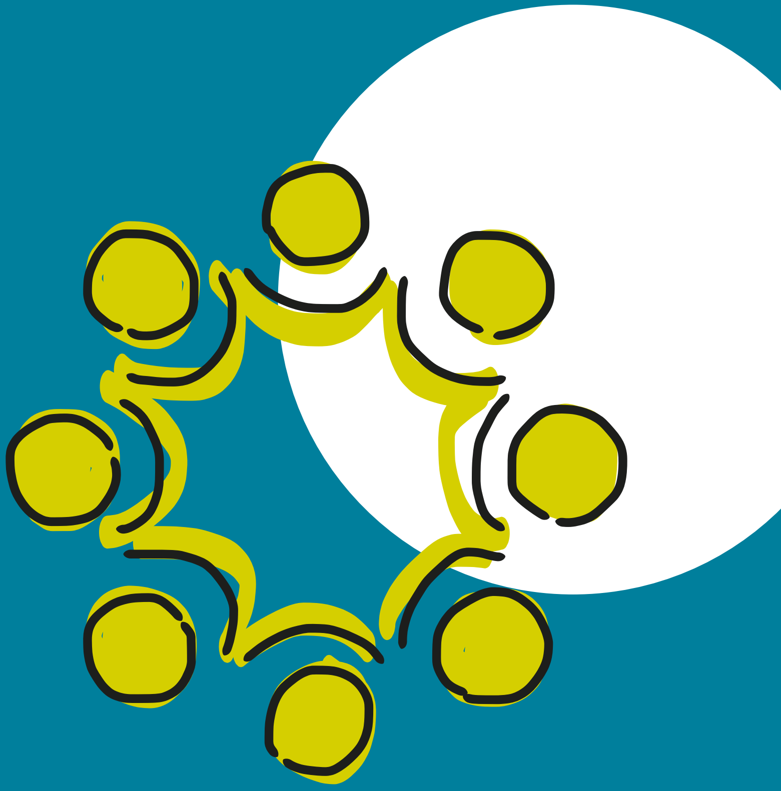
Editors
Andreas Wagner
Michael Granitzer
Christian Gütl
Per Öster
Megi Sharikadze
Stefan Voigt

Proceedings

7th International Open Search Symposium

#ossym2025

8-10 October 2025
CSC - IT Center for Science
Helsinki, Finland



open search foundation

CSC

ISSN: 2957-4935
ISBN: 978-92-9083-705-3

Impressum

Editors:

Andreas Wagner, CERN, Geneva, Switzerland
Michael Granitzer, University Passau, Germany
Christian Gütl , Graz University of Technology, Austria
Per Öster, CSC – IT Center for Science, Finland
Megi Sharikadze, Leibniz Supercomputing Centre, Munich, Germany
Stefan Voigt, Open Search Foundation, Germany

ISSN: 2957-4935
ISBN: 978-92-9083-705-3
DOI: 10.5281/zenodo.17258462

This report should be cited as:

Proceedings of 7h International Open Search Symposium #ossym2025, CSC – IT Center for Science, Helsinki, Finland and Online, 8-10 October 2025, M. Granitzer, C. Gütl, P. Öster, M. Sharikadze, S. Voigt, A. Wagner (eds). http://doi.org/10.5281/zenodo.17258462

More Information

- Conference Website on CERN Indico
  https://indico.cern.ch/e/OSSYM-2025
- Open Search Community at Zenodo
  https://zenodo.org/communities/opensearch
- Event information at the Open Search Foundation
  https://opensearchfoundation.org/en/events-osf/ossym25

Foreword

Dear readers,

In another edition of the #ossym conference series, it is our great pleasure to present the proceedings of the 7th International Open Search Symposium - #ossym25 - which takes place from 8 to 10 October 2025 in Helsinki, Finland, hosted by CSC – IT Center for Science Ltd.

In this year's symposium there are 22 accepted papers from 57 authors. The constant strong interest in open search and artificial intelligence is reflected in many inspiring contributions to topics such as "Architecture and Infrastructure", "Applications", "Alternative Search Engines", "Information Literacy", "Social Science Track", "Legal aspects", as well as "Ethics and Society".

All in all, the #ossym symposium offers a variety of formats exchange and inspirations. From scientific presentations, workshops on horizontal aspects of open search topics, to exchange with industry players and policy makers. It provides a platform for researchers engaging in the Open Web Search Initiative to present and exchange on their results with the community.

Not covered in these proceedings, nevertheless very important, are the keynote speeches featuring valuable insights into technical, governmental, community-related and ethical aspects:

- Henna Virkkunen (Executive Vice-President Tech Sovereignty, Security and Democracy of the European Commission): Welcome address,
- Aura Salla (Member of the Europen Parliament): Opening Keynote,
- Harri Ketamo (Founder & Chairman Headai): "The openness of knowledge data and its role in Future Search Solutions",
- Viivi Lähteenoja (Chief Executive Officer, The MyData Company & Doctoral Researcher at the University of Helsinki): "Searching for Trust: Reflections on Epistemic Virtue Online".

We want to express our special thanks to all authors for their sound contributions, the program committee for their valuable reviews and recommendations, all keynote and featured speakers for the inspiring insights, all sponsors for their financial support to the event, as well the team of this year's host, CSC, for all the organizational efforts. Without all these great contributions and helpful support, it would not be possible to successfully run the International Open Search Symposium.

The initial motivation for establishing the #ossym conference series was to establish a place for exchange, demonstration and inspiration on the multifaceted approaches, disciplines, angles and activities in the vibrant Open Web Search community. We are convinced that #ossym provides this place of exchange in a very fruitful and interesting manor. Every #ossym takes the Open Web Search activities and related initiatives a big step further - year after year. We look forward to meeting all participants: whether onsite or online. Your engagement, contributions and lively discussions are the foundation that the Open Search Initiative is based on and that stimulate its advancements.

In this spirit: We are very happy to already announce and look forward to the next year's gathering – #ossym2026 - to be hosted by the Germany Aerospace Center (DLR) in Berlin!

On behalf of the #ossym25 conference committee,

Michael Granitzer, Christian Gütl, Megi Sharikadze, Stefan Voigt, and Andreas Wagner

Symposium Organisation

Programme Committee

Conference Chairs

Local Chairs and Organization

Publication Chairs

Poster Session Chairs

Contents

Preface

Papers

Extended Abstracts

Appendix

Research Track Information

ARI - Architecture & Infrastructure
ASE - Alternative Search Engines
ATS - AI Tools in Science
ETS - Ethics & Society
LCO - Addressing legal challenges in Open Web Search
RLM - Retrieval Augmented Generation & Large Language Models

https://doi.org/10.5281/zenodo.17246104

ARCHITECTING THE DATASTORE FOR THE URL FRONTIER OF OPENWEBSEARCH.EU

Noor A. Fathima*[1], M. Dinzinger[2], M. Granitzer[2], A. Wagner[1]

[1] CERN, Geneva, Switzerland
[2] University of Passau, Passau, Germany

Abstract

This paper presents the architectural evolution of the URL Frontier datastore within the OpenWebSearch.eu initiative [2], transitioning from an OpenSearch-based prototype [3] to a high-throughput ScyllaDB deployment [4]. Motivated by the need for low-latency, write-optimized infrastructure to support continuous web crawling, we conducted a structured evaluation comparing OpenSearch, HBase [6], Cassandra [7], and ScyllaDB across performance, scalability, operational complexity, and infrastructure compatibility.

Our findings identified ScyllaDB as the most suitable datastore due to its shard-per-core design using the Seastar framework [8], SSD optimization, and minimal maintenance overhead. We detail the deployment process using rootless containers managed via Podman [9] and secured through Puppet-managed `nftables` [10, 11], as well as the integration with Scylla Manager for future cluster scaling [5].

