Selected Papers of #AoIR2025:
The 26th Annual Conference of the
Association of Internet Researchers
Niterói, Brazil / 15 – 18 Oct 2025

AOIR ETHICS IN ACTION: GUIDELINES, REGULATIONS, AND PRACTICES ACROSS DIVERSE GLOBAL COMMUNITIES

Michael Zimmer
Marquette University

Jessica Vitak
University of Maryland

Anna Lenhart
George Washington University

Ylva Hård af Segerstad
University of Gothenburg

Thomas Hartvigsson
University of Gothenburg

Annika Bergviken-Rensfeldt
University of Gothenburg

Thomas Hillman
University of Gothenburg

Megan Brown
University of Michigan

Andrew Gruen
Mozilla Foundation

Gabe Maldoff
Goodwin Procter

Solomon Messing
New York University

Suggested Citation (APA): Zimmer, M., Vitak, J., Lenhart, A., Hård af Segerstad, Y., Hartvigsson, T., Bergviken-Rensfeldt, A., Hillman, T., Brown, M., Gruen, A., Maldoff, G., & Mesing, S. (2025, October). *AoIR Ethics In Action: Guidelines, Regulations, and Practices Across Diverse Global Communities*. Panel presented at AoIR2025: The 26th Annual Conference of the Association of Internet Researchers. Niterói, Brazil: AoIR. Retrieved from http://spir.aoir.org.

Eli Asikin-Garmager
Wikimedia Foundation

Cameran Ashraf
Wikimedia Foundation

Leila Zia
Wikimedia Foundation

Since its inception, the Association of Internet Researchers (AoIR) has fostered critical reflection on the ethical and social dimensions of the internet and Internet-facilitated communication and interactions. The AoIR Ethics Working Committee has been committed to ensuring the AoIR Ethics Guidelines remain helpful and relevant to researchers and ethical review committees; our goal is to support ethical decision-making in internet research by keeping the guidelines responsive to emerging technologies, evolving research practices, and the diverse needs of scholarly and practitioner communities. This panel includes five papers that engage in “AoIR Ethics in Action” – demonstrating how the AoIR Ethics Guidelines and their core principles are applied, interpreted, and shaped across diverse, global contexts. The presentations explore internet researchers’ preparedness and design-making processes related to the ethics of their work; how the emerging regulatory frameworks in the U.S. and Sweden illuminate evolving challenges in research ethics; and examine how the guidelines can inform ethical decision-making in specific cases, such as web scraping and the use of Wikipedia data.

RESEARCH ETHICS PRACTICES & PREPAREDNESS: A STUDY OF AOIR & ICA
This paper reports on survey results from internet researchers in AoIR and ICA to assess their engagement with ethical guidelines and preparedness for handling ethical dilemmas in digital data collection. The findings indicate potential gaps in ethical preparedness and training, emphasizing the need for updated AoIR Ethics Guidelines and stronger institutional support for researchers dealing with pervasive data

ETHICAL CHALLENGES IN PERVASIVE DATA RESEARCH: STAKEHOLDER PERSPECTIVES AND PATHS FORWARD
This report analyzes responses to a U.S. NTIA Request for Comment on ethical guidelines for pervasive data research, identifying core tensions such as legal/definitional ambiguities, anonymization limitations, and the challenge of balancing ethical principles when performing research with large-scale datasets. It argues that researchers and civil society must proactively strengthen ethical research practices in the absence of clear government guidelines.

ETHICS IN TRANSITION: ADAPTING RESEARCH GOVERNANCE TO A PROPOSED NEW LEGAL FRAMEWORK IN SWEDEN
This contribution examines Sweden’s shift toward decentralized research ethics governance under a proposed legal reform, which will transfer responsibility for ethical review from national boards to individual institutions. It explores tensions between

standardized regulations and local adaptations, highlighting potential risks such as institutional disparities and legal uncertainties for interdisciplinary digital research.

WHEN IS SCRAPING LEGITIMATE? ETHICAL, LEGAL, ADMINISTRATIVE, AND TECHNICAL CONSIDERATIONS
This white paper presents a comprehensive framework for web scraping in social science research, examining the legal, ethical, institutional, and scientific factors that researchers must consider when scraping the web. It presents an overview of the current regulatory environment impacting when and how researchers can access, collect, store, and share data via scraping, offering researchers guidance on best practices to balance data access with responsible research ethics.

RESEARCH AND PRIVACY ON WIKIPEDIA
This white paper examines the ethical and privacy challenges of conducting research on Wikipedia, emphasizing concerns about re-identification, user anonymity, and community expectations. It proposes best practices for researchers and Wikipedia users to help navigate ethical tensions while ensuring transparency, user safety, and adherence to privacy norms.

Together, these papers illustrate the ongoing relevance of the AoIR Ethics Guidelines in navigating the complex ethical, legal, and regulatory landscape of internet research.

