Method Impossible? A Reflection on the Fragility of Methodologies And Infrastructures For Search Engine Research

Ashwin Nagappa[a], Abdul Karim Obeid[a], Sebastian Sünkler[b], Kateryna Kasianenko[a], Daniel Angus[a]

a Queensland University of Technology, Kelvin Grove, Brisbane, Australia University 1, Address, City, Index, Country
b Hamburg University of Applied Sciences, Finkenau 35,22081 Hamburg, Germany

1. Background

Searching for information is a common online activity, spanning activities from the dull and mundane to life- and society-changing (Haider & Sundin, 2019). The incorporation of Generative AI capabilities into many commercial search engines makes it increasingly important to study these search systems and practices critically (Sandvig et al., 2014). However, the unwillingness of these large commercial platforms to share data, frustrates efforts to support necessary public-interest research (Tromble, 2021). At a time when social sciences' need for reliable methods to examine search engines is arguably the most acute, scholars must instead rely on 'importing' (Marres & Weltevrede, 2013) methods not developed with humanities and social science research in mind.

The dominant research on search engines have largely treated them as technical infrastructures (Ingwersen & Järvelin, 2005), with computer science and information retrieval scholarship prioritising speed, accuracy, and relevance over the social, cultural, and political dimensions of search. Where humanities and social science research has engaged with search engines, it has tended to proceed on a case study basis. Despite their central role in shaping information about sensitive topics such as health, race, politics, and gender, search systems cannot be studied holistically due to a lack of robust, standardised methods for auditing them at a scale proportional to their societal reach. Where methodologies or tools exist, they are labour-intensive, context-dependent, and difficult to replicate across time, geography, or user profiles. Therefore, social science scholarship on search engines remains sparse due to limited access to data and opaque algorithms.

This presentation addresses that gap. Drawing on our experience developing scalable search engine research methods in Australia and Germany, we present a combination of tools — autonomous scraping infrastructure, desktop-based query tools, and SEO-derived visibility data — that together enable query variation and timely simulation of search sessions in response to ongoing events. These experiences teach us that, in centering the social sciences perspective on search engines, we have to constantly adapt, improvise, or borrow approaches from the fields of digital marketing and advertising. Hence, some of these approaches are nearly impossible or are methodological approaches that do not persist over time.

2. Reflections from Australian Search Experience (ASE) Research and Search Studies in Germany

[1]SEASON 2026, September 15–17, 2026, Hamburg, Germany
EMAIL: a.nagappa@qut.edu.au (A.1); abdul.obei@qut.edu.au (A. 2); sebastian.suenkler@haw-hamburg.de (B. 1); k.kasianenko@qut.edu.au (A. 3); daniel.angus@qut.edu.au (A.4)
ORCID: 0000-0002-4187-1393 (A. 1); 0000-0002-9838-1838 (A. 2); 0000-0001-9848-1137 (B. 1); 0000-0002-7159-5676 (A. 3), 0000-0002-1412-5096 (A.4)
DOI: https://doi.org/10.48441/4427.3731

AlgorithmWatch's automated data donation method (Krafft et al., 2017) is a highly useful approach for collecting search results from citizen scientists and was used in the first phase of the Australian Search Experience (ASE) project. The method uses a browser plugin to enable the collection of search results at scale from diverse individuals using a common query list. Consequently, the lack of variation in search queries led to almost uniform search results across all participants (Bruns et al., 2025). While data donations enabled the researchers to gather search results from real users, the project could not collect data on sensitive queries needed to examine normative issues in depth or recruit geographically and culturally diverse users.

The second phase of the ASE project applies an autonomous scraping infrastructure that simulates Australian web browsing sessions by routing network traffic of 'Seleniumwire Undetected ChromeDrivers' (UltrafunkAmsterdam, 2026) through dedicated Australian geographic 'TinyProxy' proxies (2026) operating on Amazon Web Services (n.d.) EC2 instances. This extends the work of van Hoof et al. (2024) on Agent-based techniques. In addition, the Result Assessment Tool (RAT) (Sünkler et al., 2025) was used to run desktop searches for fewer queries. This combination of tools allows us to vary queries and to run timely search simulations in response to ongoing events.

A key limitation of this approach is the brittleness of web scraping (Freelon, 2018), in this context, due to ongoing legal action against SerpApi (Prado, 2025). Against the backdrop of current AI innovation, anti-scraping measures require humans to intervene, for example, to solve CAPTCHA or validate identity. This creates a central ethical tension between compliance with platforms' Terms of Service (ToS) and the public interest in algorithmic transparency.

Building a robust search simulation that mirrors real user experiences requires a deeper understanding of information seekers' contexts. This can be supported by leveraging visibility tools developed by the Search Engine Optimisation (SEO) industry (Schultheiß et al., 2023) – a by-product of search engines themselves. To gather data efficiently, researchers can employ agile methods, including Python scripts for autocomplete, paid solutions to collect 'people also ask' queries, and SerpApi[2] for quick, multi-platform data collection.