BACKGROUND AND PRIOR WORK

The OpenWebSearch.eu project originally adopted OpenSearch [3] as the primary back-end datastore for its URL Frontier service, based on the reference implementation from the URL Frontier project [12]. This decision was guided by the need for a system that could support document indexing, querying, and distributed replication with minimal integration overhead. OpenSearch, being a well-supported fork of Elasticsearch, offered a mature ecosystem with open-source plugins, horizontal scalability, and real-time observability features that aligned with the project's early priorities, especially in monitoring and debugging the crawling process.

However, as crawling operations matured and scaled, limitations in OpenSearch began to surface. The URL Frontier, by nature, is a write-heavy system: URLs are rapidly inserted, updated, dequeued, and reprioritized as part of continuous crawling workflows. While OpenSearch performs well under read-intensive or search-centric workloads, it is not specifically optimized for high-throughput, low-latency write operations, particularly when frequent document updates are involved. This mismatch began to manifest in performance bottlenecks and operational complexity as the size of the index grew and the system moved closer to production-scale loads.

In response to these constraints, Apache HBase was proposed by a collaborator as a potential alternative. HBase [6], a column-family NoSQL database built on top of HDFS, is known for its suitability in high-ingestion, high-availability applications, particularly those requiring fast random writes and structured row-based access. In internal communications, it was noted that HBase's data model better aligns with the URL Frontier's query pattern, especially for managing state transitions and batch updates. The proposed deployment architecture mirrored the traditional HBase setup: a single Zookeeper node for coordination, one master, and multiple RegionServers for data handling.

However, a deeper evaluation of HBase revealed several operational challenges that made it incompatible with our existing infrastructure. Our environment is based on OpenStack-managed bare-metal servers with SSD storage and S3-compatible object storage—neither of which are natively compatible with the HDFS backbone required by HBase. Setting up and maintaining an HDFS cluster would introduce considerable overhead and complexity, particularly without an existing Hadoop ecosystem. While adaptations of HBase for object storage do exist, they typically come with reduced performance and limited community support, further disincentivizing this route.

As a result, we expanded our exploration to alternative NoSQL databases that could better utilize SSDs, integrate seamlessly with our infrastructure, and support the write-heavy nature of the Frontier workload. This led us to evaluate ScyllaDB [4], a drop-in replacement for Cassandra [7], but reengineered in C++ to deliver lower latency and better throughput on modern multi-core hardware. The following sections will outline how ScyllaDB emerged as the preferred solution and how its architecture better aligns with the operational and performance requirements of the URL Frontier at scale.

EVALUATION FRAMEWORK AND DESIGN CRITERIA

To guide the evaluation of candidate datastores for the URL Frontier, we conducted a structured diagnostic using a set of targeted questions across four domains: (1) Data Storage and Management, (2) Performance and Scalability, (3) Infrastructure and Resource Utilization, and (4) Data Consistency and Reliability.

These questions helped us identify the practical limitations of our OpenSearch-based setup and assess alternatives like Apache HBase and ScyllaDB. The diagnostic framework provided clarity on aspects such as write performance

[*] noor.afshan.fathima@cern.ch

https://doi.org/10.5281/zenodo.17246104

under load, SSD utilization, data partitioning strategies, and operational complexity in bare metal environments.

Based on the outcomes of this evaluation—summarized in Appendix A[1]—we identified ScyllaDB as the most suitable candidate. It addressed the primary bottlenecks around write-heavy operations and infrastructure alignment while offering improved predictability and observability for future scaling.

The diagnostic evaluation outlined above guided a comparative analysis of four candidate systems: ScyllaDB, OpenSearch, Cassandra, and HBase. The table1 below summarizes their key characteristics and how they align with our system's architectural, operational, and scalability requirements:

Based on this comparison and further internal testing, ScyllaDB emerged as the optimal datastore for the URL Frontier service. The next section details how its design aligns with our infrastructure and performance goals.

DESIGN MOTIVATION AND SYSTEM SELECTION

ScyllaDB is specifically designed to take full advantage of modern multi-core servers like the one used in our infrastructure (64-core, 256 GB RAM, 12 TB SSD). Its architectural features align closely with the requirements of the URL Frontier service:

Shard-per-Core Architecture

ScyllaDB operates using a shard-per-core architecture, creating 64 independent shards on our 64-core machine. Each shard is assigned to a dedicated core and handles a subset of the data. This isolation minimizes cache contention and cross-CPU communication, enabling high parallelism and efficient core utilization.

Memory Management

The 256 GB of RAM is partitioned across shards, allowing each to manage its memory independently. This enables efficient row-based caching and minimizes disk I/O. Unlike Java-based systems, ScyllaDB's C++ implementation avoids garbage collection pauses, leading to more consistent performance.

I/O Optimization

ScyllaDB employs an I/O scheduler tailored to SSDs and leverages the Seastar asynchronous framework. Each shard conducts non-blocking I/O operations directly, capitalizing on the high IOPS of SSDs to ensure fast read/write operations.

Networking and Client Requests

ScyllaDB's shard-aware drivers route client requests directly to the relevant shard, reducing request overhead and improving latency. This design allows balanced load distribution across all cores.

Thread and Storage Management

Each core runs a single thread, avoiding context-switching overhead. This ensures predictable latency and high throughput. With direct SSD access, each shard manages its own data, taking advantage of parallelism at both the compute and storage layers.

Benefits for Our Setup

- High Throughput: The architecture fully utilizes our 64-core machine for parallel processing.
- Low Latency: Combined effects of SSDs, shard-local caching, and async I/O provide rapid data access.
- Efficient Scaling: The system supports future growth in data volume while maintaining performance.

In summary, ScyllaDB's architectural principles and implementation make it a natural fit for our infrastructure and the demanding requirements of the URL Frontier. The selection was informed by a combination of empirical evaluation, expert advice, and practical deployment considerations. The following section will describe the deployment process and operational setup in more detail.

DEPLOYMENT AND OPERATIONAL CONSIDERATIONS

Initial Setup

ScyllaDB was installed and deployed on the dedicated URL Frontier server, a 64-core, SSD-equipped bare-metal node provisioned via CERN's OpenStack infrastructure. The deployment was configured for single-node operation initially, with horizontal scaling planned via Scylla Manager.

Networking and Security

Network access was managed using Puppet to enforce firewall rules with nftables. The local ruleset allowed container-level access on the CQL port (9201), while requests for perimeter-level access were submitted to the CERN Security team. The security group was updated to allow external services to connect to the ScyllaDB instance without exposing unnecessary surfaces.

Installation Method

Although the official documentation recommends direct installation on the host for maximum performance, we opted for a containerized deployment using Podman. This choice balanced maintainability, isolation, and consistency across environments with minimal observed performance overhead. The container was managed as a rootless systemd service, ensuring automatic start-up and persistent state across reboots.

Cluster Scaling and Future Setup

Plans are underway to expand the deployment into a multi-node Scylla cluster using Scylla Manager, which requires its own metadata store (own ScyllaDB instance) and an agent on each node. This setup will enable automated repair, backup,

[1] A full list of the evaluation questions and their corresponding answers can be found in Appendix A.

