Check for updates

Original research article

BIG DATA & SOCIETY

Big Data & Society
October–December: 1–17
© The Author(s) 2025
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517251381686
journals.sagepub.com/home/bds

Sage

Web scraping for research: Legal, ethical, institutional, and scientific considerations

Megan A Brown[1], Andrew Gruen[2], Gabe Maldoff[3], Solomon Messing[4], Zeve Sanderson[4] and Michael Zimmer[5]

Abstract

Scientists across disciplines often use data from the internet to conduct research, generating valuable insights about human behavior. However, as generative artificial intelligence relying on massive text corpora becomes increasingly valuable, platforms have greatly restricted access to data through official channels. As a result, researchers will likely engage in more web scraping to collect data, introducing new challenges and concerns for researchers. This paper proposes a comprehensive framework for web scraping in social science research for U.S.-based researchers, examining the legal, ethical, institutional, and scientific factors that we recommend researchers consider when scraping the web. We present an overview of the current regulatory environment impacting when and how researchers can access, collect, store, and share data via scraping. We then provide researchers with recommendations to conduct scraping in a scientifically legitimate and ethical manner. We aim to equip researchers with the relevant information to mitigate risks and maximize the impact of their research amid this evolving data access landscape.

Keywords

Data access, web scraping, internet law, data collection, research ethics

Introduction

In 2023, Twitter announced it would suspend no-cost academic research access to its Application Programming Interface (API), with a new pricing structure that requires researchers to pay roughly $500,000 per year for comparable access (XDevelopers, 2023).[1] In April 2023, further restrictions were imposed when Twitter deactivated developer accounts set up under its former policy, stalling numerous research projects and public initiatives that relied on Twitter data (Coalition for Independent Technology Research, 2023). In August 2024, Meta shut down CrowdTangle less than three months before the 2024 U.S. elections and in the midst of elections in many other countries around the world (CrowdTangle, 2024). This tool previously enabled researchers to access public data via an API and dashboard, offering the ability to download and analyze this data independently. Consequently, social media researchers and those who built public goods on top of that data were left with few options for data collection through official channels.

At the same time, the regulatory environment governing data scraping remains unclear. There exists a patchwork of state and federal regulations regarding both access and storage of internet data, particularly when that data involve personally identifiable information. Additionally, new data access laws being considered in the U.S. and going into effect in the E.U. expand avenues for researchers to scrape data.

[1]School of Information, University of Michigan, Ann Arbor, MI, USA
[2]Working Paper, LLC, New York, NY, USA
[3]School of Law, University of Maine, Portland, ME, USA
[4]Center for Social Media and Politics, New York University, New York, NY, USA
[5]Department of Computer Science, Marquette University, Milwaukee, WI, USA

Corresponding author:
Megan A Brown, School of Information, University of Michigan, Ann Arbor, MI 48109, USA.
Email: mgnbrown@umich.edu

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Given this backdrop, we expect to see an increase in research relying on scraping data from the web—including by researchers new to the practice—which creates a need for guidelines outlining the opportunities and risks associated with scraping. To guide researchers who may be taking on scraping projects for the first time, we propose a framework for understanding the multifaceted landscape of scraping for social science research, examining the legal, ethical, institutional, and scientific considerations of data access via scraping. Using this framework, we identify critical questions researchers must consider to ensure that research is scientifically rigorous and ethical while minimizing legal risks when using scraping as a data collection tool. We provide guidance on key questions that researchers will face with these data collection methods, recognizing that some of the challenges are immediately resolvable while others will likely require multistakeholder coordination and consensus building that leads to the creation of new norms, practices, and regulatory frameworks. We also provide a checklist for researchers to think through the considerations we outline. Taken together, the information and recommendations in this paper can advise researchers on how to navigate the complexities of scraping while maximizing the potential for ethical, high-impact, scientifically rigorous research.

What is scraping?

We define scraping to be automated data collection via the internet that captures data designed to be used and/or rendered on a web page or app. The study of the internet is rich with examples of work based on scraped datasets (e.g., see Aslett et al. (2024); Baumgartner et al. (2020); Bisbee et al. (2022); Boeker and Urman (2022); Faddoul et al. (2020); Fathoni and Priyawati (2022); Lan et al. (2021); Lange and Monscheuer (2022); Robertson et al. (2018)), but the practice goes back to nearly the beginning of the Web.

The "World Wide Web Wanderer" created in 1993 at MIT may be the earliest incidence of scraping. Researchers interested in the growth of the Web built a tool that "spider-walked" the web, following link to link to the end of a chain and recording information about the number of unique pages discovered (Gray, 1995). This approach is employed to this day in the form of "crawlers"—tools used most often by search engines to collect the contents of web pages, index them, and then rank them by search terms.

In the early days of its development, Microsoft's Bing scraped search results from Google to improve its product. When Google caught Bing, Microsoft vice president Harry Shum admitted to doing so (along with monitoring clicks on Google in its browser), arguing that everyone collectively provides the data for search, and pointing out that across technology companies, "Everyone does this…" (Singel, 2011).

Today, norms around scraping in the technology industry are broad and flexible. Google, Microsoft, Meta, OpenAI, and other tech companies large and small regularly scrape and index content from across the web continually, for profit, and without explicit permission; in the case of machine learning model builders, this practice is increasingly contentious (Grynbaum and Mac, 2023).[2] Sometimes companies profit by way of public-facing free services, which they monetize via advertising, as is the case with most search engines (e.g., Google or Bing). However, scraped data can be used more directly for profit: Companies also scrape to develop structured data they can sell—for everything from checking academic publications for plagiarism (iThenticate—Plagiarism Checking for Academic Research—Turnitin, 2023) to collecting social media photos to create facial recognition systems for police (Clearview AI—Facial Recognition, 2023) to building chatbots for which customers may directly pay (OpenAI—ChatGPT, 2024). Moreover, some companies simply sell the scraped data themselves.[3]

The practice is not limited to web crawling for search engines, artificial intelligence model builders, or enterprise data brokers. Scraping has become a primary method of data collection in many industries for both public interest work and for-profit work. Journalists regularly scrape to investigate and tell stories, arguing it is central to a functioning democracy (The Markup Staff, 2020). In other cases, activists have scraped court records to make analysis and oversight of powerful entities, such as law enforcement, easier (Police Data Accessibility Project, 2023). Despite its broad application, the ethical and legal frameworks governing scraping remain ambiguous and often contentious, underscoring the need for clearer guidelines for researchers.

How do researchers scrape data?

Scraping, while not originally a tool for research, allows researchers to collect data from a variety of internet sources. Using scraping rather than traditional data collection methods may change the questions researcher choose to answer because of the timeliness, availability, or structure of scraped data (Marres and Weltevrede, 2013). Researchers can scrape data using several mechanisms (Davidson et al., 2023; Ohme et al., 2024). Such mechanisms include traditional scraping, which involves the *creation of structure* from inconsistently or loosely structured data from a website, often by parsing HTML. Researchers can also scrape data via undocumented APIs, where researchers use the APIs that serve web content to end users, which are typically unofficial, undocumented, and not permitted for use by third-party developers. Lastly, researchers can enroll participants or "citizen scientists" in the use of a browser plugin, which collects data from a participants' web activity and relays it to the research team.

Importantly, these three access methods are distinct from the official APIs provided by websites (e.g., the Meta Content Library or Twitter's Developer API), which platforms explicitly make available for developers to interact with platform data under conditions set by the platforms. For researchers interested in more details regarding these methods, we provide details in Supplementary Material A. Each of these mechanisms for scraping—traditional scraping, undocumented API scraping, and browser plugin scraping—has different legal, scientific, ethical, and institutional factors that researchers must consider when conducting their projects, which we describe below.

Research considerations

In this section, we provide an overview of the legal, ethical, institutional, and scientific considerations around scraping. We first define the legal considerations for researchers taking on a scraping project, focusing on data access laws that impact researchers in the United States, including U.S. criminal and civil statutes, as well as laws outside the U.S. that researchers should consider. We then discuss the ethics of research using scraped data and how researchers can assess risks to subjects in the data. Given that many researchers work within university contexts, we provide an overview of institutional resources and constraints that researchers must consider when scraping data. Lastly, we discuss scientific factors regarding the validity of scraped data that should inform research design.

Legal considerations

Scraping implicates several overlapping areas of law and regulation.[4] First, behavior on the Internet is often mediated by contractual terms enforced by various private actors that create the content and virtual spaces that form the Internet. Second, scraping can intersect with various statutes and common laws that regulate intellectual property, trespassing, and computer hacking. Third, where scraping results in the collection of personal information, scraping can intersect with privacy and data protection laws that may dictate notice and choice requirements as well as technical and organizational protections that could interfere with, or at least complicate, some areas of research.

All of these areas of law continue to evolve with the growth of data-driven business models and increasing recognition of the relationship between data and economic power (Fung, 2023). To add complexity, the global nature of the internet means that different countries' laws may apply to scraping depending on various factors, such as the location of those doing the scraping, the location of the data source, and the location of any individuals whose data are scraped.

