6-14-2024

Digital Trace Data – “Wild Wild West” Practices & Ethics of Web Scraping

Nicolai Etienne Fabian
University of Groningen, n.e.fabian@rug.nl

Edin Smailhodzic
University of Groningen, e.smailhodzic@rug.nl

Abayomi Baiyere
Queens University, speak2ab@gmail.com

DIGITAL TRACE DATA – "WILD WILD WEST" PRACTICES & ETHICS OF WEB SCRAPING

TREO Paper

Nicolai Etienne Fabian, University of Groningen, Groningen, Netherlands, n.e.fabian@rug.nl
Edin Smailhodzic, University of Groningen, Groningen, Netherlands, e.smailhodzic@rug.nl
Abayomi Baiyere, Queens University, Kingston, Ontario, Canada, a.baiyere@queensu.ca

Abstract

There is a rise in computationally intensive theory development and the need for digital trace data to fuel such research. Increasingly, questions emerge on how to collect digital trace data via web scraping and application programming interfaces (APIs). However, there are no clear guidelines on the process. Therefore, challenges surrounding practices and ethics emerge which in turn hampers replicability and theory development. In our study, we systematically review a sub-sample of studies on web scraping, discuss common ethical and practical challenges, and provide recommendations for future guidelines.

Keywords: web scraping, web crawler, ethics, computationally intensive theory development

1 Introduction

The Information Systems (IS) field is increasingly seeing calls to engage in computationally intensive research (Berente et al. 2019; Miranda et al. 2022). The basis for these projects is formed by digital trace data, often coming from different online environments such as social media, online communities, or other forms of web and digital data (Boegershausen et al. 2022; Miranda et al. 2022). To put the amount of data into perspective, every minute, consumers conduct over 6 million Google searches, leave 4 million likes on Facebook, and watch 48.000 hours of videos on Twitch (Boegershausen et al. 2022; Statista 2023). This ever-increasing data treasure offers unparalleled opportunities for researchers. Yet, the guidance on how to collect and leverage trace data is akin to the "Wild Wild" where anything goes.

To collect digital trace data, researchers can make use of web scraping (web crawlers/web spiders) and/or application programming interfaces (API) to automatically collect information from websites (Tiedrich 2024). For example, researchers can interact with the Twitter API to collect Tweets or use a web scraper to collect conversations from online forums. However, the utilization of web scraping is highly sensitive in terms of practices, ethics, as well as theoretical implications (Boegershausen et al. 2022). While in the past, the IS field was at the forefront of discussing acceptable use of data stemming from online environments (Allen et al. 2006), not much has happened since then (Boyd and Crawford 2012). As such, questions of what constitutes good research practices surrounding digital trace data (e.g., public vs. private data) as well as reporting standards for this type of research are yet unanswered (Boegershausen et al. 2022). Thus, we are in dire need of a systematic overview of web scraping practices in the IS field to uncover current conventions and identify how to link this crucial pre-step to computationally intensive theory development.

Following recent advances in understanding web scraping in other fields (Boegershausen et al. 2022), we set out to systematically understand practices in the IS field. We collected and analyzed data from the leading IS journals. Our exploratory approach yielded challenging findings. As our title suggests, web scraping practices in the IS field resemble the Wild West. For example, ethical challenges of data collection (e.g., nature of data, informed consent) are barely indicated and often missing. Furthermore,

the practice of data collection often not meet expectations about replicability, potentially contributing to the replication crisis (Dennis et al. 2020). As such, choices made in the scraping process like sampling decisions or technical limitations (what data is possible to collect) are not openly discussed, which in turn harms theory development (Berente et al. 2019). In our project, we make three contributions to the literature: we (1) integrate knowledge from systematic web scraping from related fields (Boegershausen et al. 2022), (2) challenge current web scraping conventions for reporting, replicability, and ethical questions, to (3) provide recommendations on scraping practices and thereby inform the discussion on the role of digital trace data in computationally intensive theory development.

2 Background

We aim to examine articles from the leading journals in the IS field ("Senior Scholar's List of Premier Journals"), in which web scraping was a key part of the article's method and/or contribution. There are multiple terms that researchers can apply to indicate that they made use of web scraping, such as terms related to web crawling, web spiders, automated data collection, and Application Programming Interfaces (APIs) as indicated by Boegershausen et al. (2022). In total, we found 524 across the eleven journals. For this TREO, we engaged in a first exploration of the data and randomly selected a total of forty papers in MISQ and ISR (twenty each) as the first step to shed light on scraping practices and ethics within the most rigorous journals of our field. Subsequently, we systematically coded the papers on their theoretical and conceptual background as well as their practices and ethics about scraping following grounded theory methodology in reviewing the literature (Wolfswinkel et al. 2013).

3 Findings and Discussion

We organize our findings alongside the two dimensions of ethics and practices, where we discuss key current practices as well as recommend future actions based on our findings (Table 1).

