Review’ (2023) 25 New Media & Society 3165.

⁸ *ibid.*

but remain open, and subject to evolution through ongoing interpretive practices. This holds true even when taking institutional constraints such as canons of statutory interpretation and the rule of precedents seriously. Building on this, the paper introduces the dichotomy of closed versus open affordances, showing that legally shaped affordances are inherently open ended and reconfigurable, that their trajectory depends on the interpretive choices made by actors capable of influencing the development of the law, and that adopting a conservative interpretive stance at the stage when research proposals are built effectively forecloses such openness.⁹

This paper proceeds in two main steps. Section 2 highlights ambiguities in legal rules governing data access by researchers, ambiguities that make *de jure* affordances broader than computational social scientists, including digital sociologists, often perceive. Starting from the premise that legal interpretation does not arise from external factors, such as gaps in statutes, judicial confusion, or self-restraint, but rather from the very nature of the Law itself, Section 3 shows how an interpretation oriented toward researchers' interests could configure data accessibility as a *de jure* affordance in the future and illuminates the room for manoeuvre available to regulators and courts tasked with steering the conflicts between platforms and researchers, a battle reinvigorated by the adoption of the Digital Services Act (DSA)¹⁰ in the European Union (EU). This paper suggests that it is with this interpretive latitude in view that research organisations should actively support researchers engaged in platform research and fight to preserve diverse contexts of data access, an essential counter to the increasingly aggressive interpretive entrepreneurship.

Before proceeding with the analysis, it is worth clarifying three concepts—web scraping, web crawling, and data mining—in order to facilitate dialogue between computational researchers and lawyers, which this paper very much seeks to encouraged.

- **(Web)-Scraping:** the process of automatically extracting data from the web. Web-scraping (or scraping, for short - though in theory you could also be scraping data from other sources than the internet) is a method of data collection. A scraper is usually tailored to extract specific contents within specific websites (e.g. tweets on Twitter/X, or posts and usernames on Reddit) and to parse and store that content into a dataset that can later be analysed for e.g. research.
- **(Web)-Crawling:** crawling and scraping are sometimes used interchangeably, but crawling normally refers to a process runs through links to systematically browse through entire websites, typically for the purpose of indexing (e.g. the google crawler indexes the web constantly by crawling periodically through every webpage): a crawler is not targeted; it is not focused on content extraction and its scope can be to explore the entire web, whereas a scraper is usually smaller in scope and targeted to the extraction of specific content. In practice, though, crawling and scraping overlap and largely work together: crawlers need to scrape specific data (mostly hyperlinks) in order to browse through and index website.

⁹ Documentation of practices at the University of Amsterdam illustrates this institutional conservatism. Jef Ausloos and others, 'Information Law and the Digital Transformation of the University Part II. Access to Data for Research' (2023) <<https://openresearch.amsterdam/en/page/109109/information-law-and-the-digital-transformation-of-the-university>>.

¹⁰ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277, 27.10.2022, S. 1-102. See the preliminary findings of the European Commission against Tik Tok and Meta following formal proceedings launched under the DSA, including Article 40(12). European Commission, 'Commission Preliminarily Finds TikTok and Meta in Breach of Their Transparency Obligations under the Digital Services Act' (European Commission, 24 October 2025) <https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2503>.

Meanwhile, scrapers crawl parts of the content of websites before identifying the specific content or item types they want to extract.

- **Data mining (or text mining):** data mining is the process of finding patterns in data – specifically in text in the case of text mining. Data mining often hints specifically to the use of machine learning methods of analysis to identify patterns in the data in more or less supervised ways (i.e. the process of identifying these patterns can be more or less autonomously data-driven or researcher-driven). Confusingly, legal texts sometimes use the term ‘data mining’ to refer to data collection (i.e. the process of creating or extracting the raw data itself) as well as data analysis, but – at least in digital methods – data/text mining are analytical tools, not data collection tools.

2. Data Accessibility as a De Jure Affordance

2.1 Perceived v. Real Affordance

Norman adapted Gibson’s ecological psychology concept of affordances into the field of design and human–computer interaction. For Norman, affordances refer to the action possibilities that an object or interface offers to a user, as determined both by its physical properties and by the ways it can be interacted with. To be more precise, what Norman discusses in the *Psychology of Everyday Things* is essentially “perceived affordances,” which are different from real affordances.¹¹ This is because in the world of human-computer interaction, perceived affordances are much more important.¹² By introducing the concept of perceived affordances, Norman captured the critical distinction between what an object actually allows and what a user believes it allows.

This distinction is especially relevant to online platform affordances. In this context, researchers’ perceptions of affordances are directly shaped by platform governance practices, most notably their restrictive stance on data access as articulated in their terms of use and reinforced through litigation. As a result, for many researchers, the perceived affordances of online platforms with respect to data accessibility have been defined over the years less by openness than by closure.

Researchers have encountered real barriers that made or continue to make access actually fragmented, precarious, or prohibitively costly. Once-common tools such as public application programming interfaces (APIs), have been scaled back, monetised, or discontinued altogether. APIs are data infrastructures allowing platforms to give access to data pre-packaged by the platform; access to API data is conditioned by whether the platform developed an API in the first place, whether it will grant access to the API, and under what conditions. Twitter/X, for instance, replaced its free research API with a paid model that makes large-scale data collection financially unfeasible for most academic projects.¹³ Meta has wound down support for CrowdTangle, a tool that was widely used by journalists and researchers to study the circulation of information on Facebook, and rebuilt its data access service, which is now only offered to researchers affiliated with academic institutions or other not-for-profit organisations holding scientific or public interest research as a primary purpose or core

¹¹ Donald A Norman, ‘Affordance, Conventions, and Design’ [1999] interactions 38.

¹² *ibid.*

¹³ Chris Stokel-Walker, ‘Twitter’s \$42,000-per-Month API Prices Out Nearly Everyone’ [2023] *Wired* <https://www.wired.com/story/twitter-data-api-prices-out-nearly-everyone/?utm_source=chatgpt.com>; Ryan Murfieldt and others, ‘RIP Twitter API: A Eulogy to Its Vast Research Contributions’ <<https://arxiv.org/abs/2404.07340>>.

activity.¹⁴ Even when APIs are available, researchers have shown how changes to access modalities can adversely affect research.¹⁵ In some cases involving research on the impact of social media on mental health, on misinformation or on political views, platforms deliberately withdrew or restricted data access by redesigning their services.

Web-scraping publicly available content had long been understood by researchers as technically feasible and, in many cases, essential for studying digital communication and digital practices. The scraping (or more maximally the crawling) of internet data involves building programs which can navigate to and through websites, collect data from them and parse it into datasets. Functionally, scraping data does not require that the platform being scraped grants access or pre-packages data for scrapers, and scraping programs range from highly modular and sophisticated to very straightforward to build. However, because scraping platform content has increasingly been challenged as a violation of terms of service, researchers have gradually come to view these technically feasible practices as potentially legally impermissible and have adapted their methods accordingly.

Although early legal developments have been mainly US centred, they have had repercussions for the global research community and have since been followed by litigation in the EU.¹⁶ On web scraping in particular, confusion remains high in the EU.¹⁷ The high-profile *hiQ v. LinkedIn* litigation in the United States exemplifies how platforms invoked legal arguments to block data access, creating uncertainty despite some wins.¹⁸

More recent developments, however, show a willingness of judges to sanction platforms' attempt to silence researchers.¹⁹ A few commentators have noticed other positive trends. Niva Elkin-Koren, for

¹⁴ Vittoria Elliott, 'Meta Kills a Crucial Transparency Tool At the Worst Possible Time' [2024] *Wired* <https://www.wired.com/story/meta-kills-crucial-transparency-tool-worst-possible-time/?utm_source=chatgpt.com>. The Meta Content Library and API were officially launched in November 2023. Nick Clegg, 'New Tools to Support Independent Research' (*Meta*, 21 November 2023) <https://about.fb.com/news/2023/11/new-tools-to-support-independent-research/?utm_source=chatgpt.com>. See Justin Hendrix, 'The Demise of CrowdTangle and What It Means for Independent Technology Research' [2024] *Tech Policy.PRESS* <<https://www.techpolicy.press/the-demise-of-crowdtangle-and-what-it-means-for-independent-technology-research/>>.

¹⁵ Brittany I Davidson and others, 'Platform-Controlled Social Media APIs Threaten Open Science' (2023) 7 *Nature Human Behaviour* 2054. From their earliest days, platforms have maintained an ambivalent relationship with platform research. Axel Bruns, 'After the "APIcalypse": Social Media Platforms and Their Fight against Critical Scholarly Research' (2019) 22 *Information, Communication & Society* 1544.

¹⁶ Daphne Keller, 'Using the DSA to Study Platforms' (*Verfassungsblog on Matters Constitutional*, 27 October 2025) <<https://verfassungsblog.de/dsa-platforms-digital-services-act/>>; Bobby Allyn, 'Media Matters Sues Elon Musk's X over "libel Terrorism" Legal Assault' (*npr*, 10 March 2025) <<https://www.npr.org/2025/03/10/nx-s1-5323844/media-matters-elon-musk-new-lawsuit>>; Simone Ruf, 'X Prevents Research on Potential Election Interference' (*Gesellschaft für Freiheitsrechte*) <<https://freiheitsrechte.org/en/themen/digitale-grundrechte/x>>.

¹⁷ See European Commission, 'Improving Access to and Reuse of Research Results, Publications and Data for Scientific Purposes' (2024) 124–125 <<https://op.europa.eu/en/publication-detail/-/publication/77395a15-133b-11ef-a251-01aa75ed71a1>>.

¹⁸ See in particular *hiQ Labs v LinkedIn Corp* (2019) 938 F.3d 985 (9th Cir 2019); *hiQ Labs v LinkedIn Corp* (2022) 31 F.4th 1180 (9th Cir 2022). Although commercial scrapers received some positive clarifications, LinkedIn and hiQ Labs ultimately settled to avoid a ruling finding for LinkedIn on the basis of a breach of LinkedIn's User Agreement due to its scraping and creation of fake accounts. See 'hiQ and LinkedIn Reach Proposed Settlement in Landmark Scraping Case' (*Proksauer 150*, 8 December 2022) <<https://www.proksauer.com/blog/hiq-and-linkedin-reach-proposed-settlement-in-landmark-scraping-case?utm>>.

