Scraping By? Europe's law and policy on social media research access

Leerssen, Paddy; Heldt, Amélie P.; Kettemann, Matthias C.

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Digital
Communication
Research.de

Recommended citation: Leerssen, P., Heldt, A., & Kettemann, M. C. (2023). Scraping by? Europe's law and policy on social media research access. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 405–425). Digital Communication Research. https://doi.org/10.48541/dcr.v12.24

Abstract: This chapter discusses the legal aspects of researchers' access to social media data, focusing in particular on recent developments in European law. We see law as playing both an enabling and a restrictive role in facilitating platform data access. Identifying a number of shortcomings in current legislation, we argue for the creation of a sound legal framework for scholarly data research. The new Digital Services Act makes some promising first steps towards regulating programmatic data access through APIs, but many obstacles and ambiguities remain. Furthermore, a clear vision on the legal status of public interest scraping projects is still lacking. In the teeth of private ordering by global platform companies, as new gatekeepers in academic research, ensuring fair and rights-sensitive data access must be a priority for the (European) legislator.

DOI 10.48541/dcr.v12.24

Paddy Leerssen, Amélie Heldt & Matthias C. Kettemann

Scraping By?

Europe's law and policy on social media research access

1 Introduction: Research access as a regulatory problem

Over the past decade, social media research has become a point of controversy in legal and regulatory discussions. In our burgeoning platform society (Van Dijck et al., 2018), access to social media data has grown to be increasingly essential for all sorts of social science research, including the analysis of hate speech. And yet as demand grows, platforms have generally restricted their research access policies over the past decade, rather than expand them. Without clear incentives for platforms to support public interest research, they have instead tended to give precedence to user privacy and data protection concerns. Such concerns may be warranted to some extent, but also risk being exaggerated and weaponized in service of platforms' more self-interested motives in avoiding independent scrutiny of their policies (Ausloos & Veale, 2020). As tensions with platforms escalate, researchers are increasingly turning to courts and legislatures to preserve their existing data access and to demand new, legally-binding access frameworks.

This chapter discusses the legal aspects of researchers' access to social media data, focusing in particular on recent developments in European law. It follows

Cohen's (2019) observation that the law plays both a restrictive and a facilitating role for platform data access: it includes information-blocking rules that constrain data access, as well as information-forcing rules that support it. Accordingly, we start this chapter by discussing the access barriers that researchers currently face and the role of the laws in constructing them, including aspects of contract, data protection, and intellectual property. Second, we review recent legal developments with an information-forcing component, which might offer pathways towards more effective and sustainable research methods. We discuss takedown reporting requirements, GDPR data access rights, as well as recent proposals to regulate platform APIs in the Digital Services Act and related plans to draft Codes of Conduct for platform researchers.

2 How we got here: The techno-legal precarity of platform data access

As communications researchers have pointed out, the problem of platform data access exacerbated rapidly after the Cambridge Analytica scandal (e.g., Bruns, 2019; Puschmann, 2019; Freelon, 2018). In response, several platforms severely restricted researcher access through APIs, in a development described by Bruns as the "APIcalypse" and leading to what Freelon termed the "post-API age." Some platforms responded more extremely than others: for instance, Instagram shut down its research API entirely while YouTube continues to allow relatively generous access (Munger & Philips, 2020). Twitter also recently expanded its accommodations for academic researchers, including a dedicated API and access to a full archive of tweets (while at the same time, however, introducing yet more restrictions on their standard API). Still, the current situation has resulted in a drastic reduction of data access opportunities for researchers. A related concern is that differences in data access between platforms can distort research agendas, by nudging researchers towards the most open and accessible platforms. A recent literature review of research on racism and hate speech on social media supports this, showing that Twitter is "far overrepresented, especially considering its relatively small user base" (Matamoroz-Fernández & Farkas, 2021, p. 215).

Researchers have responded in various ways to this new "post-API age." Some have tried to cooperate with platforms in self-regulatory arrangements (e.g., Puschmann, 2019; King & Persily, 2019; see also Jünger in this volume), some

have introduced method innovations (Münch et al., 2021), whereas others have started to rely on platform-independent data collection methods (e.g., Freelon, 2018) and others still have adopted a "data-activist" stance with the hope of lobbying governments to regulate a privacy-compliant re-opening of APIs (Bruns, 2019). The law, including but not limited to data protection, plays an important role in each of these developments.