https://doi.org/10.5281/zenodo.17246104

Table 1: Comparison of Datastore Candidates for the URL Frontier

Feature	ScyllaDB	OpenSearch	Cassandra	HBase
Architecture	Shard-per-core, shared-nothing, based on Seastar	Distributed, document-oriented	Peer-to-peer, masterless	Master-slave, column-family on HDFS
Programming Language	C++	Java-based	Java-based	Java-based
Performance	High throughput, low latency, optimized for SSDs	Good for read-heavy and search operations	High write throughput, some latency from GC	High write throughput, higher latency from HDFS
Scalability	Linear scaling, efficient use of modern CPUs	Scales with careful shard management	Linear scaling, good for large-scale data	Scales well but requires HDFS and Zookeeper
Latency	Low latency due to C++ and direct core access	Low for read-heavy tasks, may struggle with writes	Low to moderate, affected by GC under load	Moderate, higher due to HDFS
Shard Management	Automatic, efficient shard-per-core utilization	Requires careful management to optimize	Automatic partitioning, less fine-grained	Managed with RegionServers
Caching	Built-in row-based cache, no external cache needed	External or built-in cache strategies needed	Key and row cache; external cache often needed	Block cache, may need external caching
Data Consistency	Tunable consistency	Eventual consistency, strong for specific configs	Tunable consistency	Strong consistency by default
Operational Complexity	Minimal, self-optimizing schedulers	Moderate, requires JVM tuning, shard adjustments	Requires JVM tuning, careful GC management	High, due to HDFS, Zookeeper, and RegionServers
I/O Optimization	Custom I/O schedulers for storage types	Heavily disk I/O reliant, JVM-based	Decent I/O handling, but Java limits optimization	Relies on HDFS I/O management
Use Case Fit	Real-time data ingestion, time-series, analytics	Full-text search, log analysis, data visualization	Fault-tolerant, IoT, distributed data	Bulk analytics, structured data, strong consistency
Resource Efficiency	High, avoids GC issues, direct hardware use	Moderate, JVM adds overhead	Moderate, tuning required for optimal use	Moderate, high overhead due to HDFS and Java
Global Distribution	Built-in multi-region support	Can be configured for multi-region	Good for geo-replication	Multi-datacenter possible but more complex
Administrative Effort	Low, minimal supervision needed	Moderate, needs regular tuning	Moderate, needs oversight for large clusters	High, needs dedicated management

monitoring, and performance insights. Integration and deployment of Scylla Manager are ongoing, with initial tests showing promise for future scaling and observability improvements.

Results and Early Observations

While comprehensive performance benchmarks are ongoing, early operational feedback has been highly positive. The transition to ScyllaDB significantly simplified maintenance tasks thanks to its self-tuning architecture and shard-aware design. Integration with the OWLer crawling stack was smooth, and the new datastore operates without disrupting the write-heavy demands of the Frontier service. Initial observations indicate that throughput and latency remain stable under expected load, with no degradation during integration or live crawling sessions. Future benchmarking will provide further insights into CPU utilization, storage IOPS, and overall system resilience.

Future Work and Roadmap

Several improvements are planned to extend the capabilities of the current datastore setup:

- Full Cluster Expansion: Transition from a single-node deployment to a production-grade multi-node ScyllaDB

https://doi.org/10.5281/zenodo.17246104

cluster. This will include the integration of Scylla Manager agents and a high-availability PostgreSQL metadata store.

- Automated Observability: Incorporate Prometheus and Grafana dashboards for real-time monitoring of key performance metrics, including CQL latency, disk I/O, and memory usage per shard.
- Resilience Testing: Implement fault injection tests to validate the resilience of the distributed ScyllaDB setup under node failures and high load.
- CI/CD Integration: Automate the deployment and testing of ScyllaDB containers using GitLab CI/CD pipelines to streamline updates and ensure consistency across environments.
- TTL and Retention Policies: Introduce automated TTL mechanisms for stale URLs and implement long-term storage policies to manage dataset growth efficiently.
- Dual Store Design: Explore hybrid setups where ScyllaDB handles real-time operations, while OpenSearch remains active as a searchable store for crawl metadata and indexing.
- Benchmarking Suite: Finalize a reproducible benchmarking suite to evaluate system performance under varied workloads and to inform scaling decisions moving forward.

These enhancements aim to ensure long-term sustainability, performance, and transparency in managing the URL Frontier as the OpenWebSearch.eu infrastructure continues to scale.

CONCLUSION

This paper has presented the architectural journey behind redesigning the datastore for the URL Frontier of OpenWebSearch.eu, from its initial OpenSearch-based setup to a more scalable and write-optimized ScyllaDB deployment. Through a structured evaluation framework, practical testing, and operational integration, ScyllaDB emerged as a strong fit for the project's evolving requirements.

The migration has already yielded early benefits, particularly in system maintainability and seamless integration with the crawling stack. As we move forward, ongoing improvements in scalability, observability, and resilience will further strengthen the infrastructure's role in enabling open, distributed, and ethically governed web search.

By openly documenting these design decisions and tradeoffs, we aim to support broader efforts toward building transparent and community-driven alternatives to commercial web search infrastructure.

ACKNOWLEDGEMENTS

This publication has received funding from the European Union's Horizon [1] Europe Research and Innovation Program under Grant Agreement No. 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014)

APPENDIX A: EVALUATION STUDY OF ALTERNATIVE DATASTORES

Following the non-adoption of the horizontally scaled OpenSearch cluster at CERN and subsequent requests to evaluate an alternative data store for the Frontier Service backend, this study outlines the assessment of potential replacements for OpenSearch. The initial decision to scale OpenSearch from a single node to a multi-node cluster was driven by the scope of the project, which aimed to avoid significant engineering efforts required for experimentation and deploying alternative systems. Comprehensive optimization of the OpenSearch cluster was performed, including fine-tuning configurations and leveraging CERN's substantial hardware resources to mitigate bottlenecks. This approach was deemed sufficient within the project's timeline and scope. However, due to the cited reasons that the data model of OpenSearch does not align well with the query pattern of the Frontier application and that it would remain a potential bottleneck despite horizontal scaling efforts, this assessment aims to explore alternative data store solutions. The goal is to identify a data store capable of sustaining and achieving the projected 10 TB per day crawling target while maintaining minimal operational complexity, high throughput, and low latency. The focus is on selecting a backend solution that meets the Frontier Application's performance, scalability, and reliability requirements without incurring disproportionate engineering effort for deployment and ongoing maintenance. Earlier, CassandraDB was tested by the team members but was not adopted. There is a proposal to consider HBase as an alternative, based on the assertion that its data model better aligns with the query pattern of the Frontier application. To evaluate this proposal effectively, we will analyze it using the following key assessment points, which also help with our documentation and report writing tasks.

Key Areas for Assessment: Understanding the Query Pattern of the Frontier Application Data Volume and Storage Needs Performance and Latency Requirements Scalability and Distribution Capabilities Consistency and Reliability Operational Complexity and Maintenance Caching and Performance Optimization Current and expected data model Partition Key Selection The questions related to each of the above-mentioned key areas of assessment are added as comments in this GL issue to keep the conversation organized. Please add the answers as corresponding comments.

Conclusion: Selecting a suitable alternative to OpenSearch for the Frontier Service backend requires a thorough understanding of the application's data handling patterns, performance expectations, and scalability needs. Addressing these questions will enable an informed assessment and help identify the data store that best meets the service's requirements.

https://doi.org/10.5281/zenodo.17246104

1. Understanding the Query Pattern of the Frontier Application

Question: What is the detailed query pattern of the Frontier Application?

Context: This will help determine whether the application primarily requires write-heavy operations, read-heavy operations, or a balance of both. Additionally, it will clarify if the application relies on sequential reads/writes, complex queries, or real-time data processing.

Action: Document the typical query patterns, including examples of common read and write operations and their frequency.

Response: The application is both read-heavy and write-heavy. However, the spectrum of queries is small. It basically breaks down to only three kinds of queries.

The backend persists the crawl space: this is a large set of URLs, identified by an URL ID. This URL ID is a hash of the normalized URL, hence every URL is uniquely identified by the hexadecimal string representation of this SHA-256 hash. Besides that, meta information (as map/dictionary) is stored alongside each URL ID and URL. Most important metadata field is `nextFetchDate`, which specifies the timestamp of the next planned fetch. Additionally, the metadata map also contains a set of tags (like HTML, Adult, etc.), which impact the scheduling of URLs for crawling.

To put it in a nutshell, the elements of the crawl space are uniquely identified by the URL ID and comprise several columns, namely URL, `nextFetchDate` and a static list of metadata fields. For the sake of distributing the crawl space among crawlers, the URL Frontier application divides the crawl space along the URL ID in 512 subsets/batches.

The query pattern looks as follows (three kinds of queries):

- **Scan operation over a subset of the crawl space:** This is a search request to retrieve new URLs to be fetched. The Frontier application scans over one subset/batch of the crawl space, which is ordered by the URL ID, and looks for all elements that meet certain filter criteria. The default filter criteria are: `nextFetchDate` has to be in the past (hence it is scheduled for crawling) and the element is tagged as HTML.

