Introducing Open and Reproducible Web Search for the Social Sciences

Muhammad Taimoor Khan[a], Alisa Rieger[a], Johannes Kiesel[a], Dimitar Dimitrov[a], Stefan Dietze[a,b]

[a] *GESIS – Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, Germany*
[b] *Heinrich Heine Universität Düsseldorf, Universitätsstr. 1, Düsseldorf, Germany*

1. Introduction

Many social science use cases rely on web data as information extracted from a collection of webpages acquired through searching the web on a topic of interest. For example, using fact-checked websites to detect and verify claims (Mittal et al., 2023), using subjective webpages for social media opinion mining (Pham et al., 2025), or using news articles and online forums to adapt large language models for news and social media analysis (Kmainasi et al., 2025). Webpages collected over time can be investigated for shifts in social and political behavior. For example, studying the diversity of news stories leading to the German federal election 2021 (Färber et al., 2023), studying the evolution of online news headlines over decades (Nickl et al., 2025), and analyzing the shift in search engine results on political candidates leading to the U.S. Federal election 2018 (Metaxa et al., 2019). Despite interest in using web data for social science research, the proprietary nature of mainstream search engines prevents transparent reporting of the webpages collection process as well as the reproducibility of the search results.

2. Shortcomings of Using Mainstream Proprietary Search Engines for Web Data

Given the effort and cost involved in crawling and frequently updating the index at web scale to make webpages searchable, proprietary search engines have established themselves as the de facto standard and gatekeeper for finding webpages and information on the web (Hendriksen et al., 2024). They are optimized for everyday use to navigate the web, and this focus introduces severe drawbacks for scientific use, e.g., finding webpages for a social science use case. Their proprietary and black-box nature impedes methodological transparency, reproducibility, and open access (Granitzer et al., 2024). For example, given the lack of transparency about how webpages are crawled, indexed, processed, and ranked in contemporary search engines, reproducing search results and reusing search results in research is challenging. The underlying web index of search engines is constantly being updated, and algorithmic intransparency is further increased through various means of personalization of search results. The inherent biases of the indexed data as well as the ranking algorithms are reflected in results of acute information topic queries (Gao & Shah, 2020), which is even more pronounced when there is no ideal query response that is neutral and unbiased across dimensions and communities (Lewandowski, 2019).

3. How Open Web Search for the Social Sciences Overcomes These Shortcomings

To fill this gap, we present the Open Web Search for the Social Sciences (OWS3), built on the foundations provided by the Open Web Search (OWS)[1] project that aims at establishing a transparent Open Web Index[2]. Since 2023, OWS has crawled petabytes of web data, approaching web-scale

[1] https://openwebsearch.eu/
[2] https://openwebindex.eu/

SEASON 2026, September 15–17, 2026, Hamburg, Germany
DOI: https://doi.org/10.48441/4427.3727

coverage; however, it is not a search engine but rather the foundational infrastructure to support search engines. OWS3 offers a search facility on the index, supporting its most frequent languages, i.e., English (38%) and German (8%), respectively (Fröbe et al, 2023). It aims to serve as one of the largest web archives for the social sciences, having transparent processes, reproducible query results, and enrichments contributing to studying social science use cases, e.g., named entities, political learning, and human values, etc. It uses MOSAIC[3] framework to rank the webpages, process query requests, and fetch results. The OWS3 pipeline has the following features.

- **Open crawling:** The crawling is guided by the information owners through *robot.txt* and *sitemaps*, broadly governed by legal licensing frameworks and GDPR jurisdictions (Granitzer, 2024). The crawled webpages are stored as raw *WARC* files.

- **Open processing:** It processes webpages using *Python-ISO639* to detect language, and *Resiliparse* tools to extract features from the webpages. The extracted features are provided as metadata stored in *PARQUET* files. They can be reproduced from the raw *WARC* files.

- **Open indexing:** The webpages are indexed as terabytes of inverted common index file format (CIFF) index files. They are organized as daily shards with indices along with corresponding metadata files. The shards are distributed across multiple providers. The unintentional multiple crawls of a webpage are identified through URL hash matching and are deduplicated, keeping the older copy.