¹⁹ See a battle triggered by X Corp., which sued the Center for Countering Digital Hate (CCDH), a non-profit organisation studying the spread of hateful content on social media, alleging that CCDH had deliberately scraped data from Twitter in violation of its terms of service, a U.S. District Judge dismissed the case in early 2024, striking the breach of contract and tort claims and emphasizing that it was 'evident' X had filed the lawsuit to silence critics and punish them for their speech.

example, notes the premises of a signal for finding restrictive terms of use unenforceable when they undermine the fair use doctrine.²⁰ And others, arguing that “website and services that are accessible to the general public without requiring authentication entail significantly lower legal risk,” have attempted to produce clearer guidelines for researchers, while acknowledging that the ethical and legal frameworks governing scraping remain “ambiguous and often contentious.”²¹

2.2 Closed v. Open Affordance

One fundamental aspect of the Law that it is often overlooked by researchers, perhaps particularly by non-legal scholars working with computational or digital methods, is that the Law is inherently a work in progress. Legal rules cannot anticipate every possible situation and as such are not easily translated into organisational compliance rules, especially when one leaves the rule’s core to reach its penumbra;²² rather, legal rules are characterised by what legal theorists, such as Hart, call an open texture.²³ This means that legal rules are formulated in general terms and inevitably contain areas of vagueness or indeterminacy, which must be interpreted in light of specific contexts, including the state of the art, and often require further specification to be meaningful within a particular organisational setting.

For computational researchers accustomed to working with technical systems that are rule-bound and often deterministic, this interpretive character of the Law can be difficult to grasp. Yet it is precisely this openness that makes the Law both a constraint on and a potential enabler of governed behaviour, including research activities, depending on how ambiguities are ultimately resolved on the ground for a particular context. In this sense, legally shaped affordances, such as data accessibility, are fundamentally open affordances, associated with a set of potentially divergent meanings, as opposed to closed affordances, whose meaning is unambiguous and precisely specified, leaving little room for interpretation.

Importantly, such de jure affordances are sites of struggle and thereby interpretative battles. As such, they are subject to the pressure of legal arrangements enforced by private regulators such as online platforms. Legal arrangements encompass the relational mechanisms, e.g., terms of use, through which the Law, both hard and soft, is mobilised to serve private interests. When deployed by online platforms, these arrangements operate as integral components of broader governance systems, interacting with technological design to shape user behaviour, and are one manifestation of interpretive entrepreneurship. Durkee defines interpretive entrepreneurship as “the act of developing the law by interpreting it,”²⁴ and argues that “[i]nterpretive entrepreneurs might exploit legal uncertainty to pursue business plans, (...) and change the regulatory environment along the way. Or they may shop around favorable interpretations to regulators, or publicize reputation-friendly interpretations to investors and the public. Through each mode, interpretive entrepreneurs seek to

X Corp v Center for Countering Digital Hate Inc et al [2024] US District Court, Northern District of California 3:23-cv-03836-CRB.

²⁰ Niva Elkin-Koren, ‘A2D for Researchers in Digital Platforms Bridging the Transatlantic Divide’ (23 February 2024) <<https://verfassungsblog.de/a2d-for-researchers-in-digital-platforms/?utm=>>.

²¹ Megan A Brown and others, ‘Web Scraping for Research: Legal, Ethical, Institutional, and Scientific Considerations’ (2025) 12 Big Data & Society 20539517251381686. These authors however do not engage into detailed analysis of specific legal barriers to web scraping, particularly from an EU perspective.

²² HLA Hart, *The Concept of Law* (Oxford University Press 1972).

²³ See *ibid.*

²⁴ Melissa J Durkee, ‘Interpretive Entrepreneurs’ (2021) 107 Virginia Law Review 431.

influence legal development.”²⁵ Inevitably, given the substantial resources devoted to interpretive entrepreneurship, it exerts a critical influence on perceived platform affordances.

However, and it is crucial, data accessibility, as a *de jure* affordance, cannot be reduced to interpretative entrepreneurship. Data accessibility as a *de jure* affordance is fundamentally dialogical, reflecting tensions that arise between diverging perspectives, although these tensions are not necessarily perceived as such by stakeholders leveraging the affordances and in particular computational researchers.²⁶

That data accessibility as an affordance is dialogical is true even if bespoke statutory provisions have been adopted to govern data accessibility practices, as illustrated by the provisions of the DSA governing researcher access to two types of data: publicly accessible data and enclosed data. A close reading of Article 40 DSA sheds light upon two key interpretative battles: the lawfulness of scraping and the protection of trade secrets. As a result, the DSA’s impact can vary with interpretation, making it potentially more or less progressive. Regulators must therefore exercise particular care when entering the interpretative battlefield.

2.2.1 Access to Publicly-Accessible Data

When the wording of the DSA is compared with the substantive language used by platforms in their terms of service, a first interpretative battle emerges focused upon publicly-accessible data.

Let’s start with a concrete example. X’s Terms of Service, as of 15 November 2024, state that “crawling or scraping the Services in any form, for any purpose without our prior written consent is expressly prohibited.”²⁷ These provisions leave no scope for crawling or scraping, even for the purposes of scientific research, thereby constraining researchers’ perceived ability to access data. Meta’s Terms of Service are drafted somewhat differently, allowing for more nuance, although they remain highly ambiguous. In the version effective 1 January 2025, one reads: “You may not access or collect data from our Products using automated means (without our prior permission) or attempt to access data that you do not have permission to access, regardless of whether such automated access or collection is undertaken while logged in to a Facebook account. We also reserve all of our rights against text and data mining.”²⁸

These terms of service have been adopted after the entry into applicability of the DSA²⁹ and more than four months after the designation of these actors as very large online platforms and search engines (VLOPSEs) within the meaning of Article 33 DSA. Yet, under Article 40(12) DSA, VLOPSEs are under an obligation to “give access without undue delay to data, including, where technically possible, to real-time data, provided that the data is publicly accessible in their online

²⁵ *ibid.*

²⁶ Fiona Maine and Rupert Wegerif, ‘Dialogism’ in Vlad Petre Glăveanu (ed), *The Palgrave Encyclopedia of the Possible* (Springer International Publishing 2022) <https://doi.org/10.1007/978-3-030-90913-0_239>.

²⁷ X, ‘Terms of Service’ (X, 15 November 2024) <<https://x.com/en/tos>>.

²⁸ Meta, ‘Terms of Service’ (Meta, 1 January 2025) <<https://www.facebook.com/terms>>. Compare with other platform practices described in Mozilla, ‘Public Data Access Programs - A First Look’ (2024) <https://assets.mozoprod.net/network/documents/Public_Data_Access_Programs__A_First_Look_Final_cExWfCH.pdf>.

²⁹ Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act) OJ L 277, 27.10.2022, pp. 1–102.

interface by researchers,”³⁰ when covered researchers meet certain conditions.³¹ These terms of service thus seem to adopt a very conservative reading of Article 40(12), which consists essentially in arguing that 1) VLOPSEs are merely required to give access to publicly-accessible data upon express request by researchers and 2) access to publicly-accessible platform data through crawling or scraping is not expressly permitted by Article 40(12) and therefore can be unilaterally banned by VLOPSEs.

To be clear, while data crawling or scraping are not directly regulated by Article 40(12), the provision does not expressly grant VLOPSEs the right to unilaterally restrict these activities. In fact, to determine whether VLOPSEs have the right to unilaterally restrict these activities for research purposes, additional analysis is required, which would involve considering legal rules governing intellectual property (IP), such as copyright and database rights, contract, as well as possibly other laws, such as laws restricting access to computer systems and data. What is more, the reference to real-time data in Article 40(12) could support an *a fortiori* argument, suggesting that, in principle, crawling and scraping cannot be restricted, and that, in the case of VLOPSEs, they are conditionally obliged to provide access to real-time data. Some commentators have also suggested that an extensive definition of public data should be adopted under Article 40(12) DSA to give covered researchers access to “aggregate outputs of algorithmic recommendation systems,”³² “full historical access” to public data,³³ data from platforms that make login credentials publicly available and data from huge private groups.³⁴ Moreover, these commentators emphasise the need to provide a range of access methods to meet the needs of a diverse user base, including users such as journalists who may not be familiar with computational techniques.³⁵

By demonstrating that several interpretations are possible for Article 40(12) DSA, this paper highlights its inherent indeterminacy and underscore the risk that its formulation could ultimately be counterproductive for researchers, who would then be compelled to advance a two-pronged argument: 1) IP, contract and other applicable laws do not prevent them from crawling or scraping and 2) Article 40(12) does not prevent them from crawling or scraping without rightsholders’ prior authorisation.

This paper suggests that a better way forward would be to recognise the lawfulness of good-practice platform scraping conducted for scientific research as unpacked in the following section and impose upon VLOPSEs an obligation to provide real-time access to publicly accessible data, broadly defined, and through multiple technical means.

Such an interpretative stance is particularly warranted to avoid rendering the scraping of many platforms of significant relevance, though not of the same scale as VLOPSEs, most notably digital labour platforms, unlawful, thereby perpetuating the difficulties of conducting quantitative research on digitally mediated gig work.

³⁰ Article 40(12) DSA.

³¹ Note that the category of covered researchers under Article 40(12) DSA is broader than under Article 40(4) DSA.

³² Mozilla (n 28) 26.

³³ *ibid* 11.

³⁴ Daphne Keller, ‘The Stakes of “Publicly Accessible” - Researchers’ Rights to Data under the DSA’ (*Verfassungsblog on Matters Constitutional*, 12 December 2025) <<https://verfassungsblog.de/dsa-fine-x-research-data/>>.

³⁵ Mozilla (n 28) 6.

2.2.2 Access to Enclosed Data

A second interpretative battle emerges when the wording of the DSA is compared with that of other newly introduced data-access mandates targeting enclosed data.

Under Article 40(4) DSA, VLOPSEs are obliged to provide vetted researchers with access to data, upon receiving a reasoned request from the Digital Services Coordinator (DSC) of their Member State of establishment, where the research contributes to the detection, identification, and understanding of systemic risks in the EU and/or to the development of associated mitigation measures. Article 40(5)(b) DSA gives VLOPSEs, however, the possibility to oppose the reasoned request by arguing “that 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.”³⁶

Both the DSA and the Delegated Act³⁷ adopted by the EC under Article 40(13) DSA to further specify the conditions for access under Article 40(4) DSA (Delegated Act) do not specify how the protection of confidential information, in particular trade secrets, should impact data access. The DSA is therefore less expensive and may raise more interpretative challenges than other data access regimes, in particular the data access regimes set up under the Data Act (DA)³⁸. We touch here upon the indeterminacy of the DSA once more, indeterminacy which should be addressed by interpreting the various data access regimes together in a cohesive way.

