Journal of Frontiers in Multidisciplinary Research

The Ethics of Web Scraping In Research: A Review: Investigating the Boundaries, Legal Implications, and Societal Acceptance of Web Scraping As a Data Collection Method

Iboro Akpan Essien [1*], Ayobami Oluwadamilola Adebayo [2], Afeez A Afuwape [3], Bukky Okojie Eboseremen [4], Farid Oladega [5], Olabode Michael Soneye [6]

[1] Ekhiam Ventures Limited, Port Harcourt, Nigeria
[2] Independent Researcher, Australia
[3] University of the Oulu, Finland
[4] Tecvinson AB, Malmö, Sweden
[5] University of Electronic Science and Technology of China, Chengdu
[6] Cogeco, Ontario, Canada

* Corresponding Author: **Iboro Akpan Essien**

Article Info

E-ISSN: 3050-9726
P-ISSN: 3050-9718
Volume: 04
Issue: 01
January - June 2023
Received: 02-04-2023
Accepted: 03-05-2023
Published: 07-06-2023
Page No: 529-538

Abstract

This study provides a comprehensive review of the ethical, legal, and societal aspects of web scraping as a data collection method in modern research. The primary objective is to explore the boundaries, legal implications, and societal acceptance of web scraping, aiming to guide its responsible and sustainable use. Employing a systematic literature review methodology, the study analyzes peer-reviewed articles published between 2010 and 2023, focusing on interdisciplinary perspectives. The methodology involves a detailed search strategy, strict inclusion and exclusion criteria, and a thematic synthesis of the selected literature. Key findings reveal that ethical concerns in web scraping predominantly revolve around privacy, consent, and data ownership, while legal challenges are characterized by the ambiguity of international laws and regulations. Societal acceptance varies, influenced by the perceived benefits and risks of web scraping, with public opinion playing a significant role. The future outlook suggests evolving practices and norms in web scraping, driven by advancements in AI and machine learning, yet accompanied by emerging ethical and legal complexities. The study concludes with a call for responsible and ethical web scraping practices, emphasizing the need for a balanced approach that respects privacy, adheres to legal standards, and aligns with societal values. Future research directions are identified, focusing on filling gaps in understanding the ethical, legal, and societal implications of web scraping. This study contributes to the development of informed guidelines and practices in the evolving field of web scraping, ensuring its alignment with ethical standards, legal requirements, and societal expectations.

DOI: https://doi.org/10.54660/.JFMR.2023.4.1.529-538

Keywords: Web Scraping, Ethical Implications, Legal Frameworks, Societal Acceptance

1. Introduction

1.1. The Emergence of Web Scraping in Modern Research

The advent of web scraping as a pivotal tool in modern research marks a significant evolution in data collection methodologies. Web scraping, fundamentally, is an automated process aimed at extracting large volumes of data from websites. This technique has revolutionized the way researchers approach data-driven studies, offering a new dimension to the acquisition and analysis of information from the vast expanse of the internet (Bale *et al*., 2022).

The emergence of web scraping can be traced back to the early days of the internet when the need for efficient data collection methods became apparent. With the exponential growth of online content, traditional methods of data collection proved inadequate in handling the sheer volume and diversity of data available online. Web scraping emerged as a solution to this challenge, enabling researchers to gather large datasets from various web sources efficiently and effectively (Agrawal, 2023).

The application of web scraping in research spans across various domains, including market research, social media analysis, competitive intelligence, and academic studies. Its ability to automate the extraction process has not only saved time and resources but also opened up new avenues for research that were previously unfeasible due to the limitations of manual data collection methods (Landers *et al*., 2016).

One of the key factors contributing to the rise of web scraping in research is its versatility. The technique can be adapted to scrape data from different types of websites, ranging from simple static pages to complex dynamic web applications. This flexibility allows researchers to tailor their scraping methods to suit the specific requirements of their study, ensuring the collection of relevant and high-quality data (Bale *et al*., 2022).

Moreover, the evolution of web scraping tools and technologies has played a crucial role in its widespread adoption. The development of sophisticated scraping software and libraries, such as Selenium and Python-based tools, has made it easier for researchers to implement complex scraping tasks. These tools offer advanced features like handling JavaScript, managing cookies, and bypassing anti-scraping measures, thereby enhancing the efficiency and effectiveness of the scraping process (Agrawal, 2023).

The ethical and legal considerations surrounding web scraping have also been a topic of ongoing debate in the research community. While web scraping is not inherently illegal, it raises concerns regarding privacy, consent, and data ownership. Researchers engaging in web scraping must navigate these ethical and legal boundaries to ensure compliance with relevant laws and respect for individual privacy rights (Landers *et al*., 2016).

The emergence of web scraping in modern research represents a significant advancement in data collection methodologies. Its ability to efficiently gather large volumes of data from the web has opened up new possibilities for research across various fields. As web scraping continues to evolve, it is essential for researchers to remain cognizant of the ethical and legal implications associated with its use, ensuring responsible and compliant data collection practices.

1.2 Defining Web Scraping: Techniques and Applications

Web scraping, also known as web crawling, is a pivotal technique in the digital age, enabling the automated extraction of data from websites. This process has become increasingly important in various fields, including business intelligence, data science, and cybersecurity. By extracting data from HTML text, web scraping transforms unstructured web content into structured, machine-readable formats, facilitating the efficient retrieval and analysis of information (Bhatt *et al*., 2023).

The operational mechanisms of web scraping involve several stages, starting with the identification of target websites and the specific data to be extracted. This is followed by the deployment of web scraping tools or software, which navigate the web pages, extract the required data, and store it in a structured format for further analysis. Technologies such as spidering and pattern matching play a crucial role in these processes, enabling the scraping tools to identify and retrieve relevant data efficiently (Khder, 2021).

