# The Sorrows of Scraping for Science

Amélie Heldt 2020-11-30T11:29:30

It was at least a partial answer to the demands of many scholars, when Guy Rosen, Facebook's VP for Integrity, announced that Facebook's latest Community Standards Enforcement Report would come with more data and better access: "For more details", Rosen wrote, "the full data set is here. We're now also including a CSV export of the data, to help researchers who want to run their own analyses."

Facebook giveth, but Facebook also taketh away. They often claim to champion transparency, but this stance is somewhat at odds with a Wall Street Journal story of three weeks earlier which recounted efforts by Facebook to shut down NYU's Ad Observatory. Engineering students had developed a browser extension that more than 6,000 volunteers were using to collect data on the ads they were shown on Facebook. According to the article, the letter that Facebook's lawyers had sent the researchers said: "[s]craping tools, no matter how well-intentioned, are not a permissible means of collecting information from [Facebook]."

The opinion is still divided on the merits of Facebook's action. On the one hand, platform expert Ben Thompson wrote, Facebook may actually have been obliged to stop third parties, including universities, from scraping under an arrangement made with the Federal Trade Commission after the Cambridge Analytica scandal. The platform, it seemed, was caught between a rock and a hard place: allow browser extensions and perhaps risk liability or fight the researchers and receive the wrath of academia in a situation substantially similar to that of Cambridge Analytica. Other experts have disputed Facebook's position, however, arguing that their claims misrepresent the Ad Observatory: Cory Doctorow, for instance, counters that NYU's project only collects data from volunteers, and *not*, contrary to Facebook's allegations, from other users. However, the broad rights users usually (have to) give to extensions mean that at least some data from friends is (usually) also collected, even though plug-ins "for science" can be designed in such a way that the research team does not end up receiving any non-pseudonymized personal data. The user may be granting broader permissions that theoretically allow for broader collection, but that is not what the researchers end up receiving.

That some scraping ("for science", with safeguards) is important and some scraping (for monetary gain, without the appropriate safeguards) is problematic has led to some of the biggest critics of Cambridge Analytica's scraping to now voice their opposition to Facebook's attempt to stop the scraping of data by the NYU team: "This is so outrageous", journalist Carole Cadwalladr, who had been the one to break the Cambridge Analytica story, tweeted at her almost 600,000 followers. The Knight First Amendment Institute, also housed at NYU, also complained and recalled that researchers had been demanding that Facebook amend its terms of service to establish a "safe harbour" for public-interest research and journalism as early as 2018.

# Issues in data access

While the importance of open data is not doubted by policymakers, and "demos-scraping" has its uses, providing researchers with access to data is not so straightforward.

Access to data is an overall challenge for researchers when investigating platforms and their content moderation policies and practices. Researchers need empirical evidence to ground their arguments. If platforms do not provide data, and are in fact further restricting access, researchers may feel that their only option is to get the data themselves. One method is scraping data from the platforms, both of corporate nature and from the users themselves, presupposing (if researchers are ethical) their consent. For instance, researchers can collect publicly available information about YouTube videos or Facebook posts, or they can recruit volunteers to share information about their personal newsfeeds and recommendations on these platforms. on Facebook or Instagram. These practices, which are increasingly central to public interest research, raise questions of conformity with Facebooks's terms of service, of legality in terms of copyright and data protection (as explained by S. Golla and D. Müller).

The question of who uses scraping, and toward which end, highlights the dilemma platforms find themselves in. Take the recent example of Clearview AI. The AI company scraped three billion pictures of social media users from all over the world, without their or the platforms' consent, for its facial recognition solution, which it offers (inter alia) to public prosecutors. This model has been much criticized, not least because of the commercial exploitation of personal data toward ethically dubious ends. Among the critics of Clearview AI is the Knight First Amendment Institute, who – *prima facie* – have criticized Facebook for stopping the scraping by its own researchers. Though the Knight First Amendment team would argue that its research is ethically sound, participation is voluntary, and the code used to do the scraping is open for inspection – superficially, the technique is the same. The tension thus remains: how can we preserve channels for public interest research whilst preventing large-scale privacy infringements? One possible answer, as Cornelius Puschmann, media researcher at the University of Bremen, notes is providing a company-regulated (API-based) avenue specifically for researchers who commit themselves to upholding legal and ethical standards.

# Copyright exceptions for research

Regarding the scraping of corporate data such as policies, reports, technical information, or architecture choices, researchers encounter the (legal) problematic of possibly infringing the platforms' copyright as authors of these documents (in the broader term). As to user-generated content, scraping could lead to an infringement of the users' rights as authors and creators of original content (to the extent that a post on social media is considered protected). As a subsequent right, the copyright of platforms as database owners could also be violated. In both cases, one could refer to exemptions such as §60c German Act on Copyright and Related Rights

(UrhG), which allows the replication of up to 75% of a protected work for own research purposes. However, the critical question for researchers is whether they may store the data in a repository and share it with others for research purposes. Under §60d Nr. 2 UrhG, researchers are allowed to make the data sets prepared for data analysis (so-called corpus) publicly accessible to a defined group of people for joint research, but only if the group of people is clearly definable. Although most of this data is actually publicly available, collecting and allowing other researchers to use it still is a complicated endeavour, not least because of an 'unethical' secondary use of the data (as in Cambridge Analytica).

