The Stakes of "Publicly Accessible"

Researchers' Rights to Data under the DSA

Daphne Keller

12.12.2025 doi: 10.59704/bc04ce068f5dae6c

One of the main goals of the EU's Digital Services Act (DSA) is to advance transparency about online platforms. Article 40 seeks to do so by providing researchers with access to data about Very Large Online Platforms and Search Engines (VLOPSEs). As discussed in a prior blog post, Article 40(4) establishes a slow, careful process for vetted academic researchers to access platforms' internally held data. Article 40(12) complements this – and every other DSA transparency provision – by allowing a broader array of researchers to collect or scrape information that platforms are actually displaying to users in their public interfaces. This is an important enough requirement that it has been part of several Commission investigations against platforms, and was one of three grounds for its €120 million enforcement against X.

Researchers may collect data under Article 40(12) only if platforms have already made it "publicly accessible". The universe of information available for these researchers will depend on what "publicly accessible" means. If history is any guide, platforms will oppose research by arguing that some data – despite being freely visible to anyone who visits their sites or apps – is not "publicly accessible" in the sense of Art. 40(12). Similar disputes about lawful information "access" have derailed journalism and research on both sides of the Atlantic for years. The term "public" also has a confusing array of legal meanings in contexts ranging from news reporting to insider trading to copyright.

The purpose of Article 40(12) would be defeated if similar uncertainty deterred the very research that lawmakers intended to unleash. The DSA provides other legal mechanisms to balance research goals with potentially competing policy priorities like data protection. "Publicly accessible" data for research should be defined broadly, given the DSA's language, purpose, and overall design. This post will examine the legal scope of available data under Article 40(12). It begins with legal analysis, continues with comparison to other relevant laws, and concludes with review of specific data categories for research.

I. Text and Context in the DSA

While the DSA does not define "publicly accessible", the term's ordinary meaning is simple: it refers to information that the *public* can *access*. In other words, information that platforms themselves publicly display should be available to researchers. But the DSA, which also uses similar terms like "make public", "public access", or "publicly available", offers few textual cues about what precise data is "publicly accessible". This inconsistent wording may be an artifact of hasty drafting. Article 40(12) was introduced late in the DSA's development, and negotiated in trilogue between three EU lawmaking bodies without public input. Similar definitional problems also seem endemic to laws about information access. The University of Amsterdam's Kacper Szkale reported similar issues defining "lawful access" under EU copyright law: legislation uses varied terms that "all seem related, yet they are differently formulated and some can be understood to refer to completely different things".

Article 40(12)'s basic purpose is self-evident, though. It aims to eliminate barriers – particularly platform-created barriers – to research. This portion of Article 40 was drafted after a series of high-profile platform attacks on researchers and major retrenchments or failures in platform-administered transparency efforts. Lawmakers clearly intended to expand researcher rights beyond the status quo – but also to balance competing interests and policy priorities. The same EU Council draft that created Article 40(12) also

added Article 40(8)'s requirement that researchers use data only as "necessary for, and proportionate to" legitimate research goals.

Article 40(8)'s balancing requirement protects competing interests without the need for a cramped interpretation of "publicly accessible" that would lock data away from *all* Article 40(12) researchers. Article 40(12) also fits within a larger structure of DSA rules supporting a broad interpretation of "publicly accessible".

A. The Outer Limits of "Publicly Accessible"

To be eligible for DSA research in the first place, data must come from the online interface of a regulated VLOPSE. DSA Article 3's restrictive definitions of "online platforms" and "VLOPSEs" thus, independently, bar Article 40(12) researchers from using much of the data that might colloquially be considered not "publicly accessible". They may not collect messages from non-VLOP messaging apps or private Discord servers, for example. Even private messages on Facebook were excluded from the company's VLOP services designation.

Facebook posts shared to only a few "friends" present a harder case, because those posts *are* presumably part of a VLOP service. Courts or regulators might ultimately interpret Article 40(12) to mean, for example, that posts shared with more than a specified number of people are "publicly accessible". Alternately, authorities could rely on Article 40(8) to adopt the more flexible, project-specific standards discussed below.