Web scraping techniques vary in complexity, ranging from simple scripts that can extract data from static web pages to sophisticated systems capable of interacting with dynamic websites, handling JavaScript, and even bypassing anti-scraping measures. The choice of technique depends on the complexity of the website and the nature of the data being sought. Python, with its rich set of libraries and frameworks, has emerged as a popular language for implementing web scraping projects due to its simplicity and versatility (Singrodia *et al*., 2019).

The applications of web scraping are diverse and far-reaching. In business intelligence, it is used to gather market data, track competitor activities, and monitor consumer trends. In data science, web scraping facilitates the collection of large datasets for analysis and modeling. In the realm of cybersecurity, it assists in gathering intelligence on cyber threats and vulnerabilities. Additionally, web scraping has found applications in fields such as e-commerce, healthcare, journalism, and academia, where it enables the extraction of valuable insights from the vast amounts of data available online (Bhatt *et al*., 2023).

Despite its numerous benefits, web scraping is not without its challenges. Ethical and legal issues are at the forefront of concerns surrounding web scraping. The process often raises questions about privacy, consent, and data ownership. It is crucial for practitioners to navigate these ethical and legal boundaries responsibly, ensuring that their scraping activities comply with relevant laws and respect individual privacy rights (Khder, 2021).

Therefore, web scraping is a powerful technique that has transformed the way data is collected and analyzed in the digital age. Its ability to automate the extraction of vast amounts of data from the web has made it an indispensable tool in various fields. As web scraping continues to evolve, it is essential for practitioners to remain aware of the ethical and legal implications of their activities, ensuring that they engage in responsible and compliant data collection practices.

1.3 Historical Overview: The Evolution of Web Scraping in Research

The history of web scraping is intrinsically linked to the evolution of the World Wide Web itself. As the internet transformed into a ubiquitous source of data, the need for efficient methods to extract this data became increasingly apparent. Web scraping, initially a rudimentary practice, has evolved into a sophisticated technique employed across various sectors, including business, academia, and journalism (Saurkar, Pathare, & Gode, 2018).

In the early stages of the internet, data extraction was primarily a manual process, involving simple copy-and-paste methods. This approach was not only time-consuming but also prone to errors, limiting the scope and reliability of data collection. The advent of automated web scraping techniques marked a significant shift, offering a more efficient and accurate means of data extraction. These techniques ranged from basic text grabbing and regular expression matching to more advanced methods involving HTTP programming and

HTML parsing (Saurkar, Pathare, & Gode, 2018).

The turn of the century witnessed a surge in the development of web scraping tools and technologies. This period saw the emergence of specialized web scraping software, vertical aggregation platforms, and semantic annotation recognizing tools, which significantly enhanced the capability to extract and analyze web data. The introduction of these tools democratized data access, enabling users with varying levels of technical expertise to engage in web scraping activities (Namoun *et al*., 2020).

The mid-2000s to the present day represents a period of rapid advancement in web scraping technologies. The proliferation of dynamic and complex websites necessitated the development of more sophisticated scraping tools capable of handling JavaScript, managing cookies, and circumventing anti-scraping measures. This era also witnessed the integration of machine learning and artificial intelligence in web scraping, further expanding its potential applications (Weber & Napoli, 2018).

In the realm of academia, web scraping has become an invaluable tool for researchers. It has enabled the collection of large datasets from the web, which are used in various fields such as social sciences, economics, and computer science. The ability to scrape data from online sources has opened new avenues for research, allowing scholars to conduct studies that were previously unfeasible due to data limitations (Weber & Napoli, 2018).

Journalism is another field that has been profoundly impacted by the evolution of web scraping. Journalists have leveraged web scraping to gather information from a multitude of online sources, including news websites, social media platforms, and online forums. This has enhanced the depth and breadth of journalistic research, enabling more comprehensive and data-driven reporting (Weber & Napoli, 2018).

Despite its widespread adoption, web scraping has also raised ethical and legal concerns. Issues related to privacy, consent, and data ownership have sparked debates among practitioners and policymakers. The evolution of web scraping has, therefore, been accompanied by a growing awareness of the need for responsible scraping practices that respect legal boundaries and ethical norms (Namoun *et al*., 2020).

In summary, the history of web scraping is a testament to the rapid advancements in technology and the increasing importance of data in the digital age. From its humble beginnings as a manual, error-prone process, web scraping has evolved into a sophisticated technique employed across various fields. As web scraping continues to develop, it is imperative for practitioners to balance the benefits of data extraction with the ethical and legal considerations that accompany its use.

1.4 Aim and Objectives of the Study

The aim of this study is to comprehensively analyze and understand the ethical, legal, and societal dimensions of web scraping as a data collection method in modern research. This involves exploring the boundaries of web scraping practices, their legal implications, and societal acceptance, with a focus on developing insights that can guide responsible and sustainable use of web scraping technologies.

The research objectives are;

1. To investigate the ethical boundaries of web scraping.
2. To assess societal acceptance and public perception.
3. To synthesize ethical, legal, and societal perspectives to provide a holistic view of web scraping.

2. Methodology

The methodology for this study is structured as a systematic literature review, focusing on the ethical, legal, and societal aspects of web scraping. This approach ensures a comprehensive and unbiased collection and analysis of relevant literature.

2.1 Data Sources

The primary data sources for this study include academic databases, online journals, and digital libraries. Key databases such as PubMed, IEEE Xplore, Google Scholar, and Web of Science were utilized. These sources provide access to a wide range of peer-reviewed articles, conference papers, and academic journals that cover the interdisciplinary aspects of web scraping.

2.2 Search Strategy

The search strategy involves using specific keywords and phrases related to web scraping, such as "web scraping," "data collection ethics," "legal implications of web scraping," and "societal acceptance of web scraping." Boolean operators (AND, OR) are used to combine these terms effectively. The search is conducted within the titles, abstracts, and keywords of the publications to ensure relevance.