Perhaps the most prominent effort at self-regulation in this space is Facebook's Social Science One, a partnership with US academics launched in early 2019 aiming to provide a secure and confidential access regime for researchers, who would be vetted through an independent application process (King & Persily, 2019). Unfortunately, the project was initially hamstrung by repeated delays and complications, which, according to Facebook, were the result of legal compliance concerns related to US privacy and EU data protection laws. However, many researchers did not take these claims at face value and criticized the project as an attempt to stave off binding regulation by governments with a (ultimately inadequate) promise of voluntary access (Bruns, 2019). In December 2019, the co-chairs and European advisory body issued a damning public letter expressing their frustration with the lack of progress, concluding that "we are mostly left in the dark, lacking appropriate data to assess potential risks and benefits" and expressly inviting public authorities to step in (The Co-Chairs and European Advisory Committee of Social Science One, 2019). Funders threatened to pull out of the project. This being said, the project has since then started to produce its first dataset—a database of URL information—as well as assisted in broadening and improving research access to tools such as the CrowdTangle and Ad Library APIs. However, the dataset has been criticized, due to the extensive use of "differential privacy" anonymization method that limit its accuracy and utility (mainly for qualitative research), and so have the API tools for various reasons. Access to CrowdTangle is only possible with Facebook's permission, raising questions about gatekeeping and academic freedom. Overall, then, the record is mixed at best, with some researchers more optimistic about this self-regulatory approach than others. Cornelius Puschmann, who was involved in the Social Science One project, noted: "Facebook improved access through [Social Science One] by a lot and has been very cooperative ever since" (Heldt et al., 2020; King & Persily, 2020).

Overall, self-regulatory projects such as Social Science One projects have thus moved the debate forward, but have not fundamentally reduced the impetus, at least in Europe, for more far-reaching, legally binding reforms.

Independent data collection methods have also taken flight in the "post-API age." With the help of sock-puppet accounts, crawlers or real-world volunteers using browser plugins, for example, researchers can observe platforms directly and assemble their own datasets. However, these methods face important limitations in terms of cost, sample size and bias, operating system restrictions, and so forth. Furthermore, platforms can take legal and technical actions to restrict these projects. Unauthorized data collection can potentially run afoul of many different laws, including anti-hacking laws, intellectual property, contractual restrictions in Terms of Service, and privacy and data protection laws. Indeed, researchers have reported on the complexity of data protection in this space, though compliance is certainly possible (Bodo et al., 2018). If brought to court, favorable rulings for researchers are entirely plausible or even likely based on public interest and fundamental rights defenses (see, for instance, the US ruling in *HiQ v LinkedIn*). The problem, however, is that platforms can often enforce their anti-scraping policies through extra-legal means, simply by blocking the relevant plugins or activities through technical measures and thus foreclosing the possibility for researchers to appeal to relevant constitutional defenses and public relevant interest exceptions (e.g., in data protection law). In these ways, law and technology work together to enable what Cohen (2019) terms the "de facto propertization" of platform data.

Some data scraping activities are tolerated by platforms, in what Rieder and Hofmann (2020) term "implicit acquiescence." Others are not so lucky. One notorious case involved New York University's Ad Observatory project, a collaboration between journalists and academics seeking to collect information about political advertising via a volunteer-installed browser plugin. Mere weeks before the US election, Facebook sent them a cease-and-desist letter, threatening to block the plugin if they did not comply (Horwitz, 2020). Facebook cited its Terms of Service as well as its obligations under US privacy law, which require the platform to prevent unauthorized access to user data. Critics objected that NYU's plugin only collects personal data from their volunteers, who have consented to share the data, and not from third party users, and furthermore that academic research is justified on public interest grounds (e.g., Doctorow, 2020). The broad

permissions that users usually (have to) give to extensions mean that they might be authorizing the collection of more sensitive data, even when that is not what the researchers end up collecting.

NYU has now joined forces with the Knight First Amendment Institute to challenge Facebook's actions in court, but it will likely take many years before legal certainty is obtained. More fundamentally, existing laws do not clearly explain whether or when researchers can go further than NYU's example and collect information *without* users' prior consent; something that may be particularly important in the context of hate speech, where speakers may be unlikely to volunteer their participation. Experts including the European Data Protection Board have pointed towards the many public interest exceptions in the GDPR that could possibly support research on other grounds than consent, but these questions remain clouded in uncertainty. Certainly, platforms cannot be relied on to make this determination by themselves, if only because they may lack the necessary information about the background of data scrapers. And waiting for such conflicts to make their way through the court could take decades.

It may be easy to criticize platforms for undermining public interest research, but it must be kept in mind that independent data collection also presents very real risks. The same methods used by researchers to collect data can be abused by commercial and political actors to the detriment of user privacy. In addition to the Cambridge Analytica scandal, mentioned previously, another chilling reminder is the mass scraping of facial image data by ClearView AI, used to develop (likely unlawful) facial recognition technologies. The largest social media platforms such as Facebook, Twitter, and YouTube accused ClearView AI of violating their policies. In this light, the problem is not so much that platforms restrict independent data collection, but rather that these policies are enforced across the board without an adequate public interest exception. Vetting public interest researchers, however, is a task that platforms are ill-positioned to perform, both operationally and politically. It would be a clear threat to academic freedom if platforms were responsible for deciding which researchers were permitted to study them.