B. Distinguishing Article 40(4) and Article 40(12) Data

Article 40 classifies data into one of two categories, as discussed in the previous post. Either it is "publicly accessible" to a wider array of qualified researchers under Article 40(12), or else it is available only to vetted academic researchers who navigate Article 40(4)'s intensive regulatory review process – which was designed to protect platforms' confidential, internally held data. One way to define "publicly accessible" data under Article 40(12) is by delineating which information access should require regulators and researchers to go through Article 40(4)'s weighty process.

C. Balancing under Article 40(8)

The meaning of terms like "publicly accessible" is often hotly contested under laws that lack other doctrinal tools to balance freedom of information goals with competing interests such as privacy, property, or security. U.S. research has suffered for decades as courts dithered about legal "access" under copyright law and the anti-hacking law that X has used to sue public interest researchers.

The DSA does not need to replicate these problems by shoehorning policy balancing exercises into the definition of "publicly accessible". Article 40(8)'s proportionality requirement provides a far better tool to balance research goals against other priorities. Laws like the General Data Protection Regulation (GDPR) also form a critical backdrop to the DSA, imposing their own separate obligations on researchers.

Article 40(8) requires that research protocols respect competing rights and interests. As the CJEU recently recognized, these may be specific to individual projects. A burgeoning professional literature on public data and scraping ethics can guide researchers' choices. Researchers studying healthcare fraud might, for example, scrape data from a Facebook group for cancer patients, but discard identifying information or posts that lack particular words. Data protection regulators have suggested similar approaches for other data scrapers, with limits on both what data is *collected* and how the data is subsequently *used*. This flexible, multi-stage legal assessment contrasts strikingly with rigid laws in the U.S., where information that becomes "public" may automatically lose many legal protections. In the EU, data can be "publicly accessible" and also protected from more improper uses.

Research protocols can protect an array of interests. A study about violence against women might scrape videos from YouTube and Pornhub, for example, but blur some sexual content. Among other things, this would avoid creating a trove of commercially valuable copyrighted material – protecting the interests of affected (and notoriously litigious) copyright holders in the event of a data breach.

The exact parameters of protected research under Article 40(8) remain to be seen. Logically, Article 40(12) must authorize – and, as Martin Husovec has discussed, immunize – new uses of data beyond those permitted pre-DSA. Article 40(8) can flexibly define the scope of this new legal authorization, without requiring a strict definition of "publicly accessible" that categorically precludes valid future research.

II. Competing Policy Goals

Arguments against expansive interpretations of Article 40.12 are likely to invoke legitimate competing policy goals, including data protection and intellectual property. The DSA was enacted with, and should be interpreted in light of, its own goals. But courts and regulators defining proportionate use under Article 40(8) – or, potentially, defining the scope of "publicly accessible data" – may also look to other EU laws for guidance. Importantly, support for researcher data access can be found in the DSA's intersections with many relevant areas of law.

A. Data Protection

Many of the hardest questions concern platform users' fundamental rights to privacy, data protection, and sometimes even protection from state surveillance. Examples discussed below include posts deleted by users or shared in very large "private" Facebook groups. I will argue that, for each, Article 40(8) offers the best tool to define permissible use. But reasonable minds may differ, and courts or regulators might ultimately decide that some such data is categorically not "publicly accessible" under Article 40(12).

Importantly, the rights at stake are those of the same platform *users* that the DSA was intended to protect. The European Data Protection Supervisor has expressed suspicion about platforms asserting users' data protection rights in objecting to research. Companies may not, he wrote, "escape accountability… on the pretext of safeguarding the rights of others". Respecting data protection rights does not mean honoring every user preference, however. Otherwise, as experts including Paddy Leerssen have pointed out, bad actors could exclude their posts from DSA research.

B. Intellectual Property

Platforms might argue that some data is not "publicly accessible" because of their own intellectual property rights or their restrictive licensing agreements with third parties such as music rightsholders. Since most user posts are protected by copyright, excluding content on this ground would extinguish a wide swathe of research. So, too, would a rule allowing platforms to put data off-limits simply by accepting restrictive licenses.