## Self-regulating third-party access

As one of the authors (Leerssen) explains elsewhere in more detail, Facebook has proposed an alternative to academic data scraping in the form of Social Science One, their self-regulatory data access framework. Social Science One is a partnership with US academics, launched in early 2019 with much fanfare and promised to provide a secure and confidential access regime for researchers, who would be vetted through an independent application process. 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, these claims have been called into question, as discussed further below, and many researchers did not take these claims at face value. In December 2019, the 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. Funders threatened to pull out of the project. What little data they have released has been criticized, since the extensive use of 'differential privacy' anonymization method has undermined its accuracy and utility (mainly for qualitative research).

But the picture has nuances: As 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".

On the specific topic of political advertising, which is at issue in the NYU dispute, Facebook points to its self-regulatory Ad Library as a sanctioned alternative to data scraping. But this tool has also disappointed some researchers and was criticized by academics, journalists, and regulators for its restrictions, inconsistencies, and omissions. Indeed, it was through scraping methods such as those applied by NYU that researchers were able to discover many of these flaws. Relying purely on self-disclosure seems challenging.

The Social Science One and Ad Library projects have thus moved the debate forward, but have not fundamentally reduced the impetus, at least in Europe, to add legally binding access rules.

# Regulating research access?

Research access is becoming an important theme in EU digital policy, including in the pivotal Digital Services Act. Commissioner Vestager announced in a recent speech on the Digital Services Act that "researchers, too, need to have access to data that allows us to understand how those algorithms are affecting our society […] And since those choices affect us all, that data can't be a sort of esoteric knowledge, that only a small priesthood who work for these big platforms gets to see." She even name-checked ad archives as a particular target for regulation.

A parallel development is the European Digital Media Observatory: a new centre for expertise, bringing together academics and fact-checkers on disinformation. This group has until now not been vested with any legally binding data access rights, but it has launched an initiative to develop a Code of Practice for legally compliant research access under Article 40 GDPR, starting with a Call for Comment that runs until 24 December 2020.

Regulating research access is a complex task that raises many thorny issues. Thorny, but not necessarily new. A recent AlgorithmWatch report, co-authored by one of the authors (Leerssen), shows that precedents from data access regimes in other sectors, such as medicine, can act as blueprints for secure and reliable data access – even when sensitive personal data is involved. An essential precondition is binding regulation and oversight, which is necessary to overcome conflicting incentives from regulated entities; to preserve the independence of data recipients; and to clarify the relationship to other applicable areas of law such as IP and privacy. For good reason, one can hardly imagine a self-regulated medical sector.

# Compliance, data protection and research ethics

One important issue to be addressed in data access regulation is GDPR compliance. Although platforms have often cited this as an obstacle, they have incentives to exaggerate these claims. The European Data Protection Board has rebuked platforms' objections: "It would appear therefore that the reluctance to give access to genuine researchers is motivated not so much by data protection concerns as by the absence of business incentive to invest effort in disclosing or being transparent about the volume and nature of data they control." Mathias Vermeulen treats GDPR compliance primarily as a problem of legal *uncertainty*. He argues that EU policy should clarify the current law as it relates to platform-researcher data access. One approach would be a Code of Conduct for researchers pursuant to Article 40 of the GDPR, as is currently being developed under the auspices of EDMO, which would grant researchers the right to handle platform data rights under the condition of strict research ethics principles.

With such a Code in place, the stage would be set for the regulation of data access. What shape these rules take continues to be up for debate, with more questions than answers. The foremost issue is, what data must be disclosed? And, relatedly: By which platforms? To whom and for what purposes? Commissioner Vestager

already singled out recommender systems and advertising as two important topics, but this still leaves unanswered (a) what information in particular must be disclosed on these topics, and (b), more fundamentally, whether the topics of transparency are to be defined exhaustively ex ante, or rather flexibly and ex post. Who qualifies as an eligible “researcher” is also a crucial matter: many debates tend to focus on university-affiliated academics, but broader segments of civil society, such as activism and journalism, also perform important public interest research. A spectrum of different access tiers can be envisaged, from highly selective frameworks more akin to Social Science One, to more broadly accessible tools more akin to the Ad Library and CrowdTangle. CrowdTangle has been proven especially helpful for researchers and journalists studying disinformation.

# Data access by law?

Overall, the multifaceted nature of data collected and data access needed for socially responsible research means that we probably will not be able to settle on one specific transparency rule to cover all public interest concerns. Rather, transparency can and must come in many forms.

One form could be by national law: The ongoing reform of the German Network Enforcement Act (NetzDG) has been a missed opportunity to redefine transparency obligations to include data access rights for researchers. The amended § 2 (2) NetzDG will merely oblige platforms to provide information in their transparency reports on *whether* and to *what extent* researchers were granted access to information, but it remains at the platforms’ sole discretion to grant access. While the transparency reports under the NetzDG provide high-level aggregate data on complaints and enforcement actions, these are of minimal use to researchers, since they do not offer detailed insights into individual cases. Ultimately, neither the original version of the NetzDG, nor the (currently pending) amended version are providing meaningful access to data because platforms will obey to the letter but not a bit more.

Clearly, weighing access to data and ensuring privacy is a delicate affair. The technical and ethical challenges here are compounded by pervasive legal uncertainty, which puts researchers in a precarious position and fails to prevent platforms from painting them with the same brush as malicious data hoarders. Both platforms and governments should make an effort to improve on the availability of data for research, and, to this end, clarify the law in this space. Only with meaningful and reliable access can researchers start to answer some of societies’ most pressing questions.

*The authors would like to thank Felix Victor Münch, Cornelius Puschmann, and Gregor Wiedemann for helpful comments.*