2.3 Inclusion and Exclusion Criteria for Relevant Literature.

The inclusion and exclusion criteria for relevant literature in this study were meticulously defined to ensure a focused and comprehensive review of the ethical, legal, and societal aspects of web scraping. For inclusion, the study primarily targeted peer-reviewed articles published between 2010 and 2023, ensuring the relevance and timeliness of the information. The focus was on studies that specifically addressed the ethical boundaries, legal implications, and societal acceptance of web scraping. This included articles providing empirical data, case studies, or theoretical analyses related to web scraping practices. Exclusion criteria were set to maintain the study's specificity and quality. Non-peer-reviewed articles and grey literature were excluded to ensure the academic rigor of the sources. Publications outside the specified date range were also excluded to maintain the contemporary relevance of the study. Additionally, studies not directly related to web scraping or its ethical, legal, and societal implications were omitted. This was done to keep the review focused on the core aspects of web scraping and avoid diluting the findings with peripheral or unrelated data. These criteria were instrumental in shaping a well-defined and targeted literature review, providing a solid foundation for the study's findings and conclusions.

2.4 Selection Criteria

The selection process involves a two-stage screening. Initially, titles and abstracts are reviewed to assess relevance based on the inclusion and exclusion criteria. Subsequently, full-text articles are examined for detailed analysis. The selection criteria also consider the quality of the research methodology, the relevance of the findings to the study's aim, and the diversity of perspectives presented.

2.5 Data Analysis

Data analysis involves a thematic synthesis of the selected literature. Key themes and patterns related to the ethical, legal, and societal dimensions of web scraping are identified

and categorized. The analysis focuses on extracting insights, trends, challenges, and recommendations from the literature. The findings are then synthesized to provide a comprehensive understanding of the current state of web scraping practices and to identify areas for future research.

By employing this systematic literature review methodology, the study aims to provide an in-depth and structured analysis of the complexities surrounding web scraping, contributing to the development of informed guidelines and practices in this evolving field.

3. Literature Review

3.1 Ethical Boundaries: Privacy, Consent, and Data Ownership

The ethical boundaries of web scraping, particularly concerning privacy, consent, and data ownership, have become increasingly significant in the digital age. As web scraping techniques become more sophisticated, they raise critical ethical questions that researchers and practitioners must address.

Privacy concerns are at the forefront of ethical considerations in web scraping. The process often involves extracting personal data from websites, which can include sensitive information. Singh (2016) emphasizes the importance of attaining personal privacy on the web, highlighting the need for ethical web scraping practices that protect individual privacy. This includes ensuring that data collection methods do not infringe on the personal privacy rights of individuals whose data is being scraped.

Consent is another crucial ethical issue in web scraping. The unobtrusive nature of web scraping means that individuals whose data is being collected are often unaware of and have not consented to this collection. Rudiger (2017) discusses the ethical concerns around issues of privacy violation, informed consent, and the right to withdraw in the context of corpus linguistic studies using internet resources. The paper underscores the ethical responsibility of researchers to consider the 'publicness' of internet-based information and the implications of using such data without explicit consent.

The ethical boundaries of web scraping are not only a concern for researchers but also for businesses and organizations that rely on web-scraped data. The challenge lies in balancing the benefits of web scraping for data acquisition with the ethical implications of such practices. This involves developing and adhering to ethical guidelines that respect privacy, ensure informed consent, and recognize data ownership rights.

In summary, the ethical boundaries of web scraping, encompassing privacy, consent, and data ownership, are critical considerations in the digital age. As web scraping continues to evolve, it is imperative for all stakeholders involved in this practice to engage in ethical reflection and adhere to guidelines that protect the rights and privacy of individuals. This will ensure that web scraping is conducted responsibly and ethically, aligning with broader societal values and legal frameworks.

3.2 Legal Boundaries: International Laws and Regulations

The legal landscape of web scraping is complex and multifaceted, encompassing various international laws and regulations. As web scraping practices have become more prevalent, they have increasingly intersected with legal issues, particularly in the context of international law and cross-border data flows.

One of the key legal challenges in web scraping is the issue of jurisdiction and the applicability of different national laws. Erlina and Siswandi (2020) discuss the intricacies of law enforcement in undelimited maritime boundaries from an international law perspective, highlighting the conflicts that arise when different legal systems intersect. Similarly, in the context of web scraping, the extraction of data from websites located in different countries can lead to legal ambiguities, especially when the laws of one country conflict with those of another.

The legal regime in specific regions, such as the Arctic, also presents unique challenges. Shvelidze (2019) examines the international legal issues in the Arctic region, particularly focusing on the Russian Arctic Sector. This study underscores the importance of understanding regional legal frameworks, which can have significant implications for web scraping activities, especially when they involve data sources located in geopolitically sensitive areas.

Moreover, the legal and ethical perspectives of searching for information on the web are crucial in understanding the boundaries of web scraping. Kavallieros *et al*. (2018) explore the legal and ethical requirements for technologies used in fighting cybercrime, emphasizing the need for compliance with a complex framework of EU regulations, national laws, and international policies. This research highlights the necessity of a harmonized regulatory framework, especially in cases of cross-border cooperation, which is directly relevant to web scraping practices that often span multiple jurisdictions.

The legal boundaries of web scraping are further complicated by the evolving nature of internet regulations. Issues such as data privacy, intellectual property rights, and contractual obligations between website owners and users play a significant role in determining the legality of web scraping activities. Navigating these legal complexities requires a thorough understanding of both national and international laws, as well as an awareness of the ongoing changes in legal standards and practices related to digital data and the internet.

In summary, the legal boundaries of web scraping are shaped by a diverse array of international laws and regulations. Understanding these legal frameworks is essential for ensuring that web scraping activities are conducted in a lawful manner. As the digital landscape continues to evolve, staying informed about legal developments and adapting to new regulations will be crucial for practitioners and researchers engaged in web scraping.

3.3 Societal Boundaries: Public Perception and Acceptance of Web Scraping

The societal boundaries of web scraping, particularly in terms of public perception and acceptance, are critical aspects that influence the practice and evolution of this data collection method. Understanding societal attitudes towards web scraping is essential for navigating the ethical and legal complexities associated with it.