Contractual restrictions. Social media companies and other website operators rely on agreements with their users—often titled "terms of service," "terms of use," "terms and conditions," and "end user license agreements" (collectively, "user contracts")—to protect their legal rights, including to police intellectual property interests, set acceptable guidelines for user behavior, and define dispute resolution procedures. User contracts may seek to regulate and/or prohibit scraping and other forms of automated data collection or analysis necessary for research purposes (Sobel, 2021). At the same time, as this section explains, courts have increasingly limited the power of platforms and websites to unilaterally restrict scraping through take-it-or-leave-it user contracts.

Restrictions on scraping. Most websites prohibit scraping and other automated data collection without the company's permission. Whether framed as a restriction on the automated *collection* of data from the company's services or as a prohibition on automated *access* to the services, these terms are designed to limit scraping by restricting the means by which the services or relevant data are accessed. While user contracts may restrict unauthorized data collection and other activities relevant to research in a variety of ways, provisions blocking scraping specifically distinguish "automated" from nonautomated uses of the services.

User contracts may also implicitly authorize some forms of scraping without requiring a researcher to seek the company's express permission. Most websites, including social media services that prohibit scraping in their user contracts, permit "crawling"—automated scanning that allows search engines to link to and index a website. These websites may authorize such data collection via a robots.txt file that is accessible to search engines' indexing bots. Some companies' robots.txt files permit only specifically named third parties to crawl their services, while others allow crawling by any third party while strictly limiting which elements of the service the third party may crawl. The line between scraping and crawling is not clearly defined.

Challenging restrictions on scraping. A user contract prohibiting scraping does not automatically make scraping illegal. To establish a breach, a website operator must show: (1) a valid agreement, (2) a violation of its terms, and (3) resulting harm (Meta Platforms, Inc. v. BrandTotal Ltd, 2022). Courts have recently narrowed the circumstances, under which scrapers are bound by user contracts, emphasizing how acceptance of the terms was manifested (X Corp. v. Bright Data Ltd, 2023).

User contracts generally state that users agree to the contract when the user creates an account, logs in to the service, or simply uses the website (Meta Platforms, Inc. v. BrandTotal Ltd, 2022, 1256). "Browsewrap" agreements imply user consent to terms simply by using a service. "Clickwrap" agreements require explicit user action, such as clicking a button to accept. A hybrid "sign-in-wrap" agreements state that users have consented through the affirmative action of signing up for or logging into a service.

Courts have increasingly limited the enforceability of browsewrap agreements, especially when the terms are seen as unfair or unexpected. In the scraping context, the Ninth Circuit held that, absent actual knowledge by the user that they are violating the terms, such contracts are only enforceable if (1) the website provides clear notice of the terms and (2) the user takes an action clearly indicating assent (Berman v. Freedom Financial Network, LLC, 2022). Courts are more likely to enforce user contracts when users take affirmative actions to accept the terms. However, even clickwrap and sign-in-wrap agreements may be invalid if the operator cannot show users actually viewed the terms. Courts may also decline to enforce a contract when the conduct at issue falls outside the scope of the user's agreement. For instance, a California court found a scraping restriction unenforceable where the scraping occurred while the party was not logged in, and the data were publicly accessible—even though the scraper had previously accepted the user contract (Meta Platforms, Inc. v. Bright Data Ltd, 2024).

Enforcing user contracts in research contexts can be difficult for website operators, particularly when trying to prove harm from a breach. While commercial scraping is often challenged as unfair competition, such arguments carry less weight in noncommercial settings such as research. In one case, a court found no harm where a nonprofit scraped publicly available data for advocacy purposes (X Corp. v. Center for Countering Digital Hate, Inc., 2024). Operators may still claim harms based on privacy, intellectual property, or hacking. Additionally, courts may be more receptive to public interest defenses, including freedom of expression, in research-related scraping; however, such arguments remain largely untested.

Researchers should exercise caution before concluding that they are not bound by browserwrap agreements that restrict scraping. First, in the *Berman v Freedom Financial Network, LLC* decision, the court noted that if "the website operator can show that a consumer has actual knowledge of the agreement" the terms prohibiting scraping may be enforceable (Berman v. Freedom Financial Network, LLC, 2022, 856). Courts have typically set aside antiscraping provisions in cases where the defendant scraped only information that was either publicly available or already accessible to them through normal use of the service. Courts may be less willing to permit scraping that involves content protected by technical safeguards and/or accessible only to certain classes of users.[5]

Statutory restrictions on scraping. Scraping can also raise statutory considerations. For example, the Computer Fraud and Abuse Act ("the CFAA") has been used by companies and website operators attempting to shut down scraping. The CFAA relates to scraping through a provision that prohibits "[w]hoever" from "intentionally access[ing] a computer without authorization or exceed[ing] authorized access, and thereby obtain[ing]…information from any protected computer." (18 U.S.C. §1030(a)(2)(C) n.d.). While these laws can technically impose criminal penalties on scraping, and therefore, be enforced by law enforcement, "the real area of growth has been with the CFAA's civil provisions" (Sellars, 2018), which allow affected website operators to bring claims for damages.

Website operators may bring CFAA claims to attempt to shut down data scraping activity by arguing that such scraping involves access to the website without authorization. For a CFAA claim to succeed, a website operator must show: (1) unauthorized access to the platform or exceeding authorized access and (2) loss of at least $5000 as a result of the violation (Facebook, Inc. v. Power Ventures, Inc., 2016). In the research context, the United States District Court for the District of Columbia has ruled that a researcher's violation of a website's user contract alone is insufficient to generate criminal liability if the researcher did not breach a website's technical controls in order to gain "unauthorized access" (Sandvig v. Barr, 2020).

In evaluating whether there has been "unauthorized access," courts look at the technical structure for access to content on the website (hiQ Labs, Inc. v. LinkedIn Corporation, 2022). An overview of these technical structures can be found in Supplementary Material B. For example, where scraping occurs on a website that is available to the general public—without the need for an account or any other authorization—courts have found that the CFAA does not apply (Meta Platforms, Inc. v. BrandTotal Ltd, 2022, 1261–62). For password-protected web pages, by contrast, the CFAA could apply to access without appropriate authorization (hiQ Labs, Inc. v. LinkedIn Corporation, 2022, 1199–1200), but courts have found that access to such web pages through a valid user account is sufficient to demonstrate that access was "authorized" and not regulated by the CFAA (Facebook, Inc. v. Power Ventures, Inc., 2016, 1067).[6] At the same time, accessing content through a valid user account is not a permanent shield from CFAA liability because website operators can rescind such permission (including by revoking user accounts, by setting IP address blockers, or by sending cease-and-desist letters), after which continued access could constitute a CFAA violation (Facebook, Inc. v. Power Ventures, Inc., 2016).

Intellectual property. Researchers must consider intellectual property protections, particularly copyright. In the U.S., the Copyright Act grants owners exclusive rights to use their works, subject to exceptions like "fair use." While academic research often qualifies, copyright analysis is fact-specific and context-dependent.

Not all scraped materials are protected by copyright. Copyright covers original works of authorship—not facts or raw data. However, original analysis or presentation of data may be protected. Accessing a work is not a violation; infringement occurs when protected elements are copied or used without authorization. Whether use qualifies as "fair

use" depends on: (1) the purpose (e.g., noncommercial or scholarly), (2) the nature of the work, (3) the amount used, and (4) the impact on the market. Courts emphasize how transformative the use is and whether it competes with the original. Use through unlawful access can weigh against fair use, but isn't determinative. To reduce risk, researchers should ensure their use serves a distinct scholarly purpose and is necessary to that purpose. Commentary, criticism, or analysis typically supports fair use; mere replication or substitution does not.

International laws are often similar. For example, the EU's 2019 Directive permits text and data mining for research by authorized institutions with lawful access (Directive (EU) 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market, 2019).

Privacy and data protection laws. Where scraping involves personal information, researchers may be directly regulated by privacy and data protection laws. U.S. privacy laws provide significant exceptions that are relevant to internet research, including exemptions for publicly available information, for nonprofit organizations, and for certain forms of research. Accordingly, U.S. privacy laws typically will not impose material limits on scraping for academic research purposes, provided that researchers do not circumvent technical controls or otherwise access information that individuals have intentionally made private without authorization. However, outside the U.S., many jurisdictions, including the European Union, apply broad protections to personal information that do not exclude researchers or publicly available information. These laws can apply extraterritorially and therefore may be relevant to U.S.-based researchers.

U.S. privacy laws. There is no single federal privacy law that applies across all sectors, but rather a patchwork of federal and state requirements that govern the collection and processing of personal information. At the federal level, specific laws apply to specific sectors (e.g., the Health Insurance Portability and Accountability Act applies to covered entities in the healthcare sector). Additionally, the Federal Trade Commission (FTC) has general authority to police "unfair and deceptive acts or practices," which the FTC has wielded to shape norms affecting how companies collect, use, and share personal information (Federal Trade Commission, 2024). However, this authority has not been invoked to regulate academic research (Levine, 2021).