RESEARCH ETHICS PRACTICES & PREPAREDNESS: A STUDY OF AOIR & ICA

Michael Zimmer
Marquette University

Jessica Vitak
University of Maryland

Abstract

We present results from a survey study of AoIR and ICA (International Communication Association) researchers who collect digital data about people in their work to (1) inventory their current data collection practices, (2) assess their engagement in educational activities focused on data ethics and use of existing ethics resources (such as the AoIR ethics guidelines), and (3) identify how these activities and resources can be improved to help ensure internet research is conducted ethically. These findings provide helpful insights into the ethics practices and preparedness among the diverse set of scholars who make up the internet research community spanning two organizations, provide a comparative analysis across scholarly communities to identify opportunities for broader engagement in research ethics training and outreach, and inform future updates to the AoIR Ethics Guidelines.

Introduction

The field of internet research encompasses a multidisciplinary approach to studying the internet and its impact on various aspects of persons, society, culture, technology, communication, and behavior. Internet research draws upon insights and methodologies from diverse disciplines, including but not limited to communication studies, sociology, cultural studies, anthropology, information studies, and computer science. The field has undergone a remarkable expansion, evolving from its origins in studying websites and online spaces to encompassing a wide array of digital platforms and activities, including digital cultures, gaming, mobile apps, streaming and content platforms, virtual realities, ubiquitous computing, and the like.

To engage in their work, many internet researchers collect and/or analyze digital data about people – a category of data referred to as "pervasive data" [4]. This includes (but is not limited to) data collected from websites, social media platforms, mobile devices, APIs, and related data that might be generated through digital interactions, often without data subjects even knowing it is generated or collected. Yet, each of the disciplines that make up the internet research community carries different histories and norms when it comes to research ethics and how to appropriately manage pervasive data collected about research subjects.

Further, the collection of pervasive data often does not neatly fall into the regulatory category of "human subjects research," meaning it might not be overseen by university ethics review committees, leaving internet researchers largely on their own to make decisions about ethical data collection, storage, sharing, and publication practices.

Since as early as 2002, the Association of Internet Researchers (AoIR) has helped fill this gap by publishing a set of guidelines designed to guide ethical conduct in internet research [1–3]. These guidelines serve as a framework for researchers to navigate the ethical complexities inherent in studying online phenomena and interacting with internet users. Over the course of two decades and multiple iterations, the AoIR ethics guidelines have played a crucial role in promoting ethical conduct, fostering responsible research practices, and advancing the ethical development of internet research as an interdisciplinary field.

However, to date there has been no study of the ethical practices and preparedness of internet researchers themselves. Building from our ongoing assessments of the ethical attitudes and preparedness of numerous research communities, this study inventories the current data practices of internet researchers within the AoIR and International Communication Association (ICA) communities, assesses the extent to which internet researchers have participated in educational activities focused on data ethics and make use of existing resources (such as the AoIR ethics guidelines), and solicits input on how these activities and resources can be improved to ensure their research is conducted ethically and in a way that benefits humanity as a whole.

Our research questions are:

RQ1: How do internet researchers across different disciplines collect and use pervasive data, and how do they assess any ethical issues connected to such data?

RQ2: What educational opportunities have internet researchers been provided around research ethics, and how do they assess their ability to address the ethical implications of their research?

RQ3: What new resources or educational opportunities would enhance internet researchers' ability to assess the ethical implications of their work?

Methodology

Recruitment took place in late 2023 and early 2024 after receiving approval by our ethics review board for the anonymous online survey. We recruited AoIR members through the AoIR listserv, while ICA members were reached through targeted emails to relevant divisions and the Facebook Group page for the "Communication and Technology" division. Both recruitment efforts provided access to a multidisciplinary and global population of internet researchers. We received 152 responses from AoIR and 73 responses from ICA members.

The survey asked about participants' current data collection methods and practices, any formal and informal ethical training they might have received, what tools/resources they relied on when engaging in research, their perceived level of preparedness to handle ethical issues in their research, and also collect data about their discipline, location, and related demographics. Questions also solicited participant perspectives on the appropriateness of various data collection techniques and common ethical issues related to internet research methodologies.

Findings

The survey and interview findings will provide descriptive data on the research ethics attitudes, knowledge, and training among a diverse, global population of internet researchers who rely on pervasive data. Findings will be used to develop an inventory of current practices, resources, and gaps among internet researchers seeking ethical guidance around the use of data.

Results of this study will also inform future updates to the AoIR ethics guidelines, guide the development of additional educational and outreach resources, and highlight potential opportunities for additional interventions to ensure that internet researchers across all disciplines are well-informed about the ethical considerations and implications of their work.

We have conducted similar surveys of other researcher communities who frequently rely on pervasive data, including contributors to the Conference on Neural Information Processing Systems (NeurIPS). Combined with the AoIR and ICA communities. Future work will include a comparative analysis across these scholarly communities to identify

opportunities for broader engagement in research ethics training and outreach in interdisciplinary environments.