- **Exists operation for a set of URLs:** After crawling, the crawler logs return the URL as well as discovered outlinks. For integrating these potentially new links to the crawl space, the Frontier application computes the URL ID for all these links and looks up whether these IDs are already persisted in the crawl space. Outlinks that have already been discovered can be thrown away and it is no expensive update operation necessary.

- **Update of crawled URLs and Insert of new discovered links:** The update/insert operations are most expensive among the three kinds of queries in the query pattern. Crawled URLs, which are already in the crawl space, are updated with a new `nextFetchDate` and refined meta information. Discovered links are inserted with a `nextFetchDate` in the near future and a default set of meta information.

Follow-Up Question: Is the application relying on random read/writes or sequential read/writes? Are the updates large-scale modifications or incremental changes? Given that the `nextFetchDate` determines how frequently data is updated, how is `nextFetchDate` determined?

Response: **Read/write pattern:** The read operations are sequential with respect to the URL ID, thus it is a Scan over the crawl space, which is ordered by URL ID, retrieving new URLs to be crawled next. The write operations are completely random.

Scale of updates: The updates on the Frontier applications are ongoing, in order to extend our crawling from StormCrawler-only to more crawlers. In the Deliverable D1.2, I framed it as "Stream-based processing" (StormCrawler) and "Batch-based processing" (others). When implementing the Batch-based processing, one aspect became clear to me, which will—hopefully—increase performance significantly.

The backend has to handle both read-heavy and write-heavy querying. Hence, the URLs have to be persisted in a shallow way, as an ordered list with a hash-based identifier. Consequently, all read operations can be sequential, and furthermore the number of write operations can be decreased by employing `exists` operations plus cheap `insert` operations instead of expensive `update` operations.

OpenSearch actually allows cheap `exists` operations, yet it has two shortcomings: It is made for Search requests, but not for Scans. These require pagination; a lot of data has to be loaded into RAM, which makes these sequential reads expensive. Beyond that, the latency of write operations seems to not scale well, as it is directly dependent on the underlying index structures. I hope that HBase behaves differently in this regard. By definition, it only has a single index (like e.g. Cassandra, as well) and thus data access and (data manipulation) should be realized in a simpler, yet more performant way.

How is `nextFetchDate` determined? It is determined by a software component of the URLFrontier called Scheduler. Basically, it adds two weeks to the current time, plus some minor adaptations to prioritize certain content (if the web page content has changed since the last crawl, the `nextFetchDate` is sooner; if it is Adult content, it is later, etc.)

Clarification Request: When updates are performed on the Frontier Application, are they typically large-scale modifications affecting multiple records or incremental changes affecting individual records?

https://doi.org/10.5281/zenodo.17246104

If I have understood correctly based on the answer to the next question too, then:

- Since the `nextFetchDate` is updated on an ongoing basis, with adjustments influenced by real-time changes to the content and other factors, this implies that updates are generally incremental rather than large-scale modifications.
- The updates seem to be frequent and distributed over time, rather than occurring in periodic bulk updates, due to the dynamic nature of scheduling based on content changes and prioritization rules.

Response: Yes, every update comprises incremental changes. So every write operation updates the URL items that were crawled (so 1) and inserts newly discovered links (I try to keep this number small, so 0–8).

However, one could collect these update operations as a bulk and send a bulk operation against the backend. This "bulking" could be implemented in custom logic (for HBase, I haven't intended to do this so far) or, for the case of OpenSearch, it is already implemented in the logic of the Java client library.

2. Data Volume and Storage Needs

Question: What is the current and projected volume of data stored and processed per day? **Context:** Establishing data volume requirements is essential to ensure that the new data store can handle current loads and scale efficiently to be able to crawl 10 TB/day. **Action:** Review current data metrics and perform projections based on growth patterns.

3. Performance and Latency Requirements

Question: What are the current performance benchmarks (e.g., write latency, read latency) that must be maintained or improved? **Context:** Identifying performance metrics will help evaluate which data stores meet or exceed these benchmarks under similar or greater loads. **Action:** Analyze existing performance data and define acceptable latency thresholds for both read and write operations. **Response:** Update operations are most expensive and their latency has been the bottleneck in the OpenSearch setup. In optimal case, the costs of an Update operation should be agnostic to the size of the crawl space. However, for OpenSearch, as the index was growing, all read and write operations became more expensive. The write latency has been a bigger problem as the read latency, as the stream-based processing of the StormCrawler is unfortunately not robust enough to handle congestions resulting from slow updating of the crawl space.

4. Scalability and Distribution Capabilities

Question: How well can the candidate data stores scale horizontally and distribute data across nodes? **Context:** The ability to scale without significant operational overhead or performance loss is crucial for supporting high-volume data ingestion and processing. **Action:** Compare the scalability characteristics of each alternative (e.g., linear scaling, horizontal scaling support) against OpenSearch.

5. Consistency and Reliability (urgent)

Question: What level of consistency is required by the Frontier Application, and can the alternative data stores provide this? **Context:** The Frontier Application may have specific requirements for strong, eventual, or tunable consistency. The choice of data store should align with these needs to maintain data integrity. **Action:** Clarify the consistency model needed and assess each candidate's ability to provide it, including their mechanisms for data replication and failure recovery. **Response:** I'd say the Frontier application requires strong consistency for update operations. After an update of a crawled URL with a new nextFetchDate, it has to be guaranteed that it will not be read with the old - now invalid - nextFetchDate again, meaning that it is crawled twice within a short period of time. In reality, eventual consistency would however probably be also okay as a scan over a subset/batch of the crawl space presumably takes longer than the eventual consistency to realize. For insert operations of newly discovered links, eventual consistency is okay.

6. Operational Complexity and Maintenance

Question: What is the operational overhead associated with maintaining the data store? **Context:** Ease of maintenance, monitoring, and scaling is important to minimize downtime and manual intervention. **Action:** Evaluate the complexity of setting up, managing, and scaling each data store, considering factors like configuration, monitoring, and required expertise.

7. Caching and Performance Optimization

Question: Does the data store have built-in caching mechanisms, or will it require external caching solutions to meet performance needs? **Context:** Caching capabilities can greatly impact the efficiency of read-heavy workloads and reduce latency. **Action:** Assess the need for built-in vs. external caching solutions for each candidate and their impact on overall performance.

8. Current and Necessary Data Model (not urgent)

Question: What is the current and expected data model that is necessary? **Context:** Understanding the expected data model is crucial to determine whether HBase or any other alternative is a better fit for the Frontier application. This involves identifying how data is structured, accessed, and updated within the application. The analysis should focus on whether the data model aligns with the application's query patterns, data relationships, and workload characteristics (e.g., write-heavy or read-heavy operations). **Action Points:**

- Review and document the current data model used by the Frontier application.
- Identify key data attributes and relationships critical for the application's functionality.

https://doi.org/10.5281/zenodo.17246104

- Analyze the compatibility of the expected data model with HBase and compare it with OpenSearch and other potential data stores.

9. Partition Key Selection (urgent)

Question: What is the current partition key? **Context:** Understanding how the partition key was selected is vital for assessing the current system's effectiveness and identifying potential performance bottlenecks or hotspots. The partition key plays a critical role in data distribution, load balancing, and system performance across nodes or shards. Analyzing the current partition key choice helps ensure data is evenly distributed, preventing nodes or regions from being overwhelmed. **Action Points:**

- What partition key is currently being used for the Frontier Application, and why was this choice made?
- Can you provide examples of how the current partition key affects data distribution across nodes or shards?
- Have you observed any hotspots or imbalances in data distribution related to the chosen partition key? If so, what measures have been considered to address these?