These incidents underscore the fundamental precarity of developing research methods and tools for platform services. Whether relying on self-regulatory data-sharing arrangements, independent plugins, or tools built on platform APIs, researchers operate at the pleasure of platforms who maintain at all times the technical and legal power to alter, restrict or shut down entirely their access—and

who may do so at the slightest threat of legal or political risk. According to Rieder and Hofmann (2020), this techno-legal precarity requires an institutional response, focused on creating more dependable modes of access:

> A common characteristic of the data collecting projects mentioned above is their ephemeral, experimental, and somewhat amateurish nature. While this may sound harsh, it should be obvious that holding platforms to account requires 'institution-building,' that is, the painstaking assembly of skills and competence in a form that transposes local experiments into more robust practices able to guarantee continuity and accumulation. (p. 23)

This institution-building, according to Rieder and Hofmann (2020), would need to be paired with regulatory measures aimed at enhancing the "observability" of platform, for instance by regulating platform APIs: "The main goal, here, is to develop existing approaches further and to make them more stable, transparent, and predictable" (p. 22). Such demands bring us to recent debates in European law, where governments have increasingly sought to impose information-forcing rules on platforms. These rules may help to create the conditions for more robust and dependable data access frameworks and institutions to develop, although, as will be discussed below, these are early days still.

3 Regulating research access: Recent developments in European law

This section provides an overview of legislative and regulatory initiatives that enable access to platform data for research purposes. As will be shown, current efforts are both disparate and initial. With few exceptions, it concerns drafts and proposals rather than in-force measures. We start with one of the most widespread types of transparency regulation, content moderation reporting, followed by discussions of GDPR data access rights, the API-related rules from the Digital Services Act proposal, and the European Digital Media Observatory's proposal for a Code of Conduct.

3.1 Mandatory content moderation reporting (in the NetzDG and elsewhere)

One of the most common modes of data access regulation is the so-called "Transparency Report": the periodical, public reporting of aggregate data about content moderation actions. This practice originates in self-regulation, where it has long served as a rallying point for civil society initiatives such as Ranking Digital Rights and the Manila Principles. Over the past decade, platforms have gradually begun to concede to these demands and release transparency reports, which have gradually grown in scope and detail (Keller & Leerssen, 2020). In recent years, governments in Europe and elsewhere have sought to regulate transparency reporting practices.

Transparency reporting obligations can be found in numerous laws and proposals. The majority focus on moderation related to hate speech and related topics, including Germany's Netzwerkdurchsetzungsgesetz (NetzDG), France's Loi Avia, Austria's Communications Platform Law (Fischer et al., 2020), and the EU's proposed Terrorist Content Regulation. The EU's recent Digital Services Act proposal also includes expansive transparency reporting rules, with escalating levels of disclosure applied based on the size of the platform service.

Most of these instruments have not yet passed into law and/or entered into force, with the exception of the NetzDG. In force since January 1, 2018, the NetzDG offers insights into the practical impact and utility of transparency reporting regulation. Thus far, eight different platforms have released one or more semi-annual reports under this framework: Twitter, Reddit, Facebook, TikTok, Change.org, Jodel, Google/YouTube, and Soundcloud.

Overall, the response from researchers to this data has been muted at best and dismissive at worst (Suzor et al., 2019). Researchers' critiques of NetzDG transparency reporting are several (Heldt, 2019). Most fundamental, however, is the criticism that aggregate data offered by transparency reports leaves researchers without content-level insights into particular cases. As a result, researchers are unable to independently assess platforms' content classifications, and thus to determine the quality of content moderation decisions and its impacts on various groups. For instance, the fact that Google has removed x pieces of content due to hate speech between June and September 2020 does not tell us whether this content concerned, for instance, white supremacy, radical Islam, or some other variant of hate speech; whether it targeted its victims based on gender, race, or

some other protected category; whether the removed content was classified correctly (i.e., false positives); how much non-removed content was reviewed but ultimately left up (i.e., false negatives); and so forth. All of these questions require, at a minimum, access to the actual content at issue (Keller & Leerssen, 2020) and to the practices in use when enforcing content standards against hate speech.

A related criticism is that content removal reporting cannot be assessed meaningfully without robust indicators of the overall prevalence of this content across the platform. For instance, Facebook might report a bi-annual increase in hate speech removals of 15 percent, suggesting an improved detection rate. Even assuming that Facebook's classifications are correct (which we cannot, as discussed above), the opposite could still be true if overall prevalence of hate speech posts simultaneously increased by over 15 percent. In a bid to address these concerns, Facebook has since November 2020 become the first platform to publish prevalence estimates regarding hate speech (Kantor, 2020), though robust comparisons over time are difficult to make since comparable data is lacking and the special situation of an increase in automation in content governance during the COVID-19 crisis caused changes in platform moderation practices (see also Ahmad in this collection).