At the state level, many states have passed consumer privacy laws, such as the California Consumer Privacy Act (CCPA). As of this writing, nineteen states have passed consumer privacy laws modeled to varying degrees on the CCPA. Most, but not all, of these state privacy laws include important exemptions for nonprofit entities. Additionally, state laws vary with respect to the definition of "publicly available" data, which the laws govern. We elaborate on these dimensions in Supplementary Material C.

To the extent that these U.S. laws apply to scraping for research purposes, researchers generally can comply by maintaining public-facing privacy policies that describe their collection, use, and disclosure of personal information, and by permitting individuals to gain access with a copy of their personal information on request. Researchers may also be subject to other requirements from their respective institutions and research teams regarding how data are collected, analyzed, and stored, such as to implement appropriate security to protect personal information and to apply "data minimization" standards by collecting and storing only information that is reasonably necessary for the intended research purposes. Some of these laws also provide individuals with the right to request that researchers delete personal information, but in many cases, a researcher may have grounds to override the request under various exceptions.

Data protection laws in other countries. Privacy and data protection laws outside the U.S. may be implicated where research impacts individuals in other countries or where research teams have a presence in those countries. For example, the Personal Information Protection Law in China, the General Data Protection Regulation (GDPR) in the EU, and the General Data Protection Law in Brazil all regulate the use of data about individuals within their jurisdictions (though frequently with provisions accounting for academic research). Because the EU GDPR has become a model for data protection law around the world (and is known as one of the world's strictest data protection regimes), we use the GDPR as an example of how such laws may impact data scraping for research.

Researchers may be subject to requirements under GDPR if they are either (1) based in the EU or (2) are "monitoring the behavior" of individuals in the EU, such as by tracking or evaluating any individual's activities in the EU over time. Key provisions in the GDPR require entities to establish a "legal basis" for processing personal information, to provide privacy notice to individuals whose data the business processes, and to offer those individuals certain privacy rights regarding the data, among other internal accountability and documentation requirements. These requirements can pose challenges to researchers who are attempting to scrape data. However, the GDPR provides important exemptions and relaxed requirements for researchers.

First, researchers can demonstrate that data scraping is lawful under the "legitimate interests" legal basis if it does not create undue risks for individuals when compared to the potential benefits of the research. Researchers can mitigate risk and satisfy this legal basis by (1) adhering to ethical and methodological standards, (2) appropriately safeguarding the data—including through technical security controls and through privacy-enhancing techniques such as pseudonymization—and (3) minimizing the data collected

to only what is strictly necessary (European Data Protection Board, 2024). Second, while the GDPR applies heightened protections to certain sensitive data categories, such as data relating to race, ethnicity, religion, sexuality, health, and political beliefs, research that complies with applicable laws usually is exempt from additional requirements. Third, researchers generally are exempt from requirements to notify individuals about the collection of their data for research purposes where providing such notice would require disproportionate effort or undermine research aims (*GDPR Article 14(5)(b)*, n.d.). Lastly, researchers usually are exempt from data subject rights to the extent that those rights limit legitimate research activities (e.g., where individuals opt-out and skew research results) (European Data Protection Supervisor, 2020). When relying on these exemptions, the law puts the onus on researchers to demonstrate how the proposed research design appropriately addresses privacy risks for the data subjects, including appropriate technical and organizational measures (*GDPR Article 89(1)*, n.d.). We provide further detail on these requirements in Supplementary Material D.

Data access laws. Data access laws can support the legal basis for scraping publicly available data. The E.U.'s 2022 Digital Services Act (DSA) requires certain "very large online platforms" and "very large online search engines" to provide researchers access to publicly accessible data through their interfaces—including potentially via scraping—when the research concerns systemic risks in the E.U. (*DSA Article 40(12)*, n.d.). This access extends to nonprofit-affiliated researchers, not just academic institutions.

The DSA also allows "vetted researchers" to access nonpublic platform data, subject to approval by national regulators (Digital Services Coordinators) and specific conditions, such as institutional affiliation, independence, data protection safeguards, and a public research purpose (*DSA Article 40(4)*, n.d.; *DSA Article 40(8)*, n.d.). The DSA's provisions are not limited to E.U.-based researchers, though E.U.-based applicants may be prioritized. Some platforms voluntarily offer research access to users in any jurisdiction. More details regarding the DSA's data access provisions can be found in Supplementary Material E.

In contrast, the U.S. lacks laws granting researchers access to platform-held data. While privacy laws allow the use of personal data for academic research, they do not compel disclosure by private entities. Proposed legislation, such as the 2023 Platform Accountability and Transparency Act, aimed to change this by empowering the NSF to authorize researcher access under strict safeguards (The Platform Accountability and Transparency Act, LYN23256 1RR, 2023). Similar proposals include the Social Media Data Act and the Kids Online Safety Act (Kids Online Safety Act, LYN22092 2SF, 2022; Social Media Disclosure and Transparency of Advertisements Act of, 2021, 2021).

Summarizing the law of scraping. The patchwork of legal requirements that could apply to research involving scraping means that there is no one-size-fits-all analysis, and researchers will need to consider the specific facts of each research proposal to assess their legal risks. In general, scraping that involves breaking into online spaces that are not otherwise available to the public will create higher legal risks than scraping only publicly accessible spaces. Privacy considerations will also require researchers to develop plans to appropriately safeguard any personal information they collect and to evaluate and take steps to limit any risks to individuals resulting from the research. Additionally, given the importance of data access for sound digital policymaking, legislators—led by the E.U.—are starting to propose rules to mandate access to data held by certain platforms for specified public policy purposes. Taking these together, it is important for researchers to consider the access methods, scope of data collected, and privacy minimization procedures of data stored as they navigate uncertainty around data access as the regulatory environment rapidly shifts.

Ethical considerations

The ethics of scraping is not easy to reason about using blanket rules. Within the U.S., the regulatory framework that guides research ethics—45 C.F.R. § 46, known as the "Common Rule"—provides a robust set of protections for research subjects focused on principles of respect for persons, beneficence, and justice. These principles become operationalized through the formal ethical review of research protocols that must address basic tenets of research ethics, such as protecting the privacy of subjects, obtaining informed consent, maintaining the confidentiality of any data collected, and minimizing harm.

Yet, with the rise of Internet-based research, traditional ways of assessing the ethics of research falter. Relying on social media data has prompted considerable debate over core issues such as what constitutes "public" data, whether informed consent is necessary when dealing with "found data," and even at what stage computational research becomes human subjects research requiring particular ethical protection (Metcalf and Crawford, 2016; Pater et al., 2022).

One key consideration is consent: the question of whether the subjects of research using scraped data understand they are being studied and have agreed to such data collection and use. Conventional research on human subjects typically involves direct interaction, and thus informed consent can be attained with relative ease. However, just as observing people in public is generally not considered human subjects research, institutional review boards (IRBs) typically take the view that studying content publicly available on the internet is not considered human subjects research requiring consent. Complicating this further is the

lack of consensus among different university IRBs on whether and how to review protocols relying on publicly available online data (Vitak et al., 2017; Zimmer and Chapman, 2020).

Meanwhile, organizations such as the Association of Internet Researchers (AoIR) have offered ethics guidelines (Franzke et al., 2020), and ethics scholars have pushed for researchers engaging with public data to reflect critically on how their data collection might impact users and their communities (Shilton et al., 2021; Zook et al., 2017) and to be mindful of the original context in which users might be posting data online when considering the appropriateness of scraping data for other uses (Zimmer, 2018).

Central to these recommendations is assessing whether subjects expected that their actions may be observed by researchers: Fiesler and Proferes (2018) show that few users of Twitter (now X) are aware that their public tweets are used by researchers. Children are even less likely to fully understand that their data may be used for research, and special precautions should be taken when scraping data from sources that include content created by minors. Research ethics scholars have further pointed to the need to be particularly mindful when collecting online data on sensitive topics or from vulnerable populations (Fiesler et al., 2016; Klassen and Fiesler, 2022).

Of course, it may be exceptionally difficult to obtain informed consent when collecting platform data, particularly when those data comprise social interactions and social ties between people, or "relational data." For example, social graph and engagement data scraped from social networking sites might implicate secondary users beyond those who might be "consenting" to the collection of their data. Further, images or tags might identify other users who are not able to provide consent.

In addition to addressing consent at the point of data collection, researchers must also grapple with the ethical implications of how scraped data may be used downstream. As these datasets are increasingly fed into machine learning models and other computational tools, they may be repurposed in contexts far removed from the original research intent, such as facial recognition, predictive policing, or behavioral profiling (Crawford and Paglen, 2021; Scheuerman et al., 2021). These applications can carry profound societal risks, especially when they perpetuate bias, enable surveillance, or impact already vulnerable communities (Paullada et al., 2021). Responsible researchers should anticipate these possibilities and take proactive steps to mitigate harm. This may include limiting the scope of data shared, specifying licensing restrictions, maintaining clear documentation about the dataset's intended use (Gebru et al., 2021), and avoiding public release when the risk of misuse is high. Ethical reflection should encompass the entire data lifecycle—not just collection and analysis but also how data may be interpreted, repurposed, or operationalized by others (Shilton et al., 2021; Zook et al., 2017).