Social norms and their evolution play a significant role in shaping public perception. Hoek, Edwards, and Waa (2022) examine the shift from social acceptance to societal disapproval in the context of smoking. This transition of norms provides a useful analogy for understanding how societal attitudes towards web scraping might evolve. As with smoking, initial indifference or acceptance of web scraping could shift towards disapproval if its negative implications, such as privacy breaches or unethical data use, become more

prominent in public discourse.

The acceptance of web scraping is also influenced by cultural and ethical considerations. Kutalek *et al*. (2020) explore perceptions around the collection of body fluids for research on the persistence of the Ebola virus, offering insights into how cultural norms and ethical concerns can affect the acceptance of certain research practices. Similarly, in web scraping, societal acceptance can vary based on cultural attitudes towards privacy, data ownership, and the perceived benefits of the practice. In some cultures, web scraping might be seen as a valuable tool for innovation and research, while in others, it could be viewed as an intrusion into personal privacy.

The societal boundaries of web scraping are not static; they are shaped by ongoing dialogues and debates about technology, privacy, and ethics. As public awareness of data privacy issues increases, so too might skepticism and resistance towards practices like web scraping. This evolving landscape requires practitioners and policymakers to engage actively with public concerns, ensuring that web scraping is conducted transparently and ethically, with due consideration for societal norms and expectations.

In summary, public perception and acceptance of web scraping are influenced by a complex interplay of factors, including societal norms, cultural attitudes, and ethical considerations. As the practice of web scraping continues to grow, understanding and respecting these societal boundaries will be crucial for its sustainable and responsible use.

3.4 Case Studies: Boundary-Testing Scenarios in Web Scraping

Web scraping, as a method of data collection, has been applied in various fields, leading to boundary-testing scenarios that provide valuable insights into its capabilities and limitations. Examining specific case studies helps in understanding how web scraping is utilized in different contexts and the ethical, legal, and practical challenges it poses.

One notable application of web scraping is in the field of psychology, where it is used for the automatic extraction of big data from the internet. Landers *et al*. (2016) present a comprehensive overview of theory-driven web scraping, demonstrating its potential in psychological research. This case study highlights the process of collecting massive datasets from webpages, which can include tens of thousands of variables. The approach is grounded in substantive theory, ensuring that the data collected is relevant and meaningful for specific research questions. This case study exemplifies how web scraping can be tailored to meet the unique needs of a research domain, albeit with considerations for ethical and legal compliance.

These case studies collectively demonstrate the diverse applications of web scraping across different fields. They highlight the importance of considering ethical, legal, and practical aspects in the implementation of web scraping projects. From enhancing research capabilities in psychology to improving educational experiences in geoscience and streamlining information access through chatbots, web scraping proves to be a versatile tool. However, each case also brings to light the need for careful planning and adherence to ethical and legal standards to ensure the responsible use of web scraping technologies.

In summary, the exploration of these case studies reveals the expansive utility of web scraping in various domains. It underscores the importance of boundary-testing scenarios in understanding the full potential and limitations of web scraping as a data collection method. As web scraping continues to evolve, these case studies provide valuable lessons and benchmarks for future applications in different sectors.

4. Legal Implications and Societal Acceptance

4.1 Legal Challenges and Implications of Web Scraping

Web scraping, while a powerful tool for data collection, presents a myriad of legal challenges and implications that must be carefully navigated. The legal landscape surrounding web scraping is complex, involving issues of copyright, privacy, and data protection, among others.

Li *et al*. (2023) explore the potential of licensing as a means to mitigate the negative implications of commercial web scraping. The rise of AI models and the large-scale scraping of web content have led to significant legal disputes, highlighting the need for clearer guidelines and frameworks. Licensing is proposed as a solution to address content creators' concerns, ensuring responsible data reuse. However, the effectiveness of licensing in the context of web scraping remains a subject of debate, with questions about the specific terms and the sociotechnical environments required to facilitate its use at scale.

Speckmann (2021) delves into the possibilities, limitations, and ethical and legal challenges of web scraping. The paper underscores the importance of considering the legal implications of scraping activities, particularly when dealing with large datasets obtained from the internet. The legal landscape is further complicated by the varying laws and regulations across different jurisdictions, making it challenging for practitioners to ensure compliance.

The legal challenges of web scraping are not limited to intellectual property and privacy concerns. They also encompass contractual issues, such as the terms of service of websites, which often prohibit automated data extraction. Violating these terms can lead to legal disputes and potential liabilities for individuals and organizations engaged in web scraping.

Moreover, the evolving nature of internet regulations and the increasing focus on data protection laws, such as the General Data Protection Regulation (GDPR) in the European Union, add another layer of complexity. Web scrapers must be cognizant of these regulations and ensure that their activities do not infringe upon the data rights of individuals, particularly when handling personal data.

The legal challenges and implications of web scraping are multifaceted and require careful consideration. Practitioners must stay informed about the legal landscape and adapt their practices accordingly. The potential of licensing as a mitigating strategy offers a promising avenue, but its practical implementation and effectiveness in addressing the legal concerns of web scraping remain areas for further exploration and discussion.

4.2 Societal Acceptance: Public Opinion and Ethical Considerations

The societal acceptance of web scraping is a complex issue, influenced by public opinion and various ethical considerations. As web scraping becomes more prevalent in various sectors, understanding its impact on society and the ethical implications is crucial.

Archer *et al*. (2021) explore the implementation of eHealth in

low-resource countries, highlighting the challenges posed by limited resources, infrastructure, and ethical considerations. While this study focuses on eHealth, it sheds light on broader issues relevant to web scraping, such as the importance of ethical practices in technology implementation. The study emphasizes that user characteristics, perceived privacy, and security play significant roles in the acceptance of technology. These factors are directly applicable to web scraping, where concerns about privacy and data security can influence public opinion and acceptance.