The DA, the Digital Markets Act (DMA)³⁹ and the DSA all introduce new data access mandates from which researchers can benefit. While Article 40 DSA is built specifically for researchers, the DA expressly foresees that the new data access mandate can benefit researchers. Recital 33 DA specifies that “[a] third party to whom data is made available may be a natural or legal person, such as a consumer, an enterprise, a research organisation, a not-for-profit organisation or an entity acting in a professional capacity.”⁴⁰ The DMA on the other hand seem to adopt a very broad concept of third party authorised by an end user or a business user, without breaking it down with examples.

The DA is particularly interesting as it extensively covers the situation in which the data covered by the data access mandate could comprise trade secrets. Recital 31 specifies that the DA “requires data holders to disclose certain data to users, or third parties of a user’s choice, even when such data qualify for protection as trade secrets.”⁴¹ Said otherwise, “[d]ata holders cannot, in principle, refuse a data access request under this Regulation solely on the basis that certain data is considered to be a trade secret, as this would subvert the intended effects of this Regulation.”⁴²

³⁶ Article 40(5)(b) DSA.

³⁷ Commission Delegated Regulation (EU) of 1.7.25 supplementing Regulation (EU) 2022/2065 of the European Parliament and of the Council by laying down the technical conditions and procedures under which providers of very large online platforms and of very large online search engines are to share data with vetted researchers C(2025) 4340 final.

³⁸ Regulation (EU) 2023/2854 of the European Parliament and of the Council of 13 December 2023 on harmonised rules on fair access to and use of data and amending Regulation (EU) 2017/2394 and Directive (EU) 2020/1828 (Data Act) OJ L, 2023/2854, 22.12.2023.

³⁹ Regulation (EU) 2022/1925 of the European Parliament and of the Council of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act) OJ L 265, 12.10.2022, pp. 1–66.

⁴⁰ Recital 33 DA.

⁴¹ Recital 31 DA.

⁴² Recital 31 DA.

With this said, if trade secrets are present within the covered data, the data may be specifically protected. Recital 31 continues by stating that “data holders should identify trade secrets prior to the disclosure, and should have the possibility to agree with users, or third parties of a user’s choice, on necessary measures to preserve their confidentiality, including by the use of model contractual terms, confidentiality agreements, strict access protocols, technical standards and the application of codes of conduct.”⁴³ It is only when there is no agreement on necessary measures to preserve information confidentiality or when the user/third party fails to implement such measures that the data holder can withhold or suspend the sharing of data identified as trade secrets, as long as it notifies both the user/third party and the competent authority.

What is more, but this should happen only in exceptional circumstances, Recital 31 DA foresees that “a data holder who is a trade secret holder should be able, on a case-by-case basis, to refuse a request for the specific data in question if it is able to demonstrate to the user or to the third party that, despite the technical and organisational measures taken by the user or by the third party, serious economic damage is highly likely to result from the disclosure of that trade secret.”⁴⁴ Serious economic damage still under Recital 31 “implies serious and irreparable economic loss.”⁴⁵

In its treatment of trade secrets, Article 40 DSA is less clear. On the one hand, Recital 97 DSA states in rather generic terms that “consideration of the commercial interests of providers should not lead to a refusal to provide access to data necessary for the specific research objective pursuant to a request under this Regulation,”⁴⁶ and expressly refers to the possibility of implementing data vaults to grant access to data without prejudice to the Trade Secret Directive.⁴⁷ Article 40(2) DSA, on the other hand, acknowledges that DSCs and the European Commission “shall take due account of the rights and interests of the providers of very large online platforms or of very large online search engines and the recipients of the service concerned, including the protection of personal data, the protection of confidential information, in particular trade secrets, and maintaining the security of their service.”⁴⁸ Article 40(5) DSA echoes Article 40(2) DSA in the context of an objection to a reasoned request issued by the DSC of establishment. When objecting, data providers can point to significant vulnerabilities in the protection of confidential information, including trade secrets.

The Delegated Act does not go much further than the DSA. As a rule of thumb, Recital 21 clarifies that “commercial considerations should not be considered as a ground to automatically refuse access to requested data but rather as a ground to modify the means of providing access to the data, which may result in imposing additional data security and confidentiality requirements.”⁴⁹ However, Article

⁴³ Recital 31 DA.

⁴⁴ Recital 31 DA.

⁴⁵ Recital 31 DA. This recital adds that “[s]uch a substantiation should be based on objective elements, demonstrating the concrete risk of serious economic damage expected to result from a specific data disclosure and the reasons why the measures taken to safeguard the requested data are not considered to be sufficient. ”

⁴⁶ Recital 97 DSA.

⁴⁷ Directive (EU) 2016/943 of the European Parliament and of the Council of 8 June 2016 on the protection of undisclosed know-how and business information (trade secrets) against their unlawful acquisition, use and disclosure, OJ L 157, 15.6.2016, S. 1-18.

⁴⁸ Article 40(2) DSA.

⁴⁹ Trade secrets should fall within the broader bucket of commercial considerations as trade secrets are not a subject matter that is protected through exclusive rights.

9(3)-(4) does not articulate the different steps of the process for taking into account trade secrets and protective measures.

As a result, while the DA seems to focus upon exceptional economic consequences, the DSA seems to focus upon risk exposure and could therefore appear less restrictive than the DA for similar categories of data. Yet, it is hard to explain why big players such as VLOPSEs should find it easier to object to the sharing of trade secrets than other types of players such as data holders under the DA when readily available data is at stake.⁵⁰ In addition, for gatekeepers subject to an obligation to give access to both provided and generated data under the DMA⁵¹ there is no governed process to object to the sharing of trade secrets to third parties, which do not even have to be vetted in any specific way.⁵² It would thus be strange for the DSA to be more demanding than the DMA for overlapping categories of actors when provided and generated data is at stake. These two reasons should therefore lead DSCs to stress the importance of Article 12(3) of the Delegated Act and interpret the reference to “the likelihood and severity of harm resulting from these alleged significant vulnerabilities”⁵³ in a rather restrictive way. Of note, and this is an important consideration for interpreting the DA, there is even an argument that the codification produced through the DA fails to reflect a consensus that raw data should never be protected as trade secrets.⁵⁴

Once again, this paper shows that several interpretations are possible for Article 40(5) DSA and highlights its inherent indeterminacy. This paper suggests that trade secret considerations used to refuse access should be systematically rejected each time covered data amount to provided and observed data. What is more, data providers should be required to precisely identify trade secrets before objecting to reasoned data access requests.

3. Data Accessibility for Researchers

Interpreting the law is never purely mechanical or formalistic; it is always normatively guided. Interpreters weigh which principles or policies justify a particular application, which is a choice about the interest to be defended. This is true even if there are several recognised methods for interpreting legal texts. It is thus a mistake to think that legal interpretation is fundamentally about linguistics, although it certainly helps if a coherent set of concepts is used to express legal rules.

Leveraging prospective legal methods to substantiate data accessibility affordances available to researchers implies a two-step process: identifying researchers’ data needs and interpreting the Law to meet these needs without falling into the trap of blind adherence to the data imaginary.⁵⁵ In essence, this means ensuring that researchers are placed in a position to develop robust methods for internet data-collection and analysis.

⁵⁰ Articles 4, 5 DA. Readily available data is defined under Article 2(17) DA as “product data and related service data that a data holder lawfully obtains or can lawfully obtain from the connected product or related service, without disproportionate effort going beyond a simple operation.”

⁵¹ Article 6(9)-(10) DMA.

⁵² This has been described by some commentators as an unintentional omission.

⁵³ Article 12(3) Delegated Act.

⁵⁴ See in fine Josef Drexler, ‘Designing Competitive Markets for Industrial Data - Between Propertisation and Access’ (2017) 8 Journal of Intellectual Property, Information Technology and Electronic Commerce Law 257.

⁵⁵ David Beer, *The Data Gaze - Capitalism, Power and Perception* (Sage Publications Ltd 2018).

3.1 Identifying Researchers' Data Needs

To interpret data-access rules in light of researchers' interests, it is essential to articulate what those interests concretely entail. Researchers require not only access to datasets but also the ability to engage in a variety of practices that make meaningful inquiry possible.

First, statistically representative data collection is fundamental.⁵⁶ Researchers need to gather data systematically and at scale, often by means of automated tools, to perform inquiries on representative samples which can be meaningfully extrapolated to broader populations. These means allowing researchers to collect large, dynamic, and otherwise inaccessible bodies of information, which are indispensable for producing analyses that are generalisable, situated and meaningful. Without systematic, scalable collection, research risks bias, narrow selection, or missing crucial subsets of the population.

Second, and more generally, data quality assessment is a central concern for meaningful inquiry. Researchers must be able to evaluate the coverage and completeness of the data they use, but also its consistency and accuracy.⁵⁷ This entails access not only to raw data, but also to accompanying metadata, documentation of collection processes, and information about platform-side constraints or omissions (such as missing fields or omissions of certain populations, or disappeared data such as e.g. deleted posts). In studies leveraging real world data, a deep understanding of data quality issues, such as missingness is essential.⁵⁸ Importantly, data quality is often understood as “fitness for use,” which makes approaching data from the perspective of the data user essential.⁵⁹

Third, exploratory access is crucial. Research typically begins with open-ended exploration of datasets to discover patterns, anomalies, or researchable questions.⁶⁰ This requires flexible access that goes beyond predefined queries or fixed data extracts. In other words, researchers should be able to ask new questions of the data, filter, subset, aggregate, cross-tabulate, and maybe recombine dimensions. Access mechanisms that provide only rigid, or highly limited subsets via narrowly defined pre-queries, constrain the capability to explore and thereby limit the scope of inquiry and the emergence of serendipitous discovery.

⁵⁶ Gevaert et al., for example, discuss challenges related to sampling platform workers. J Gevaert and others, ‘Strategies for Surveying Platform Workers: Lessons from a Belgian Case Study’ (2025) 59 *Quality & Quantity* 211. See on representativeness of Twitter data Andreas Petutschnig and others, ‘Evaluating the Representativeness of Socio-Demographic Variables over Time for Geo-Social Media Data’ (2021) 10 *ISPRS International Journal of Geo-Information*.

⁵⁷ See for examples of data quality criteria mentioned in the literature, Maxi Kindling and Dorothea Strecker, ‘Data Quality Assurance at Research Data Repositories’ [2022] *Data Science Journal*. For digital social research specifically, see Jessica Daikeler and others, ‘Assessing Data Quality in the Age of Digital Social Research: A Systematic Review’ (2025) 43 *Social Science Computer Review* 943.

