Selected Papers of #AoIR2024:
The 25th Annual Conference of the
Association of Internet Researchers
Sheffield, UK / 30 Oct - 2 Nov 2024

FROM ia_archiver TO OpenAI: THE PASTS AND FUTURES OF AUTOMATED DATA SCRAPERS

Katie MacKinnon
University of Copenhagen

Emily Maemura
University of Illinois Urbana-Champaign

Generative AI tools have become widespread over the past two years with the introduction of chatbots like ChatGPT, Copilot, and Gemini, and text-to-image services like DALL-E, Midjourney and Stable Diffusion. Alongside the spread of these tools, there is an increasing awareness that the datasets used to train AI are scraped from materials openly available on the web, with no attribution, compensation, or recourse for those content creators whose materials are collected by automated scrapers or crawlers. One project reacting to data scrapers is Nightshade, designed for artists who are concerned with the unauthorized use of their visual works in AI training datasets. Its creators describe Nightshade as a tool that “transforms images into ‘poison’ samples,” and therefore does not “rely on the kindness of model trainers, but instead associates a small incremental price on each piece of data scraped and trained without authorization” (Nightshade, n.d.). While data scraping practices related to the development of AI have recently come under scrutiny, there is a longer history of using and responding to crawlers across many industries. Before interventions like Nightshade, the robots.txt exclusion protocol was developed as the primary way to govern the behavior of crawlers, and has been widely used and discussed in the field of internet studies for decades (Thelwall & Stuart, 2006; Elmer, 2009). In this paper, we present a history of the protocol’s development and critique its use as a proxy for consent in widespread data scraping.

Thirty years ago the use of automated crawlers for accessing websites was being hotly debated. The early web had grown exponentially from an estimated 159,000 hosts registered in 1989, to over 2 million by 1993 (Coffman & Odlyzko, 1998). The web could no longer be navigated by directory listings and webrings; instead search engines indexed the ever-growing number of web pages using crawlers (aka wanderers, spiders or robots) to discover new material. Yet, these crawlers scraped data and wreaked

Suggested Citation (APA): MacKinnon, K. & Maemura, E. (2024, October). *From ia_archiver to OpenAI: The Pasts and Futures of Automated Data Scrapers*. Paper presented at AoIR2024: The 25th Annual Conference of the Association of Internet Researchers. Sheffield, UK: AoIR. Retrieved from http://spir.aoir.org.

havoc for website creators, making multiple requests in quick succession, preventing access by human users, and eating up precious bandwidth that was in limited and costly supply. A discussion arose on the WWW-Talk mailing list for how to manage and mediate access by these bot-based crawlers (Koster, 1994 February 25). The resulting proposal was the Robots Exclusion Protocol (REP), a plain text file that website creators could post whose machine-readable syntax defined a set of rules for crawlers to follow, indicating what content was allowed or 'disallowed' for access by robots (Koster, 1994). Despite its name, REP is not a technical standard managed by the Internet Engineering Task Force. Instead, REP is widely described as a 'gentlemen's agreement,' an honor policy among website owners and crawler operators, i.e., REP can be overridden technically as crawlers can simply ignore the rules listed in a robots.txt file. In 2024, robots.txt is framed as a relic of a lost era, described as "a handshake deal between some of the earliest pioneers of the internet to respect each other's wishes and build the internet in a way that benefitted everybody" (Pierce, 2024). In this modern retelling, where the internet was once governed by honor and good will, it is now merely an arena of theft.

While REP had not risen to public consciousness until recent discussions of AI and ML datasets, it has been more frequently discussed in web archiving contexts. Archival crawlers share their origins with indexers and scrapers of the early web, and Heritrix (a crawler widely used in libraries and archives) developed the Internet Archive (IA) is based upon technology from founder Brewster Kahle's web traffic analytics company Alexa Internet (acquired by Amazon in 1999). Incidentally, Alexa's ia_archiver bot is the source of much of the early web data accessed via the Wayback Machine. Technologically, the Heritrix archival crawler reduces the REP to a single crawler setting to "obey" or "ignore." Yet the checkbox choice to 'ignore' robots.txt directives is entangled with the legal, social, and cultural position of archives institutions. National web archives often ignore REP since they view their legal authority as overriding REP restrictions. Additionally, IA announced in 2017 that they intended to ignore robots.txt files after determining they hinder access to the archive and that the files "are geared toward search engine crawlers [and] do not necessarily serve our archival purposes (Graham, 2017). Ogden (2020, 2022) explores justifications for ignoring the REP at IA in more detail, and highlights the extreme position of Archive Team's Jason Scott who argues "If you don't want people to have your data, don't put it online" (Scott, 2011). Archives are largely seen as a public service, and their role in providing access to otherwise-unavailable historical web content is used to justify mass scraping without processes of permission, offering only opt-out removal requests. We argue here that web archives should never be exempt from adhering to web creators' intentions for their data and a renewed interest in their position is warranted as web archives data is increasingly being made available to ML models and methods (Baack, 2024; van Strien, 2023).