Other ethical considerations when scraping data include understanding whether one's data collection practices place an undue load on websites, affecting performance or causing downtime. An ethical approach to data scraping would generally involve being mindful of the impact on the platform's resources and using techniques such as rate limiting and respectful crawling behavior to minimize disruption.

Institutional considerations

As scraping involves significant ethical, legal, and scientific considerations, researchers must navigate diverse challenges and potentially severe risks that likely fall outside of their areas of expertise. In other contexts where such risks appear, universities institutionalize such considerations in order to provide clear guidance, establish uniform processes, and mitigate detrimental outcomes to individuals and the institution. However, given the complexity of issues introduced by scraping, it is often left to the researchers to understand and navigate the various institutional actors that can inform their decision-making. The multifaceted nature of the institutional support required when scraping, and the fact that the support structures do not typically work together, creates risk for researchers. This section aims to provide guidance for researchers to ensure they are connected to the most important institutional actors.

Institutional review boards. In the U.S., the primary method through which ethical considerations are formalized is the IRB. Institutional Review Boards were established to protect human research subjects by ensuring that research aligns with federal regulations and guidelines developed to protect human subjects of research, first articulated in the Belmont Report[7] and later codified in U.S. law by way of the Federal Policy for the Protection of Human Subjects, colloquially referred to as the Common Rule (45 CFR Part 46).[8]

In the context of reviewing data scraping research protocols, a challenge for IRBs is determining the level of risk associated with scraping digital trace or social media data. If data are publicly available (e.g., posts by a public X account) or have already been collected for other purposes, IRBs generally rule the study either exempt[9] or as not engaged in human subjects research. In the "exempt" determination, the IRB assesses that the research falls under one of six exempt categories[10] and is thus governed by the ethical principles of the Common Rule, but not by the legal statutes. In the "not human subjects research" determination, the IRB determines that the researcher is not obtaining information about living humans through interaction or intervention, or not obtaining identifiable private information, and thus falls outside of the Common Rule.

There are two primary ways IRBs differ when evaluating scraping projects. First, IRBs themselves differ between universities, a pattern that is long-standing and well-documented

in meta-scientific research (Goldman and Katz, 1982). To our knowledge, there is no systematic study of how different IRBs evaluate scraping but anecdotal evidence suggests there is wide variation across university IRBs.

Second, because scraping different websites carries differential risks and benefits, institutions may focus on diverse considerations when evaluating a scraping effort for research. For example, scraping public Twitter profiles introduces different risks to user privacy than does scraping private WhatsApp groups, while scraping extremist social media platforms presents different public benefits than scraping GoFundMe pages for medical fundraising. Institutional Review Boards must assess the risks and benefits of research projects based on, among other factors, the granularity and identifiability of the data, the pre- or post-collection processing methods, the data storage architecture, the data sharing plan for publication or replication, and the scholarly or public importance of the research outputs.[11]

Researchers will benefit from an understanding of IRBs generally[12] and their university's IRB specifically. Institutional Review Boards should not be seen as a gatekeeper for engaging in ethical research but rather a partner in ensuring that research aligns with requisite legal, institutional, and ethical standards.

Legal review. The IRB's legal remit is to ensure that research aligns with laws and regulations that govern human subjects' research. However, as previously discussed, scraping (potentially) introduces criminal and civil risks under a patchwork of state and federal laws—the relevance of which for academic scraping continues to be adjudicated by the courts. Generally speaking, IRBs do not have the expertise to evaluate and will not consider the legal implications of scraping. Researchers need to understand both the legal risks associated with a particular project and the support the institution might provide if legal action is taken against them. These questions are especially exigent given that individual researchers, including students or staff under the supervision of faculty, may face legal risk if they engage in scraping.

A university's Office of General Counsel (OGC) is tasked with providing legal guidance to ensure compliance with the university's legal obligations and inform researcher decision-making. Researchers can—but are typically not obligated to—engage with their OGC to evaluate the risks associated with a particular project and get their institution's perspective on those risks. However, it is important that researchers are aware that OGCs exist to represent and protect the university itself. Thus, a researcher undertaking scraping activities that may introduce acute legal risk may want to engage external counsel. Because the "law of scraping" remains unsettled—and the level of legal risk varies considerably depending on the precise study design—it will almost never be possible to find a path to "zero" legal risk without impeding research aims. Effective counsel can help researchers understand how to minimize those risks, particularly by staying within the bounds of practices that have become commonplace among other actors.

Technical review. While social scientists have increasingly used data scraping and computational methods in research, they have not always kept pace with state-of-the-art methods for collection and privacy-preserving data management (Hemphill et al., 2021). Data management plans are required by IRBs during the review process, but IRBs are generally not responsible or equipped to provide technical guidance on the state-of-the-art, thus, the onus is on the researcher to ensure technical rigor when collecting data via scraping.

Most notably, the ethical and legal risk of any social science research project depends on the technical rigor of data collection, storage, processing, and release. For example, secure data management systems with established access controls and deletion protocols limit the chances of data leakage, which lowers the ethical and legal risks of scraping for research. Some universities—especially large, well-funded research institutions—have research IT units that can provide support on key data management questions and provide access to helpful tools and systems. In some cases, universities may provide computing infrastructure that facilitates secure data storage. In contexts where institutional resources are not available, researchers may need to reach out to colleagues in disciplines—such as medicine, public health, and computer science—where technical questions around secure and privacy-preserving data collection have been central to research for some time.

Risks may be reduced further by relying on expert centers of excellence, either within a researcher's own institution or those available more broadly. For example, many researchers studying digital platforms have hosted data with the Inter-university Consortium for Political and Social Research at the University of Michigan (ICPSR About the Organization, n.d.) or contributed data to the Media and Democracy Data Cooperative (MDDC About, n.d.).

What all this suggests is that researchers should not rely solely on the standard IRB approval process when planning research relying on scraping. Instead, they should work to ensure they have adequate legal protections and that their projects are as technically rigorous as possible, often through the institutional channels discussed above.

Scientific considerations

In addition to the ethical, legal, and institutional considerations we highlighted above, researchers must consider scientific issues when scraping data. Many of the issues we highlight in the following section are not unique to scraping as a data collection methodology but scraping may

exacerbate or ameliorate some of the concerns that exist when researchers use APIs or other data collection means.

Poor sampling frames. Perhaps the single largest scientific challenge that researchers face when scraping data is that of their sampling frame. Researchers often have no way of generating random samples based on key units (e.g., users, content, etc.) when they scrape or use an undocumented API, nor do they have the ability to understand the representativeness of the sample they do collect.[13] For example, a researcher scraping results from an in-product search function (e.g., searching X or Facebook for #covid) cannot confirm that all users or all content is included. Perhaps more importantly, researchers often have no way of understanding what the sample they've just collected represents compared to the entire corpus. Further complicating the analysis of such scraped data, duplicate content may be difficult to identify, and search results may have passed through moderation filters that are subject to change daily.

Other technical methods of collection are subject to similarly severe sampling frame issues. Using a browser plugin, for example, may result in significant selection bias—users who understand what a plugin is, how to install it, and what, exactly, their consent is covering are surely not representative of the general public. While these sampling issues can be mitigated to some extent with compensation, thus broadening the populations that are willing to share their data with researchers, it still does not allow for researchers to get a representative sample of a given service's users, as only the company knows that information. Thus, even efforts to collect a nationally representative sample may still not have an accurate sampling frame.

The key takeaway from these sampling frame issues is that the overall representativeness of research based on scraped data is extremely difficult to confirm. For example, Tromble et al. (2017) and González-Bailón et al. (2014) find significant variation in Twitter datasets depending on the sampling frame and API endpoints used. With scraped data, however, researchers face a higher level of uncertainty surrounding the sampling frame and data returned, as there is typically no documentation regarding what researchers can expect to reliably collect, as is the case with APIs. Given this limitation, we recommend that researchers carefully consider any claims they make with regard to the broad applicability of their findings.

Missing data. Another scientific issue with scraped data is that some may be missing from collection in a nonrandom way. Because the data being collected is what is presented to a user, it is likely to be optimized for that purpose instead of completeness (Wu and Taneja, 2021). Further, platforms may not generate logs accurately; meaning actions the platforms say users took may not be recorded. This is particularly true for data purporting to capture "views," which are often sampled in ways that may result in bias to minimize cost or are optimized for other purposes (Herrman and Isaac, 2016; TikTok, 2023).