Another problem is that Facebook undermined the functioning of NetzDG by making their complaint mechanism difficult to access. This has had the effect of discouraging users from submitting complaint, such that Facebook received significantly fewer complaints relative to its size. Since the NetzDG transparency obligations only cover formal notices submitted within its framework, this reporting can paint a distorted picture by omitting content moderation practices initiated under platforms' self-regulatory flagging systems. Facebook in particular was removing significantly more content based on these self-regulatory systems than under the official framework, but the same problem also applies to other platforms and their self-regulatory flagging mechanisms. German authorities have fined Facebook for its practices, and recently proposed amendments to the NetzDG would require platforms to make their NetzDG complaint mechanisms easily accessible. More fundamentally, the problem remains that most takedown reporting rules may fail to capture the totality of moderation actions undertaken by the platform.

Of course, transparency reports have some (limited) utility in tracking trends in content moderation over time. For instance, NetzDG transparency reports

give a high-level view on how much data is removed, which removal grounds are triggered most frequently, and so forth. Indeed, Facebook's transparency reports under NetzDG provided empirical support for the critique that their implementation of this law discouraged users from submitting complaints, by showing that they received substantially fewer than Twitter and Google.

As of May 2020, the German government is amending the NetzDG. The legislator has acknowledged the need for researchers to access data in order to better understand platform practices, but unfortunately this finding was not put into practice. The legislator could have added an access to data provision for research purposes, but the amended version of § 2 (2) NetzDG only stipulates an obligation to report on whether and to what extent relevant insights were granted to members of the scientific and research community. It does not specify *how* researchers will get these "relevant insights" or impose any obligation on platforms to provide them.

Another proposed amendment is to add a new section to § 2 (2), which requires platforms to disclose the use of automation for content moderation purposes, regardless of whether the content was removed because it was considered unlawful or because of a violation of the platform's own content rules. This information could be valuable to further understand how hate speech is detected by platforms, although the information provided here is likely to remain of a rather general nature.

3.2 Copyright

In general, copyright law is rather perceived as an obstacle in the overall attempt to gather third-party data—even for research purposes. But new reforms are underway to relieve some of these constraints. Researchers might infringe the platforms' rights when collecting policies and documents. Recently, legislators have recognized the need to re-adapt to the new possibilities for research and innovation via digital technologies. In 2017, Germany passed a provision for text and data mining in order to bring copyright law in line with the needs of the information society. Under § 60d (1) German Copyright Act, one may collect and duplicate automatically and systematically data in order to create a corpus for research purposes. Similarly, Article 3 of the Digital Single Market Directive makes it mandatory for Member States to provide for an exception allowing text

and data mining "for the purposes of scientific research." The provision does not, however, provide access to data in itself. Instead, the scope of application is restricted to works to which researchers have "lawful access." In Germany, for instance, scraping might infringe the platforms' exclusive rights to reproduce, distribute and publicly reproduce under Section 87b (1) German Copyright Act when third-parties repeatedly and systematically reproduce the "database." However, this restriction will, generally, not affect researchers because of the non-commercial nature of their action.

3.3 GDPR data subject rights

The GDPR does not only block data access; it can also force data access by virtue of its transparency provisions. The GDPR offers a number of data access rights regarding personal data held by the platforms, including the right of access, the right to data portability, and the right to an explanation regarding automated decision making. These rights are granted to data subjects, rather than researchers per se, but Ausloos and Veale (2020) demonstrate that they can nonetheless be repurposed as research tools by enlisting data subjects as volunteers. Their work explores some of the ethical considerations involved and outline a number of use-cases, including research into content moderation, online tracking, the use of biosensors, and digital labor issues. They do not address hate speech research in particular beyond the general issue of content moderation, and further exploration of use-cases in this space would likely be fruitful.

In theory, other user-facing rights could potentially also be retooled for research purposes. For instance, in the context of self-regulation, researchers have crowdsourced the explanations that Facebook offers their users regarding their microtargeted advertisements under their "Why Am I Seeing This" feature, in order to gain insights into targeting practices (WhoTargetsMe, 2020). Rules and proposals for user-facing information rights abound under European law, including the rules on recommender systems in Article 30 of the Digital Services Act. For the most part, however, these rules focus on easy-to-digest, broadly understandable explanations for a general audience, which may only offer marginal benefits to specialized researchers (Leerssen, 2020).

3.4 Digital Services Act: Data access for "vetted researchers"

Perhaps the most significant data access proposal for hate speech research access regulation is Article 31 of the EU's newly-proposed Digital Services Act. Titled "Data access and scrutiny," this article authorizes local platform regulators, so-called "Digital Services Coordinators" (DSC), to compel platforms above a certain size to disclose relevant data to "vetted researchers." The DSA has not yet been finalized. Our discussion focuses on the text of European Commission's original proposal of 15 December 2020.