The ethical considerations associated with web scraping are multifaceted. They include concerns about privacy, data ownership, and the potential misuse of scraped data. The perception of web scraping in society is often shaped by how these ethical issues are addressed. For instance, scraping personal data without consent can lead to public backlash and decreased trust in organizations that employ such practices.

Public opinion on web scraping is also influenced by the perceived benefits and risks associated with the practice. In sectors where web scraping is used for beneficial purposes, such as healthcare or academic research, it may be more readily accepted. However, in scenarios where web scraping is used for intrusive advertising or data exploitation, public opinion is likely to be negative.

The study by Archer *et al*. (2020) further underscores the importance of considering the user experience in technology implementation. In the context of web scraping, this translates to understanding how the collected data impacts end-users and society at large. Ensuring that web scraping practices do not harmfully affect individuals or communities is vital for maintaining societal acceptance.

Therefore, societal acceptance of web scraping is closely tied to public opinion and ethical considerations. The perception of web scraping in society hinges on how it is used, the measures taken to protect privacy and data security, and the overall impact on individuals and communities. As web scraping continues to evolve, it is imperative for practitioners to engage in ethical practices and consider the societal implications of their activities to maintain public trust and acceptance.

4.3 Comparative Analysis: Legal and Societal Perspectives across Different Regions

The legal and societal perspectives on various issues can vary significantly across different regions, influenced by cultural, historical, and legal factors. A comparative analysis of these perspectives provides valuable insights into how different societies and legal systems approach similar challenges.

Hirsch (2021) discusses the sociological perspectives on international tribunals, highlighting how these tribunals are influenced by and affect the socio-cultural patterns of the communities they are embedded in. This study underscores the importance of understanding the social environment in which legal institutions operate. International tribunals, much like the practice of web scraping, must navigate the complex interplay of legal norms and societal expectations. The study illustrates how tribunals interact with local communities and the need to maintain legitimacy within these communities, a concept that is also relevant in the context of web scraping, where societal acceptance can significantly impact its practice.

Głowacki *et al*. (2021) provide a comparative analysis of the rule of law and its social reception in Germany and Poland, exploring how societal perceptions influence the functioning of the market economy. This study is particularly relevant for understanding how societal attitudes towards legal norms, including those related to data privacy and intellectual property (key aspects of web scraping), can vary between countries. The research shows that societal perceptions of the rule of law can contribute to economic insecurity or, conversely, reduce transaction costs and build trust in the state. These findings highlight the importance of societal acceptance and trust in legal systems, which can also apply to the acceptance of web scraping practices.

Bi (2023) examines corporate obligations in anti-sexual harassment in the workplace, comparing Chinese and U.S. laws. This comparative study sheds light on how different legal frameworks address common societal issues. The analysis of Chinese and American approaches to regulating workplace sexual harassment provides insights into how societal values and legal traditions shape legal responses to societal problems. Similarly, in the context of web scraping, societal values and legal traditions in different regions can influence how this practice is regulated and perceived.

The comparative analysis of legal and societal perspectives across different regions reveals significant variations in how societies and legal systems approach common issues. From international tribunals to corporate obligations and the rule of law, these studies demonstrate the impact of cultural, historical, and legal contexts on societal attitudes and legal norms. Understanding these differences is crucial for global practices like web scraping, as it highlights the need for context-specific approaches that respect local legal traditions and societal values.

4.4 The Future of Legal and Societal Norms in Web Scraping

The future of legal and societal norms in web scraping is a dynamic and evolving landscape, shaped by technological advancements, legal developments, and societal attitudes. Understanding the trajectory of these norms is crucial for anticipating how web scraping will be practiced and regulated in the future.

Gold and Latonero (2018) delve into the ethical and legal considerations for web crawling and scraping, highlighting the ambiguous regulatory environment largely governed by online social norms. As web scraping technologies become more advanced, the legal frameworks and societal norms surrounding them are increasingly challenged. The authors emphasize the importance of addressing user privacy in the balance, as courts begin to tackle issues raised by web crawlers. This perspective suggests that future legal norms in web scraping will likely focus more on protecting individual privacy and data rights, reflecting broader societal concerns about data security and personal privacy.

Kostenko *et al*. (2023) explore the concept of the metaverse and its implications for legal norms, particularly in the context of criminal law. The study proposes the development of comprehensive electronic jurisdiction based on the latest legislation to regulate social relations in the metaverse. This research indicates that as digital environments like the metaverse become more prevalent, legal norms will need to adapt to address the unique challenges they present. This adaptation may include new forms of jurisdiction and legal codes specifically designed for virtual environments, which could also influence how web scraping is regulated in such spaces.

Donald (2020) examines the network effects of legal systems and their impact on global legal development. The study

highlights how law, originating in local environments, can be transmitted globally, affecting how legal norms are adopted and adapted in different contexts. This analysis is pertinent to web scraping, as it suggests that legal norms around web scraping could spread and evolve through network effects, influenced by the adoption of these norms in influential jurisdictions. The study also points to the potential for legal norms to become disconnected from local societal needs, underscoring the importance of ensuring that legal developments in web scraping remain responsive to societal attitudes and concerns.

In summary, the future of legal and societal norms in web scraping is likely to be characterized by greater emphasis on privacy protection, the adaptation of legal frameworks to new digital environments like the metaverse, and the global transmission of legal norms through network effects. As these developments unfold, it will be important for practitioners, policymakers, and researchers to stay informed and engaged with the evolving legal and societal landscape surrounding web scraping.

5. Discussion

5.1 Synthesizing Ethical, Legal, and Societal Perspectives.

The synthesis of ethical, legal, and societal perspectives is essential in understanding and addressing the multifaceted implications of practices like web scraping. These perspectives provide a comprehensive view of the challenges and opportunities presented by technological advancements and their integration into society.

Couture *et al*. (2022) explore the ethical implications of artificial intelligence in population health and the public's role in its governance. Their study emphasizes the importance of citizen participation in the governance of AI, reflecting broader societal implications. This approach resonates with the need for public engagement and ethical governance in web scraping practices. The study suggests that involving citizens in the governance process can lead to more ethically sound and socially accepted practices, highlighting the significance of public opinion and ethical considerations in shaping legal and societal norms.