In addition, the researcher may run into timeouts and/or rate limits when attempting to collect data from an API, which may result in missing entries. Generally speaking, APIs do not respond to queries by delivering data in random order—rather, they tend to deliver data in the way most convenient for the developer, which is often the order in which the cases were created. This creates a strong correlation with time, meaning collections with gaps due to rate limiting can miss data concentrated around particular times, further imperiling the representativeness of the resulting data.

Temporal instability of platform data. Platforms may update timeline algorithms, follower recommendation systems, user interfaces, features, affordances, and more without any announcement to the users or observers of the platform. These changes can inhibit data collection by breaking pipelines constructed to collect data in a particular format that no longer exists. They can also inhibit data collection by introducing unobserved and unexplained variation in the data collection process. This contributes to overall challenges with the temporal validity of datasets created or findings drawn from those datasets (Munger, 2023).

When researchers scrape data, they must contend with the fact that the platform may change at any time. Similarly, many platforms constantly use A/B testing to evaluate changes to projects, meaning the appearance and content on a platform can vary from account to account. While this is not unique to scraping, the potential for platforms to change in ways that halt data collection altogether is higher with scraping than with APIs. With data collection from APIs, researchers (and other data consumers) can collect data in a predictable, structured way, and the format of the data returned generally does not change without significant notice. However, when platforms change their web interfaces or underlying APIs, it is generally without notice and can halt data collection that relies on parsing specific fields.

Construct validity of data. Often, API-based data collection will result in a mismatch between core concepts and operationalized measurement. Consider using impressions as a measure of media consumption. While they indicate that content was loaded on a device that belongs to the end user, they often do not actually measure whether it appeared on a user's screen, much less whether the user actually read it.

This issue becomes even more complicated when we attempt to measure what the "modal" user experience may be on a given website. Many APIs, including some from Google Search and YouTube, deliver results without personalization, which differ from what "logged in" users actually see. Results using queries supplied by researchers may be different from how users actually search for content or

information on platforms, further limiting the construct validity of such results. Such cases require the use of browser plug-ins connected with real users to appropriately measure what is happening to users' on-platform, but these methods can run afoul of platform Terms of Service, and in some cases have resulted in legal action by platforms (Ortutay, 2021).

Reproducibility of research using scraped data. Reproducibility is a known challenge in Internet studies (Demir et al., 2022). Challenges with reproducibility stem from several features of Internet research. The first is that platforms are temporally instable, which we discussed above. As the underlying data may change after research is conducted (e.g., from a platform banning an account or a website removing content), research may become unreproducible because the data is no longer present. Furthermore, scraping pipelines to collect data may break as new antiscraping techniques are implemented and code becomes incompatible with newer versions of APIs and websites, further contributing to a lack of reproducibility. While it may be impossible to fully replicate a pipeline because of website changes and the temporal instability of platform data, researchers can thoroughly document their scraping and analysis pipelines and release data where it is ethical to do so to increase the reproducibility of results.

Recommendations

Different data access methods necessitate different legal, scientific, ethical, and institutional considerations. Importantly, the same data collected in different ways may carry different implications. In the following section, we make recommendations for a set of best practices that U.S.-based researchers can follow (and adapt to their contexts). In addition to immediate recommendations for how to conduct research using scraped data right now, we also provide recommendations for researchers to improve how we take on data scraping projects in the long term.

While the legal and, to some extent, institutional recommendations are necessarily specific to the United States, the ethical and scientific considerations apply to research and researchers regardless of their geography.

Legal

As legal norms surrounding scraping continue to evolve, the path forward for researchers will require case-by-case analysis. Factors such as the method of data collection, the terms of applicable user contracts, the nature of the services and data being scraped, and the affected jurisdictions will all bear on the legal risks that researchers face when scraping for research purposes.

Nonetheless, legislative developments and recent cases evaluating scraping offer the following guiding principles for researchers looking to gather data via scraping in the near term.

- Websites and services that are accessible to the general public without requiring authentication entail significantly lower legal risk.
- Researchers should consider the potential harms that scraping may pose for individual users and the platforms, as material damages can increase the likelihood and value of any claim against the researcher.[14]
- Researchers should collect the minimum data necessary for conducting research and have a credible and documented process for securing the data and protecting privacy. Researchers can analyze the privacy impact of their proposed research through a Data Protection Impact Assessment (see Supplementary Material D), which they can use to identify appropriate safeguards for protecting privacy and security.
- Tools and mechanisms that platforms and search engines make available to researchers to address E.U. requirements under the DSA may allow researchers in the U.S. to gain more streamlined access to certain types of data, particularly for research addressing "systemic risks."

In this evolving landscape of regulation, researchers, IRBs, and OGCs will need to be comfortable with some degree of legal uncertainty for the foreseeable future. The goal of legal analysis should be identifying strategies to manage legal risks rather than eliminating legal risk entirely. Artificial intelligence companies and other data-driven businesses have developed business models that use scraped data as input, in spite of lingering uncertainties. Excess caution in the face of evolving and uncertain legal norms will lead research data collection practices to fall behind emerging commercial norms, which will slow or block research activities that have the potential to confer wider benefits to the public at large.

Ethical

The rapid growth of projects that rely on scraping large datasets for social and online platforms is testing the ethical frameworks and processes used by researchers and ethical review boards to ensure the protection of human subjects from harm. Our key recommendation for researchers navigating the ethics of scraping data is to "remember the human" (Fiesler et al., 2024) and engage in critical "reflexivity" (Shilton et al., 2021) regarding any possible power dynamics that might emerge between the researcher and the users and communities whose data are being scraped. Further, researchers must take steps to acknowledge that the "publicness" of data is not the only factor to consider regarding the ethical dimensions of scraping. Rather, we urge researchers to again take a more reflective

stance and assess the contextual appropriateness (Zimmer, 2018) of collecting, analyzing, and publishing data scraped from a particular platform. This is particularly important when data might be related to sensitive topics or collected from vulnerable communities.

Along with consulting any available IRB or local ethical expertise, numerous resources are available to guide researchers on the ethical complications of scraping data (Buchanan and Zimmer, 2021), including guidelines published by the AoIR (Franzke et al., 2020; Markham and Buchanan, 2012) and the Pervasive Data Ethics for Computational Research (PERVADE) Data Ethics Tool (PERVADE, 2023). Researchers may also consider new collective governance frameworks such as data cooperatives and data trusts (Duncan, 2023), which can incorporate wider viewpoints into how researchers collect, manage, and share scraped data.

Institutional

As we highlight in "Institutional Considerations" section, the institutionalization of the ethical, legal, and technical considerations aims to limit risks for researchers, in part by leveraging the expertise that exists across institutional contexts. A key challenge for researchers is determining who to engage, when, and toward what ends. The IRB will ultimately be responsible for ensuring that a project conforms to internal policies and legal standards for research. Put simply, we recommend that researchers engage their IRB early and often. An IRB should be able to provide guidance on how it understands the risks associated with scraping, and connect researchers to other resources from across the university that can help mitigate those risks. Those resources include IT services, libraries, and general counsel. For researchers engaging in scraping for the first time—or for the first time at a particular institution—colleagues can also act as a resource for understanding how scraping is viewed within that particular environment and how to best navigate the specific actors who are involved in institutionalizing the considerations discussed above.

Researchers may also benefit from engaging external resources to navigate the various risks associated with scraping. Institutional Review Boards are designed to consider the specific legal and ethical considerations associated with human subjects' research, which (as discussed in "Institutional Considerations" section) has limitations within the context of scraping public and semipublic digital trace data. However, they may not have expertise in considering the legal and technical dynamics associated with scraping, and not every institution will have the resources for navigating these considerations. In these contexts, researchers could consider engaging external groups. While these resources will be specific to the particular project, researchers could consider engaging:

- Archival organizations, such as the Social Media Archive, which may be able to provide technical advice and infrastructure for projects that involve scraped data.[15]
- Professional or disciplinary organizations, which may be able to provide resources with best practices for scraping.
- External legal organizations, such as the Knight First Amendment Institute, which may be able to provide legal advice and counsel for projects that involve scraping.[16]

Building institutional capacity requires recognizing the unique challenges faced by early-career researchers (ECRs) involved in scraping. Early-career researchers often lack the support and knowledge available to faculty, even though they frequently conduct scraping themselves. As Brown et al. (2024) note, "ECRs experience heightened precarity… but can carry the bulk of the research labor." This work entails risk. Universities should review policies to protect ECRs, and tenured faculty should advocate for stronger safeguards. Administrators and advisors should proactively share available resources, especially in departments where scraping is common. External organizations must also ensure their resources are accessible and clearly communicated to ECRs. Finally, ECRs should engage stakeholders to understand risks and protections and advocate for themselves within their institutions.

Scientific

Data access methods have broad implications for the scientific validity and utility of data collected. As we highlight in "Scientific Considerations" section, researchers must contend with concerns regarding data availability, sampling frames, missing data, temporal instability of data, and construct validity of the data. As a result, extraordinary care must be taken in order to ensure the validity of scientific findings that result from scraped data. Three recommendations stand out.