References

[1] Ess, Charles and Steve Jones. 2002. *Ethical decision-making and Internet research: Recommendations from the AoIR ethics working committee*. Association of Internet Researchers. Retrieved from https://aoir.org/reports/ethics.pdf

[2] franzke, aline, Anja Bechmann, Michael Zimmer, and Charles Ess. 2020. *Internet Research: Ethical Guidelines 3.0*. Association of Internet Researchers. Retrieved from https://aoir.org/reports/ethics3.pdf

[3] Annette Markham and Elizabeth Buchanan. 2012. *Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0)*. Association of Internet Researchers. Retrieved from http://aoir.org/reports/ethics2.pdf

[4] Shilton, Katie, Emanuel Moss, Sarah A. Gilbert, Matthew J. Bietz, Casey Fiesler, Jacob Metcalf, Jessica Vitak, and Michael Zimmer. 2021. Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research. Big Data & Society, 8(2). https://doi.org/10.1177/20539517211040759

ETHICAL CHALLENGES IN PERVASIVE DATA RESEARCH: STAKEHOLDER PERSPECTIVES AND PATHS FORWARD

Anna Lenhart
George Washington University

Abstract

The widespread collection of digital data through online services, known as pervasive data, has become a critical resource for researchers studying human behavior, public health, and societal trends. However, the use of pervasive data raises significant ethical concerns related to privacy, consent, data security, and regulatory compliance. In response, the U.S. National Telecommunications and Information Administration (NTIA) issued a Request for Comment (RFC) on the potential development of national ethical guidelines for pervasive data research. The RFC sought public input on key issues, including barriers to data access, legal and regulatory challenges, and mitigation strategies for ethical risks. This report analyzes the 50 stakeholder responses to the RFC, identifying five core tensions, including legal ambiguities, the challenge of balancing ethical principles with large-scale datasets, and the limitations of anonymization. While government-issued guidelines remain uncertain, stakeholders can take proactive steps, such as training researchers, piloting alternative review processes, conducting legal analysis, and improving platform designs to enhance user autonomy and ethical data sharing. By strengthening ethical research practices now, civil society and academic institutions can lay the groundwork for more informed and effective future regulations, ensuring that pervasive data research serves the public interest while respecting individual rights.

The Rise of Pervasive Data and Its Ethical Challenges

Over the last few decades, new technologies such as web-based monitoring tools, content delivery networks, education technology, Internet of Things devices, mobile sensors, streaming services, search engines, online marketplaces, social media platforms, AI systems, etc., have created troves of data that can be used to inform important research in the public interest.

This data has been described as “pervasive data” (Shilton et al, 2021), and researchers have leveraged pervasive data to better understand human behavior, societal forces, public health, and the impact of the technology that surrounds us. These insights are essential for informing policy in the digital age, and researchers and organizations have called for ethical guidelines to help ensure this work is done responsibly. However, the risks to the rights and welfare of individuals associated with the use of pervasive data for research are nuanced and context-specific (Shilton et al, 2021).

The NTIA's Request for Comment on Ethical Guidelines

The recognition of both the importance of research with pervasive data and the privacy and risks to individuals has been an ongoing conversation among academics, civil society organizations, legislators, and international governments. In December 2025, the National Telecommunications and Information Administration (NTIA), a U.S. agency tasked with advising the President on telecommunications and information policy issues, published a public request for comment (RFC) regarding Ethical Guidelines for Research using Pervasive Data (NTIA, 2024).

The RFC prompted public feedback on whether national ethical guidelines should be developed for research using pervasive data, addressing both possible benefits and drawbacks. It sought input on how to define pervasive data, the barriers to accessing it, and what types of data would be most valuable for public interest research. Key ethical concerns addressed in the RFC included consent and autonomy, privacy risks, and mitigation strategies throughout the research lifecycle—from data acquisition to dissemination. The RFC also prompted feedback on legal and regulatory considerations, the role of existing ethical frameworks, and how guidelines should evolve with technological advancements. Ultimately, it was designed to get concrete feedback from various stakeholders on the potential for establishing a structured, responsible approach to using pervasive data in research while balancing ethical, legal, and practical concerns.

Fifty responses to the RFC were received from stakeholders across various industries and sectors, and all responses are publicly available as part of the U.S. regulatory process.

Key Stakeholder Perspectives

This report aims to use the perspectives shared through the RFC commenting process as a starting point to consider: a) what are the most salient points of tension among stakeholders regarding government issued ethics guidelines for researchers using pervasive data? And b) what actions (writing, organizing, piloting, convening, etc) can non-government stakeholders take over the next few years to both improve ethical practices in the field and increase the chance that government attempts (whenever they may be) to draft such guidelines are successful?

Our work is not an attempt to write guidelines or best practices, as other non-governmental organizations have and continue to make progress in the efforts. Additionally, this report is not an analysis of the challenges researchers face when trying to obtain pervasive data, again many organizations and councils are working on this. Rather, seek to analyze the public responses themselves to provide additional insight into how key stakeholders frame the ethical challenges of research – and attempts to regulate research – relying on pervasive data.