Challenge	Current practice	Recommendation
1. Ethics (scraping): Nature of data (public vs. private)	Not reporting or only in passing	Make clear that data is either (1) public, (2) scraping is allowed in terms and conditions/cookie policy, or (3) the owner gave consent
2. Practices (scraping)	Unclear reporting of practices (what data and variables, when, how)	Provide the code for the scraper to show variables collected and specify the time of collection
3. Practices (API)	No reporting of API affordances and limitations	Make clear what the API allows to collect, what limitations are present, and which version was used
4. Choice of trace data	Not reporting why data is useful for the purpose in light of alternatives	Make choices for data usage explicit and report "feasibility" as a key choice

Table 1. Summary of challenges, current practices, and recommendations

The ethical and legal implications of web scraping methods are rarely considered (either covered in passing or not at all for most cases) in our sub-sample. For example, there is a great deal of ambiguity surrounding the ethical issues when authors fail to state clearly if the data was public or private or if permission of any kind was gained from individuals whose data was scraped. Interestingly, the exceptions to this are somewhat older studies (Allen et al. 2006) where authors argue why collecting/scraping data from online communities can be ethical (Moon and Sproull 2008). Otherwise, only one study in our sample explicitly discusses ethics (Benjamin et al. 2019), in the context of scraping from the darknet. Other studies' lack of substantial ethical consideration highlights our community's need for a more comprehensive and open ethical framework. We recommend that research engaging in scraping clarify that digital trace data is (1) either "*public*" (Boegershausen et al. 2022) and therefore scraping is ethically not an issue, or (2) that scraping is allowed by the terms and conditions or cookie policy of the website (Jiang et al. 2022), or (3) that the owner gave permission for data collection.

Regarding web scraping practices, we see a somewhat ambiguous approach as many studies provide little details in their descriptions of the scraping process which harms replication efforts (both APIs and

scraper). For web scrapers, rarely information is provided on what data was collected (e.g., next to the variables used in the study), how and over what time the data was collected (e.g., websites dynamically change/employ countermeasures to scraping), and especially what data might be missing (e.g., technically not possible to collect). Moreover, in our sub-sample, no author reported the limitations of the API. For example, what data is available/not available through the API (e.g., some APIs restrict data collection based on age, place, and nature of data), what's the access limit of the API (quantity of data), as well as which version of the API was used (e.g., API access policies is subject to constant change). Thus, decisions taken in the process (selection, exclusion, rationale), are either not explained at all or less than ideally described and thereby hamper replication of API research and in turn fuel the replication crisis (Dennis et al. 2020). Lastly, "technical feasibility" (Boegershausen et al. 2022) or "ease of access" in non-scraping studies is frowned upon in the community. Though, scraping studies often suffer from these limitations (e.g., one data provider is easy to scrape, whereas another is hard/impossible to scrape). Thereby, we suggest clear recommendations on making the scraping process more transparent (Table 1), to outline choices made, and also how these choices impact subsequent theory development (Miranda et al. 2022). In this regard, we also hope to provoke a discussion on web scraping practices in our field and provide preliminary guidelines to facilitate the process of best practice development.

References

Allen, G. N., Burk, D. L., and Davis, G. B. 2006. "Academic Data Collection in Electronic Environments: Defining Acceptable Use of Internet Resources," *MIS Quarterly* (30:3), 599–610.

Benjamin, V., Valacich, J. S., and Chen, H. 2019. "DICE-E: A Framework for Conducting Darknet Identification, Collection, Evaluation with Ethics," *MIS Quarterly* (43:1), 1–22.

Berente, N., Seidel, S., and Safadi, H. 2019. "Data-Driven Computationally Intensive Theory Development," *Information Systems Research* (30:1), 50–64.

Boegershausen, J., Datta, H., Borah, A., and Stephen, A. T. 2022. "Fields of Gold: Scraping Web Data for Marketing Insights," *Journal of Marketing* (86:5), 1–20.

Boyd, D., and Crawford, K. 2012. "Critical Questions for Big Data," *Information, Communication & Society* (15:5), 662–679.

Dennis, A. R., Brown, S. A., and Wells, T. M. 2020. "EDITOR' S COMMENTS Replication Crisis or Replication Reassurance : Results of the IS Replication Project," *MIS Quarterly* (44:3), iii–x.

Jiang, Y., Ho, Y. C., Yan, X., and Tan, Y. 2022. "What's in a 'Username'? The Effect of Perceived Anonymity on Herding in Crowdfunding," *Information Systems Research* (33:1), 1–17.

Miranda, S. M., Berente, N., Seidel, S., Safadi, H., and Burton-Jones, A. 2022. "Computationally Intensive Theory Construction: A Primer for Authors and Reviewers," *MIS Quarterly* (46:2), iii–xviii.

Moon, J. Y., and Sproull, L. S. 2008. "The Role of Feedback in Managing the Internet-Based Volunteer Work Force," *Information Systems Research* (19:4), 494–515.

Statista. 2023. "Media Usage in an Internet Minute as of December 2023." (https://www.statista.com/statistics/195140/new-user-generated-content-uploaded-by-users-per-minute/).

Tiedrich, L. 2024. "The AI Data Scraping Challenge: How Can We Proceed Responsibly." (https://oecd.ai/en/wonk/data-scraping-responsibly).

Wolfswinkel, J. F., Furtmueller, E., and Wilderom, C. P. M. 2013. "Using Grounded Theory as a Method for Rigorously Reviewing Literature," *European Journal of Information Systems* (22:1), 45–55.