⁵⁸ Hythem Sidky and others, ‘Data Quality Considerations for Evaluating COVID-19 Treatments Using Real World Data: Learnings from the National COVID Cohort Collaborative (N3C)’ (2023) 23 *BMC Medical Research Methodology* 46.

⁵⁹ Richard Y Wang and Diane M Strong, ‘Beyond Accuracy: What Data Quality Means to Data Consumers’ (1996) 12 *Journal of Management Information Systems* 5.

⁶⁰ In its seminal work on exploratory data analysis, Tukey states that “[e]xploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone – as the first step.” John W Tukey, *Exploratory Data Analysis* (Addison-Wesley Publishing Company 1977) 3.

Fourth, iterative engagement is essential for robust, data-driven inquiry.⁶¹ Data-driven inquiry is rarely linear: researchers refine their questions and methods as they proceed, which means they need to query and re-query data, test hypotheses, and adapt their approaches as new insights emerge. This requires not only initial access but sustained and adjustable access, including the capacity to revisit data, adjust inclusion criteria, refresh collection, correct errors, and maybe follow up new leads. Without such iterative engagement, research may be brittle, unable to respond to emergent issues, or unable to validate and replicate results across iterations.

Fifth, in addition to robust theoretical frameworks and flexible analytical tools, researchers require conditions that enable them to document, verify, and reproduce their analyses.⁶² This need underscores the importance of stable access protocols and sufficiently granular data that can support replication and validation across studies. Ensuring that datasets are accessible and reproducible is essential for maintaining the integrity of digital research, especially in contexts where platforms continuously evolve and modify their interfaces or data structures.

Considering the growing demand for high-quality analysis, it is crucial to recognise that researchers require multiple data access contexts, each offering distinct advantages and limitations. Three of these data access contexts are illustrated in Table 1. A fourth form, sometimes presented as a solution to overcome challenges raised by web scraping is data partnership, i.e., ad hoc privileged access for specific research teams who had successfully negotiated with platforms.⁶³ Given its low level of repeatability, this approach is not included in the table.

One typical access context is API-based access, which provides structured, and well-documented channels for retrieving data, often formatted in machine-readable structures such as JSON or XML. APIs are efficient, particularly for repeated queries, and they facilitate systematic data collection that aligns with best practices in computational research. APIs vary in function: some enable near real-time data access, while others require data to be retrieved through periodic downloads. However, APIs are not without constraints. They often impose rate limits, exclude certain metadata, or restrict access to historical content. For example, a recent study of TikTok's Research API under the DSA revealed substantial data gaps, making it difficult to rely solely on API data for comprehensive auditing or transparency research.⁶⁴ For instance, some accounts were excluded from the research API,⁶⁵ authorised Tik Tok videos were not accessible although they are public, the same was true for

⁶¹ The importance of iteration has led some data science researchers to propose "Progressive Data Science as a novel knowledge discovery paradigm where progressiveness is inherent in every step of the data science process, and ensuring success in such a novel paradigm requires the concerted effort from various research communities." Cagatay Turkey and others, 'Progressive Data Science: Potential and Challenges' <<https://arxiv.org/abs/1812.08032>>.

⁶² On the importance of peer review of data publication methodologies and peer review of data see e.g., Catherine Jones and others, 'Citation and Peer Review of Data: Moving Towards Formal Data Publication' (2011) 6 *International Journal of Digital Curation*.

⁶³ E.g., the partnership between Facebook and the controversial Social Science One initiative, or the partnership announced in early 2024 between Meta and the Centre for Open Science. For a critique of data partnerships see Bruns (n 15).

⁶⁴ "In sum, our experiments show that TikTok's research API can not serve as a source for research that requires up-to-date, complete, and consistent information." Carlos Entrena-Serrano and others, *TikTok's Research API: Problems Without Explanations* (AI Forensics 2025) 17 <<https://aiforensics.org/work/tk-api>>. See also Daniela Alvarado Rincón, Ognjan Denkovski and Salvatore Romano, 'Unpacking TikTok's Data Access Illusion' [2025] *Tech Policy.PRESS* <<https://www.techpolicy.press/unpacking-tiktoks-data-access-illusion/>>.

⁶⁵ Entrena-Serrano and others (n 64) 10.

videos that are posted as ads, the metadata was missing in approximately one out of every eight videos previously obtained through data donations.⁶⁶

API-based access is one method platforms use to control how data is accessed and used, allowing them to set strict permissions, audit usage, and enforce a range of governance rules. An alternative approach is to provide access through secure processing environments, such as data vaults or clean rooms, which operate as controlled query environments setting restrictions upon data extraction modalities,⁶⁷ analytics tools, query types,⁶⁸ query depths,⁶⁹ join rules⁷⁰...

⁶⁶ Entrena-Serrano and others (n 64).

⁶⁷ Note that it is not possible to prevent all modalities of extractions, as per definition if querying is allowed results will be displayed to the user.

⁶⁸ Query type defines what kind of query is allowed.

⁶⁹ Query depth defines how granular or detailed a query can be.

⁷⁰ Join rules are controls on how datasets can be linked

Table 1: Three complementary data access contexts

Data-access context	Platform control level	Key characteristics	Main advantages	Main limitations
Direct holder-to-user data flows (e.g., through APIs or secure processing environments)	High	Structured and tidy data, documented, machine-readable access to selected data via platform-controlled endpoints; subject to permissions, quotas, and governance frameworks	Assuming application process is straightforward and that the API isn't discontinued: Efficient, stable; large-scale and computationally parsimonious data collection; pre-prepared tidy and structured fields, consistent across queries	Platform defines scope, rate limits, fields returned; may exclude historical data and critical metadata; subject to risk of sudden policy changes (e.g. introduction of access fees) and API discontinuation, limiting terms of use (e.g. data can't be shared or made open-access, which limits reproducibility)
Web Scraping / Crawling	Low	Automatic extraction of publicly browsable content, including dynamic elements from platform interfaces, independent of API constraints	Captures content "as seen" by users; can expose discrepancies between API output and public interface; supports exploratory and iterative research; enables historical capture	Technical difficulty rapidly escalates with technical barriers (e.g. CAPTCHAs); can be resource-intensive; risk of incomplete access if content is hidden from interface
Data Donations	Medium to High	Users voluntarily share their own data (feeds, ads, engagement traces); platform may mediate. Often done through asking users to install data collection plug in on their browser, or – more rarely and more labour intensive, by supporting them in doing data access requests and sharing the data with the researchers.	Reveals user-side perspective; can uncover undisclosed data collection; enhances transparency of personal experiences. Integrates direct and informed user consent.	Dependent on volunteer participation; securing user participation and developing the setup and skills required is often highly labour intensive; selection bias to user self-selection for participation; requires privacy-preserving infrastructures; limited to what users can access or export

Secure processing environments are usually recommended for particularly sensitive data but have been used to give access to publicly accessible data. Meta's Content Library built as a web-based tool is often described as an example of a clean room as it prevents researchers from downloading raw data, although it only offers access to publicly accessible content.⁷¹ Tik Tok's Virtual Compute Environment (VCE) is also one example of a clean room built for publicly accessible data, where data cannot be downloaded and all queries must be executed through the VCE interface.⁷² Both these environments have been heavily criticised by researchers.⁷³ One reason in particular warrants emphasis: the VCE provides access only to aggregated results, not to individual-level data, and is therefore far more restrictive than web-scraping.⁷⁴

Restricted APIs and secure processing environments thus represent two distinct models for controlling access to data. They are two instances of a broader 'direct holder-to-user data flows' approach. Restricted APIs allow data to leave the platform, but only in predefined and limited forms, with controls such as field restrictions, rate limits, and logging ensuring that only authorised data is retrieved. By contrast, secure processing environments keep data within a controlled query environment and permit only approved computations, e.g., returning aggregate outputs rather than raw data. Secure processing environments offer a higher-assurance model for preventing unauthorised use or re-identification.

Another vital approach is web scraping, which enables researchers to extract publicly browsable data that APIs may or may not omit. This method allows for comparisons between what platforms make available through official APIs and what is accessible through their public-facing interfaces. Scraping also supports exploratory and iterative research by capturing all available public information on a page, or interface-specific content that might otherwise remain inaccessible and may make it possible to create historical or time-series datasets through repeated data collection.

Finally, data donations represent a complementary modality that addresses a critical gap left by API and scraping-based methods.⁷⁵ Data donations, where users voluntarily share their personal data, such as their feed content or advertisement exposure, offer valuable insights into the user-side perspective, or the results of personal data access requests put to the platforms. These enable researchers to study with the granularity of the individual level how algorithms curate individual feeds, how targeted advertising operates, how online interactions are shaped or how work and fares are allocated and have changed over time.⁷⁶ They may as well help reveal undisclosed data collection practices.⁷⁷

⁷¹ See Philip Mai, 'A First Look at Meta's New Content Library and Content Library API' (*SocialMediaLab.ca*, 25 March 2024) <<https://socialmedialab.ca/2024/03/25/a-first-look-at-metas-new-content-library-and-content-library-api/?utm=>>.

⁷² See Alvarado Rincón, Denkovski and Romano (n 64).

⁷³ *ibid*; Entrena-Serrano and others (n 64).

⁷⁴ "[T]he restriction to aggregated results is fundamentally incompatible with how meaningful research on social media platforms works." Alvarado Rincón, Denkovski and Romano (n 64).

⁷⁵ On the potential of data donations see Valerie Hase and others, 'Fulfilling Data Access Obligations: How Could (and Should) Platforms Facilitate Data Donation Studies?' (2024) 13 Internet Policy Review.

⁷⁶ Reuben Binns and others, 'Not Even Nice Work If You Can Get It; A Longitudinal Study of Uber's Algorithmic Pay and Pricing' <<https://arxiv.org/abs/2506.15278>>.

⁷⁷ 'Press Release: Reversing.Works Investigation Exposes Glovo's Data Privacy Violations, Marking a Milestone for Worker Rights and Technology Accountability' (*Reversing.Works*, 22 November 2024) <<https://reversing.works/posts/2024/11/press-release-reversing-works-investigation-exposes-glovo-data-privacy-violations-marking-a-milestone-for-worker-rights-and-technology-accountability/>>.

Importantly, contrary to what their name might suggest, data donations are not a blank cheque. Data protection law continues to apply to research processing activities for as long as the data remains personal.

Together, these three access contexts, direct holder-to-user data flows (ranging from APIs to secure processing environments), web scraping, and data donations, form a comprehensive ecosystem that supports rigorous, transparent, and reproducible digital research. Taken together, these methods enable triangulation: direct holder-to-user data flows provide structured ‘official’ data; data scraping capture the public-facing and dynamic dimensions of platforms; and data donations offer insight into different types of personalised experience. By combining these approaches, researchers can cross-validate findings, identify gaps or inconsistencies in platform-provided data, and ultimately generate analyses that are more comprehensive, robust, and representative.