First, researchers should develop deliberate sampling strategies rather than relying on convenience samples. For example, The Markup's Citizen Browser recruited a national panel of Facebook users before installing scraping tools, enabling them to reweight data for population inference and appropriately caveat their findings (The Citizen Browser Project—Auditing the Algorithms of Disinformation—The Markup, 2020). Other studies have used identifier-based sampling to create random samples of content on platforms such as Twitter and YouTube (McGrady et al., 2023; Siegel et al., 2021).

Second, researchers must clearly caveat claims based on scraped data. Personalization and recommendation systems can distort what content appears, introducing biases correlated with unobserved variables. Additionally, the most consumed content may not be the most easily scraped,

and consumption metrics can be unreliable (Herrman and Isaac, 2016; TikTok, 2023). Often, it is difficult or impossible to know what data is missing and why, complicating any effort to correct for bias.

Third, researchers should be cautious about generalizing findings. Because scraping rarely provides complete visibility into underlying systems, it is hard to verify how representative a dataset is. For instance, scraped Twitter search results may be incomplete and not generalize to other platforms with different audiences and features. Public data about platform users are often highly aggregated, making debiasing particularly challenging.

Rather than tossing out any research where researchers are unable to get a representative sampling frame, we recommend that researchers carefully caveat their claims based on the attributes of the dataset they are using. In the other example, where researchers used identifier-based methods to create random samples of users and/or posts, researchers could caveat that their findings are based on an identifier-based sample, and there may be unknown variation in how platforms assign identifiers. Researchers could also consider developing well-defined subsamples (e.g., a group of important political actors, celebrities, news networks, or other groups of interest) and collect full data. This path would typically yield a smaller dataset than a random sample of all posts or all posts with a given set of keywords, but a well-defined sampling frame could reduce uncertainty surrounding unknown biases in the dataset and could offer a clear basis for researchers to make claims about generalizability (or lack thereof). Because of the limitations of scraping, research based on scraped data should be seen as the first step on a longer journey toward understanding online behavior, with perhaps larger and better samples or additional data sources.

While actionable steps will be specific to the particular research project, we recommend that researchers consider:

- Collaborating with researchers with experience scraping data or using scraped data in their research
- Careful consideration of whether the question at hand ought to be answered using scraping or via some other approach
- An assessment of the literature to ensure the sample needed to answer the right set of questions, especially in the context of scraping data from platforms in which recommendation systems govern content discovery and distribution
- Foregoing convenience sampling or snowball sampling in favor of clearly defining a limited subpopulation that the data will describe up front (and sampling from it)
- Conducting a variety of tests to understand and refine the data pipeline, identify and correct errors and sources of missing data, and deeply understand the representativeness of the dataset being used for research
- Thoroughly documenting all scraping pipelines, using version control where appropriate, and publishing code and metadata where possible to increase transparency of the research process

A researcher's checklist

The primary motivation for this paper is to assist researchers in conducting public interest internet research. To that end, while we cannot provide a prescriptive set of recommendations that removes all legal, ethical, institutional, and scientific concerns, we can provide a clear list of considerations for which we recommend that researchers have written responses prior to embarking upon a scraping project. Further, we believe that documenting a researcher's responses to each of the considerations below will be useful throughout the research process, particularly when engaging with institutional actors, as it will help all stakeholders align on the facts and practices of the specific scraping instance. Finally, we recommend that researchers use the information generated in this checklist to generate (with the support of their university, if available) a "communications plan" for how to publicly discuss their activity, particularly in cases where they face backlash for scraping. We present this checklist for researchers in the supplementary materials.

Discussion

In the longer term, researchers can and should advocate for protections for researchers using scraped data for public interest research when such research is ethically and scientifically rigorous. For researchers, this can include communicating with policymakers and the public regarding the importance of research on the Internet, as well as demonstrating high ethical rigor to encourage trust in scientists (Dommett and Tromble, 2022). Additionally, researchers can engage with coalitions such as the Coalition for Independent Technology Research[17] to advocate for policies that protect researchers.

We recommend that researchers look for opportunities to engage across the research community to formalize data security and privacy standards for research involving personal information scraped or otherwise obtained from online sources. Data privacy and security standards may ultimately provide a foundation for future data access rights. Researchers may also look toward new collective governance frameworks for research data, allowing broader input from communities impacted by research into the data collection and governance process.

To ensure the ethics of research based on scraped data, steps must be taken to fully address the conceptual gaps that persist when research that relies on scraped data often falls outside formal ethical review structures and the

educational gaps that might prevent researchers from fully understanding the ethical complexities of scraped data. One path forward is the continued updates to regulatory policy that governs the protection of human subjects in research activities—such as the Secretary's Advisory Committee to the Office for Human Research Protections 2013 publication of new recommendations to the Office of Human Research Protections—to consider emerging ethical concerns with Internet-based research.[18] This body should consider the ethical aspects of the increased use of scraped data and provide updated recommendations.

Academic disciplines and publishers have been increasingly engaging in more direct consideration of ethical dimensions of their methodologies. For example, in 2020, the Neural Information Processing Systems conference required that all submissions include a statement addressing broader societal impacts or potential harmful consequences of the research, and instituted an ethical review process for papers flagged by reviewers who saw potential ethical concerns. More research bodies could embrace these forms of self-governance and establish community norms for the ethical use of scraped data. Ethical gaps can be further addressed with longer-term development of enhanced educational resources for both IRBs and researchers alike, incorporating the ethics of scraped data within Responsible and Ethical Conduct of Research programs and related research methodology curricula.

Institutional support for research using scraping can focus on four areas. First, legal issues, both domestic and international, remain central. Institutions should ensure stakeholders such as IRBs and general counsel understand relevant laws to guide researchers and manage risk. New structures—such as data ethics advisory boards or integrating technical and legal reviews into IRB processes—can help address these questions. However, institutional legal resources often cannot provide individual counsel, so external legal organizations could fill this gap. For example, the Knight First Amendment Institute offers legal support to researchers studying social media platforms.[19] Though limited in scope, this can serve as a model for similar initiatives. Second, disciplinary associations and university departments should expand training beyond scraping methods to cover data security, risk mitigation, and ethical and legal considerations. These trainings may also help researchers align with journal expectations, which often vary widely in assessing the risks of scraped data. Third, while each project poses unique challenges, many issues are common across scraping research. Centralizing resources and best practices would improve security, reduce risk, and create efficiencies. Archives like the Social Media Archive can play this role, and funders should support such efforts to benefit a broad range of researchers. Finally, institutional strategies should be designed to address the specific risks faced by ECRs.

Beyond the scope of any one paper, as the field of internet-based research continues to expand and scraping becomes a more commonly used practice to collect data, it is increasingly important to establish standardized practices around the creation of data through scraping. Scientific research benefits from methodologies that ensure data integrity and representativeness, similar to the research practices developed to ensure the reliability of polling. Establishing a formal set of guidelines for scraping can enhance the quality and reliability of research data. This could be achieved through research dedicated to the development of best practices for scraping. Creating empirically driven best practices for scraping could ensure that findings based on scraped data meet rigorous scientific standards. In addition to setting standards and best practices for data collection, researchers should develop specialized training and courses that cater to the evolving needs of internet research, particularly within disciplines where many researchers conduct research on the internet. As this area grows, educational institutions play a large role in equipping future researchers with the necessary skills and knowledge. This can include not only training in scraping methods but also in data analysis, ethics, and the application of scraped data in various fields of study.

Conclusions

In this article, we present a comprehensive framework designed to guide researchers through the complex landscape of legal, ethical, institutional, and scientific considerations essential for scraping projects. We examine pertinent case law and U.S. codes, offering an overview of the current legal environment. Additionally, we explore the ethical dimensions of internet research, directing readers to key resources that encourage critical reflection on their proposed studies. We further discuss the institutional support available to researchers, highlighting both internal mechanisms—such as IRBs, Research IT teams, and OGCs—and external experts such as the Social Media Archive at ICPSR and the Knight First Amendment Institute. Moreover, we address scientific concerns related to the use of scraped data, emphasizing the importance of assessing external validity and the generalizability of datasets, as well as the limitations of the conclusions that can be drawn. Overall, this paper equips researchers and institutional stakeholders with the necessary tools and knowledge to conduct scientifically robust research via scraping, ensuring the pertinent legal, ethical, institutional, and scientific factors are thoroughly considered.

Acknowledgments

The authors thank Josephine Lukito, Jonathan Nagler, and Joshua A. Tucker for their feedback on this manuscript.

ORCID iDs

Megan A Brown https://orcid.org/0000-0002-1338-8054
Andrew Gruen https://orcid.org/0009-0006-6516-9730

Gabe Maldoff https://orcid.org/0009-0003-0867-4771
Michael Zimmer https://orcid.org/0000-0003-4229-4847

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Supplemental material

Supplemental material for this article is available online.