A preliminary analysis of the RFC responses highlight five key tensions including: conflicts and ambiguities within the existing legal landscape (including voluntary frameworks), challenges government-issued guidelines may face balancing the need for consistency and flexibility to account for contextual factors, balancing the ethical call to protect human autonomy with the opacity and massiveness of datasets used in research with pervasive data, and challenges with anonymization and participatory research practices.

The RFC responses also include several actions the academic and civil society can take to continue progress towards trusted ethics and privacy practices in research that uses pervasive data and improve the knowledge base that eventual government-issued guidance may rely on. Actions range from training researchers and review boards, piloting alternative review processes and publishing case studies, increasing public awareness of internet research, conducting legal analysis to compare guidelines with data protection regulations around the globe, design, test and advocate for better platform designs to support autonomy (choice screens), and better tools for users to donate data to researchers.

Next Steps: Advancing Ethical Research Practices

The insights gathered from the RFC responses highlight both the complexity and urgency of developing ethical guidelines for pervasive data research. While government-issued guidelines may take time to materialize, stakeholders—including researchers, civil society organizations, and technology companies—can take immediate steps to strengthen ethical practices. Our analysis helps highlight possible next steps, including improving researcher training, piloting alternative review processes, increasing public awareness of internet research ethics, and developing platform tools to enhance user autonomy can help build a stronger foundation for responsible data use. By proactively addressing these challenges, non-governmental actors can shape the future of ethical pervasive data research and contribute to a regulatory landscape that balances innovation, privacy, and public interest.

References

NTIA (National Telecommunications and Information Administration) (2024). Ethical Guidelines for Research Using Pervasive Data. 89 FR 99844. https://www.federalregister.gov/d/2024-29064

Shilton, K., Moss, E., Gilbert, S. A., Bietz, M. J., Fiesler, C., Metcalf, J., Vitak, J., & Zimmer, M. (2021). Excavating awareness and power in data science: A manifesto for trustworthy pervasive data research. *Big Data & Society, 8*(2). https://doi.org/10.1177/20539517211040759

ETHICS IN TRANSITION: ADAPTING RESEARCH GOVERNANCE TO A PROPOSED NEW LEGAL FRAMEWORK IN SWEDEN

Ylva Hård af Segerstad
University of Gothenburg

Thomas Hartvigsson
University of Gothenburg

Annika Bergviken-Rensfeldt
University of Gothenburg

Thomas Hillman
University of Gothenburg

Introduction

As the landscape of internet research evolves, ethical guidelines, policies, and laws remain crucial for ensuring responsible and accountable scholarship. However, the implementation of ethics policies is neither uniform nor static—it is subject to continual negotiation among researchers, institutions, regulators, and ethical review bodies (Carniel et al., 2023; Israel, 2020). This paper contributes to exploring how emerging regulatory frameworks impact research ethics in action, using Sweden's proposed new research ethics law as a case study of ethical governance in transition.

Sweden's reform of the Ethical Review Act (2003:460, 2003)—set to be replaced in 2026—marks a significant shift towards a decentralized model of ethical oversight. Under the proposed law, responsibility for ethical assessments will shift from national ethics boards to individual research institutions, affecting the governance of digital and internet research, social science methodologies, and interdisciplinary studies. While the reform is intended to reduce administrative burdens and increase research flexibility, it raises new ethical and procedural challenges, including institutional disparities, legal ambiguities, and the risk of inconsistent application.

Research Ethics Governance: Standardization vs. Local Adaptation

At the core of the transition proposed for the Swedish ethical review system is the longstanding tension between universalist and contextualist approaches to research ethics: A universalist approach seeks to establish standardized ethical frameworks that apply equally across disciplines and institutions, ensuring consistency and reducing the risk of ethical misconduct (Evanoff, 2004). A contextualist (or relativist) perspective argues that ethical standards must remain flexible and adaptive, allowing institutions and researchers to tailor ethical frameworks to specific research contexts, methodologies, and disciplinary needs. While there are overarching similarities in how the governance of research ethics is organized, the ways in which policies are interpreted and applied vary across institutions, disciplinary fields, and individual researchers. This diversity becomes particularly evident when new regulations and procedures are introduced, as different academic fields and institutional cultures bring distinct perspectives on how such frameworks should be organized. This research seeks to explore these tensions and examine how research ethics policies are negotiated and adapted in practice and can be better supported infrastructurally. Our case is based in the Swedish context, currently mobilizing and organizing for a new research ethics law Offices (2024). In particular, we are interested in hybrid research institutional settings, working in inter-disciplinary fields, often in the intersection of social sciences and other disciplinary fields and with applied research, and where research ethics is more challenging and less institutionally developed (Mathur et al., 2019). The focus is both highly relevant and timely, and from under researched from a disciplinary and institutional perspective of tensions of standpoints and disciplinary and institutional diversity.

The proposed Swedish law moves toward a more contextualist model - which is more in line with the approach to research ethics advocated by the ethical guidelines developed by AoIR - granting research institutions autonomy in developing their own ethical guidelines and conducting internal ethics reviews for low-risk studies. While this shift aligns with calls for greater disciplinary sensitivity, it also shifts responsibility to local ethical review boards and introduces concerns about fragmentation, legal uncertainty, and disparities between institutions.