Many of the details of this article will likely change, since this concerns a first draft proposal with a long and controversial legislative process ahead of it. As of mid-2021, however, the scope of Article 31 is relatively restrictive in terms of its subject matter as well as eligible researchers. In terms of its subject, Article 31 only applies to research conducted for purposes of risk assessments related to the platform service, including but not limited to the following: (a) the dissemination of illegal content, (b) effects on fundamental rights including privacy and freedom of expression, and the rights of the child, and (c) inauthentic usage of the service, "with an actual or foreseeable negative effect on the protection of public health, minors, civic discourse, or actual or foreseeable effects related to electoral processes and public security" (Articles 31 and 26(1)). This scope clearly enables research into hate speech, but may cut off other fields of inquiry.

For researchers to qualify as "vetted," they must be "affiliated with academic institutions, be independent from commercial interests, have proven records of expertise in the fields related to the risks investigated or related research methodologies, and shall commit and be in a capacity to preserve the specific data security and confidentiality requirements corresponding to each request." The restriction to academic institutions risks excluding NGOs and other third parties, unless they partner with vetted academics with a view to gaining access. To comply with the requirement of data security, researchers will likely be required to produce a data management plan demonstrating, at a minimum, GDPR compliance and perhaps the observance of other ethical or scientific standards. At present, the details of these rules remain unspecified, but the European Commission is tasked with developing guidance to ensure compliance with the GDPR. An interesting question is how this standard-setting activity will interact with other

delegated rulemaking and standard-setting ongoing in this space, including the Research Code of Conduct in production at EDMO discussed below.

Article 31 also contains ambitious but as of yet unspecified rules about disclosure formats: subparagraph 3 requires that platforms "shall provide access to data [...] through online databases or application programming interfaces, as appropriate." This clause seems to respond to ongoing debates about the governance of research APIs outlined above. Yet, it leaves many questions open as to how and when APIs or databases would be "appropriate"—again, matters for further standard setting by regulators. The provision does signal, however, that the DSA proposal envisages broadly accessible forms of data-sharing and not merely singular data grants to individual research groups; in some cases, "where appropriate," authorities might demand that data is made available programmatically to a broader pool of researchers. It could arguably provide the basis for regulators to expand and improve existing self-regulatory efforts, such as Facebook's CrowdTangle and Twitter's academic research API, and enable monitoring and scrutiny by larger sets of (vetted) researchers in real-time. The current limitations on 'vetted researchers' could however pose an obstacle to creating truly inclusive resources.

Another blind spot is Article 31 (6) DSA: according to the current proposal, platforms shall have a right to request an exemption whenever they do not have access to data. Because platforms are supposed to act against illegal content under Article 14 and 15 DSA, it might not be available for later research. That is a problem raised by journalists and prosecutors investigating war crimes: once the platforms remove the content, it is almost impossible to retrieve it (or highly dependent on the platforms' goodwill). If this is not policed properly, important material for the study of hate speech and other illegal phenomena, as well as the gatekeeping function of platforms, might be destroyed. This same data may also be an important ingredient in the training of AI tools for the detection of hateful content.

Notably absent from Article 31 is a procedure for researchers to petition either platforms or regulators for access. In the current draft, it seems, access depends on the initiative of the regulator. There is a risk here that researcher access becomes subservient to the goals and aims of regulatory investigations, instead of setting its own scientific agenda. To preserve academic freedom in this regime, regulators would ideally devise independent and objective procedures to vet and prioritize researchers and their projects.

3.5 Digital Services Act: Mandatory ad archive APIs

The DSA proposal also contains specific data access rules related to online advertising in Article 30. Microtargeted online advertising has been the subject of many controversies and policy concerns, including the dissemination of hate speech through channels that are difficult for third parties to trace or respond to (e.g., Wong, 2020). Here too, platforms above a certain size are required to provide some programmatic access to relevant researchers via an API. The requirements here are significantly more detailed than the generic data access framework of Article 31 outlined above.

This rule builds on existing self- and co-regulatory practices, currently known as "ad archives" or "ad libraries," which have already been implemented in some form by most major advertising platforms and are increasingly subject to regulatory requirements in Europe and elsewhere (Leerssen et al., 2019). Ad archives may be valuable for hate speech research because they allow researchers to trace the use of hate speech (and other speech) within ad ecosystems and their interaction with non-ad content.

The DSA largely mirrors these existing practices in requiring that the following information is made available: the content of the ad, the name of the ad buyer, the advertising period, the total number of views, and demographic information about the audience reached. Existing self-regulatory practices for advertising continue to exhibit many errors and shortcomings (Leerssen et al., 2019), and these new binding rules may provide an impetus for platforms to invest in more rigorous implementations.