Article 40's policy goals are, in any case, well-aligned with EU policy under the 2019 Copyright in the Digital Single Market Directive (CDSM). That law protects non-commercial researchers' "text and data mining", including "any automated analytical technique[s] aimed at detecting and analyzing text and data in digital form in order to generate information which includes but is not limited to patterns, trends, and correlations". Copyright experts have suggested that Article 40(12) research should presumptively qualify for CDSM protections. The two laws may also be complementary in other ways. The CDSM gives researchers *legal authorization* to use data without infringing copyright, but no clear *access mandate* to obtain it. The DSA's requirement for VLOPSEs to "give access" to data can help fill that gap.

C. Platforms' Terms of Service

Some platforms have historically argued that researcher scraping violates their terms of service. As an argument against Article 40(12) research, this sounds absurd: VLOPSEs should not be able to nullify DSA obligations by the "one weird trick" of amending their unilaterally-imposed contract terms. The Commission's decisions initiating proceedings against AliExpress and Temu and its enforcement against X reinforce this conclusion. In all three cases, the platforms' terms of service prohibited scraping. This likely violated Article 40(12), as the Commission explained, because it "failed to give qualified researchers access to the data that is publicly accessible". In other words, contractually restricting access did *not* prevent data from being "publicly accessible" under Article 40(12). It merely violated platforms' obligations to "give access".

Uncertainty about the validity of contractual restrictions on "accessible" data has been a major hindrance to research under other laws. The CDSM, for example, specifically provides that contractual provisions claiming to bar authorized research "shall be unenforceable". But the CDSM only authorizes use of data when researchers have "lawful access" in the first place. As Szkalej has explained, this leads to circular questions about whether contracts – or technical measures like blocking researchers' IP addresses – can make data no longer lawfully accessible. Remarkably similar disputes about contracts and lawful "access" plague journalists and researchers in the U.S. The DSA's research agenda should not founder on these same rocks.

III. Examples of Potentially Public Data

Some categories of relevant data for DSA researchers could arguably be classified as not "publicly accessible" under Article 40(12) – but could, alternately, be subject to more nuanced proportionality limits under Article 40(8). In most cases, I believe the latter is preferable.

A. Data from platforms that require a login, but make login credentials publicly available

In 2023, X temporarily made all posts, including those marked as public, visible only to logged in users. The general public could still create new accounts and automatically be approved to view the posts. Could adding this trivial login barrier make data no longer "publicly accessible" to researchers under Article 40(12)?

For this data, the core policy question is about users' rights and expectations. If users intended for their posts to be publicly available without a login, platforms should not have the power to unilaterally withdraw them from DSA research. That said, users' expectations of privacy will vary across platforms and even kinds of research. Where there is more serious reason to believe that users expected a login system to prevent the general public from seeing their posts, those posts might be considered not "publicly accessible". Alternately, the same consideration could shape more nuanced, project-specific, rules under Article 40(8).

Unusually, the DSA itself offers textual clues about this 'posts-behind-a-login' scenario. A close examination, however, illustrates the problems with relying on imprecise DSA language to answer questions of this sort. The rest of this subsection engages in rather tortured textual exegesis, with inconclusive results. Readers who don't enjoy that kind of thing may wish to skip ahead.

The DSA does not define "publicly accessible", but does define something similar-sounding: "dissemination to the public". That term means "making information available, at the request of the recipient of the service who provided the information, to a potentially unlimited number of third parties". Recital 14 expands on this, specifically addressing situations "where access to information requires registration or admittance to a group" of platform users. That data still counts as "disseminated to the public" when users are "automatically registered or admitted without a human decision or selection". If this Recital about publicly *disseminated* information also applied to publicly *accessible* information under Article 40(12), then the X posts would be available for researcher use.

That said, there are reasons not to equate the two terms. The DSA primarily uses the word "dissemination" in articles unrelated to research or users' rights, often in provisions defining different platforms' tiers of DSA obligations. A "hosting service" becomes a more heavily regulated "online platform" if it "disseminates information to the public", for example. And online platforms graduate to the most heavily regulated VLOP category based on the number of recipients to whom content is "disseminated". It seems unlikely that DSA rules drafted to bring platforms in scope of the regulation would simultaneously be intended to define the balance between research and user privacy. Proportionality analysis under Article 40(8) seems like a better legal tool than this strained textual analysis.