The Swedish Case: From National Regulation to Institutional Autonomy

Sweden's current research ethics framework has faced extensive criticism, particularly from researchers in internet studies, social sciences, and the humanities, who argue that existing regulations are overly rigid and biomedical-centric, failing to accommodate ethical complexities of qualitative research, particularly in digitally saturated or data-driven contexts (Buchanan, 2017; Markham, In press). The existing system is administratively burdensome, requiring ethical approval even for studies involving publicly available or anonymized data. Furthermore, legal punishments to breaches of ethical review requirements are often disproportionate to the offense.

In response to these challenges, the Swedish government launched an inquiry in 2023, leading to a proposal for a new Research Ethics Act, likely to come into effect 1 January 2026. In a report (von Essen & Niklewski Nilsson, 2024), the inquiry proposed a number of changes to the existing ethical review act, notably the following:

Exempting certain low-risk research from mandatory ethics review—including studies using publicly available data or data from consenting adults, provided they pose minimal risk to personal integrity.

Decentralizing ethical review, shifting responsibility from national ethics boards to individual research institutions, which will establish internal guidelines and review procedures.

Reforming oversight mechanisms, transitioning Sweden's Ethics Review Appeals Board (ÖNEP) from a punitive authority to a compliance-support function, reducing legal sanctions for researchers.

Clarifying ethical review requirements, maintaining strict oversight for high-risk research (e.g., studies involving physical interventions, biomedical data, or vulnerable populations).

In November 2024, the report was sent out for consideration to relevant consultation bodies in Sweden with the request for opinions on the proposal. While the reform is intended to reduce administrative burdens and increase research flexibility, it raises new ethical and procedural challenges, including institutional disparities, legal ambiguities, and the risk of inconsistent application. The main points of criticism and concern regarding the proposed new ethics law for research involving human subjects that have been raised during the consultation include:

- Unclear Implementation and Increased Bureaucratic Burden
- Lack of Uniformity in Ethical Standards
- Interpretation Issues ("Minimal Risk" and Ethical Exemptions)
- Challenges for Multidisciplinary and International Research
- Educational and Competency Gaps
- Potential for Legal and Ethical Ambiguities
- Concerns About Research Integrity and Public Trust

Internet Research and the Future of Ethical Review

This study explores the transition proposed for the Swedish ethical review system, examining how research ethics policies are negotiated and adapted in practice. Specifically, it investigates how hybrid research institutional settings—which operate across interdisciplinary fields, at the intersection of social sciences and other disciplines, and within applied research contexts—mobilize and organize in response to these changes. The study is both highly relevant and timely, given the ongoing restructuring of ethical governance in Sweden and the broader implications for research autonomy, institutional responsibility, and disciplinary diversity. Despite its significance, this area remains under-researched, particularly from a disciplinary and institutional perspective, where tensions between universalist and contextualist standpoints and between standardized regulations and local adaptations shape how research ethics is interpreted and implemented. By focusing on institutional and disciplinary diversity, this study contributes to a broader understanding of how research ethics evolves within complex and dynamic academic environments.

Sweden's transition toward localized research ethics governance also highlights broader debates on ethical oversight in internet research. It raises critical questions for the AoIR community, particularly regarding the role of institutional autonomy in shaping ethical governance, the risks and benefits of context-sensitive ethical frameworks in digital research and the balance between ethical accountability and academic freedom in emerging regulatory models.

References

Lag om etikprövning av forskning som avser människor [Act concerning the ethical review of research involving humans], (2003).

Buchanan, E. (2017). Internet research ethics: Twenty years later. *Internet research ethics for the social age: New challenges, cases, and contexts*, xxix-xxxiii.

Carniel, J., Hickey, A., Southey, K., Brömdal, A., Crowley-Cyr, L., Eacersall, D., Farmer, W., Gehrmann, R., Machin, T., & Pillay, Y. (2023). The ethics review and the humanities and social sciences: disciplinary distinctions in ethics review processes. *Research Ethics*, *19*(2), 139-156.

Evanoff, R. (2004). Universalist, Relativist, and Constructivist Approaches to Intercultural Ethics. *International Journal of Intercultural Relations - INT J INTERCULT RELAT*, *28*, 439-458. https://doi.org/10.1016/j.ijintrel.2004.08.002

Israel, M. (2020). Organizing and contesting research ethics: The global position. *Handbook of research ethics and scientific integrity*, 51-65.

Markham, A. (In press). Ethics. In H. Mork Lomell & M. Kaufman (Eds.), *Handbook on Digital Criminology* (pp. pp forthcoming). De Gruyter Press.

Mathur, A., Lean, S. F., Maun, C., Walker, N., Cano, A., & Wood, M. E. (2019). Research ethics in inter-and multi-disciplinary teams: Differences in disciplinary interpretations. *PLoS ONE*, *14*(11), e0225837.