Yatsenko (2022) discusses the gnoseological nature of the formation and formation of social responsibility of higher education institutions in the context of structural and innovative transformations. The focus on professional responsibility, autonomy, and the principle of freedom of choice and responsibility in providing services and research underlines the importance of social responsibility in adapting to technological changes. This perspective can be applied to web scraping, where higher education institutions play a crucial role in researching, teaching, and implementing responsible practices that align with societal and legal expectations.

Therefore, synthesizing ethical, legal, and societal perspectives is crucial for understanding and navigating the complex landscape of practices like web scraping. The integration of these perspectives can guide the development of responsible practices that respect individual rights, adhere to legal standards, and align with societal values. As technology continues to evolve, this synthesis will be essential in shaping the future of web scraping and ensuring its positive contribution to society.

5.2 Navigating the Complex Interplay of Ethics, Law, and Public Opinion

Navigating the complex interplay of ethics, law, and public opinion is increasingly critical in the digital age, especially in practices like web scraping. This interplay shapes the boundaries of what is permissible, ethical, and socially acceptable in the use of technology.

Floridi (2021) discusses the transition from self-regulation to hard law in the digital industry, highlighting a significant shift in how ethical issues are addressed. Initially, the digital industry relied on self-regulation to formulate and adhere to ethical codes and standards. However, this approach has proven insufficient, leading to a shift towards legal compliance, particularly in the European Union, where legislation is rapidly evolving to keep pace with technological advancements. This shift underscores the growing importance of legal frameworks in guiding ethical practices in the digital realm, including web scraping.

Ariff *et al*. (2021) explore the ethics and integrity in social media, emphasizing the societal impact of technological advancements. The widespread acceptance of social media in society brings with it ethical challenges, particularly in the misuse of technology. This study highlights the need for ethical and integral use of social media, reflecting broader concerns about the ethical use of technology, including web scraping. The paper suggests that societal acceptance of technology is contingent upon its ethical use, underscoring the importance of aligning technological practices with societal values and ethical standards.

Navigating the interplay of ethics, law, and public opinion in the context of web scraping and other digital practices requires a balanced approach. Legal compliance, ethical integrity, and societal acceptance are intertwined, each playing a crucial role in shaping the responsible use of technology. As the digital landscape continues to evolve, understanding and addressing these complex relationships will be key to ensuring that technological advancements are aligned with ethical standards, legal requirements, and societal expectations.

5.3 Recommendations for Harmonizing Legal and Ethical Practices

Harmonizing legal and ethical practices, especially in the context of rapidly evolving technologies like web scraping, is essential for ensuring responsible use and societal trust. This harmonization requires a careful balancing of technical possibilities, legal frameworks, and ethical considerations.

Geneviève *et al*. (2018) in their study on the European Influenzanet Consortium, emphasize the importance of ethical approaches in the collection, processing, and analysis of participants' information. The study highlights the need for harmonization of ethical approaches, especially in cross-border research collaboration. This is particularly relevant for web scraping, where data is often collected from multiple jurisdictions with varying legal and ethical standards. The recommendation is to develop common ethical guidelines that respect individual privacy and data protection laws, ensuring that web scraping practices are both legally compliant and ethically sound.

Luscombe, Dick, and Walby (2021) discuss the challenges of

navigating technical, legal, and ethical hurdles in web scraping within the social sciences. They highlight the importance of algorithmic thinking in the public interest, suggesting that web scraping should be guided by principles that prioritize societal benefits and ethical integrity. The authors recommend a multi-disciplinary approach that involves stakeholders from various fields to address the complex challenges posed by web scraping. This approach would ensure that legal and ethical considerations are integrated into the technical development and application of web scraping tools.

Black (2016) explores the use of the World Wide Web as a complex data set for digital humanities research, addressing the legal and ethical questions surrounding web scraping. The study provides insights into how web scraping can contribute to humanities research while distinguishing it from commercial and malicious activities. Black recommends addressing the complex concerns surrounding web scraping by developing clear guidelines that outline the legal and ethical boundaries of its use. This includes ensuring compliance with copyright laws, respecting the privacy of individuals, and maintaining the integrity of the websites from which data is scraped.

In conclusion, harmonizing legal and ethical practices in web scraping requires a collaborative and interdisciplinary approach. Developing common ethical guidelines, engaging in algorithmic thinking that prioritizes public interest, and establishing clear legal and ethical boundaries are key recommendations for responsible web scraping. As the digital landscape continues to evolve, these recommendations will guide practitioners, researchers, and policymakers in ensuring that web scraping is conducted in a manner that is legally sound, ethically responsible, and socially beneficial.

5.4 The Role of Policy Makers and Researchers in Shaping Future Norms.

The evolving landscape of web scraping in research necessitates a collaborative effort between policy makers and researchers to shape future norms. This collaboration is crucial in addressing the ethical, legal, and societal challenges posed by web scraping technologies.

Choi *et al*. (2016) emphasize the importance of bridging the gap between science and policy, particularly in the health sector. Their international survey of scientists and policy makers in China and Canada highlights the need for effective communication and collaboration between these two groups. This approach is equally relevant in the context of web scraping, where policy makers and researchers must work together to develop regulations and guidelines that balance technological advancement with ethical and legal considerations. The study suggests that involving both parties in the early stages of policy development can lead to more informed and effective regulations.

Peez (2022) provides insights into the contributions and blind spots of constructivist norms research in international relations. The study underscores the importance of systematic evidence and gap analysis in understanding large-scale trends and shaping future research agendas. In the realm of web scraping, this approach can guide policy makers and researchers in identifying areas that require further investigation and regulation. By systematically analyzing existing research and identifying gaps, policy makers can develop more targeted and effective policies, while researchers can focus their efforts on areas that need further exploration.