Response: The prior OpenSearch-based backend uses a hash of the PLD (Paid-Level Domain; domain) as partition key. The hash is an integer and further taken modulo NUM_BATCHES, which has been 400. So we have 400 partitions and each partition is one index (and each index has one shard, so 400 OpenSearch shards). The hash of the PLD is not perfectly evenly distributed, but close enough for our case. Plus it is ensured that URLs of the same domain are in the same partition. However, when moving to batch-based processing, this is not necessary anymore. The only identifier is the URL ID, and read operations are either sequential scans over the range of URL IDs or exists/get operations looking for single URL IDs. Data partitions can be arbitrary as long as URL IDs are persisted sequentially. There has not been any hotspots or imbalances in data distribution.

REFERENCES

[1] Horizon Europe, `https://research-and-innovation.ec.europa.eu/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-europe_en`

[2] OpenWebSearch.eu Consortium. *OpenWebSearch Project Overview*. [Online]. Available: `https://openwebsearch.eu`

[3] OpenSearch Project. *OpenSearch Documentation*. [Online]. Available: `https://opensearch.org/docs`

[4] ScyllaDB Inc. *ScyllaDB Documentation*. [Online]. Available: `https://docs.scylladb.com`

[5] ScyllaDB Inc. *Scylla Manager Overview*. [Online]. Available: `https://docs.scylladb.com/operating-scylla/manager`

[6] George, L. *HBase: The Definitive Guide*. O'Reilly Media, Inc., 2011.

[7] Lakshman, A., Malik, P. "Cassandra: A decentralized structured storage system," in *ACM SIGOPS Operating Systems Review*, vol. 44, no. 2, pp. 35–40, 2010.

[8] ScyllaDB Inc. "Seastar: High performance server-side application framework," [Online]. Available: `https://www.scylladb.com/2017/02/14/seastar-high-performance-application-framework`

[9] Red Hat. *Podman Documentation*. [Online]. Available: `https://podman.io`

[10] The Netfilter Project. *nftables wiki*. [Online]. Available: `https://wiki.nftables.org`

[11] Puppet Inc. *Puppet Documentation*. [Online]. Available: `https://puppet.com/docs`

[12] OpenSearch implementation of the URL Frontier `https://github.com/PresearchOfficial/opensearch-frontier/`

https://doi.org/10.5281/zenodo.17228339

EXTRACTING AND UTILIZING STRUCTURED DATA FROM THE OPEN WEB INDEX

L. Caspari*, M. Dinzinger, J. Mitrovic, M. Granitzer, University of Passau, Passau, Germany

Abstract

Structured data is a valuable source of information that can be found on many web pages and can be extracted efficiently during crawling. It is often encoded in the form of JavaScript Object Notation for Linked Data (JSON-LD) or Microdata using schema.org definitions for entities such as FAQ pages or addresses, allowing efficient parsing and extraction of data. The OpenWebSearch.EU (OWS) [Hendriksen et al.(2024)] project, which publicly releases crawled web data on a regular basis, is a useful source for fresh structured data, as published datasets contain specific columns for JSON-LD and Microdata encountered during crawling. In this paper, we present initial statistics on the occurrence of structured data in the OWS datasets, focusing on the presence of certain entities in schema.org, namely Frequently Asked Questions (FAQs), opening hours, phone numbers, and addresses. Additionally, we discuss two practical application scenarios of the extracted data. In our first use case, in line with our previous work [Dinzinger et al.(2025)], we demonstrate how FAQ data can be used to construct multilingual Q&A-style datasets, which can be used to train large language models (LLMs) for tasks like question answering or retrieval. In our second case, we show the potential of structured data to enrich map applications and improve user experience. These use cases exemplify the value of structured data and demonstrate the benefits of its systematic extraction and integration into real-world applications.

INTRODUCTION

Structured data provides an important source of information about webpages that can easily be parsed and ingested by downstream applications like search engines or map providers. It can be used to enrich the results shown to users, i.e. by displaying the address and opening hours of a shop. Structured data is specified by webmasters using schema.org definitions of entities like addresses, opening hours or FAQs. While it can be specified in a variety of formats, JSON-LD and Microdata have emerged as popular choices [Volpini et al.(2024)]. As defining information in these formats requires additional effort from webmasters, the data is generally of high quality and can easily be extracted during crawling. However, many downstream applications require the extracted data to be fresh, necessitating frequent revisits of webpages. An important source for fresh structured data is the OpenWebSearch project [Granitzer et al.(2024)], which releases crawled JSON-LD and Microdata as part of their regularly released datasets. Thus, the project presents an important resource for publicly available and fresh structured data. To better understand the prevalence of JSON-LD and Microdata within the OWS index, we analyze datasets from five days in February 2025 with a total size of 1.6TB. We find that more than half (54.2%) of the crawled webpages use JSON-LD or Microdata. However, when attempting to extract specific schemas, the ratio drops significantly, i.e. phone numbers can only be found on 1.8% of webpages.

Given these extracted schemas, we establish two use cases for structured data, namely leveraging FAQ pages to construct a Q&A dataset that can be used to train or evaluate LLMs on question answering or retrieval tasks, and extracting FAQs, opening hours, addresses and phone numbers to enhance map applications. Our use cases demonstrate the potential of structured data in downstream applications and the importance of having publicly available and fresh data to enable researchers and companies to build upon the wealth of information contained within. The code to reproduce our results is available on GitHub[1].

The remainder of this paper is structured as follows: After looking at related work, the methodology section gives an overview of the schema definitions we focus on and explains our extraction approach. The following section contains statistics about the occurrence of Microdata and JSON-LD as well as of specific schemas like addresses or phone numbers. We then detail our two use cases for the extracted data before discussing limitations and problems with data quality. Finally, we conclude our paper with a summary of our findings and future extensions of our work.

RELATED WORK

Using schema.org[2] classes to represent information about entities like restaurants, events or products has become increasingly prevalent since the introduction of schema specifications in 2011 [Brinkmann et al.(2023), Volpini et al.(2024)]. Established by Google, Bing, Yahoo and Yandex to offer webmasters a unified way of defining structured data, it provides information in machine-readable and widely accepted formats, with JSON-LD and Microdata being common ways in which structured data is made available. Microdata is an extension of the HTML5 specification[3] and is thus added directly to the HTML tags themselves, whereas JSON-LD is specified within one set of script tags[4], making it easy to extract. While both formats have seen increasing adoption in the past years with JSON-LD being present on 41% of webpages and Microdata on 26% in 2024 [Volpini et al.(2024)], their usage patterns differ. As analyzed by Volpini et al., JSON-LD is most

* laura.caspari@uni-passau.de

[1] `https://github.com/padas-lab-de/owi-sdm`

[2] `https://schema.org/`

[3] `https://html.spec.whatwg.org/multipage/microdata.html`

[4] `https://json-ld.org/`

https://doi.org/10.5281/zenodo.17228339

commonly used for organization data, local businesses and product listings, whereas Microdata often specifies webpage structure or site navigation. Apart from the imbalance of schema usage between the different formats, adoption of schema annotations also varies depending on the domain with a much higher usage of entities like products or local businesses [Brinkmann et al.(2023)] than entities related to educational resources [Navarrete et al.(2019)].

The higher prevalence of structured data increases the value of extracting its content for downstream applications. Apart from using structured data to increase search visibility [Recalde et al.(2021)], it can also be extracted to provide training data for machine learning models [Peeters et al.(2020), Dinzinger et al.(2025)]. While the usage of schema-based annotations requires additional effort by webmasters and thus is generally of high quality, applications using structured data still need to filter out low quality samples. For instance, a high percentage of schema.org dataset annotations do not describe actual datasets [Alrashed et al.(2021)], drastically limiting the usability of this schema for dataset search. Similarly, certain properties of common entities like products, e.g. the product ID or category, are seldom filled [Brinkmann et al.(2023)].

METHODOLOGY

The following paragraphs offer a general introduction into the specification of structured data and define the exact schema classes that are of interest. Subsequently, an overview of the extraction pipeline is provided along with the format in which extracted data is stored.

Defining Structured Data with Schemas

In our context, structured data is specified using entities defined by the schema.org type hierarchy. For each entity, e.g. an FAQ page, schema.org defines the properties and its type, which are specified in a key-value-based manner. While there are various ways of specifying structured schema data, this paper will focus on JSON-LD and Microdata, which are part of the datasets published by OWS. Due to our current use cases, we will specifically consider schemas for defining phone numbers[5], addresses[6], opening hours[7] and FAQ pages[8].

