Delegated Regulation on data access provided for in the Digital Services Act

Comment of Daphne Keller
Director, Program on Platform Regulation, Stanford Cyber Policy Center
May 22, 2023

Observations on the State of Play

I write in my personal capacity as a scholar of platform regulation and as a technology lawyer with considerable experience within platforms. I direct the Program on Platform Regulation at Stanford’s Cyber Policy Center, and previously served as Associate General Counsel at Google. I have worked on platform regulation for over twenty years. My work on platform transparency includes U.S. Congressional testimony, blog posts, and other written publications, and participation in numerous events and discussions with researchers, civil society, platforms, and representatives of governments.

This recitation of credentials is offered in the hopes that it will bolster the somewhat impressionistic observations with which this submission begins. From what I can see, the discussion of platform transparency is in its very early days. Maturation that might otherwise have been expected over the past several years was delayed by the Covid pandemic. The result is a discussion that is far more fragmented and evolving than that encountered by experts in most other aspects of platform regulation. The global academic and civil society dialog about transparency bears little resemblance to the much more mature discussion and shared understanding of relevant issues that informed many other DSA provisions, such as the notice and action rules. Those were shaped by decades of experience, litigation, scholarship, policymaker inquiries, and civil society engagement. No comparable foundation exists for platform transparency rules. In some areas, experts have made considerable progress, particularly in devising sophisticated measures to reconcile transparency and data protection. But overall, we are very much still figuring out how this can all work.

The practical and policy questions raised by platform transparency mandates are vast. Transparency rivals or exceeds the rest of intermediary liability in its complexity. Reaching sound and informed conclusions will require analysis that cuts across areas of expertise ranging from data science to surveillance, and from misinformation to machine learning. Experts who might make important contributions from these and other fields have not necessarily been part of the same discussions to date. Some have participated within specific silos or pockets of practice and expertise, unaware that the others exist. To the extent that this consultation yields participation from across relevant fields, organizations, and regions, it may in itself provide a
revealing new survey, showing who is thinking about the mechanics and deeper policy issues surrounding platform transparency.

This larger conversation about platform transparency, and the expectations and practices under the DSA, can and should evolve in the coming years. Article 40 provides an important part of the framework for that to happen. In order to realize Article 40’s tremendous potential, and to some extent “future-proof” it, the Delegated Act should build in the flexibility to make this future growth possible. To the extent possible, regulators and stakeholders involved in implementation should have opportunities to revisit and update the rules and processes for data disclosure. My remaining and more specific comments are drafted with that backdrop in mind.

**Article 40.12 and Scraping**

Article 40.12, which spells out rules for researcher access to data that is publicly accessible in platforms’ online interfaces, is a critical backstop not only to vetted researcher access under Article 40.4, but to the entire DSA. As I read it, Article 40.12 unambiguously supports the extraction of data and information from publicly available webpages, also known as “scraping,” by a broader scope of researchers than those vetted under Article 40.4. It also presumably supports data access for those same researchers via application program interfaces (APIs).

Scraping is a standard information retrieval mechanism which underlies much day-to-day activity on the Internet. Perhaps most famously, it is the necessary first step in assembling web search indexes like those provided by Google or Bing. Scraping underlies an assortment of other web businesses, as well as public interest efforts like Pro Publica’s journalism tools and reporting. Like any other tool for information transmission, it can be used for good and bad purposes. In the U.S., Meta has been particularly aggressive in opposing third party scraping, successfully eliminating a would-be provider of social media interoperability and threatening academic researchers. One of its recent lawsuits was reportedly against a firm Meta itself had hired to scrape third party sites.

Scraping has been essential to real-world platform research in the past, and will continue to be under the DSA. Without the ability to assess the content that platforms actually show to users, public and regulatory understanding of platform behavior – and DSA compliance – will be hobbled. Scraping is a rare form of transparency that does not depend on the very platforms who are being studied to generate information or act as gatekeepers. As such, it is uniquely resistant to problems caused by platforms’ own errors and inadvertent or deliberate manipulation of data. Scraping allows researchers to see the world, or at least the very important world of content made publicly available by platforms, as it actually exists.
For researchers who engage in scraping, Article 40.12 does two things. First, it immunizes them against lawsuits by platforms, who have otherwise been known to bring claims ranging from breach of contract to copyright law infringement against scrapers. Second, it gives researchers their own mechanism to compel platforms to permit scraping, suspending technical barriers that might otherwise frustrate research.