We also see remarkable differences compared to self-regulatory standards. The most significant change by far is that the DSA's rule applies to *all* advertisements sold on the service, whereas platform projects have been far more narrowly targeted to (varying definitions) of political campaign and issue ads. This broader approach covering all ads has been endorsed by many researchers and activists, who objected that platforms failed to reliably define and detect political ads—thus creating sampling problems and undermining the research utility of their data—and that non-political, commercial ads also deserve scrutiny. The new approach leaves it to researchers themselves to define and operationalize their own interest categories.

The metadata about advertisements required by the DSA proposal also differs on two points. First, the DSA is more expansive in that it also requires that platforms disclose their *targeting criteria* for each ad: "whether the advertisement was intended to be displayed specifically to one or more particular groups of recipients of the service and if so, the main parameters used for that purpose." Again, this change responds to widespread criticism from researchers about the lack of such data in the existing databases (Leerssen et al., 2019). Platforms have objected that disclosing targeting criteria may run afoul of user privacy, which may indeed place limits on the documentation of Facebook's custom audience targeting methods, but is not evidently compelling for other aspects of targeting. Furthermore, the requesting of "main parameters" suggest that platforms will not have to be exhaustive in their documentation. Thus, the further interpretation and implementation of this rule remains subject to debate. In January 2020, only one month after the DSA was proposed, Facebook did announce that it would be making targeting data available on a limited basis to academic researchers in the US, related to the US elections. We cannot assess at this time what the value of these disclosures will be, but the lessons learned here will certainly be instructive for the future of Article 30 DSA.

Second, the DSA also takes a large step *backwards* by omitting advertisement spending data. Spending data has been standard inclusion in all self-regulatory ad libraries (albeit in general ranges rather than precise amounts), and it remains unclear why it has been omitted here.

As noted, Article 30 DSA requires large platforms to disclose their ad archive data through public APIs, enabling programmatic access by researchers as well as other third parties. It should be noted here that Facebook's existing Ad Library API has been criticized extensively by researchers, due to inconsistency, performance issues and bugs, and a lack of user-friendliness (Mozilla, 2019; Rosenberg, 2019). This is another failure mode for ad archive regulation, which might require further regulatory standard-setting to address. An alternative approach would be to demand that platforms disclose their data to an independent third party, which would be entrusted with designing and operating an effective researcher API. For instance, the EU's Data Governance Act Proposal provides "Data Altruism Organisations" (chapter IV) that would "lead to the establishment of data repositories" and "facilitate cross-border data use" (Nr. 36 of the DGA's explanatory memorandum). Such registered third-parties would be subject to strict transparency obligations

and specific requirements under Article 19, making them trusted intermediaries for a general interest data access.

3.6 The EDMO Code of Conduct

A final development worth noting is the push to develop a Code of Conduct for researchers handling platform data, spearheaded by the Commission-funded European Digital Media Observatory (EDMO). This procedure is based on Article 40 of the General Data Protection Regulation, which allows stakeholders involved in the processing of personal data to design voluntary codes specifying GDPR compliance methods in their particular field of activity. These Codes can then be approved by Data Protection Authorities (DPAs) in order to create legal certainty about the requirements of data protection law (which can otherwise be highly general and ambiguous). EDMO's mandate is largely centered on combating disinformation, but they have already announced that their Code of Conduct initiative is not intended to be limited to this subject matter. It therefore bears relevance to other fields of social media research, including the analysis of hate speech.

EDMO's proposal, like most discussed here, is at an early stage: their Article 40 Working Group was officially announced in November 2020, with an official call for comments soliciting input from relevant stakeholders. The Working Group now has the task of processing these comments and further specifying their approach.

Since Article 40 GDPR merely serves to clarify existing law, it cannot create new obligations on platforms to share data with researchers or other third parties, beyond what they voluntary commit to when signing up for the Code of Conduct.

One role the Code could play is to clarify how data protection law should apply in new data-sharing arrangements such as the DSA access frameworks outlined above. For instance, the European Commission could draw on an academic Code of Conduct in assessing who qualifies as a "vetted researcher," and evaluating their data management plans under Article 31 DSA's data access framework. A related role that a Code of Conduct might play is clarifying when and how independent data collection efforts comply with the GDPR—a matter which continues to raise legal uncertainty for researchers and platforms alike. By creating a procedure to certify the GDPR compliance of independent data collection projects, the Code could help to operationalize public interest exceptions without forcing platforms

to act, as Mathias Vermeulen puts it, "as de facto gatekeepers who decide on the validity of specific research proposals and methods" (Vermeulen, 2020, p. 21). Such an institutionalized, vetted approach has the advantage of greater accountability for both platforms and data recipients, although an overly bureaucratic access procedure could discourage buy-in from researchers and may risks privileging certain forms of research over others.