Brooks and Guy (2020) discuss the relevance of the European Union (EU) for health law and policy, illustrating how health has been gradually 'Europeanised' despite being a traditionally national competence. This study highlights the role of policy makers in shaping research agendas and the importance of adapting to changing legal landscapes. Similar dynamics are at play in the field of web scraping, where legal and ethical norms are continually evolving. Researchers in this field must be aware of these changes and adapt their practices accordingly, while policy makers should consider the implications of new technologies on existing legal frameworks.

In conclusion, the role of policy makers and researchers in shaping future norms in web scraping is pivotal. Effective collaboration, systematic analysis of trends and gaps, and adaptation to evolving legal landscapes are key to ensuring that web scraping practices are ethical, legal, and socially acceptable. As web scraping continues to advance, the interplay between policy development and research will be crucial in navigating the challenges and opportunities it presents.

6. Conclusions

The study has highlighted the multifaceted nature of web scraping, encompassing ethical boundaries, legal implications, and societal acceptance. Key findings indicate that ethical concerns primarily revolve around privacy, consent, and data ownership. Legally, web scraping is situated in a complex landscape with varying international laws and regulations, often leading to ambiguities in compliance and enforcement. Societally, the acceptance of web scraping is influenced by its perceived benefits and risks, with public opinion playing a crucial role in shaping its application in different contexts.

Looking ahead, the practice of web scraping is expected to evolve with advancements in technology, particularly in AI and machine learning. These developments promise to enhance the efficiency and capabilities of web scraping tools but also bring forth new ethical and legal challenges. The future outlook suggests a trend towards more sophisticated, automated, and ethically conscious web scraping practices, necessitating continuous adaptation to changing legal frameworks and societal expectations.

The study underscores the importance of moving towards more responsible and ethical web scraping practices. This involves a balanced approach that respects individual privacy, adheres to legal standards, and aligns with societal values. The integration of ethical considerations into the technical development and application of web scraping tools is crucial. Additionally, transparency in practices and clear communication about the purposes and benefits of web scraping can enhance its societal acceptance.

Future research should aim to address the existing gaps in understanding the ethical, legal, and societal implications of web scraping. This includes exploring public perceptions, developing ethical guidelines, examining the legal and economic impacts, and understanding the interplay between technology, society, and law. Research should also focus on the development of frameworks and best practices that ensure the responsible use of web scraping technologies in various domains.

Finally, this study provides a comprehensive overview of the

current state of web scraping, highlighting the need for a harmonized approach that balances technological advancements with ethical integrity and legal compliance. As web scraping continues to evolve, it is imperative for stakeholders to engage in continuous dialogue and collaboration to ensure its responsible and sustainable development.

7. References

1. Agrawal HG. Web Scraping and its Applications. International Journal of Scientific Research in Engineering and Management. 2023;8(1). DOI: 10.55041/ijsrem26601.
2. Ansgar K. Ethics considerations for Corpus Linguistic studies using internet resources [Internet]. 2015. Available from: https://doi.org/10.6084/M9.FIGSHARE.1494732.V1
3. Archer N, Lokker C, Ghasemaghaei M, DiLiberto D. eHealth implementation issues in low-resource countries: model, survey, and analysis of user experience. J Med Internet Res. 2021;23(6):e23715. DOI: 10.2196/23715.
4. Ariff NZZM, Kassim NBA, Shoid MSM, Baharuddin K. Ethics and Integrity in Social Media. International Journal of Academic Research in Business and Social Sciences. 2021;7(5):310-8. DOI: 10.6007/ijarbss/v11-i5/9775.
5. Bale AS, Ghorpade N, Kamalesh S, R. S., R. R., R. S. Web Scraping Approaches and their Performance on Modern Websites. In: 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC); 2022 Aug 17-19; Coimbatore, India. IEEE; 2022. p. 956-9. DOI: 10.1109/ICESC54411.2022.9885689.
6. Bi J. Corporate Obligations in Anti-Sexual Harassment in the Workplace: A Comparison Based on Chinese and U.S. Laws. Communications in Humanities Research. 2023;15:216-21. DOI: 10.54254/2753-7064/15/20230747.
7. Bhatt C, Bisht A, Chauhan R, Vishvakarma A, Kumar M, Sharma S. Web Scraping Techniques and Its Applications: A Review. In: 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT); 2023 Aug 11-12; Dehradun, India. IEEE; 2023. p. 1-8. DOI: 10.1109/CISCT57197.2023.10351298.
8. Black ML. The World Wide Web as complex data set: Expanding the Digital Humanities into the twentieth century and beyond through Internet research. International Journal of Humanities and Arts Computing. 2016;10(1):95-109. DOI: 10.3366/ijhac.2016.0162.
9. Brooks E, Guy M. EU health law and policy: Shaping a future research agenda. Health Econ Policy Law. 2021;16(1):1-7. DOI: 10.1017/S1744133120000274.
10. Christian Y, Erline M. Web-Based Chatbot with Natural Language Processing and Knuth-Morris-Pratt (Case Study: Universitas International Batam). JST (Jurnal Sains dan Teknologi). 2022;11(1):132-41. DOI: 10.23887/jst-undiksha.v11i1.43258.
11. Choi BC, Li L, Lu Y, Zhang LR, Zhu Y, Pak AW, *et al*. Bridging the gap between science and policy: an international survey of scientists and policy makers in China and Canada. Implement Sci. 2015;11(1):1-11. DOI: 10.1186/s13012-016-0377-7.
12. Couture V, Roy MC, Dez E, Laperle S, Bélisle-Pipon JC. Ethical implications of artificial intelligence in population health and the public's role in its governance: perspectives from a citizen and expert panel. J Med Internet Res. 2023;25:e44357. DOI: 10.2196/44357.
13. Donald DC. Legal System Network Effects and Global Legal Development. Notre Dame J Int'l Comp L. 2020;10:267. DOI: 10.2139/ssrn.3444159.
14. Erlina EC, Siswandi C, Gusman RA. Law Enforcement Issues and Regulations in Undelimited Maritime Boundaries: An International Law Perspective. Lentera Hukum. 2020;7(1):1-16. DOI: 10.19184/EJLH.V7I1.16643.
15. Floridi L. The end of an era: from self-regulation to hard law for the digital industry. Philos Technol. 2021;34:619-22. DOI: 10.1007/s13347-021-00493-0.
16. Geneviève LD, Wangmo T, Dietrich D, Woolley-Meza O, Flahault A, Elger BS. Research ethics in the European Influenzanet Consortium: scoping review. JMIR Public Health Surveill. 2018;4(4):e9616. DOI: 10.2196/publichealth.9616.
17. Głowacki K, Hartwell CA, Karunska K, Kurczewski J, Botsch E, Göhring T, *et al*. The Rule of Law and Its Social Reception as Determinants of Economic Development: A Comparative Analysis of Germany and Poland. Law and Development Review. 2021;14(2):359-400. DOI: 10.1515/ldr-2021-0043.
18. Gold Z, Latonero M. Robots welcome: Ethical and legal considerations for web crawling and scraping. Wash JL Tech Arts. 2017;13:275.
19. Hoek J, Edwards R, Waa A. From social accessory to societal disapproval: smoking, social norms and tobacco endgames. Tob Control. 2022;31(2):358-64. DOI: 10.1136/tobaccocontrol-2021-056574.
20. Hirsch M. Introduction: Sociological Perspectives on International Tribunals. Temple Int Comp Law J. 2021;34(2). DOI: 10.2139/SSRN.3761071.
21. Kavallieros D, Chalanouli C, Kokkinis G, Panathanasiou A, Lissaris E, Leventakis G, *et al*. Searching for crime on the web: Legal and Ethical perspectives. In: 2018 International Conference on Cyber Security and Protection of Digital Services (Cyber Security); 2018 Jun 11-12; Glasgow, UK. IEEE; 2018. p. 1-8. DOI: 10.1109/CyberSecPODS.2018.8560689.
22. Khder MA. Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. International Journal of Advances in Soft Computing & Its Applications. 2021;13(3):144-68. DOI: 10.15849/ijasca.211128.11.
23. Kostenko O, Zhuravlov D, Dniprov O, Korotiuk O. Metaverse: Model criminal code. Baltic Journal of Economic Studies. 2023;9(4):134-47. DOI: 10.30525/2256-0742/2023-9-4-134-147.
24. Kutalek R, Baingana F, Sevalie S, Broutet N, Thorson A. Perceptions on the collection of body fluids for research on persistence of Ebola virus: A qualitative study. PLoS Negl Trop Dis. 2020;14(5):e0008327. DOI: 10.1371/journal.pntd.0008327.
25. Landers RN, Brusso RC, Cavanaugh KJ, Collmus AB. A primer on theory-driven web scraping: Automatic extraction of big data from the Internet for use in psychological research. Psychol Methods.