Article 40.12 is wise in providing scraping rights for publicly accessible data to a much broader range of researchers than those who can be vetted under Article 40.4. Allowing a broader diversity of researchers, with assorted backgrounds, skills, and interests to carry out research is consistent with the old maxim of open source software development: with enough eyeballs, all bugs are shallow. Problems that may not be solved by a smaller team may become tractable through attention from more diverse and distributed sources.

The Delegated Act should spell out that researchers acting under Art. 40.12 are legally immunized against anti-scraping claims from platforms, and that platforms may not impede their work. At the same time, it should not provide prescriptive rules that might, intentionally or not, narrow the scope of covered research or researcher rights. Article 40.12 is an effective backstop to other DSA provisions precisely because it is open-ended and flexible for unforeseen future uses. It should remain so.

This is not to say that other laws provide no limits on scraping. Clearly non-DSA areas of law such as the GDPR may do just that, and provide legal guardrails for anyone – researchers or otherwise – who engages in scraping. Equally clearly, some kinds of scraping are socially beneficial, including for reasons that go far beyond research. Among other things, scraping can promote or provide mechanisms for competition and innovation. It is precisely because scraping is implicated in such broad areas of law and policy, beyond that addressed by Article 40 or the DSA, that the Delegated Act should avoid setting new restrictions that may have unintended consequences – either through direct impact on researchers, or by establishing indirect precedent or models for other areas of law.

In particular, the Delegated Act should not create new gatekeepers with power to limit scraping, or authority to define a new and more restrictive universe of eligible researchers under Article 40.12. Regulators or independent advisory bodies may productively serve as intermediaries, supporting researchers and helping platforms identify scrapers whose work is likely to be protected under Article 40.12. They should not, however, be vested with power to prevent, impede, or become necessary sources of permission for other scraping-based research.
Formulating Reasonable, Actionable, and Productive Data Requests Under Article 40.4

Article 40 sets forth a skeleton process for coordination between researchers, platforms, and Digital Services Coordinators. In order for Article 40 to reach its real potential, it will be important to devise more detailed mechanisms for that coordination to take place. The goal of this coordination should be to identify (1) what relevant data platforms actually have, in what formats, (2) how well the data will serve legitimate research purposes, and (3) what risks or downsides will accompany the disclosure and use of the data. With this information in hand, it will be possible to make better assessments of the data necessary for, and proportionate to, the purposes of particular research. It is not realistic to expect Digital Services Coordinators to possess the expertise in data science or other fields to make these determinations in many cases. Independent advisory bodies of the sort identified in Article 40.13 may have important roles in filling this gap.

One difficulty in this process comes from the lack of shared understanding about what data platforms have, and what queries or data requests will in fact generate information that can answer researchers’ questions. This can be a major problem even for platform employees trying to gather internally held data, as I detailed in a recent blog post. Identifying the right definitions of key metrics, and the relevant data, is often a very iterative process. It may require the expertise of multiple platform employees, and multiple failed attempts, before reliable methods for data collection can be established. A number of examples of sources of error or inconsistency in data gathering, which I compiled with the input of many experienced professionals, is here. These examples are from the world of content moderation, but similar small-but-critical questions will arise in other areas of research. Researchers and DSCs will be frustrated, and the purposes of Article 40 will be thwarted, without a mechanism to facilitate dialog or set standards and expectations to ease this process.

Another major difficulty involves identifying the tradeoffs involved in a researcher’s data request. Most obviously, if a researcher seeks personally identifying information about platform users, serious questions about privacy and data protection will arise. Critically, individual users may be identifiable even from data sets that are supposedly “anonymous” or “pseudonymous,” for reasons I explore in detail here. This renders the participation of privacy and data experts essential in vetting projects under Article 40.4. Another set of tradeoffs involves platforms’ trade secrets, confidential strategies for combating spam or other harms, or security practices – including practices that relate to state surveillance or industrial espionage. Researchers may of course be contractually bound to maintain confidentiality for such information. Academic research organizations may provide the resources, expertise, ERB review, and institutional muscle to maintain best practices. At the end of the day, though, measures of this sort have limits. Mistakes will be made, and information will leak. As one example, some researchers will
inevitably later become government employees in countries around the world, or go to work for platforms’ competitors, rivals, or business partners. (As someone who has worked at both Google and Stanford, I have seen this researcher-to-industry-or-government career arc play out hundreds of times.) These individuals will unavoidably, and perhaps unconsciously, bring with them know-how or confidential information gleaned from data disclosed under Article 40.