Notes

1. Researcher API access previously provided 10 million tweets per month. The Basic access tier announced on March 29, 2023, provides 10 thousand tweets for $100 per month; the Pro tier provides one million tweets for $5,000 per month; and the small enterprise access provides 50 million tweets for $42,000 per month, making Twitter cost-prohibitive to most researchers.
2. Most large companies do not index websites if the robots.txt file opts out. Robots.txt was codified in 1994 by Koster and is only now going through a standards codification process via the Internet Engineering Task Force's Request for Comments process (Koster et al., 2022)
3. [3]Some of the highest-profile legal cases involving scraping have focused on simply selling scraped data. For example, *hiQ Labs, Inc v. LinkedIn Corporation*, one of the only federal appellate decisions to address scraping, involved access hiQ's access to LinkedIn data to create a "people analytics" product, which it marketed and sold to business customers (*hiQ Labs, Inc. v. LinkedIn Corporation*, 2022).
4. When, in 1996, Judge Frank Easterbrook was asked to present on "The Law of Cyberspace," he questioned whether it was appropriate to talk of a "cyberspace" or "Internet" law at all and he likened such a concept to a "law of the horse"—where there is no specific law of horses but instead many considerations (Easterbrook, 1996). Like Internet law, scraping is situational and implicates disparate areas of law, but common themes and trends emerge that can guide researchers through the maze of legal complexity.
5. In this context, researchers' deploying other users to assist in scraping—such as through the installation of browser extensions that can scrape within their user accounts—could lead to breach of contract claims, at least with regard to those users who have accepted the platform's user agreement.
6. This case held that Power Ventures did not access Facebook without authorization when users initially gave Power permission to access Facebook data through their accounts. Rather, Power Ventures only violated CFAA when they continued to access Facebook data after Facebook sent a cease and desist and blocked Power Ventures IP addresses from accessing Facebook's servers.
7. See https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html
8. See https://www.hhs.gov/ohrp/regulations-and-policy/regulations/common-rule/index.html
9. See https://www.ecfr.gov/current/title-45/subtitle-A/subchapter-A/part-46
10. Helpfully elaborated on by Michigan State University at https://hrpp.msu.edu/help/required/exempt-categories.html
11. Again, while the focus of this paper is on U.S. university-based researchers, some international work may be relevant. The Report of the European Digital Media Observatory's working group on Platform-to-Researcher Data Access (European Digital Media Observatory, 2022) offers GDPR-specific guidance that explicitly focuses on each of these issues for nonscraped data.
12. See, for example https://admindatahandbook.mit.edu/book/latest/index.html
13. To be clear, this could also be the case with official APIs, but at least with these tools, there are often documented descriptions of the API's implementation, which allows data consumers to better understand the biases introduced. Undocumented APIs and scraping typically lack such documentation, which makes sources of bias in the data unknowable.
14. Where scraping does not result in any harm, even if courts find a breach of user contract, a platform's remedies often are limited to blocking future scraping rather than monetary penalties.
15. For more information, see https://socialmediaarchive.org/
16. For more information, see https://knightcolumbia.org/
17. See https://independenttechresearch.org/
18. Available at https://www.hhs.gov/ohrp/sites/default/files/ohrp/sachrp/mtgings/2013%20March%20Mtg/internet_research.pdf
19. https://knightcolumbia.org/content/knight-institute-to-provide-new-legal-support-to-researchers-studying-online-platforms

References

18 U.S.C. §1030(a)(2)(C). (n.d.) (Access to Computer Systems Without Authorization).

Aslett K, Sanderson Z, Godel W, et al. (2024) Online searches to evaluate misinformation can increase its perceived veracity. *Nature* 625(7995): 548–556.

Baumgartner J, Zannettou S, Keegan B, et al. (2020) The pushshift reddit dataset. *Proceedings of the International AAAI Conference on Web and Social Media* 14(1): 830–839.

Berman v. Freedom Financial Network, LLC. (2022) 30 F.4th. (9th Cir.).

Bisbee J, Brown M, Lai A, et al. (2022) Election fraud, YouTube, and public perception of the legitimacy of President Biden. *Journal of Online Trust and Safety* 1(3). https://doi.org/10.54501/jots.v1i3.60.

Boeker M and Urman A (2022) An empirical investigation of personalization factors on TikTok. *WWW '22: Proceedings of the ACM Web Conference 2022*: 2298–2309. https://doi.org/10.1145/3485447.3512102.

Brown MA, Lukito J, Pruden ML, et al. (2024) Making academia suck less: Supporting early career researchers studying harmful content online through a feminist ethics of care. *New Media & Society* 14614448241303999.

Buchanan E and Zimmer M (2021) Internet research ethics. The Stanford encyclopedia of philosophy. Retrieved 2021-01-12, from http://plato.stanford.edu/entries/ethics-internet-research/.

Clearview AI—Facial Recognition. (2023) Retrieved 2023-06-21, from https://www.clearview.ai.

Coalition for Independent Technology Research. (2023) Letter: Twitter's New API Plans Will Devastate Public Interest Research. Retrieved 2024-05-27, from https://independenttechresearch.org/letter-twitters-new-api-plans-will-devastate-public-interest-research/.

Crawford K and Paglen T (2021) Excavating AI: The politics of images in machine learning training sets. *Ai & Society* 36(4): 1105–1116.

CrowdTangle. (2024) Important Update to CrowdTangle | March 2024 | CrowdTangle Help Center. Retrieved 2024-03-20, from http://help.crowdtangle.com/en/ articles/9014544-important-update-to-crowdtangle-march-2024.

Davidson BI, Wischerath D, Racek D, et al. (2023) Platform-controlled social media APIs threaten open science. *Nature Human Behaviour* 7(12): 2054–2057.

Demir N, Große-Kampmann M, Urban T, et al. (2022) Reproducibility and replicability of web measurement studies. In: *Proceedings of the ACM Web Conference* 2022 (pp. 533–544).

Directive (EU) 2019/790 of the European Parliament and of the Council on Copyright and Related Rights in the Digital Single Market. (2019) https://eur-lex.europa.eu/eli/dir/2019/790/oj. (OJ L 130, 17.5.2019, p. 92–125).

Dommett K and Tromble R (2022) Advocating for platform data access: Challenges and opportunities for academics seeking policy change. *Politics and Governance* 10(1): 220–229.

Duncan J (2023) Data protection beyond data rights: Governing data production through collective intermediaries. *Internet Policy Review* 12(3): 1–22.

Easterbrook FH (1996) *Cyberspace and the Law of the Horse.* Vol 1996, pp. 207–208. Chicago, IL: University of Chicago Legal Forum. https://chicagounbound.uchicago.edu/uclf/vol1996/iss1/7.

European Data Protection Board. (2024, May 23) Report of the Work Undertaken by the ChatGPT Taskforce. Retrieved from https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.

European Data Protection Supervisor. (2020, Jan 6) A preliminary opinion on data protection and scientific research. Retrieved from https://www.edps.europa.eu/sites/default/files/publication/20-01-06_opinion_research_en.pdf.

European Digital Media Observatory. (2022, May 31) Report of the European digital media observatory's working group on platform-to-researcher data access, Annex 4—Compendium of EU Member State Laws. Retrieved from https://edmo.eu/wp-content/uploads/2022/02/Report-of-the-European-Digital-Media-Observatorys-Working-Group-on-Platform-to-Researcher-Data-Access-2022.pdf.

Facebook, Inc. v. Power Ventures, Inc. (2016) 844 F.3d 1058. (9th Cir.).

Faddoul M, Chaslot G and Farid H (2020) A longitudinal analysis of YouTube's promotion of conspiracy videos. arXiv. https://doi.org/10.48550/arXiv.2003.03318.

Fathoni AN and Priyawati D (2022) Mapping of underdeveloped areas based on research frequency utilizing distributed web scraping and web GIS. *International Journal for Disaster and Development Interface* 2(2): 275–291.

Federal Trade Commission. (2024, March 4) FTC cracks down on mass data collectors: A closer look at avast, X-Mode, and InMarket. Retrieved from https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/03/ftc-cracks-down-mass-data-collectors-closer-look-avast-x-mode-inmarket.

Fiesler C and Proferes N (2018) Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4(1). https://doi.org/10.1177/2056305118763366.

Fiesler C, Wisniewski P, Pater J, et al. (2016) Exploring ethics and obligations for studying digital communities. In: *GROUP '16: Proceedings of the 2016 ACM international conference on supporting group work*, pp.457–460. https://doi.org/10.1145/2957276.2996293.

Fiesler C, Zimmer M, Proferes N, et al. (2024) Remember the human: A Systematic review of ethical considerations in reddit research. In: Proceedings of the ACM on human-computer interaction, 8(GROUP). https://doi.org/10.1145/3633070.

Franzke AS, Bechmann A, Zimmer M, et al. (2020) *Internet Research: Ethical Guidelines 3.0*. Association of Internet Researchers. https://aoir.org/reports/ethics3.pdf.

Fung B (2023, Mar) DOJ will hire more data experts to scrutinize digital monopolies, Antitrust chief says—CNN business. Cable News Network. Retrieved from https:// www.cnn.com/2023/03/06/tech/doj-data-experts/index.html.