B. Data from huge "private" groups

The largest "private" Facebook groups include millions of people, dwarfing most physical public spaces. Researchers and whistleblowers see these

groups as staging grounds for larger public trends, making them relevant for DSA research.

Posts to "public" Facebook groups can be seen without an account; posts to "private" groups are visible only to members. If we analogized public "accessibility" to public "dissemination", as discussed above, then Recital 14 would apply: researchers could scrape groups if members were "automatically registered or admitted without a human decision or selection". This seems like a relevant factor in discerning users' expectations or intentions, but hardly the only one. Again, Article 40(8) proportionality analysis would allow for more tailored rules.

C. Posts deleted by platforms

Posts that platforms delete after researchers obtain copies are highly relevant for DSA research, which often focuses on content moderation. They also present particularly thorny legal issues, particularly for content that platforms considered unlawful to publish. Researchers are likely to have affirmative legal obligations to report and delete child sexual abuse material, in particular. Content that may be defamatory, copyright infringing, or supportive of terrorism, however, is different. Both the DSA and existing non-DSA laws may well permit ongoing research uses of that content, and even allow excerpts to appear in published findings.

DSA drafters were clearly aware of this concern. They addressed it in a different transparency provision, Article 39, which requires VLOPSESs to host public advertising archives that exclude content that violated the law or platforms' own rules. Omitting that exclusion requirement in Article 40's rules for qualified researchers seems purposeful, and suggests that data can be "publicly accessible" based on *past* availability on platform's services.

D. Posts deleted by users

Posts deleted by platform *users* can be different. Users both expect to be able to delete posts and have rights to do so under laws including data protection and copyright. Researchers who collect posts may, knowingly or unknowingly, retain copies after users have deleted them from platforms. Deleted posts can sometimes be important, documenting publicly significant information that political leaders, high profile trolls, or influencers sought to hide.

Such posts should not be excluded from research use simply because they are not *currently* "publicly accessible". That interpretation would also eliminate posts moderated by platforms, contrary to the DSA's likely goals. It would also make maintenance of clearly-legal research datasets very difficult. Instead, both Article 40(8) and the GDPR's background rules suggest that research protocols must be designed to account for users' deletion interests.

E. "Public" posts from users who reasonably expected few viewers

Some social media users post "publicly", but realistically expect that only a few people will see their posts. Many of the reported 48% of Twitter users with fewer than ten followers, per early research on the social network, probably fell in this category. Parents who post videos of a baby's first steps for distant family on YouTube, similarly, probably never expect their video to have an afterlife in a European research center. The median YouTube video has only 39 views, so many videos likely fall in what we might call a contextually private category.

In principle, content with a sufficiently low view-count or follower-count might be deemed "non-publicly accessible" under the DSA. But efforts to avoid collecting data on such bases often fail for technical reasons beyond researchers' control. For both this practical reason and the policy reasons discussed above, classifying such efforts as Article 40(8) proportionality measures makes more sense. That would also allow more nuanced approaches, like the tiered researcher access rules that Former CrowdTangle CEO Brandon Silverman suggests for "meaningfully public data", the focus on "inferred user expectations of privacy" suggested by Mozilla, or the ethical guidelines for "unpermissioned research" proposed by the University of Amherst's Ethan Zuckerman and Ryan McGrady in forthcoming work.

IV. Conclusion

Researchers who are qualified and eager to begin work under DSA Article 40(12) may be deterred by uncer-

tainty about what data counts as "publicly accessible". That definition should be broad in light of Article 40's pro-research purpose, the DSA's overall structure, and its inclusion of alternate means to balance competing interests.

This text is a cross-posting with Tech Policy Press.

ABOUT THE AUTHOR

Daphne Keller

Daphne Keller is the Director of Platform Regulation at Stanford Law School's Program in Law, Science & Technology.Daphne Keller is the Director of Platform Regulation at Stanford Law School's Program in Law, Science & Technology.

SUGGESTED CITATION

Keller, Daphne: The Stakes of "Publicly Accessible": Researchers' Rights to Data under the DSA, Verfassungsblog, 12.12.2025, doi: 10.59704/bc04ce068f5dae6c.

Licensed under CC BY-SA 4.0.

Langzeitarchiviert bei IntRechtDok unter: doi.org/10.17176/20251214-172117-0.