Assessing these risks, how they may be mitigated, and whether they are “necessary for, and proportionate to” the purposes of research, will be a complex undertaking. So, too, will be defining the relevant data sets. The bare process laid out in Article 40 – in which researchers identify required data and DSCs vet research before any dialog with the platform, which must then respond within fifteen days – would, standing alone, be a recipe for frustration on all sides. In many cases it would impair the actual research that Article 40 should unlock.

A number of mechanisms might alleviate these problems. Many could be facilitated by independent bodies of the sort anticipated in Article 40. A DSC vetting process that meets the requirements of the DSA could also be supplemented, for example, by optional steps and dialog with platforms, independent bodies, or other relevant organizations and stakeholders. A DSC might look to the existence or results of such a process in assessing whether a researcher’s application makes the requisite showing of proportionality and necessity under 40.8(e).

Similarly, DSCs could encourage platforms to go beyond the bare requirements of the DSA, by identifying other concerns besides those listed in 40.5(a) and (b) in responding to data access requests or seeking amendments. The input that DSCs are willing to consider or seek from researchers, platforms, and other entities in vetting applications should not be limited strictly to the information listed in Article 40 itself. Research under 40.4 is meant to contribute to the “detection, identification and understanding of systemic risks” under Article 34. The values and goals identified in that Article -- ranging from the prevention of gender-based violence to the protection of freedom of expression and information -- are necessarily relevant in assessing an application under 40.8.

An example of a mechanism to facilitate near-term realization of Article 40’s potential comes from my Stanford colleague Nate Persily. It is informed by Dr. Persily’s extensive and frustrating work on Social Science One. DSCs could prioritize approval of an initial large batch or batches of carefully vetted data, ensuring that researchers whose questions can be answered using that data have a relatively swift and simple path to approval for new research. Persily suggests a “who saw what, when” data set similar to that from Social Science One. This could serve a dual purpose. First, it would enable an immediate, administratively simple roll-out of valuable research projects. Second, the process of vetting this data and research could serve as a stress test, and inform later vetting processes.
More broadly, I believe that Article 40 and the Delegated Act can best succeed by fostering the early creation of shared tools and data sets (APIs or ad libraries, for example), as well as the shared forums for discussion that will allow the research community to identify priorities and data needs. Wise development of these foundational resources can unlock more and better research than would be achieved with equivalent investment in ad hoc or isolated projects, and can complement those more individualized projects.

Informed investment in the underlying infrastructure to support sound research will be important, given the ultimately limited resources available to academics and other researchers. To the extent that the European Union itself can provide funding to build foundational tools and resources to support an array of future research projects, such spending will be a sound priority. Similarly, it will be important for universities, grantmaking foundations, and other funders of academic research to have the wherewithal to identify viable and promising research projects. Funders should have the tools and information necessary to vet projects early in their decision-making process. They should not learn about issues such as those discussed in this filing only after resources have been allocated. Like everything else about platform transparency, this will be a learn-as-we-go process. The Commission can facilitate that process through support for shared resources, and a flexible and pragmatic Delegated Act.

**Making Surveillance Capitalism Worse**

As a veteran of many years of platform regulation, I am firmly of the “if you build it, they will come” school of thought. Tools developed for one purpose are inevitably sought out and deployed for others. Anticipating those future uses should be part of responsible technology development.

A concern about platform transparency is that work platforms do to provide access to data and information will also create new mechanisms and resources for use in ad targeting or other practices often decried as “surveillance capitalism.” It may also render data more readily accessible for [law enforcement](https://www.law.com) in the United States and around the world, reducing both legal and practical arguments that platforms might otherwise have had against providing users’ personal information to those governments.

The problem is not that transparency mandates will cause platforms to *collect* new user data. It is that transparency mandates may cause platforms to *organize* that data, and make it usable, in ways they would not otherwise have bothered to do. This will be particularly true for VLOPs beyond the few famously data-driven companies, such as Meta. Once a platform has invested the employee time and resources necessary to aggregate, package, and verify data, be it for
researchers under Article 40 or transparency reports under Article 24, it might as well use the same data for other purposes.

Experts I have talked to are divided about how serious a problem this is. My own conclusions are not firm enough to support a strong recommendation. At minimum, however, both the Commission and DSCs should act in awareness of this question in vetting research proposals for data that exists, but has not previously been assembled or organized in the requested fashion.