Figure 1 shows an excerpt of the aforementioned schemas in JSON-LD which were encountered when crawling the webpage of a Subway store located in Seattle. The structured data contains important information about the store which can be used to enrich downstream applications.

Extracting and Merging Schemas

While structured data is a valuable resource, it is not available for every webpage. Therefore, we first use owilix[9] to download OWS datasets from five different days in February 2025, which contain files in Parquet format[10]. Specifically, we use the datasets published on the 19th and 21st-24th of February. As the Parquet files include specific columns for Microdata and JSON-LD, we subsequently filter out all entries for which both columns are empty and only use columns that are of interest to us, reducing the initial size of 1.6TB by a factor of four. We then apply our extraction code on the filtered data to obtain FAQs, opening hours, addresses and phone numbers contained in the structured data, with the extracted information being saved to Parquet files. As the structure of the data is quite dependent on the schema, we store each in a separate Parquet file with the exception of phone numbers and addresses, which are merged together. The resulting files organized per day are available for download from our MinIO instance[11].

```
{
  ...
  "telephone": "(425) 614-3256",
  "address": {
    "@type": "PostalAddress",
    "addressCountry": "US",
    "addressLocality": "Bellevue King",
    "addressRegion": "WA",
    "postalCode": "98007",
    "streetAddress": "1410 156th Ave NE"
  },
  "openingHours": ["Mo 08:00-22:00", "Tu 08:00-22:00", "We
    08:00-22:00", "Th 08:00-22:00", "Fr 08:00-22:00", "Sa
    09:00-22:00", "Su 09:00-22:00"]
  "@type": "FAQPage",
    "mainEntity": [{
      "@type": "Question",
      "name": "How can I place a Subway Catering order?",
      "acceptedAnswer": {
        "@type": "Answer",
        "text": "To place an order, visit us online at
          catering.subway.com or call your local restaurant."
      }
      ...
    }]
  ...
}
```

Figure 1: An excerpt of JSON-LD extracted from the page of a Subway store in Seattle.

DATA EXPLORATION

To get an initial idea about how often structured data appears in the OWS crawls, we analyze the filtered datasets from February 2025 and find that 54.2% of webpages contain microdata or JSON-LD. However, as shown in Table 1, this number quickly drops when looking at a specific schema. In fact, all schemas we are interested in occur on less than 2.1% of webpages.

Taking a closer look at the individual schemas, we also observe that a significant number of them are malformed or contain invalid data when applying simple sanity checks. To ensure some basic quality of the extracted data, we ignore entries that only contain empty values. Furthermore, for phone numbers and opening hours, we ensure that the ex-

[5] https://schema.org/telephone

[6] https://schema.org/address

[7] https://schema.org/openingHours

[8] https://schema.org/FAQPage

[9] https://opencode.it4i.eu/openwebsearcheu-public/owi-cli

[10] https://parquet.apache.org/

[11] https://console.share.innkube.fim.uni-passau.de/browser/public/ows-extracted%2F

https://doi.org/10.5281/zenodo.17228339

Table 1: Occurrence of specific schemata in OWS datasets.

	Schema Name	#	%
☎	Telephone	5,352,078	1.78
📍	Address	6,121,713	2.04
❓	FAQPage	1,645,691	0.55
🚪	OpeningHours	2,210,530	0.73
⌛	OpeningHoursSpecification	1,639,338	0.54

tracted string contains at least one digit. Figure 1 illustrates that these simple measures already lead to a large number of discarded entries, showing that many webmasters struggle with obliging to the schema format or insert empty or invalid values. A manual inspection of parts of the extracted data further revealed that while most entries contain sensible information, some webmasters used unhelpful default values, i.e. "question" and "answer" in extracted Q&A pairs. Another issue with data quality is posed by entries that only contain partial information, i.e. an address that only mentions the city, but not the street address of the entity. Further processing of the extracted data to ensure high quality thus poses an important but non-trivial task for our multilingual data.

USE CASES

In the following sections, we describe two real-world use cases of structured data. The first use case, the extraction of FAQ-style annotations to build a Q&A dataset, has already been implemented. The second use case of extracting structured data to enrich map applications is a work in progress in collaboration with Murena[12], a company that provides deGoogled and privacy preserving smartphones and cloud services.

FAQ Dataset

FAQ pages represented in structured data provide an interesting resource for building Q&A datasets. Their natural separation into questions and answers makes it easy to leverage them for question answering tasks. Furthermore, as the FAQ page schema requires an answer to be specified as either accepted or suggested, the schema contains an implicit relevance signal which can be extracted to make the dataset usable for retrieval tasks. Our recent work [Dinzinger et al.(2025)], in which we built a large-scale multilingual retrieval dataset by extracting FAQ page schemas from data provided by the Web Data Commons (WDC) project [13], clearly demonstrates the use of FAQ-style structured data. Furthermore, we show that multilingual FAQs can be used to build bilingual corpora for a large number of language combinations. Both WebFAQ retrieval[14] and WebFAQ bitext[15] are available on HuggingFace and as part of the Massive Text Embedding Benchmark (MTEB) [Muennighoff et al.(2023)] python package.

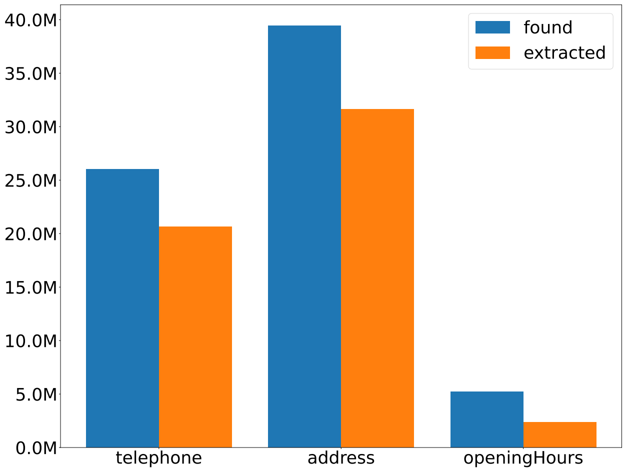


Figure 2: The number of found and extracted entries per schema in millions.

While the WDC dumps provide a large resource of natural Q&A data, they are updated only on a yearly basis, thus likely containing many stale FAQs. The regularly published OWS datasets can alleviate this problem by providing fresh data for crawled web pages. We therefore apply the procedure developed to generate WebFAQ on the OWS data, extracting around 9.95 million Q&A pairs across the five days. To build a multilingual retrieval corpus, we perform language classification on the extracted Q&A pairs using FastText [Joulin et al.(2016)]. Figure 3 shows the distribution of the 10 most common languages found in the extracted FAQ data. While English unsurprisingly occurs most often, we also extract a large number of Q&A pairs for other languages like German, Spanish or French. Similarly to WebFAQ, the FAQs extracted from OWS data are available as a collection of multilingual retrieval datasets on HuggingFace[16].