Similarly, Article 35 DSA proposes "the drawing up of codes of conduct at Union level to contribute to the proper application of this Regulation, taking into account in particular the specific challenges of tackling different types of illegal content and systemic risks, in accordance with Union law, in particular on competition and the protection of personal data." Finally, these Codes might also be a venue for platforms, in light of the mounting public pressure, to make certain data access commitments, including proactive data-sharing with compliant researchers as well as non-interference with compliant data scraping projects. Overall, a key question remains the interaction between the GDPR and DSA codes of conduct in this space; whether EDMO will choose to focus on supporting and facilitating the DSA's (future) access rules, or rather to create an independent, GDPR-based framework of its own.

4 Outlook: First steps taken, long read ahead for responsive API regulation

Clearly, these are heady times for the regulation of research access. Quite suddenly it has become a hot topic for lawyers and policymakers—hot, if not overheated. The result has been a spate of different proposals and initiatives, some more promising than others. Many of these plans are still at an extremely early stage, and may still take years to come to fruition. But experience shows that the early stages of drafting are often pivotal, since it is then that concepts, frames, ideas can become anchored in legislative minds and texts. All the more important, therefore, for communications researchers and other social scientists to involve themselves in these discussions and demand rulemaking that actually responds to their research needs.

If European policymakers were to listen more closely to the research community, they might for instance realize that their recurrent emphasis on aggregate

takedown reporting rules, without insights into the underlying content, may be somewhat misplaced. Such rules continue to proliferate in various instruments, despite offering a rather minimal benefit to the scientific understanding of the topics they regulate, including hate speech. At the same time, policymakers still lack a clear policy vision on what many researchers find most urgent: tools to study the actual spread of harmful content, and the substance of what is ultimately being flagged and removed. Or indeed, on academic research unconstrained by governments' particular interests or agendas. A clear stance on the status of independent scraping projects has also not emerged yet, and efforts to regulate APIs are still in their infancy. National laws fail to protect researchers against overbroad Terms of Service that jeopardize good-faith research efforts, despite the significant public interests often implicated in this activity. Collecting the pictures of the January 6, 2021, attacks on the Capitol through scraping the social media app Parler, for instance, has been an invaluable source for public interest-based reporting.

While the DSA is still in the making, it is encouraging to see that it contains a clear statement in favor of mandatory procedures for researcher data access, including the regulation of automated disclosure via public databases and APIs. Also promising are the DSA's rules on Ad Archive APIs, the Commission's backing of a GDPR Code of Conduct, and new experimentation with data subject rights as a tool for researchers.

Whilst these efforts appear well-intentioned, the devil remains in the details. Regulating the design of APIs in particular is a complex and relatively unprecedented issue, raising questions as to whether governments will be up to the task. To ensure that researchers' access to user data via APIs is GDPR-compliant, compliance-by-design solutions could be explored. One possibility is pseudonymized/anonymized data outputs, which could eliminate the need for substantial vetting procedures for certain APIs. Recent developments in self-regulation, such as Facebook's attempts at differential privacy, seem to point in this direction. Approaches that allow access to more sensitive data would likely require more extensive vetting procedures, at the possible cost of scalability and uptake amongst researchers. Generally speaking, reproducibility and reliability of the data produced remains a concern.

Perhaps the most feasible approach, at least in the short term, might be to develop certification schemes or safe harbors to protect independent scraping

efforts from restrictive platform policies; this issue is not currently addressed in any relevant legislation, but the EDMO Code of Conduct and other GDPR standard setting could already be an important first step towards creating greater certainty in this space, so that ethical and privacy-conscious research, in compliance with researchers' special duties of care, is not restricted unnecessarily. There is no doubt that privacy and academic research can be reconciled, but particularly in sensitive areas such as hate speech, safeguard procedures are crucial to prevent abuse and preserve the rights of users and victims. Just like ethical tests for medical trials or trials involving humans, data use audits might have to precede large-scale API uses by scientists.

In the longer term, however, there is no way around establishing a clear and sound legal framework for scholarly data access; independent scraping is not enough, and there is a clear need—and political will—to also regulate API access and data grants. The more social interaction happens in the digital sphere, subject to the private ordering of global platform conglomerates, the more should legislators protect the lawful access to research data.

Paddy Leerssen is a PhD candidate in Information Law at the University of Amsterdam, Netherlands, and a non-resident fellow at the Stanford Center for Internet and Society, USA.

Amélie Heldt is a researcher at the Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg, and associated with the Humboldt Institute for Internet and Society, Berlin, Germany. https://orcid.org/0000-0002-1910-9925

Matthias C. Kettemann is senior researcher at the Leibniz Institute for Media Research | Hans-Bredow-Institut, Hamburg, Germany. He is research group leader at the Humboldt Institute for Internet and Society, Berlin, Germany, and the Sustainable Computing Lab at the Vienna University of Economics and Business, Austria. https://orcid.org/0000-0003-1884-6218

References

Ausloos, J., & Veale, M. (2020). Researching with data rights. *Technology and Regulation,* 136-157. https://doi.org/10.26116/techreg.2020.010

Bodo, B., Helberger, N., Irion, K., Zuiderveen Borgesius, K., Moller, J., ..., & de Vreese, C. (2018). Tackling the algorithmic control crisis: The technical, legal, and ethical challenges of research into algorithmic agents. *Yale Journal of Law & Technology, 19*(1), 133–180.