3.2 Interpreting the Law to Meet Researchers’ Data Needs

Speculative interpretation, as a prospective legal method, considers possible (as opposed to realistic) readings of a rule in light of the interest(s) to be defended. The very concept of ‘realistic’ readings is itself problematic, as it obscures the normative stance it presupposes, which may consist in assigning greater weight to commercial interests over public interests.

As explained above, diversity of access context is fundamental to support exploration, coverage and reliability of platform research.⁷⁸ This consideration should inform the interpretation of legal rules governing researchers’ access to platform data and support three further interpretive conclusions related to secure processing environments, web scraping and data donations.

3.2.1 Secure processing environments

The impact of commercial interests is clear in the European Commission’s initial vision for secure processing environments under Article 40(4) of the DSA. Unlike restricted APIs, these environments allow platforms to maintain detailed oversight of researchers’ investigations, including real-time monitoring of all queries executed on the data.⁷⁹ A systematic and proactive use of prospective legal methods could help clarify what good-practice access to data in this context looks like and the implications of operating secure processing environments.

The draft Delegated Act included a bespoke definition of secure processing environments that differed slightly from the definition in the Data Governance Act (DGA),⁸⁰ assuming that such environments would be operated by data holders or by third parties acting on their behalf.⁸¹ As noted in an earlier contribution that helped inform a response to the consultation on the draft Delegated Act,⁸² this approach is not optimal for safeguarding researchers’ independence, particularly when trusted research environments already host and manage similar environments. This prompted the European

⁷⁸ See also Keller (n 16); Entrena-Serrano and others (n 64).

⁷⁹ See e.g., Mozilla (n 28); Keller (n 16).

⁸⁰ Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) OJ L 152, 3.6.2022, pp. 1–44.

⁸¹ See Article 2(19) of the draft Delegated Act (“the secure processing environment provided by or on behalf of the data provider as defined in Article 2, point (20), of Regulation (EU) 2022/868.”)

⁸² Sophie Stalla-Bourdillon, ‘Three Key Points on the Delegated Act: How to Preserve Researcher Autonomy under Article 40 DSA?’ (*The DSA Observatory*, 29 November 2024) <<https://dsa-observatory.eu/2024/11/29/digging-into-the-delegated-act-how-to-preserve-researcher-autonomy-under-article-40-dsa/>>.

Commission to revise its definition, reverting to the one in the DGA and allowing for the possibility of a secure processing environment operated by a research organisation.⁸³

However, what cannot be resolved by revising the text of the Delegated Act is the practical reality: VLOPSEs operate secure processing environments in a highly restrictive manner, often imposing their own methods and systems even for non-sensitive data, before researchers have the opportunity to demonstrate that they can provide protective environments themselves. It is therefore on the regulators to adopt a functional definition of secure processing environments and look for alternatives beyond platforms' own settings.

What has not changed significantly in the final version of the Delegated Act is the substance of Article 15 and its provision relating to data management requirements such as archiving, storage, refresh and deletion requirements, and limitations upon the use of standard analytical tools, that may hinder the research performance unless these requirements or limitations are explicitly listed in the reasoned request.

What DSCs should understand is that secure processing environments provide multiple mechanisms to control and constrain researchers' work, and these mechanisms extend beyond the formal limits imposed on analytical tools. For example, restrictions may include query depth, join rules, minimum aggregation, limits on query frequency, and enforced output suppression⁸⁴ or noise injection.⁸⁵ Even if a researcher has access to powerful analytical tools, these embedded controls set strict limits on data use and should therefore be carefully scrutinized even if they are not expressly mentioned in Article 15 of the Delegated Act. An interpretation aligned with researchers' interests should therefore adopt a broad understanding of "analytical tools" and apply analogous reasoning to comparable constraints.

3.2.2 Web scraping

Web scraping has attracted significant attention in the wake of large language model (LLM) training. Intense debate has emerged over the lawfulness of such practices, particularly in light of two key areas of EU law: intellectual property, especially copyright and the *sui generis* database right, and data protection. Both legal domains are subject to extensive regulation at the EU level, where efforts toward harmonisation have been especially vigorous.

For the purposes of this discussion, it is crucial to distinguish between web scraping conducted for artificial intelligence (AI) training and that undertaken for platform research. When this distinction is made, the analysis becomes more straightforward and tends to weigh in favour of researchers. Platform research may be generally understood as the systematic collection, analysis, and

⁸³ Article 2(20) DGA ("physical or virtual environment and organisational means to ensure compliance with Union law, such as Regulation (EU) 2016/679, in particular with regard to data subjects' rights, intellectual property rights, and commercial and statistical confidentiality, integrity and accessibility, as well as with applicable national law, and to allow the entity providing the secure processing environment to determine and supervise all data processing actions, including the display, storage, download and export of data and the calculation of derivative data through computational algorithms").

⁸⁴ Output suppression refers to mechanisms that prevent or withhold the release of query results, typically when outputs fall below predefined thresholds or risk disclosing sensitive information, thereby limiting what researchers can ultimately observe or extract from the data.

⁸⁵ Noise injection refers to techniques that deliberately introduce random perturbations into query results in order to reduce the risk of re-identification or inference about individual data subjects, at the cost of reduced accuracy or precision.

interpretation of data related to the operation, content, users, and societal effects of digital platforms, typically for the purpose of advancing scientific understanding or informing public policy.

As noted above, legal uncertainty surrounds web-crawling and scraping practices, prompting various commentators to call for greater clarity on the issue.⁸⁶ This paper suggests that a more systematic use of prospective legal methods could help shift perceptions of the current landscape.⁸⁷

Data Protection

From a data protection perspective, it is worth noting the analysis produced by the European Digital Media Observatory (EDMO), which is specifically devoted to the challenge of platform-to-researcher data access.⁸⁸ The findings of the EDMO working group make it clear that data protection law is not, in itself, an insurmountable barrier to researchers' access to data, although compliance with data protection law is often mentioned by researchers asked to identify legal issues arising when trying to access data held by third parties,⁸⁹ and platforms are using data protection law as a reason for restricting access.⁹⁰

The EDMO report notes that "[t]he lower the expectation of privacy, the lower the risk level. In general, data subjects are less entitled to expect privacy in relation to data they have put in the public domain, such as publicly available social media posts."⁹¹ But here a clear distinction should be made between provided data and derived data or to use the terms of the report, raw data and inferences. As a result, "[t]he assessment must distinguish between raw data (e.g., copies of posts, profile information, engagement data) that may be in the public domain and any inferences drawn from that data through research, which by definition will not be public."⁹² The risk assessment framework endorsed by the EDMO working group lies upon two types of considerations: the expectations of data subjects and the potential impact on data subjects' rights and freedoms of the processing.⁹³ Data minimisation⁹⁴ and pseudonymisation⁹⁵ appear to be two key safeguards for mitigating risks. Importantly, this does not mean that query type restrictions such as 'researchers are only allowed to ask for aggregates' are justified in all cases. A spectrum of data access modalities is warranted.

⁸⁶ See e.g., Mozilla (n 28); Keller (n 16).

⁸⁷ It is particularly telling that Mozilla reports in 2024 on both explicit and implicit permissions for data scraping. Mozilla (n 28) 4.

⁸⁸ EDMO and Institute for Data, Democracy and Politics, 'Report of the European Digital Media Observatory's Working Group on Platform-to-Researcher Data Access' (31 May 2022) <<https://edmodprod.wpengine.com/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf>>.

⁸⁹ Ausloos and others (n 9) 3.

⁹⁰ See Ram Iyer, 'EC Finds Meta and TikTok Breached Transparency Rules under DSA' [2025] *Tech Crunch* <<https://techcrunch.com/2025/10/24/ec-finds-meta-and-tiktok-breached-transparency-rules-under-dsa/>>.

⁹¹ EDMO and Institute for Data, Democracy and Politics (n 88) 77.

⁹² *ibid* 78.

⁹³ *ibid* 76.

⁹⁴ *ibid* 83. Data minimisation essentially means tailoring the amount of data to be processed to the specified and limited purposes for which the data is processed. For a recent interpretation of the principle of data minimisation in the context of data processing for the purposes of concluding transport contracts see *Mousse* [2025] CJEU C-394/23, ECLI:EU:C:2025:2.

⁹⁵ EDMO and Institute for Data, Democracy and Politics (n 88) 84. Pseudonymisation essentially means reducing re-identification risks through masking identifying data points, particularly direct identifiers to start with. For a more technical definition see EDPB, 'Guidelines 01/2025 on Pseudonymisation' (2025) <https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-012025-pseudonymisation_en> accessed 6 February 2025.

Although Article 40 DSA is excluded from the scope of the recent guidelines on the interplay between the DSA and the GDPR released by the European Data Protection Board (EDPB),⁹⁶ there is no reason to believe that the EDMO report does not reflect good practice. Although divergences of approach towards the secondary use of personal data for scientific research purposes persist between member states,⁹⁷ guidance produced by data protection authorities do not moot researcher activities, although some seem to have a preference for APIs,⁹⁸ and others seem to favour non-commercial and targeted scraping.⁹⁹

The rather cautious stance taken by data protection supervisory authorities could be explained in large part by the intensive web crawling that is happening for AI training, which again should be distinguished from platform research, although such a reality is not always well perceived by computational researchers.

Looking at more recent guidance produced by the French Supervisory Authority (CNIL), one already sees some signs of relaxation. Very interestingly, the CNIL includes in its list of examples illustrating legitimate interests the use of data published on blogs, forums, social networks, and other similar sources for the production of statistics incorporating pseudonymisation and anonymisation mechanisms, to measure opinions on a topic, person, or specific product (trend detection, correlations).¹⁰⁰ The recent CNIL recommendations on AI are also worth mentioning. As regards special categories of personal data, CNIL writes that in the context of web scraping for AI training, if an organisation incidentally and unintentionally processes sensitive data despite taking appropriate precautions and without intending to collect it, this residual processing is not considered unlawful.¹⁰¹ An a priori argument could be advanced in support of platform research.

⁹⁶ EDPB, ‘Guidelines 3/2025 on the Interplay between the DSA and the GDPR’ (2025) Version 1.1 para 6 <https://www.edpb.europa.eu/our-work-tools/documents/public-consultations/2025/guidelines-32025-interplay-between-dsa-and-gdpr_en>.