- **Ranking and Interface:** OWS3 uses BM25 for ranking webpages entirely on relevance, excluding recency, popularity and other external factors. The query requests are processed and served using *Lucene* through the MOSAIC framework to support API calls as well as web interface access. The OWS3 interface equips users to download search results, which can be shared, reported, and reproduced as research artifacts.

4. Interactive Session (90 min)

In this interactive session, we discuss with participants possible use cases for webpages and their enrichments in social science research. Furthermore, we will discuss approaches for transparency and reproducibility of OWS3 for scientific use and demonstrate the OWS3 prototype.

Moderators: Muhammad Taimoor Khan, Alisa Rieger, Johannes Kiesel

Target Audience: Researchers collecting, analyzing, or developing tools/methods for web data

Time planning:

15 mins	Introduction to Open Web Search for the Social Sciences
60 mins	World Café
15 mins	Feedback and final thoughts

Questions for World Café:

[3] https://mosaic.ows.eu/

SEASON 2026, September 15–17, 2026, Hamburg, Germany
DOI: https://doi.org/10.48441/4427.3727

- What types/domains of webpages (e.g., news articles, fact-checking pages) can be of interest to social science research?
- How often do the webpages of interest to social science research be re-crawled to track their changes over time?
- Which of the OWS metadata4 are important for social science research to present for the searched webpages and their snapshots?
- Which enrichments are required for social science research to understand and filter on webpages, e.g., political leaning, human values, sentiment analysis, named entities, etc.?
- What are useful interface features that will improve downloading, sharing and reproducing search results with OWS3 for social science research?

5. References

Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2016). Using the wayback machine to mine websites in the social sciences: A methodological resource. Journal of the association for information science and technology, 67(8), 1904-1915.

David, A., & Huurdeman, H. (2014). Web archive search as research: Methodological and theoretical implications. Alexandria, 25(1-2), 93-111.

Färber, M., Schwade, J., & Jatowt, A. (2023). Measuring Variety, Balance, and Disparity: An Analysis of Media Coverage of the 2021 German Federal Election. arXiv preprint arXiv:2308.03531.

Fröbe, M., Reimer, J. H., MacAvaney, S., Deckers, N., Reich, S., Bevendorff, J., ... & Potthast, M. (2023, July). The information retrieval experiment platform. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2826-2836).

Gao, R., & Shah, C. (2020). Toward creating a fairer ranking in search engine results. Information Processing & Management, 57(1), 102138.

Granitzer, M., Voigt, S., Fathima, N. A., Golasowski, M., Guetl, C., Hecking, T., ... & Zerhoudi, S. (2024). Impact and development of an open web index for open web search. Journal of the Association for Information Science and Technology, 75(5), 512-520.

Hendriksen, G., Dinzinger, M., Farzana, S. M., Fathima, N. A., Fröbe, M., Schmidt, S., ... & Stein, B. (2024, March). The open web index: Crawling and indexing the web for public use.

SEASON 2026, September 15–17, 2026, Hamburg, Germany
DOI: https://doi.org/10.48441/4427.3727

In European Conference on Information Retrieval (pp. 130-143). Cham: Springer Nature Switzerland.

Kmainasi, M. B., Shahroor, A. E., Hasanain, M., Laskar, S. R., Hassan, N., & Alam, F. (2025, April). LlamaLens: Specialized multilingual llm for analyzing news and social media content. In Findings of the Association for Computational Linguistics: NAACL 2025 (pp. 5627-5649).

Lewandowski, D. (2019). The web is missing an essential part of infrastructure: An open web index. Communications of the ACM, 62(4), 24-24.

Metaxa, D., Park, J. S., Landay, J. A., & Hancock, J. (2019). Search media and elections: A longitudinal investigation of political search results. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-17.

Mittal, S., Sundriyal, M., & Nakov, P. (2023, December). Lost in translation, found in spans: Identifying claims in multilingual social media. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 3887-3902).

Nickl, P., Moussaïd, M. and Lorenz-Spreen, P., 2025. The evolution of online news headlines. Humanities and Social Sciences Communications, 12(1), pp.1-13.

Pham, V. T., Qu, L., Li, Z., Sharma, S., & Haffari, G. (2025, July). SURVEYPILOT: an Agentic Framework for Automated Human Opinion Collection from Social Media. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 4397-4422).

SEASON 2026, September 15–17, 2026, Hamburg, Germany
DOI: https://doi.org/10.48441/4427.3727