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

Cohen, J. (2019). *Between truth and power: The legal constructions of informational capitalism.* Oxford University Press.

Doctorow, C. (2020, November 20). Facebook is going after its critics in the name of privacy. *Wired.* https://www.wired.com/story/facebook-is-going-after-its-critics-in-the-name-of-privacy/

Fischer, G., Ketteman, M. C., & Rachinger, F. (2020). Così fan tutte: Some comments on Austria's draft communications platforms act (Graz Law Working Paper No 05-2020). https://doi.org/10.2139/ssrn.3731593

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

Heldt, A. (2019). Reading between the lines and the numbers: an analysis of the first NetzDG reports. *Internet Policy Review, 8*(2). https://doi.org/10.14763/2019.2.1398

Heldt, A., Kettemann, M., & Leerssen, P. (2020, November 30). The sorrows of scraping for science: Why platforms struggle with ensuring data access for academics. *Verfassungsblog*. https://verfassungsblog.de/the-sorrows-of-scraping-for-science/

Horwitz, J. (2020, October 23). Facebook seeks shutdown of NYU research project into political ad targeting. *Wall Street Journal*. https://www.wsj.com/articles/facebook-seeks-shutdown-of-nyu-research-project-into-political-ad-targeting-11603488533

Kantor, A. (2020, November 19). Measuring our progress combating hate speech. *Facebook.* https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech/

Keller, D., & Leerssen P. (2020). Facts and where to find them: Empirical research on Internet platforms and content moderation. In N. Persily & J. Tucker (Eds.), *Social media and democracy: The state of the field and prospects for reform* (pp. 220–251). Cambridge University Press.Formularende

King, G., & Persily, N. (2019). A new model for industry-academic partnerships. *PS: Political Science and Politics, 53*(4), 703–709. https://doi.org/10.1017/S1049096519001021

King, G., & Persily, N. (2020, February 13). Unprecedented Facebook URLs dataset now available for academic research through Social Science One. *Social Science One*. https://socialscience.one/blog/unprecedented-facebook-urls-dataset-now-available-research-through-social-science-one

Leerssen, P. (2020). The soap box as a black box: Regulating transparency in social media recommender systems. *European Journal of Law and Technology, 11*(2). https://doi.org/10.2139/ssrn.3544009

Leerssen, P., Ausloos, J., Zarouali, B., Helberger, N., & de Vreese, C. (2019). Platform ad archives: Promises and pitfalls. *Internet Policy Review, 8*(4). https://doi.org/10.14763/2019.4.1421

Matamoros-Fernández, A., & Farkas, J. (2021). Racism, hate speech, and social media: A systematic review and critique. *Television & New Media, 22*(2), 205–224. https://doi.org/10.1177/1527476420982230

Mozilla (2019, March 28). Facebook and Google: This is what an effective ad archive API looks like. *The Mozilla Blog*. https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like

Munger, K., & Phillips, J. (2020). Right-wing YouTube: A supply and demand perspective. *The International Journal of Press/Politics*. Advanced online publication. https://doi.org/10.1177/1940161220964767

Münch, F. V., Thies, B., Puschmann, C., & Bruns, A. (2021). Walking through Twitter: Sampling a language-based follow network of influential Twitter accounts. *Social Media + Society, 7*(1). https://doi.org/10.1177/2056305120984475

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society, 22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

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

Rosenberg, M. (2019, July 25). Ad tool Facebook built to fight disinformation doesn't work as advertised. *The New York Times.* https://www.nytimes.com/2019/07/25/technology/facebook-ad-library.html

Suzor, N. P., Myers West, S., Quodling, A., & York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *International Journal of Communication*, 13, 1526–1543.

The Co-Chairs and European Advisory Committee of Social Science One (2019, December 11). Public statement from the Co-Chairs and European Advisory Committee of Social Science One. *Social Science One*. https://socialscience.one/blog/public-statement-european-advisory-committee-social-science-one

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

Vermeulen, M. (2020). The keys to the kingdom. Overcoming GDPR-concerns to unlock access to platform data for independent researchers. Draft paper. https://doi.org/10.31219/osf.io/vnswz

WhoTargetsMe (2020). Our research. https://whotargets.me/en/our-research/

Wong, J. C. (2020, January 29). One year inside Trump's monumental Facebook campaign. *The Guardian*. https://www.theguardian.com/us-news/2020/jan/28/donald-trump-facebook-ad-campaign-2020-election?CMP=Share_iOSApp_Other