⁹⁷ See the findings of EDPB, ‘Study on the Secondary Use of Personal Data in the Context of Scientific Research’ (2025) <https://www.edpb.europa.eu/our-work-tools/our-documents/other/study-secondary-use-personal-data-context-scientific-research_en?utm_source=chatgpt.com>.

⁹⁸ See e.g., ‘Concluding Joint Statement on Data Scraping and the Protection of Privacy’ (2024) <https://www.priv.gc.ca/en/opc-news/speeches-and-statements/2024/js-dc_20241028/>. After issuing the Initial Statement, the co-signatories shared a copy with industry: Alphabet Inc. (YouTube), ByteDance Ltd. (TikTok), Meta Platforms, Inc. (Instagram, Facebook and Threads), Microsoft Corporation (LinkedIn), Sina Corp (Weibo), and X Corp. (X, previously Twitter) to get comment on how they comply with the expectations outlined in the document. It is clear that one major concern underlying this statement is unlawful scraping.

⁹⁹ This seemed to be the initial position of the Dutch Supervisory Authority, although the guidance has now disappeared from its website. See Barnea Jaffa Lande, ‘The Netherlands: New Data Scraping Guidelines’ (*Lexicology*, 13 May 2024) <<https://www.lexicology.com/library/detail.aspx?g=9461bde9-6145-409a-8d18-d15cbe4cfca1>>. Compare with Garante, ‘Web Scraping Ed Intelligenza Artificiale Generativa - Nota Informativa’ (2024) <<https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/10020316>>.

¹⁰⁰ CNIL, ‘Recommandations - Réutilisation de Données Publiées Sur Internet’ (2024) 11 <https://www.cnil.fr/sites/cnil/files/2024-06/recommandations_reutilisateurs_donnees_publiees_sur_internet.pdf>.

¹⁰¹ CNIL, ‘La Base Légale de l’intérêt Légitime : Fiche Focus Sur Les Mesures à Prendre En Cas de Collecte Des Données Par Moissonnage (Web Scraping)’ (CNIL, 19 June 2025) <<https://www.cnil.fr/fr/focus-interet-legitime-collecte-par-moissonnage>>.

Copyright and database sui generis right

In reality, the Directive on Copyright in the Digital Single Market (CDSM)¹⁰² appears more problematic than the General Data Protection Regulation (GDPR) for researchers, at least after having reviewed what appears to be emerging as the prevailing view.¹⁰³

Article 3 CDSM setting forth a mandatory exception to both copyright and database sui generis right for reproductions and extractions¹⁰⁴ is usually presented as supporting the work of researchers “when conducting text- and datamining after *lawfully* obtaining access to copyrighted works”¹⁰⁵ or databases protected by sui generis rights. The prevailing interpretation of Article 3(1) CDSM therefore seems to be that the text and data mining (TDM) exception is only triggered in cases in which researchers have lawful access to protected subject matters. Crucially, lawful access is understood to exclude situations in which contractual and unilateral terms restrict TDM for scientific research purposes.¹⁰⁶ In other words, it is deemed to exclude situations in which terms of service/terms of use restrict automated reproduction of the data available on the platform for secondary use, including for scientific research purposes, although automated reproduction (e.g., triggered by browsers) is obviously necessary for viewing the content in the first place, i.e., for primary use.

Such a ‘realistic’ interpretation of the law, which has been criticised by several commentators already particularly in the light of AI development,¹⁰⁷ is problematic for a number of reasons, not least because it risks unnecessarily chilling public-interest scientific research. In reality, as with most sets of legal rules intended to govern technology-mediated practices, legal uncertainty permeates both the domain and effects of Article 3 CDSM and its national transpositions.¹⁰⁸ In other words, the concept of *lawful access* is tainted with ambiguities, and no well-established test has emerged in case law yet. The key question for the purpose of this paper is to determine when a computational researcher

¹⁰² Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC OJ L 130, 17.5.2019, pp. 92–125.

¹⁰³ See for a review of the literature on the topic Giorgos Vrakas, ‘A Literature Review of “Lawful” Text and Data Mining’ (2024) 4 Open Research Europe.

¹⁰⁴ This precision is extremely important in particular in the case of AI development for scientific purpose as explained in Eleonora Rosati, ‘Copyright Exceptions and Fair Use Defences for AI Training Done for “Research” and “Learning,” or the Inescapable Licensing Horizon’ [2025] European Journal of Risk Regulation 1. The right to communication to the public is not impacted by this exception.

¹⁰⁵ Ausloos and others (n 9) 54; Martin RF Senftleben, ‘Study on EU Copyright and Related Rights and Access to and Reuse of Data’ (2022) Independent Expert Report for the European Commission 43 <https://www.ivir.nl/publicaties/download/KI0822205ENN.en_.pdf>. See also Kacper Szkalej, ‘The Paradox of Lawful Text and Data Mining? Some Experiences from the Research Sector and Where We (Should) Go from Here’ (2025) 74 GRUR International 307; Jonathan Griffiths, Tatiana Synodinou and Raquel Xalabarder, ‘Comment of the European Copyright Society Addressing Selected Aspects of the Implementation of Articles 3 to 7 of Directive (EU) 2019/790 on Copyright in the Digital Single Market’ (2023) 72 GRUR International 22.

¹⁰⁶ Senftleben (n 105) 43. For a definition of scientific research under the CDSM see Rosati (n 104).

¹⁰⁷ See e.g., Thomas Margoni and Martin Kretschmer, ‘A Deeper Look into the EU Text and Data Mining Exceptions: Harmonisation, Data Ownership, and the Future of Technology’ (2022) 71 GRUR International 685; Thomas Margoni, ‘Saving Research: Lawful Access to Unlawful Sources under Art. 3 CDSM Directive?’ (*Kluwer Copyright Blog*, 22 December 2023) <<https://copyrightblog.kluweriplaw.com/2023/12/22/saving-research-lawful-access-to-unlawful-sources-under-art-3-cdsm-directive/>>. See also Tatiana Eleni Synodinou, ‘Lawfulness for Users in European Copyright Law: Acquis and Perspectives’ (2019) 10 JIPITEC <<https://www.jipitec.eu/jipitec/article/view/243>>; Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, ‘Text and Data Mining: Articles 3 and 4 of the Directive 2019/790/EU’ (2019); Rossana Ducato and Alain Strowel, ‘Ensuring Text and Data Mining’ (2021) 43 remaining issues with the EU copyright exceptions and possible ways out 322; Szkalej (n 105).

¹⁰⁸ In France, for example, the TDM exception for scientific research is implemented in Articles L. 122-5-3 and following of the Intellectual Property Code.

leveraging data mining techniques and technologies while scraping publicly accessible platform data can be considered to have lawful access to protected subject matters.

The short and honest answer to this question is that we do not know for sure and that although several arguments would justify answering the question by stressing the need for rightholders' prior authorisation, other reasonable arguments in favour of researchers scraping without any rightholders' prior authorisation can also be put forward.

Following CJEU case law decided prior to the adoption of the CDSM, a problem indeed arises when the protected subject matter has been made available through a website or app without the authorisation of the rightholder.¹⁰⁹ In this case, the source appearing on the website or app and upon which the website or app provider does not hold any right is said to be unlawful, although strictly speaking it is the reproduction and the communication to the public without prior authorisation that is unlawful. As a matter of principle, no one has the right to use and reuse a protected subject matter without obtaining the prior authorisation of the rightholder. This is a classical lawful-source problem.¹¹⁰

However, there is an argument that limiting the secondary use of a protected subject matter that is made publicly accessible when no fair compensation is foreseen, e.g., through terms of use that prohibit TDM for scientific research purposes, should not be too readily analogised to a lawful-source problem. The lawful-source problem should only arise when there is no authorisation for the primary use of the subject matter and the secondary use raises fair compensation concerns as there is a risk that a secondary market for that use arises.¹¹¹ A distinction should thus be drawn between merely augmenting a primary use, through TDM for scientific research purposes for example, for which there is no new public involved,¹¹² and decoupling a primary use into two distinct uses, for instance through the production of private copies, for which a new public is latent.

Taking a step back, one major concern behind ensuring a broad scope to rightholder economic exclusive rights lies in ensuring that they are not deprived of a fair compensation.¹¹³ Yet, it is difficult to argue that rightholders are harmed when researchers access a protected subject matter that is made publicly accessible by adhering to the access modalities set by the rightholder for primary use except for TDM for scientific research purposes, and then later on disseminate the research output without reproducing and communicating to the public substantial part of the protected subject matter.

What is more, and this line of argument is only relevant for the database *sui generis* right, recent CJEU Case Law stresses the importance of pursuing financial gains, or better to adversely affect the rightholder's investment by "depriving that person of income intended to enable him or her to redeem

¹⁰⁹ See the line of cases in the wake of *ACI Adam BV and Others v Stichting de ThuisKopie and Stichting Onderhandeligen ThuisKopie vergoeding* [2014] CJEU Case C-435/12, ECLI:EU:C:2014:254.

¹¹⁰ Interestingly, in this case the only copyright a platform might claim is on the combination of content, assuming the combination is original.

¹¹¹ Unhelpfully, the InfoSoc does not expressly explain "against what, in particular, the need to ensure a high level of protection (as follows from recital 9) should be balanced so that a 'fair balance' can actually be safeguarded" when exceptions are carved out. Szkalaj (n 105).

¹¹² The situation is different in case of secondary use for AI model training and release.

¹¹³ In civil-law systems such as France, copyright holders also hold moral rights, and in particular a moral right to divulgation under Article L. 121-2 of the Intellectual Property Code, which states that the author alone has the right to disclose her work and to determine the manner and conditions for disclosure.

the cost of that investment”¹¹⁴ to characterise infringement in the first place (before even checking whether an exception applies). There is thus space to argue that scientific research that is not directly geared towards financial gains does not amount to an infringement of a sui generis database right.

All in all, the concept of lawful source was developed to address a specific set of issues in copyright law: how to delineate the scope of fair compensation owed to rightholders under the private copy exception,¹¹⁵ and how to define the scope of the right of communication to the public.¹¹⁶ These issues are different from the question at hand, namely, what effect the TDM exception for scientific research has on the right of reproduction, for which no fair compensation is foreseen.