But using a professional SEO infrastructure introduces its own empirical limitations. These products are proprietary 'black boxes' that are better suited to digital marketing than to academic rigour. They filter, combine, and rank data based on business needs, which makes the basic algorithmic results less clear. Also, relying on third-party commercial APIs (like SerpApi) for academic research simply shifts the problem of fragility to another area, since these proxy services are also likely to be blocked or sued by platforms.

3. Conclusion

Our projects in Australia and Germany aim to audit search engines with the same rigour as that applied to social media platforms, while building reusable methods for future research. In the process, we encounter a deeper contradiction: while Google crawls and indexes the web at scale, it increasingly restricts others from doing the same. We have only achieved partial methodological stability because auditing search engines now means operating within an active contest over access to information itself. Therefore, in this presentation, we argue that studying search engines from a humanities and social science perspective demands adaptive, often improvised techniques that frequently draw on digital marketing and advertising infrastructures. These methods are not neutral; they shape both what can be observed and the kind of insights research can yield (Marres & Weltevrede, 2013). Our findings suggest that, despite persistent narratives of openness and

[2] SERP is short for "Search Engine Results Page". SerpApi is a commercial entity that provides real time search engine results pages to its customers.

transparency, auditing contemporary information ecosystems remains structurally difficult and resource-intensive — a condition that reflects the web's status as a tightly governed "walled garden" rather than any single deliberate act of obstruction.

4. Acknowledgements

The authors would like to acknowledge that this work was supported by the Australian Research Council (ARC) for the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005).

5. References

Amazon Web Services, Inc. (n.d.). Amazon Elastic Compute Cloud (Amazon EC2) [Cloud computing service]. https://aws.amazon.com/ec2/

Bruns, A., Angus, D., Nagappa, A., Kasianenko, K., Obeid, A. K., Weinbrand, S., & Tweedie, B. (2025, September 25). Assessing recommendation diversity in search results: Approaches using data donations and artificial personas. SEASON 2025: Conference of the Search Engines and Society Network. https://snurb.info/node/3422

Freelon, D. (2018). Computational Research in the Post-API Age. Political Communication, 35(4), 665–668. https://doi.org/10.1080/10584609.2018.1477506

Haider, J., & Sundin, O. (2019). Invisible Search and Online Search Engines: The Ubiquity of Search in Everyday Life. Taylor & Francis.

Ingwersen, P., & Järvelin, K. (2005). *The turn: Integration of information seeking and retrieval in context*. Springer.

Krafft, T. D., Gamer, M., Laessing, M., & Zweig, K. A. (2017). Filterblase geplatzt? Kaum Raum für Personalisierung bei Google-Suchen zur Bundestagswahl 2017 (1. Zwischenbericht). AlgorithmWatch. https://algorithmwatch.org/de/filterblase-geplatzt-kaum-raum-fuer-personalisierung-bei-google-suchen-zur-bundestagswahl-2017

Marres, N., & Weltevrede, E. (2013). SCRAPING THE SOCIAL?: Issues in live social research. Journal of Cultural Economy, 6(3), 313–335. https://doi.org/10.1080/17530350.2013.772070

Prado, H. D. (2025, December 19). Why we're taking legal action against SerpApi's unlawful scraping. Google. https://blog.google/innovation-and-ai/technology/safety-security/serpapi-lawsuit/

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *"Data and Discrimination: Converting Critical Concerns into Productive Inquiry," a Preconference at the 64th Annual Meeting of the International Communication Association*.

Schultheiß, S., Lewandowski, D., Mach, S. von, & Yagci, N. (2023). Query sampler: Generating query sets for analyzing search engines using keyword research tools. PeerJ Computer Science, 9, e1421. https://doi.org/10.7717/peerj-cs.1421

Sünkler, S., Lewandowski, D., Schultheiß, S., & Yagci, N. (2025). Result Assessment Tool (RAT): Empowering search engine data analysis. PeerJ Computer Science, 11, e2962. https://doi.org/10.7717/peerj-cs.2962

TinyProxy Development Team. (2026). Tinyproxy (Version X.X.X) [Computer software]. GitHub. https://github.com/tinyproxy/tinyproxy

Tromble, R. (2021). Where Have All the Data Gone? A Critical Reflection on Academic Digital Research in the Post-API Age. Social Media + Society, 7(1), 2056305121988929. https://doi.org/10.1177/2056305121988929

UltrafunkAmsterdam. (2026). undetected-chromedriver (Version X.X.X) [Computer software]. GitHub. https://github.com/ultrafunkamsterdam/undetected-chromedriver

van Hoof, M., Trilling, D., Moeller, J., & Meppelink, C. S. (2024). It matters how you google it? Using agent-based testing to assess the impact of user choices in search queries and algorithmic personalization on political Google Search results. Journal of Computer-Mediated Communication, 29(6), zmae020. https://doi.org/10.1093/jcmc/zmae020