A new law on ethical requirements and ethical review of research involving humans, (2024). https://www.regeringen.se/rattsliga-dokument/departementsserien-och-promemorior/2024/10/ds-202421/

von Essen, U., & Niklewski Nilsson, E. (2024). *En ny lag om forskningsetiska krav på och etikprövning av forskning som avser människor* (Departementsserien, Issue.

WHEN IS SCRAPING LEGITIMATE? ETHICAL, LEGAL, ADMINISTRATIVE, AND TECHNICAL CONSIDERATIONS

Megan Brown
University of Michigan

Andrew Gruen
Mozilla Foundation

Gabe Maldoff

Goodwin Procter

Solomon Messing
New York University

Michael Zimmer
Marquette University

Abstract

Scientists across disciplines often use data from the internet to conduct research, generating valuable insights about human behavior. However, as generative AI 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 amidst this evolving data access landscape.

Introduction

Web scraping has become an essential tool for social science research, particularly in response to increasing platform restrictions on data access. Platforms such as Twitter (now X) and Meta have imposed significant financial and procedural barriers to accessing public data through official channels, compelling researchers to turn to alternative data collection methods (XDevelopers, 2023; CrowdTangle, 2024). While web scraping provides valuable insights, it raises numerous legal, ethical, and scientific questions. This paper outlines a structured framework for responsible web scraping, emphasizing key considerations that researchers must address.

Legal Considerations

Legal frameworks surrounding web scraping are fragmented and evolving. Researchers must navigate multiple layers of legal restrictions, including contractual terms of service, federal and state statutes, and international data protection regulations. Platforms often include prohibitions against scraping in their terms of service, yet courts have scrutinized the enforceability of such agreements, especially when data is publicly accessible (hiQ Labs, Inc. v. LinkedIn Corporation, 2022). The Computer Fraud and Abuse Act (CFAA) is one of the primary laws governing unauthorized access to computer systems, and recent court rulings indicate that scraping publicly available data does not constitute a CFAA violation, whereas accessing password-protected or restricted data may lead to legal consequences (Sandvig v. Barr, 2020).

Privacy and data protection laws further complicate the landscape. While U.S. laws such as the California Consumer Privacy Act (CCPA) impose requirements on the collection and use of personal data, international regulations like the European Union's General Data Protection Regulation (GDPR) mandate stricter protections. Researchers must ensure compliance by establishing a legal basis for data processing. The EU's Digital Services Act (DSA) has also introduced mechanisms that provide researchers with rights to access public platform data, though similar legal frameworks have yet to emerge in the U.S. As a result, researchers must remain vigilant in monitoring regulatory developments and ensure that their data collection methods comply with evolving legal standards.

Ethical Considerations

The ethical implications of web scraping revolve around privacy, consent, and the responsible use of collected data. One of the fundamental concerns is informed consent, as researchers often collect data without direct participant approval. Unlike traditional human subjects research, where consent is explicitly obtained, web scraping involves analyzing content that users may not have anticipated would be used for research. Studies have shown that many social media users are unaware that their publicly available posts may be utilized by researchers (Fiesler & Proferes, 2018). Furthermore, vulnerable populations, including marginalized groups, require additional ethical scrutiny to ensure their data is not used in ways that could cause harm (Zimmer, 2018).

Another ethical concern is the need to preserve the context in which data was originally shared. Decontextualizing social media posts or other online content can lead to misinterpretation and misrepresentation (Shilton et al., 2021). Ethical frameworks recommend that researchers critically reflect on the potential consequences of their work and strive to maintain the integrity of the original data. Additionally, web scraping should be conducted in a manner that does not place undue strain on platform resources. Researchers should implement best practices, such as rate limiting and responsible crawling behaviors, to minimize disruption to online services (Buchanan & Zimmer, 2021).

Institutional Considerations

Navigating institutional policies is a critical aspect of conducting research involving web scraping. Institutional Review Boards (IRBs) play a central role in assessing whether a research project involving scraped data constitutes human subjects research. However, different IRBs may interpret regulations in varying ways, leading to inconsistencies in approval processes (Vitak et al., 2017). Researchers must proactively engage with IRBs to understand institutional guidelines and ensure compliance with ethical and legal standards.

Legal counsel within universities, typically housed within the Office of General Counsel (OGC), can provide guidance on the potential risks associated with web scraping. However, OGCs primarily serve to protect the institution rather than individual researchers, necessitating external legal advice in cases of high legal exposure. Technical review is another crucial consideration, as proper data management practices, including secure storage and privacy-preserving techniques, are essential for

maintaining compliance. Research IT support teams, where available, can offer assistance in implementing best practices for secure data handling.

Scientific Considerations
Web scraping introduces methodological challenges that can affect the validity and reliability of research findings. One of the most significant scientific concerns is sampling bias. Because researchers do not always have control over the structure of the data they collect, scraped data may not be representative of broader populations (González-Bailón et al., 2014). Sampling strategies must be carefully designed to mitigate bias and ensure that findings accurately reflect the subject of study.