Enriching Map Applications

Apart from the FAQPage schema serving as a valuable starting point for Q&A datasets, the schemas we have extracted can also serve as a valuable resource for (non-)commercial map applications. To this end, we are currently collaborating with Murena, in an effort to enhance the data provided by OpenStreetMap[17]. While OpenStreetMap provides useful information like addresses or opening hours for points of interest, driven by a community of human mappers that contribute the data, certain parts of this information like the opening hours of a shop might change too frequently to be kept up to date. This can lead to undesirable situations if users rely on incorrect data, e.g. if they choose to visit a shop just to find that the opening hours are outdated and the shop has already closed for the day. To alleviate this

[12] https://murena.com/

[13] https://webdatacommons.org/

[14] https://huggingface.co/datasets/PaDaS-Lab/webfaq-retrieval

[15] https://huggingface.co/datasets/PaDaS-Lab/webfaq-bitexts

[16] https://huggingface.co/datasets/PaDaS-Lab/owi-faq-retrieval

[17] https://www.openstreetmap.org

https://doi.org/10.5281/zenodo.17228339

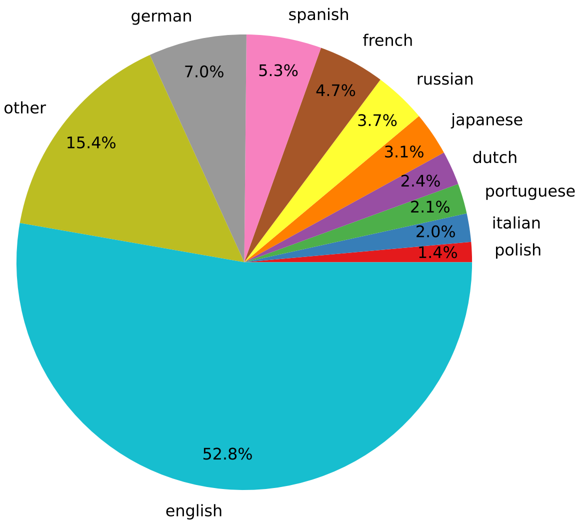
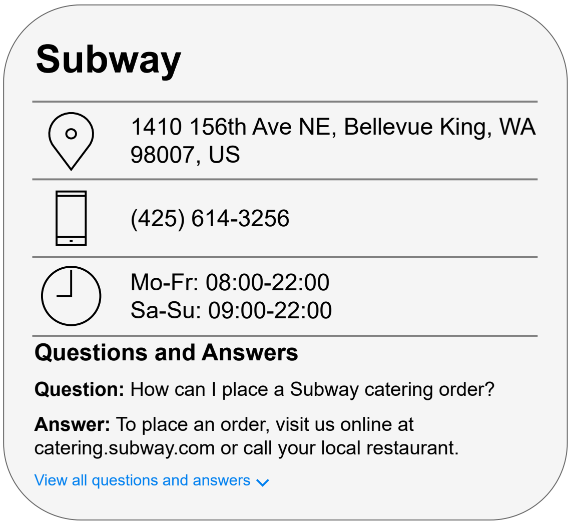


Figure 3: Language distribution for the 10 most common languages on FAQ pages.



Subway

1410 156th Ave NE, Bellevue King, WA
98007, US

(425) 614-3256

Mo-Fr: 08:00-22:00
Sa-Su: 09:00-22:00

Questions and Answers

Question: How can I place a Subway catering order?

Answer: To place an order, visit us online at catering.subway.com or call your local restaurant.

Figure 4: The data extracted for a Subway store in Seattle and how it could be presented to users.

problem, we aim to crawl and extract information available in structured data for specific URLs that Murena is interested in on a regular basis. As an initial test, we crawled 10,547 URLs representing points of interest in the area of Seattle and extracted phone numbers from 122 (1.2%) webpages, addresses from 326 (3.1%), FAQs from 290 (2.7%) and opening hours from 711 (6.7%). Although the absolute number of extracted schemas remains low, they can still contribute valuable and fresh information for a large number of locations. As an example, Figure 4 demonstrates how the data extracted from the JSON-LD partially shown in Figure 1 can be presented to users. The data was extracted from the webpage of a Subway store in Seattle and clearly contains information that would benefit a map application.

LIMITATIONS

While extracting information from structured data seems straightforward at first glance, working with real-world data has proven to be more challenging. One such challenge is posed by the schema definitions themselves. For instance, there are two different ways to specify opening hours, namely using the openingHours[18] schema that provides the information as a dictionary with a defined set of keys or as a simple text as shown in Figure 1. Another common issue are missing values for some fields, fields containing placeholder values or data not conforming to the specified schema.

As our main focus lies on extracting the information, we address the first problem by implementing extractors specific for each schema type and storing the information in separate columns of the output Parquet files. Thus, we leave it to downstream applications to merge data from different schemas describing the same entity. While we apply simple sanity checks to the extracted data like checking if the schema contains only empty strings or whether dates or phone numbers contain at least one digit, doing comprehensive filtering on a multilingual corpus is a non-trivial task. As such, we do not apply any complex filtering techniques on the extracted data to ensure its semantic validity. Additionally, we are unable to verify the correctness or freshness of the extracted data, i.e. if a phone number found on the page of a shop really belongs to it and whether the number is still up to date. Thus, downstream applications wishing to use the extracted information will likely have to implement additional filtering techniques on top of our data to ensure high quality.

CONCLUSION

Structured data has proven to be an easily extractable and valuable resource for various application scenarios. In this work, we focused on analyzing and extracting certain types of Microdata and JSON-LD from datasets provided by the OWS project. We found that while structured data is available on more than half of the crawled webpages, the occurrence of specific schemas like addresses or opening hours is much less common. Nevertheless, we demonstrate the usefulness of the extracted data in two application scenarios, first generating a question answering and retrieval dataset using the FAQPage schema, and then providing additional information like opening hours, addresses and phone numbers for points of interest, which can be used to enrich map applications.

As we believe that providing information extracted from structured data is of general interest, we plan to integrate the extraction mechanism as a regular step in the OWS preprocessing pipeline and create a new collection index for extracted structured data. This would allow interested parties to download only the extracted data instead of the much

[18] `https://schema.org/openingHours`

https://doi.org/10.5281/zenodo.17228339

larger standard OWS datasets, as well as to update information on points of interest on a regular basis without having to set up their own extraction pipelines. We will also expand our work with Murena to crawl more points of interest and provide the extracted information as part of the new collection index.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, `https://doi.org/10.3030/101070014`).

REFERENCES

[Alrashed et al.(2021)] Tarfah Alrashed, Dimitris Paparas, Omar Benjelloun, Ying Sheng, and Natasha Noy. 2021. Dataset or Not? A Study on the Veracity of Semantic Markup for Dataset Pages. In *The Semantic Web – ISWC 2021*, Andreas Hotho, Eva Blomqvist, Stefan Dietze, Achille Fokoue, Ying Ding, Payam Barnaghi, Armin Haller, Mauro Dragoni, and Harith Alani (Eds.). Springer International Publishing, Cham, 338–356.

[Brinkmann et al.(2023)] Alexander Brinkmann, Anna Primpeli, and Christian Bizer. 2023. The Web Data Commons Schema.org Data Set Series. In *Companion Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 136–139. `https://doi.org/10.1145/3543873.3587331`

[Dinzinger et al.(2025)] Michael Dinzinger, Laura Caspari, Kanishka Ghosh Dastidar, Jelena Mitrović, and Michael Granitzer. 2025. WebFAQ: A Multilingual Collection of Natural Q&A Datasets for Dense Retrieval. arXiv:2502.20936 [cs.CL] `https://arxiv.org/abs/2502.20936`

[Granitzer et al.(2024)] Michael Granitzer, Stefan Voigt, Noor Afshan Fathima, Martin Golasowski, Christian Guetl, Tobias Hecking, Gijs Hendriksen, Djoerd Hiemstra, Jan Martinovič, Jelena Mitrović, Izidor Mlakar, Stavros Moiras, Alexander Nussbaumer, Per Öster, Martin Potthast, Marjana Senčar Srdič, Sharikadze Megi, Kateřina Slaninová, Benno Stein, Arjen P. de Vries, Vít Vondrák, Andreas Wagner, and Saber Zerhoudi. 2024. Impact and development of an Open Web Index for open web search. *Journal of the Association for Information Science and Technology* 75, 5 (2024), 512–520. `https://doi.org/10.1002/asi.24818`

[Hendriksen et al.(2024)] Gijs Hendriksen, Michael Dinzinger, Sheikh Mastura Farzana, Noor Afshan Fathima, Maik Fröbe, Sebastian Schmidt, Saber Zerhoudi, Michael Granitzer, Matthias Hagen, Djoerd Hiemstra, Martin Potthast, and Benno Stein. 2024. The Open Web Index. In *Advances in Information Retrieval*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer Nature Switzerland, Cham, 130–143.