Going further, textual arguments also support distinguishing the lawful-source problem from the lawful-access problem under Article 3(1) CDSM. Although the drafters of the CDSM have been well aware of the concept of lawful source,¹¹⁷ which has been part of the EU acquis for several years, they have chosen to adopt a different notion, i.e., lawful access, in the DSM Directive.¹¹⁸

As a result, there is a reasonable argument that lawful access under Article 3 CDSM means complying with the modalities for access set by the rightholder, as long as they do not prohibit TDM for scientific research purposes. Therefore, when protected content is made publicly accessible for primary use, either freely, or through an open access policy or a subscription and researchers adheres to the access modalities, they should be able to perform TDM for scientific research purposes. Of note, Recital 14 CDSM makes it clear that persons covered by subscriptions “should be deemed to have lawful access.”¹¹⁹

Importantly, approaching legal access within the meaning of Article 3 DSM in this way preserves the *raison d’être* of both Article 3 and Article 7 CDSM.¹²⁰ Article 3 CDSM does not foresee that rightholders have the possibility to reserve the use of their works or other subject matters for the purposes of TDM for scientific research purposes, meaning they cannot opt out of scientific research TDM through unilateral terms or said otherwise restrictions upon scientific research TDM set through terms of use are ineffective. What is more, Article 7(1) CDSM makes restrictive contractual terms unenforceable, meaning rightholders cannot opt out of scientific research TDM through contract either.

Ultimately, giving a specific meaning to lawful access under Article 3 CDSM should not be seen as problematic, because there are already two different definitions of lawful access under the CDSM.

¹¹⁴ *SIA “CV-Online Latvia” v SIA “Melons”* [2021] CJEU Case C-762/19, ECLI:EU:C:2021:434. para. 37.

¹¹⁵ *ACI Adam BV and Others v Stichting de ThuisKopie and Stichting Onderhandeligen ThuisKopie vergoeding* (n 109) para 54.

¹¹⁶ *GS Media BV v Sanoma Media Netherlands BV and Others* [2016] CJEU Case C-160/15, ECLI:EU:C:2016:644 [49].

¹¹⁷ See e.g., CJEU C-435/12 *ACI Adam BV and Others v Stichting de ThuisKopie et others*, 10 April 2014, ECLI:EU:C:2014:ACI Adam BV and Others v Stichting de ThuisKopie and Stichting Onderhandeligen ThuisKopie vergoeding (n 109) para 54. (“A private copying levy system, such as that at issue in the main proceedings, which does not, as regards the calculation of the fair compensation payable to its recipients, distinguish the situation in which the source from which a reproduction for private use has been made is lawful from that in which that source is unlawful, does not respect the fair balance referred to in the preceding paragraph.”). The concept of lawful source should also be linked to the concept of lawful use, which as specified by Recital 33 Infosoc should mean the following: “A use should be considered lawful where it is authorised by the rightholder or not restricted by law.”

¹¹⁸ Margoni (n 107).

¹¹⁹ Recital 14 CDSM.

¹²⁰ Margoni stresses the importance of interpreting exceptions “in a way that safeguards the effectiveness of the provision.” Margoni (n 107).

Indeed, under Article 3 CDSM, rightholders cannot reserve the use of their works or subject matters whereas they are allowed to do so under Article 4 CDSM.

What is not often acknowledged is that the classical “lawful source” problem is less significant in a researcher-to-platform relationship, since if the source is unlawful, the platform would be subject to takedown obligations. What is more, if we genuinely care about researchers, they should not be required to verify whether the stakeholder presumed to have the authority to grant access to the subject matter actually holds that authority. After all, one line of cases decided by the CJEU sets forth a presumption of lawfulness of the source when the subsequent user is not driven by financial gains, irrespective of the role played by the user, i.e., whether she is a researcher or not.¹²¹ There are good reasons to give researchers an even stronger status. Such a status is warranted to allow researchers to investigate unlawful activities, which is of significant public interest.¹²²

Although Article 3(3) CSDM provides that “[r]ightholders shall be allowed to apply measures to ensure the security and integrity of the networks and databases where the works or other subject matter are hosted,”¹²³ these concerns should not be used as pretext to moot TDM for scientific research purposes, and in any case it is on rightholders to make the case for the implementation of such measures.¹²⁴ What is more Article 17(9) CSDM which provides that “Member States shall ensure that users have access to a court or another relevant judicial authority to assert the use of an exception or limitation to copyright and related rights”¹²⁵ should be stressed, as well as the obligation for these service providers to inform users that they can use subject matters under exceptions provided for in EU law.

Ultimately, even if under Article 7(2) CSMD Article 3 CSMD is subject to the three-step test of Article 5(5) of the older 2001 Information Society Directive (InfoSoc),¹²⁶ TDM for the purposes of scientific research should not trigger major concerns, particularly where the scraping is tailored to the research question, designed to ultimately retain only what is necessary to address it, and is likely to be scrutinised through an ethics approval process. The fact that platforms themselves may offer licenses for data analysis should be considered carefully in the context of platform research since, as explained above, platform-provided data (APIs, etc...) are rarely good substitutes for scraping.¹²⁷ Furthermore, in the case of computational platform research, there is no concern about diminished sales or interference with the normal exploitation of the protected subject matter. In light of the limited volume

¹²¹ The *GS Media* line of case includes a presumption of lawfulness when the subsequent user is not driven by financial gains. *GS Media BV v Sanoma Media Netherlands BV and Others* (n 116) para 55.

¹²² Margoni (n 107). Margoni explores the potential of interpreting content freely available online as referred to in Recital 14 as content that is freely available online without the consent of the right holder or absent any other legal basis and thus covering content that is freely available online without the consent of the right holder or absent any other legal basis.

¹²³ Article 3(3) CSDM.

¹²⁴ What is more, there is thus an argument that Article 3(3) measures are not technological measures within the meaning of Article 6(3) InfoSoc since rightholders can not restrict TDM for scientific research purposes. However, rightholders still have the option to implement effective technological measures to restrict reproduction of the protected subject matter under Article 6 InfoSoc. In this case, Member States are under an obligation to ensure that rightholders give beneficiaries of the exception the means of benefiting from that exception when they have lawful access to the protected subject matter.

¹²⁵ Article 17(9) CSDM.

¹²⁶ Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society OJ L 167, 22.6.2001, pp. 10–19.

¹²⁷ For the use of such an argument in the context of AI training see Rosati (n 104).

of such research and actual attempts to chill legitimate inquiry, asserting an unreasonable loss of income for the rightsholder would be counterintuitive to say the least.

It is also worth noting that scraping for scientific research would not typically generate any material disruption or strain for the platform scraped. Partly, this is because scrapers need only extract specific content, all the while avoiding platforms' anti-bot and anti-scraping mechanisms – which can be extensive.¹²⁸ Hence, scrapers are typically designed to be both targeted and to resemble human users. There are also best research practices for ethical scraping;¹²⁹ those include further steps to actively limit any possible strain on the websites scraped and associated servers by making the scraper as parsimonious as possible (scraping only once or at sporadic and necessary intervals; scraping the minimum amount of data required for the research), and building it to mimic human users as closely as possible (e.g. build in randomised lag between queries to avoid overloading the servers and to avoid triggering anti-bot detection). The assumption that scraping (especially when following best practice) will be disruptive of platforms' infrastructure is neither here nor there; most scrapers go completely unnoticed; by and large, they pose no more threat or disruption to platforms' maintenance or functioning than any regular human user. Especially in the context of constant and systematic internet-wide crawling by bots (including e.g. indexing bots like google bots), if a platform is going to be overloaded by a little research scraper, its design must be considered incredibly unsound.

3.2.3 Data donations

There have already been calls to incentivise platforms to create data-donation repositories.¹³⁰ However, research organisations should be even more proactive and actively support data donation practices by embracing prospective legal methods and leveraging complementary data access mandates, such as those introduced by the DMA and the DA.

As noted above, both the DMA and the DA introduce new data access mandates, which could invigorate data donations to research organisations, acting as third parties authorised by users or end users to receive data directly from data holders/providers and where applicable real-time data.¹³¹ These mandates are often framed as extensions of the data portability right set out in Article 20 GDPR and are compared to it on that basis,¹³² although they obviously comprise two layers: access and export.

Even if the scope of the portability prerogatives found in the post-GDPR regulations do not precisely overlap with the scope of Article 20 GDPR,¹³³ these prerogatives appear more powerful in several

¹²⁸ Kieron Turk, Sergio Pastrana and Ben Collier, 'A Tight Scrape: Methodological Approaches to Cybercrime Research Data Collection in Adversarial Environments' *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (Institute of Electrical and Electronics Engineers 2020).

¹²⁹ Aline Shakti Franzke and others, 'Internet Research: Ethical Guidelines 3.0.' (Association of Internet Researchers 2020) <<https://aoir.org/reports/ethics3.pdf>>; Elizabeth A Buchanan and Michael Zimmer, 'Internet Research Ethics' in Edward N Zalta and Uri Nodelman (eds), *The Stanford Encyclopedia of Philosophy* (2025) <<https://plato.stanford.edu/archives/fall2025/entries/ethics-internet-research/>>.

¹³⁰ See e.g., Mozilla (n 28) 27.

¹³¹ See Article 6(9) DMA ("and including by the provision of continuous and real-time access to such data."); Article 6(10) DMA ("with effective, high-quality, continuous and real-time access to, and use of, aggregated and non-aggregated data, including personal data"); Article 4(1) DA ("where relevant and technically feasible, continuously and in real-time"); Article 5(1) DA ("where relevant and technically feasible, continuously and in real-time").

¹³² See e.g., Barbara Lazarotto, 'The Right to Data Portability: A Holistic Analysis of GDPR, DMA and the Data Act' (2024) 15 *European Journal of Law and Technology*.

¹³³ This is true in particular for the DA which only covers readily available data as already mentioned.

ways. First, they benefit both natural persons and business users. Second, they are not restricted to personal data. Third, when personal data is covered, they are indifferent to the legal basis underpinning the primary processing. Fourth, they impose the building of a direct pipeline between data holders/providers and authorised parties, i.e., a technical and organisational arrangement that enables data to flow directly from the data holder/provider to a third party without discretionary filtering, or case-by-case control by the entity holding the data. Fifth, the formulation of the new data access mandates is at times strengthened by their explicit focus on both data and metadata, and by the requirement that metadata and data exports should support interpretability and usability.¹³⁴ As explained by the Hamburg Supervisory Authority, these data access mandates do not create alternative but cumulative data access and portability rights, even if these rights are not duplicative.¹³⁵

When Article 40 DSA is viewed in the broader context of the DMA, Article 6 of the DMA appear to complement Article 40 DSA and should therefore help researchers either in identifying the missing pieces of the data catalogues populated by VLOPSEs, since data catalogues are not required to be exhaustive and as such do not bind or limit applicant researchers in their data access applications,¹³⁶ or in expanding their understanding of data flows, in particular data flows spanning across services.