Another challenge is data completeness. Missing data due to platform restrictions, rate limits, or technical failures can introduce systematic biases that may compromise the validity of research conclusions (Wu & Taneja, 2021). Platforms frequently modify their interfaces and algorithms, sometimes without notice, which can affect data collection processes and lead to inconsistencies over time (Munger, 2023). Construct validity is also a concern, as researchers must ensure that their measurements accurately capture the intended variables. Many online interactions, such as content impressions, may not necessarily indicate meaningful engagement, raising questions about the validity of research conclusions based on such metrics.

Recommendations
To navigate the complexities of web scraping, researchers should adopt best practices that balance legal compliance, ethical considerations, institutional engagement, and scientific rigor. It is essential to focus on scraping only publicly accessible data to minimize legal risks while staying informed about evolving regulations, particularly under the EU's DSA. Ethical research practices should include conducting thorough risk assessments, minimizing data collection to what is necessary, and implementing safeguards for vulnerable populations.

Engaging with institutional stakeholders such as IRBs, legal counsel, and technical support teams is critical to ensuring compliance with internal policies and mitigating risks. Moreover, researchers must prioritize scientific rigor by clearly defining sampling strategies, accounting for missing data, and transparently reporting methodological limitations. To guide researchers, we provide a set of critical questions researchers should consider to ensure that research is scientifically rigorous and ethical while minimizing legal risks when using scraping as a data collection tool.

References
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

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

González-Bailón, S., Wang, N., Rivero, A., Borge-Holthoefer, J., & Moreno, Y. (2014). Assessing the bias in samples of large online networks. *Social Networks, 38*, 16–27. https://doi.org/10.1016/j.socnet.2014.01.004

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

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

Vitak, J., Proferes, N., Shilton, K., & Ashktorab, Z. (2017). Ethics Regulation in Social 31 Computing Research: Examining the Role of Institutional Review Boards. *Journal of Empirical Research on Human Research Ethics, 12*(5), 372–382. https://doi.org/10.1177/15562646177252

Wu, A. X., & 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. https://doi.org/10.1177/1461444820933547

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

RESEARCH AND PRIVACY ON WIKIPEDIA

Eli Asikin-Garmager
Wikimedia Foundation

Michael Zimmer
Marquette University

Cameran Ashraf
Wikimedia Foundation

Leila Zia
Wikimedia Foundation

Introduction

Wikipedia is one of the most widely used online resources, with over 15 billion page views per month. It is a cornerstone of digital knowledge and an essential reference for research across various disciplines. Each year, researchers cite Wikipedia in over 130,000 academic papers, with at least 500 studies focusing directly on the platform itself. While Wikipedia's openness has been instrumental in advancing knowledge, it also introduces significant ethical and privacy challenges, especially when research involves analyzing user behavior, contributions, and community interactions.

The unique nature of Wikipedia means that research on the platform must balance the pursuit of knowledge with the responsibility of protecting contributors' privacy. Wikipedia editors operate under varying degrees of anonymity and in diverse political, social, and legal contexts. Some contributors openly share personal details, while others rely on strict anonymity due to potential risks such as political persecution, harassment, or professional repercussions. The transparency of Wikipedia's edit histories, user discussions, and metadata further complicates privacy considerations. Researchers must navigate these challenges carefully to ensure that their work does not inadvertently expose contributors to harm.

This paper provides guidance for navigating ethical tensions while conducting research with Wikipedia data. We offer recommendations about what to pay attention to and how to navigate some of the questions we expect Wikipedia contributors and researchers face when interfacing with or conducting research projects.

Ethical Challenges in Wikipedia Research

Privacy is a fundamental ethical concern in research, particularly when it involves human subjects (U.S. Department of Health & Human Services, 1979; franzke et al, 2020). While research on Wikipedia might not fit traditional human subject research paradigms, various ethical issues still arise due to its unique nature.

One core ethical challenge is the assumption that publicly available data is inherently fair game for research (Zimmer, 2018). Although Wikipedia's policies allow open access to data, many contributors do not expect their activities to be monitored or analyzed in ways that could identify them or expose sensitive information. One of the key risks in Wikipedia research is re-identification—the ability to infer personal details about an editor based on patterns in their contributions. Even if a researcher does not directly publish identifiable information, the aggregation and cross-referencing of data with external sources can expose user identities. Studies have demonstrated that it is possible to predict personal traits such as gender, political affiliation, or geographic location based on editing patterns alone (Rizoiu et al, 2016). This concern grows as machine learning models improve, making it easier to potentially correlate Wikipedia activity with other publicly available datasets.

Wikipedia's privacy challenges extend beyond edit histories. User discussions on talk pages, participation in community forums, and username choices all provide additional data points that can contribute to personal identifiability. For example, researchers analyzing discussions on contentious topics may unintentionally expose editors involved in politically sensitive debates, putting them at risk in their home countries.