2016;21(4):475. DOI: 10.1037/MET0000081.
26. Li H, Vincent N, Jernite Y, Merrill N, Benjamin JJ, Tarkowski A. Can Licensing Mitigate the Negative Implications of Commercial Web Scraping?. In: Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing; 2023. p. 553-5. DOI: 10.1145/3584931.3611276.
27. Luscombe A, Dick K, Walby K. Algorithmic thinking in the public interest: navigating technical, legal, and ethical hurdles to web scraping in the social sciences. Qual Quant. 2022;56(3):1023-44. DOI: 10.1007/s11135-021-01164-0.
28. Namoun A, Alshanqiti A, Chamudi E, Rahmon MA. Web Design Scraping: Enabling Factors, Opportunities and Research Directions. In: 2020 12th International Conference on Information Technology and Electrical Engineering (ICITEE); 2020 Oct 6-8; Yogyakarta, Indonesia. IEEE; 2020. p. 104-9. DOI: 10.1109/ICITEE49829.2020.9271770.
29. Peez A. Contributions and blind spots of constructivist norms research in international relations, 1980–2018: A systematic evidence and gap analysis. Int Stud Rev. 2022;24(1):viab055. DOI: 10.1093/isr/viab055.
30. Rüdiger S, Dayter D. The ethics of researching unlikeable subjects: Language in an online community. Appl Linguist Rev. 2017;8(2-3):251-69. DOI: 10.1515/applirev-2016-1038.
31. Saurkar AV, Pathare KG, Gode SA. An overview on web scraping techniques and tools. International Journal on Future Revolution in Computer Science & Communication Engineering. 2018;4(4):363-7.
32. Singh L. Data Ethics—Attaining Personal Privacy on the Web. In: Ethical Reasoning in Big Data: An Exploratory Analysis. Springer; 2016. p. 81-90. DOI: 10.1007/978-3-319-28422-4_7.
33. Singrodia V, Mitra A, Paul S. A Review on Web Scrapping and its Applications. In: 2019 International Conference on Computer Communication and Informatics (ICCCI); 2019 Jan 23-25; Coimbatore, India. IEEE; 2019. p. 1-6. DOI: 10.1109/ICCCI.2019.8821809.
34. Shvelidze S. Russian Arctic Sector Boundaries: The International Issues of Legal Regime in the Arctic Region. Economics and Culture. 2019;16(1):108-16. DOI: 10.2478/jec-2019-0012.
35. Speckmann F. Web Scraping: A Useful Tool to Broaden and Extend Psychological Research. 2021;229(4). DOI: 10.1027/2151-2604/a000470.
36. Weber MS, Napoli PM. Journalism history, web archives, and new methods for understanding the evolution of digital journalism. In: Journalism History and Digital Archives. Routledge; 2020. p. 74-93. DOI: 10.1080/21670811.2018.1510293.
37. Yatsenko VV. Gnoseological Aspects in Building Social Responsibility of Higher Education Institutions in the Context of Structural and Innovative Transformations. Journal of Strategic Economic Research. 2022;3:18–23. DOI: 10.30857/2786-5398.2021.3.2.