Gebru T, Morgenstern J, Vecchione B, et al. (2021) Datasheets for datasets. *Communications of the ACM* 64(12): 86–92.

Goldman J and Katz MD (1982) Inconsistency and institutional review boards. *JAMA* 248(2): 197–202.

González-Bailón S, Wang N, Rivero A, et al. (2014) Assessing the bias in samples of large online networks. *Social Networks* 38: 16–27.

Gray M (1995) Measuring the Growth of the Web. Retrieved 2023-06-21, from https://www.mit.edu/people/mkgray/growth/.

Grynbaum MM and Mac R (2023, December) The times sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work. The New York Times. Retrieved 2024-03-20, from https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Hemphill L, Hedstrom ML and Leonard SH (2021) Saving social media data: Understanding data management practices among social media researchers and their implications for archives. *Journal of the Association for Information Science and Technology* 72(1): 97–109.

Herrman J and Isaac M (2016) The online video view: We Can Count It, but Can We Count on It? The New York Times. Retrieved from https://www.nytimes.com/2016/10/03/business/

media/the-online-video-view-we-can-count-it-but-can-we-count-on-it.html (accessed 18 April 2024).

hiQ Labs, Inc. v. LinkedIn Corporation. (2022) 31 F. 4th 1180. (9th Cir.).

ICPSR About the Organization. (n.d.) Inter-university consortium for political and social research. Retrieved 2023-05-18, from https://www.icpsr.umich.edu/web/pages/about/.

iThenticate—Plagiarism Checking for Academic Research—Turnitin. (2023) Retrieved 2023-06-21, from https://www.turnitin.com/products/ithenticate.

Kids Online Safety Act, LYN22092 2SF. (2022) 117th Congress. Retrieved from https://www.blumenthal.senate.gov/imo/media/doc/kids_online_safety_act_-_bill_text.pdf.

Klassen S and Fiesler C (2022) "This Isn't Your Data, Friend": Black Twitter as a Case Study on Research Ethics for Public Data. *Social Media + Society* 8(4). https://doi.org/10.1177/20563051221144317.

Koster M, Illyes G, Zeller H, et al. (2022, September) Robots Exclusion Protocol (No. 9309). RFC 9309. RFC Editor. Retrieved from https://www.rfc-editor.org/info/rfc9309 DOI: 10.17487/RFC9309.

Lan H, Sha D, Malarvizhi AS, et al. (2021) COVID-Scraper: An open-source toolset for automatically scraping and processing global multi-scale spatiotemporal COVID-19 records. *IEEE Access* 9: 84783–84798.

Lange M and Monscheuer O (2022) Spreading the disease: Protest in times of pandemics. *Health Economics* 31(12): 2664–2679.

Levine S (2021) Letter from acting director of the bureau of consumer protection Samuel Levine to Facebook. Retrieved from https://www.ftc.gov/blog-posts/2021/08/letter-acting-director-bureau-consumer-protection-samuel-levine-facebook.

Markham A and Buchanan E (2012) Ethical decision-making and internet research: recommendations from the aoir ethics working committee (Version 2.0) (Tech. Rep.). Association of Internet Researchers. Retrieved from http://aoir.org/reports/ethics2.pdf.

Marres N and Weltevrede E (2013) Scraping the social? Issues in live social research. *Journal of Cultural Economy* 6(3): 313–335.

McGrady R, Zheng K, Curran R, et al. (2023) Dialing for videos: A random sample of YouTube. *Journal of Quantitative Description: Digital Media* 3. https://doi.org/10.51685/jqd.2023.022.

MDDC About. (n.d.) Media and democracy data cooperative. Retrieved 2023-05-18, from https://mddatacoop.org/about/.

Meta Platforms, Inc. v. BrandTotal Ltd. (2022) 605 F.Supp.3d. (N.D. Cal.).

Meta Platforms, Inc. v. Bright Data Ltd. (2024) 2024 WL 251406. (N.D. Cal., Jan. 23).

Metcalf J and Crawford K (2016) Where are human subjects in big data research? The emerging ethics divide. *Big Data & Society* 3(1). https://doi.org/10.1177/2053951716650211.

Munger K (2023) Temporal validity as meta-science. *Research & Politics* 10(3): 20531680231187271.

Ohme J, Araujo T, Boeschoten L, et al. (2024) Digital trace data collection for social media effects research: APIs, data donation, and (screen) tracking. *Communication Methods and Measures* 18(2): 124–141.

OpenAI—ChatGPT. (2024) Retrieved 2024-09-10, from https://openai.com/chatgpt/.

Ortutay B (2021) Facebook shuts out NYU academics' research on political ads. AP News. Retrieved from https://apnews.com/article/technology-business-5d3021ed9f193bf249c3af158b128d18.

Pater J, Fiesler C and Zimmer M (2022) No humans here: Ethical speculation on public data, unintended consequences, and the limits of institutional review. *Proceedings of the ACM on Human-Computer Interaction* 6(GROUP): 1–13.

Paullada A, Raji ID, Bender EM, et al. (2021) Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns* 2(11. doi:10.1016/j.patter.2021.100336

PERVADE. (2023) PERVADE Data Ethics Tool. https://pervade.umd.edu/pervade-data-%20ethics-tool/ (accessed 23 May 2024).

Police Data Accessibility Project. (2023) Retrieved 2023-06-21, from https://www.pdap.io.

Robertson RE, Jiang S, Joseph K, et al. (2018) Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW): 1–22.

Sandvig v. Barr. (2020) Civ. Action No. 16-1368. (D.D.C., March 28).

Scheuerman MK, Hanna A and Denton E (2021) Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW2): 1–37.

Sellars A (2018) Twenty years of web scraping and the computer fraud and abuse act. *Boston University Journal of Science & Technology Law* 24: 372–376.

Shilton K, Moss E, Gilbert SA, et al. (2021) Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research. *Big Data & Society* 8(2). https://doi.org/10.1177/205395172110407.

Siegel AA, Nikitin E, Barbera P, et al. (2021) Trumping hate on twitter? Online hate speech in the 2016 U.S. *Election Campaign and its Aftermath. Quarterly Journal of Political Science* 16(1): 71–104.

Singel R (2011) Google Catches Bing Copying; Microsoft Says 'So What?'. Retrieved 2023-03-24, from https://www.wired.com/2011/02/bing-copies-google/.

Sobel BLW (2021) A new common law of web scraping. *Lewis & Clark Law Review* 25: 147. https://law.lclark.edu/live/files/31605-7-sobel-article-251pdf.

Social Media Disclosure and Transparency of Advertisements Act of 2021. (2021) 117th Congress. Retrieved from https://trahan.house.gov/uploadedfiles/social_media_data_act_bill_text.pdf.

The citizen browser project—auditing the algorithms of disinformation—The Markup. (2020) Retrieved 2024-03-20, from https://themarkup.org/citizen-browser.

The Markup Staff. (2020) Why Web Scraping Is Vital to Democracy. Retrieved 2023-06-21, from https://themarkup.org/news/2020/12/03/why-web-scraping-is-vital-to-democracy.

The Platform Accountability and Transparency Act, LYN23256 1RR. (2023) 118th Congress.Retrieved from https://www.coons.senate.gov/imo/media/doc/text_pata_117.pdf.

TikTok. (2023) Video Play Reporting Metrics. Retrieved 2024-04-18, from https://ads.tiktok.com/help/article/video-play.

Tromble R, Storz A and Stockmann D (2017) We don't know what we don't know: when and how the use of Twitter's public APIs biases scientific inference. Available at SSRN 3079927. http://dx.doi.org/10.2139/ssrn.3079927.

Vitak J, Proferes N, Shilton K, et al. (2017) Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics* 12(5): 372–382.

Wu AX and Taneja H (2021) Platform enclosure of human behavior and its measurement: Using behavioral trace data against platform episteme. *New Media & Society* 23(9): 2650–2667.

X Corp. v. Bright Data Ltd (2023) No. 3:23-cv-03698 (N.D. Cal. July 26).

X Corp. v. Center for Countering Digital Hate, Inc. (2024) 2024 WL 1246318. (N.D. Cal., March 25).

XDevelopers. (2023) Announcing new access tiers for the twitter api. Retrieved 2024-05-27, from https://devcommunity.x.com/t/announcing-new-access-tiers-for-the-twitter-api/188728.

Zimmer M (2018) Addressing conceptual gaps in big data research ethics: An application of contextual integrity. *Social Media + Society* 4(2). https://doi.org/10.1177/2056305118768300.

Zimmer M and Chapman E (2020) Ethical review boards and pervasive data research: Gaps and opportunities. *AoIR Selected Papers of Internet Research* 2020. https://doi.org/10.5210/spir.v2020i0.11369.

Zook M, Barocas S, Boyd D, et al. (2017) Ten simple rules for responsible big data research. *PLoS Computational Biology* 13(3): e1005399.