Unlike platforms such as Twitter or Facebook, which have established guidelines for academic use of their data, Wikipedia does not have a centralized ethical framework for researchers to follow. While some guidance exists—such as the Wikimedia Foundation's Terms of Use and the Universal Code of Conduct—these do not specifically address the ethical nuances of academic research. As a result, researchers often rely on general ethical principles from their institutions, such as the Common Rule in the United States, or guidelines from scholarly organizations like the Association of Internet Researchers (AoIR). However, these frameworks do not always account for the unique tensions that arise when studying a collaborative knowledge platform like Wikipedia.

Possible Tensions Between Researchers and the Wikipedia Community

We identify numerous points of possible tension between researchers and the Wikipedia community. Many Wikipedia editors are unaware that their activities are subject to research, while researchers might lack awareness of the values, norms, and expectations that govern Wikipedia's editorial community.

A common source of tension arises when researchers collect and analyze data in ways that violate community expectations of privacy. For example, while Wikipedia's edit histories are publicly available, contributors may not expect their contributions to be aggregated in large-scale studies, particularly when the research reveals patterns that could be used to infer personal details. This disconnect has led to conflicts between researchers and the Wikimedia community, sometimes requiring intervention from Wikipedia administrators or the Wikimedia Foundation's Llegal team.

Another consideration is the handling of usernames in research publications. Some researchers choose to anonymize usernames or use pseudonyms, recognizing that usernames—while public—carry significant social and reputational weight within the

Wikipedia community. Others argue that usernames should be treated like real names if they are central to research findings. This debate mirrors broader discussions in research ethics about when and how to anonymize online identities, especially in studies involving politically sensitive topics or vulnerable populations.

Researchers might also face challenges in engaging with the Wikipedia community. Unlike traditional research subjects, Wikipedia contributors are part of a self-governing collective with established norms around transparency, neutrality, and collaboration. When researchers fail to communicate their intentions or seek community input, they risk alienating contributors and undermining trust in academic research.

Recommendations for Researchers and Wikipedia Contributors

To foster ethical research practices and reduce tensions between researchers and Wikipedia contributors, our paper makes several recommendations.

Recommendations for Researchers

1. *Adhering to Wikipedia Policies*: Researchers must comply with Wikipedia's core policies, including its Universal Code of Conduct, Terms of Use, and privacy protections against doxxing. These policies emphasize user safety, informed consent, and the need to respect anonymity.

2. *Communicating with the Wikipedia Community*: Researchers should create public project pages on metawiki to disclose their study's objectives, methodologies, and affiliations. This transparency allows Wikipedia contributors to assess potential privacy risks before engaging with researchers.

3. *Protecting User Anonymity and Learning about Community Norms and Values*: Researchers should avoid publishing usernames without explicit consent, paraphrase quotes from user discussions instead of providing direct attributions, and ensure data analysis does not enable de-anonymization of contributors. Researchers should understand key characteristics of project participants, including geographic distribution and local context in which they participate.

4. *Managing Ethical Trade-Offs*: While research should be transparent, the ethical obligation to protect users' privacy should take precedence. If publishing certain findings could expose editors to harm, researchers should consider alternative reporting methods, such as aggregating data instead of providing individual case studies.

Recommendations for Wikipedia Contributors:

1. *Understanding Privacy Risks*: Wikipedia editors should recognize that their public edit histories can be analyzed to infer personal traits. To minimize risks, they can consider using pseudonymous usernames, limiting personal disclosures on user pages, and be mindful of editing patterns that could reveal sensitive affiliations.

2. *Navigating Researcher Interactions:* Editors should critically evaluate research projects before participating, ensuring they understand the researcher's institutional affiliations and ethical oversight. If concerned about privacy, they can request additional information or decline participation.

3. *Reporting Violations*: If an editor believes their privacy has been compromised due to research activities, they can report concerns to Wikipedia administrators or Wikimedia's legal team, request oversight removal of sensitive data, and contact the Wikimedia Human Rights team for additional protections as needed.

Conclusion

Wikipedia research presents both opportunities and ethical challenges. While academic studies contribute to our understanding of online collaboration, knowledge production, and digital communities, they also introduce risks to contributor privacy. Prior to this research, the lack of a standardized ethical framework for Wikipedia research had led to tensions between scholars and the Wikipedia community, highlighting the need for clearer guidelines and stronger privacy protections.

Through this white paper and its guidance, we seek to foster greater transparency, ethical awareness, and dialogue between researchers and Wikipedia contributors, and thereby advance knowledge while respecting user privacy.

References

franzke, aline shakti, Bechmann, Anja, Zimmer, Michael, Ess, Charles and the Association of Internet Researchers (2020). "Internet Research: Ethical Guidelines 3.0." https://aoir.org/reports/ethics3.pdf

Rizoiu, Marian-Andrei, Lexing Xie, Tiberio Caetano, and Manuel Cebrian. (2016) "Evolution of privacy loss in wikipedia." In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pp. 215-224.

U.S. Department of Health & Human Services. (1979). "The Belmont report: Ethical principles and guidelines for the protection of human subjects of research." Office for Human Research Protections. https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/index.html

Zimmer, Michael. (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