The process by which data scrapers seek permission to crawl a site through REP cannot engage with how the "human" is entangled with data; people and their bodies are variously attached to data and data afterlives (Cifor et al., 2020; Ebling, 2022). Since the internet facilitates "ways of being and forms of information production and flow that challenge basic definitions around data protection," (Markham, Tiidenberg & Herman, 2018) data scrapers that are programmed to collect without context and consideration are engaged in extractive data practices by design. If we consider how data collection is the “point at which the context, purpose and consent of data use are formally agreed”, we can see how REP, which works from a premise in which agreements around collection can be made between a web owner and a collector, captures “one moment in the middle of a whole series of decisions that determine the power structures under which data is collected” (Benjamin, 2021).

Within the broader field of critical data studies, there have been several recent interventions into the ways in which the collection of data is understood, critiqued, and re-imagined in feminist terms. Yet the dominant tools and mechanisms for web data collection, like data scrapers used for both web archives and the development of AI, are based on the conception of the internet as a mountain of data that’s sitting, waiting, available to be acted upon, extracted and put to use. We want to counter this recent narrative that “the basic social contract of the web is falling apart” (Pierce, 2024), and instead argue that data extractive infrastructures have always been at work over the past 30 years of the web. Paired with some of the work on critical archival theory, we aim to find new ways for web archives and modes of collection to become unbound from the “capitalist logics of data extraction” upon which they’re currently built (Theilen et al., 2021).

References

Baack, S. (2024). *Training Data for the Price of a Sandwich: Common Crawl’s Impact on Generative AI*. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/

Benjamin, G. (2021). What we do with data: a performative critique of data 'collection'. *Internet Policy Review, 10*(4). https://doi.org/10.14763/2021.4.1588

Coffman, K., & Odlyzko, A. (1998). The size and growth rate of the Internet. First Monday. https://doi.org/10.5210/fm.v3i10.620

Ebeling, M. F. E. (2022). *Afterlives of data: Life and debt under capitalist surveillance*. University of California Press.

Elmer, G. (2009). Robots.txt: The Politics of Search Engine Exclusion. In J. Parikka & T. D. Sampson, *The Spam Book: On viruses, porn, and other anomalies from the dark side of digital culture* (pp. 217–227). Hampton Press.

Graham, M. (2017, April 17). *Robots.txt meant for search engines don’t work well for web archives*. Internet Archive Blogs. https://blog.archive.org/2017/04/17/robots-txt-meant-for-search-engines-dont-work-well-for-web-archives/

Jones, M. L., Kaufman, E., & Edenberg, E. (2018). AI and the Ethics of Automating Consent. *IEEE Security & Privacy*, *16*(3), 64–72. https://doi.org/10.1109/MSP.2018.2701155

Koster, M. (1994, February 25). *Important: Spiders, Robots and Web Wanderers*. Post on WWW-Talk, accessed via archived version https://web.archive.org/web/20131029200350/http://inkdroid.org/tmp/www-talk/4113.html

Koster, M. (1994). *A Standard for Robot Exclusion*. The Web Robots Pages. http://www.robotstxt.org/orig.html

Markham, A. N. (2018). Afterword: Ethics as Impact—Moving From Error-Avoidance and Concept-Driven. Models to a Future-Oriented Approach. *Social Media + Society, 4*(3). https://doi-org/10.1177/2056305118784504

Markham, A. N., Tiidenberg, K., & Herman, A. (2018). Ethics as Methods: Doing Ethics in the Era of Big Data Research—Introduction. *Social Media + Society*, *4*(3), 2056305118784502. https://doi.org/10.1177/2056305118784502

Nightshade: What Is Nightshade?. (n.d.). Retrieved March 1, 2024, from https://nightshade.cs.uchicago.edu/whatis.html

Ogden, J. R. (2020). *Saving the Web: Facets of Web Archiving in Everyday Practice* [Phd, University of Southampton]. https://eprints.soton.ac.uk/447624/

Pierce, D. (2024, February 14). *The rise and fall of robots.txt*. The Verge. https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders

Scott, J. (2011, May 10). *Robots.txt*. Archiveteam. https://wiki.archiveteam.org/index.php?title=Robots.txt

van Strien, D. (2023, May 10). *Getting Started with Machine Learning and GLAM (Galleries, Libraries, Archives, Museums) Collections.* Internet Archive Blogs. https://blog.archive.org/2023/05/10/getting-started-with-machine-learning-and-glam-galleries-libraries-archives-museums-collections/

Theilen, J. T., Baur, A., Bieker, F., Quinn, R. A., Hansen, M., & Fuster, G. G. (2021). Feminist data protection: An introduction. *Internet Policy Review*, *10*(4). https://policyreview.info/articles/analysis/feminist-data-protection-introduction

Thelwall, M., & Stuart, D. (2006). Web crawling ethics revisited: Cost, privacy, and denial of service. *Journal of the American Society for Information Science and Technology, 57*(13), 1771–1779. https://doi.org/10.1002/asi.20388