Comparing the material scope of the DSA and the DMA it appears that “[w]hile the DSA never covers as big the services in the infrastructure layer because they cannot qualify as online platforms or search engines, the DMA covers big services across all three layers.”¹³⁷ To take the example of the company Alphabet, VLOPSEs under the DSA comprise Google App Store, Google Maps, Google Shopping, Google Play, Youtube and Google Search, while under the DMA core platform services comprise Google Maps, Google Shopping, Google Play, Youtube and Google Search as well as Android Mobile, Google Advertising, and Google Chrome.¹³⁸

What is more, the DMA includes a useful mechanism for discovery by the European Commission, i.e., Article 15 DMA, which obliges gatekeepers to “submit to the Commission an independently audited description of any techniques for profiling of consumers that the gatekeeper applies to or across its core platform services listed in the designation decision pursuant to Article 3(9).” The fact that Article 15 DMA is conceived as a cross-regulation discovery instrument is explicitly acknowledged by Article 15 itself which imposes an obligation upon the Commission to “transmit that audited description to the European Data Protection Board.”

¹³⁴ See Articles 4 and 5 DA.

¹³⁵ The Hamburg Commissioner for Data Protection and freedom of Information, ‘The Data Act as a Challenge for Data Protection’ (2025) Unofficial Translation <<https://media.licdn.com/dms/document/media/v2/D4D1FAQEmj3e5qjxBw/feedshare-document-pdf-analyzed/B4DZaBCCcOGgAc-/0/1745921523685?e=1747267200&v=beta&t=jqbh3NDfwsx1-QG-GwTL04QR5JS92w9R-ChU7jRpZmM>>. For a commentary, see Sophie Stalla-Bourdillon and Bárbara da Rosa Lazarotto, ‘On the Frictions Between the GDPR and the Data Act: A Closer Look at the Hamburg SA’s Position Paper’ (2025) 2025 Privacy and Data Protection 3.

¹³⁶ Recital 8, Delegated Act.

¹³⁷ Martin Husovec, ‘General Report Topic II - FIDE Report on the Digital Services Act and the Digital Markets Act - EU Digital Economy: General Framework (DSA/DMA) and Specialised Regimes’ (2025) 29 <https://www.fide-europe.org/xms/files/Katowice2025/REPORTS/T2/T2_EU_Digital_Economy_General_Report.pdf?utm_source=chatgpt.com>.

¹³⁸ Compare the following two lists: European Commission, ‘Gatekeepers’ (*European Commission*) <https://digital-markets-act.ec.europa.eu/gatekeepers_en>; European Commission, ‘Supervision of the Designated Very Large Online Platforms and Search Engines under DSA’ (*European Commission*) <<https://digital-strategy.ec.europa.eu/en/policies/list-designated-vlops-and-vloses>>.

Not only does the DMA complement the DSA, but the DA can also support the development of a valuable source of intelligence through its own data access mandate. This is because connected products such as smartphones are in scope for the data access mandates introduced by the DA.¹³⁹ Smartphones mediate and control a wide range of digital interactions and services. As a result, they function as key gateways for many different data streams.

A clearer understanding of the data produced directly by a smartphone and its operating system helps distinguish and interrogate the opaquer data flows created by applications. Device- and operating system-level data form the baseline: sensor outputs (for example, GPS or accelerometer data), battery and performance metrics, network signal information, and system-level logs exist simply because the phone operates as a connected product. When this baseline is known, it becomes easier to identify additional data streams that apps generate or infer. Mapping the foundational layer makes deviations into app-specific tracking more visible, clarifies which actor determines the creation of which data, and supports regulatory analysis by separating inherent device data from data that arises through app logic or user interaction.¹⁴⁰

What is more, the DA is particularly interesting in its emphasis on metadata for interpretability and usability, which is often missing from data donation exports.¹⁴¹ Yet, as stressed by the Hamburg Supervisory Authority, the DA access and portability mandates should usefully influence the GDPR ones,¹⁴² despite confusing case law on the inclusion of metadata in responses to data access requests under Article 15 GDPR.¹⁴³

Importantly, cross-research-organisation initiatives may offer a way forward by creating data intermediaries specialised in supporting data donations. These intermediaries could benefit from an accelerated pathway, which is not explicitly provided for under the Data Governance Act (DGA) however.¹⁴⁴ The DGA takes the opposite approach, building trust by imposing a strict regulatory regime on a limited set of data intermediaries.

¹³⁹ See Recital 14 DA (“Connected products that obtain, generate or collect, by means of their components or operating systems, data concerning their performance, use or environment and that are able to communicate those data via an electronic communications service, a physical connection, or on-device access, often referred to as the Internet of Things, should fall within the scope of this Regulation, with the exception of prototypes.”) See also European Commission, ‘Frequently Asked Questions - The Data Act’ (2025) Version 1.3 <<https://digital-strategy.ec.europa.eu/en/library/commission-publishes-frequently-asked-questions-about-data-act>>. There does not seem to be clarity as regards whether computers would qualify for the definition of connected products. As per Article 2(5) DA, are excluded products “whose primary function is [...] the storing, processing or transmission of data on behalf of any party other than the user.”

¹⁴⁰ By way of example, it has been shown that third-party apps can read the data generated by embedded sensors without any permission. Xing Liu and others, ‘Discovering and Understanding Android Sensor Usage Behaviors with Data Flow Analysis’ (2018) 21 World Wide Web 105. See also Hansoo Lee, Joonyoung Park and Uichin Lee, ‘A Systematic Survey on Android API Usage for Data-Driven Analytics with Smartphones’ (2022) 55 ACM Comput. Surv. <<https://doi.org/10.1145/3530814>>.

¹⁴¹ Entrena-Serrano and others (n 64). Particularly problematic is the fact that platforms frequently change data donation formats in ways that reduce the interpretability and usability of data.

¹⁴² The Hamburg Commissioner for Data Protection and freedom of Information (n 135). See also Stalla-Bourdillon and da Rosa Lazarotto (n 135).

¹⁴³ Stalla-Bourdillon and da Rosa Lazarotto (n 135).

¹⁴⁴ Regulation (EU) 2022/868 of the European Parliament and of the Council of 30 May 2022 on European data governance and amending Regulation (EU) 2018/1724 (Data Governance Act) OJ L 152, 3.6.2022, pp. 1–44.

The simplification agenda triggered by the European Commission through the proposed Digital Omnibus¹⁴⁵ may, however, undermine, or at least slow down, progress in data donations. This is because the proposed amendments to the GDPR and the DGA seek to increase the conditionality of the GDPR data access right and shifts away from the explicit emphasis on data cooperatives, which are intended to facilitate the exercise of data subject rights.¹⁴⁶

4. Conclusion

The paper examines the concept of affordances in the context of online platforms, drawing on Gibson's ecological theory and Treem and Leonardi's taxonomy of social media affordances. It focuses specifically on data accessibility for researchers and civil society actors, highlighting how both technical design and governance rules shape perceptions and practices of data access. The paper argues that prospective legal methods can help reveal gaps between *de facto* (practical) and *de jure* (legal) affordances and proposes ways to bridge them. By distinguishing between closed and open affordances, it emphasises that legally shaped affordances, such as data accessibility, are not fixed but evolve through interpretation by actors such as regulators, courts, and research organisations.

The paper contrasts perceived versus real affordances of data access on platforms, showing that restrictive terms of service, platform-driven changes and platform-initiated litigation have led many researchers to view access as limited or legally risky. It argues that data accessibility as a *de jure* affordance is inherently open-ended and dialogical, with legal rules characterised by “open texture” and only partially shaped by interpretive entrepreneurship. This openness creates both constraints and opportunities, depending on how ambiguities in law and platform governance are interpreted. Within this framework, the paper shed light upon two key interpretative battles: the lawfulness of web scraping and the protection of trade secrets.

Moreover, the paper emphasises that legal interpretation is normative, guided by the principles and policies that interpreters prioritise rather than formalistic rules. In the context of researchers' access to platform data, applying prospective legal methods involves two steps: first, identifying researchers' concrete data needs; and second, interpreting the law to accommodate these needs, moving beyond realistic readings to explore possible interpretations through a speculative process. Because researchers require systematic, representative, high-quality data collection, exploratory and iterative engagement, and mechanisms for documenting and reproducing analyses, the paper argues that meeting these needs necessitates multiple access contexts, control planes, data scraping, and data donations, each offering distinct advantages and limitations.

The paper highlights the importance of limiting the constraints set upon research work within secure processing environments, of moving away from a realistic and thereby conservative assessment of data scraping and of bridging the different chapters of the EU digital rulebook, including the DSA,

¹⁴⁵ European Commission, ‘Proposal for a Regulation of the European Parliament and of the Council Amending Regulations (EU) 2016/679, (EU) 2018/1724, (EU) 2018/1725, (EU) 2023/2854 and Directives 2002/58/EC, (EU) 2022/2555 and (EU) 2022/2557 as Regards the Simplification of the Digital Legislative Framework, and Repealing Regulations (EU) 2018/1807, (EU) 2019/1150, (EU) 2022/868, and Directive (EU) 2019/1024 (Digital Omnibus)’ (2025) COM/2025/837 final <<https://digital-strategy.ec.europa.eu/en/library/digital-omnibus-regulation-proposal>>.

¹⁴⁶ Bárbara da Rosa Lazarotto, Sophie Stalla-Bourdillon and Pablo Trigo Kramcsák, ‘Giving EU Data Governance Law a Second Life? From Rebranding to Real-World Impact’ (2025) Brussels Privacy Hub Working Paper <<https://researchportal.vub.be/en/publications/giving-eu-data-governance-law-a-second-life-from-rebranding-to-re/>>.

DMA, DA, and GDPR, that establish layered access and portability rights. The paper thus suggests that by supporting stackable data-access mandates, research organisations can strengthen platform studies, enhance transparency and reproducibility of research practices, and help advance a critique of the dominant data imaginary, which should confirm that as much as platform research is important, it is crucial to enforce as quickly as possible red flags.

Ultimately, this paper invites regulators and courts to counter interpretative entrepreneurship, with regulators having a very important proactive role to play. This paper then calls on research organisations to invest in support mechanisms that extend beyond conservative compliance guidance and foster cross-organisational collaboration. This paper also calls on legal researchers to adopt a more systematic approach to legal methods, one that explicitly states their normative stance and accounts for the full range of possible legal interpretations, without reflexively equating legal indeterminacy with unmanageable complexity.

Acknowledgements:

The authors would like to thank Thomas Margoni, Alina Trapova and Carlos Alejandro Entrena Serrano for fruitful discussions and all researchers who shared their feedback on research-institution processes. All remaining errors are the authors' own.