[Joulin et al.(2016)] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. `https://doi.org/10.48550/ARXIV.1612.03651`

[Muennighoff et al.(2023)] Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. arXiv:2210.07316 [cs.CL]

[Navarrete et al.(2019)] Rosa Navarrete, Lorena Recalde, Carlos Montenegro, and Sergio Luján-Mora. 2019. Analyzing Embedded Semantic with JSON-LD and Microdata for Educational Resources in Large Scale Web Datasets. In *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*. 1133–1138. `https://doi.org/10.1109/CSCI49370.2019.00214`

[Peeters et al.(2020)] Ralph Peeters, Anna Primpeli, Benedikt Wichtlhuber, and Christian Bizer. 2020. Using schema.org Annotations for Training and Maintaining Product Matchers. In *Proceedings of the 10th International Conference on Web Intelligence, Mining and Semantics* (Biarritz, France) *(WIMS 2020)*. Association for Computing Machinery, New York, NY, USA, 195–204. `https://doi.org/10.1145/3405962.3405964`

[Recalde et al.(2021)] Lorena Recalde, Rosa Navarrete, and Fernando Pogo. 2021. Making Open Educational Resources Discoverable: A JSON-LD Generator for OER Semantic Annotation. In *2021 Eighth International Conference on eDemocracy & eGovernment (ICEDEG)*. 182–187. `https://doi.org/10.1109/ICEDEG52154.2021.9530872`

[Volpini et al.(2024)] Andrea Volpini, Jarno van Driel, Ryan Levering, Nurullah Demir, and James Gallagher. 2024. *Structured data*. HTTP Archive, Chapter 3. `https://doi.org/10.5281/zenodo.14065771`

ISBN: 978-92-9083-705-3 Open Search Symposium 2025 - #ossym2025 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.17233862

AUTOMATING SYSTEMATIC REVIEWS: API-POWERED BIBLIOGRAPHIC DATA RETRIEVAL MODULE FOR NEUTRINOREVIEW

E. Sandner*[1,3], I. Ilicic[3], U. Sharma[1, 4], I. Jakovljevic[1], A. Simniceanu[2],
L. Fontana[2], A. Henriques[1], A. Wagner[1], C. Gütl[3]
[1]CERN, 1211 Geneva, Switzerland
[2]WHO, 1211 Geneva, Switzerland
[3]Graz University of Technology, 8010 Graz, Austria
[4] University of Delhi, Delhi, India

Abstract

Systematic reviews are the gold standard for synthesizing research evidence but are highly time- and resource-intensive, often requiring months or even years to complete. While existing review management systems provide support for the screening phase, early steps such as literature retrieval typically require external execution, creating inefficiencies and potential for error. This paper presents a proof-of-concept implementation of an automated data retrieval and deduplication module integrated into the NeutrinoReview platform. The module supports multi-database querying, metadata normalization, and duplicate resolution through a similarity-based algorithm. To assess its performance, a controlled user study compared task completion times with Rayyan, a widely used review tool. Seven novice participants performed retrieval and deduplication workflows for four medical reviews using both systems. Results showed that NeutrinoReview reduced completion time by an average of 75%. These findings highlight the potential of automation to significantly reduce human workload in the early stages of systematic reviews. While NeutrinoReview serves as a proof of concept, the demonstrated efficiency gains underscore the value of integrating robust retrieval and deduplication modules into current and next-generation review management tools to enhance timeliness, consistency, and reliability in evidence synthesis.

INTRODUCTION

A systematic literature review is a method for identifying, evaluating, and synthesizing all research relevant to a specific question, topic, or phenomenon of interest. By integrating findings from all potentially relevant studies on a given question, a systematic review (SR) provides the most reliable methodology for drawing evidence-based conclusions [1]. Consequently, SRs hold a central role in medical research and practice, where they inform evidence-based decision-making and the development of clinical guidelines [2]. SRs are equally important in the context of primary research. Conducting a systematic review of the existing evidence prior to initiating a new study is critical for ensuring its quality and relevance [3]. Comprehensive knowledge of prior studies helps identify research gaps and formulate meaningful questions that warrant further investigation. Moreover, insights from earlier work support the optimal design of new studies [4, 5]. However, the rigor of the process makes SRs highly time- and resource-intensive. Completing a single SR typically takes several months and, in some cases, even years [6–8].

Because SRs are both time- and resource-intensive, their lengthy process often fails to meet the needs of decision-makers, particularly in contexts where rapid evidence syntheses are required to inform urgent decisions or where research resources are limited. Tools to reduce the human workload in systematic reviews are available. Most review management systems primarily support the study selection process by streamlining and (semi-)automating the screening tasks [9]. However, the initial search typically has to be conducted externally, after which candidate studies are imported into the tool. Deduplication is generally well supported, although not all tools provide simple one-click functionality and instead rely on more complex options. The extent to which further streamlining of the search and deduplication phases could translate into measurable time savings for human reviewers has not yet been systematically investigated. Therefore, this paper presents a proof-of-concept implementation of a data retrieval and deduplication module for review management systems. In an experimental study, the time required to perform these two steps using the proposed module was compared with the traditional workflow in an established review tool.

The results demonstrate that the proposed module reduces the human workload for these tasks by 75%. The findings underscore the importance of developing robust data retrieval systems and integrating them into existing or next-generation review management tools.

BACKGROUND AND RELATED WORK

Conducting a systematic review typically begins with the development of a project protocol, which outlines the research objectives and provides a detailed roadmap for executing the review. Central to this protocol is the search strategy, which defines both the literature sources to be included and the search strings that will be applied to retrieve potentially relevant studies. After the protocol has been finalized and, where appropriate, published in a registry such as PROSPERO[1], the literature search is conducted separately

* elias.sandner@cern.ch

[1] `https://www.crd.york.ac.uk/prospero/`

https://doi.org/10.5281/zenodo.17233862

for each database. Typically, this process involves accessing each database's search engine via a web browser. The search string defined in the protocol is entered into the search interface, the search is executed, and the results are then exported and downloaded. Once all searches have been completed, the resulting files must be merged and deduplicated in preparation for the subsequent screening phase. During screening, the eligibility of each study is first assessed based on the title and abstract, with those deemed eligible then undergoing a full-text evaluation. After eligible studies have been identified, relevant information must be extracted, the risk of bias assessed through critical appraisal, and the findings synthesized into a manuscript.

While study selection, data extraction, and critical appraisal have been identified as highly time-intensive tasks, literature search has been found to require comparatively less time [10]. However, since executing the search and transferring files between tools are tasks that do not require human judgment and follow standardized procedures, automating them can save time, reduce the risk of manual errors, and ensure consistency across the review process.

While several tools exist to support researchers in conducting systematic reviews, they typically focus on the more time-consuming phases of the process. EPPI-Reviewer [11] and DestillerSR [12], two tools primarily designed to support users during the literature screening phase, also include integrated search functionalities. These search capabilities are limited to PubMed, allowing users to directly retrieve studies from this database, while studies from other sources require manual upload. Other widely used screening tools, such as Rayyan [13] and Covidence [14], rely exclusively on manual uploads. Each of these four tools provides a deduplication feature for the uploaded bibliographic data.

In the broader context of the research project within which this study was conducted, a prototype for a new systematic review tool called NeutrinoReview is developed. Its conceptual architecture and vision are described in [15], and the 5-tier algorithm [16] as well as the Cal-X algorithm [17] have been integrated as LLM-based screening mechanisms. Without a data retrieval and deduplication module, searches must be performed via the web interfaces of the selected libraries, deduplication must be carried out separately, and the merged, deduplicated records must then be uploaded to NeutrinoReview.

It is hypothesized that minimizing tool fragmentation through an integrated solution streamlines the workflow and reduces the time and effort spent adapting to different environments. However, to the best of our knowledge, the extent of time savings provided by an automated search feature has not yet been investigated.

METHODOLOGY

This paper introduces a open-source data retrieval and deduplication module for review management systems, engineered to reduce the procedure of multi-database retrieval and deduplication into a streamlined operation. This functionality is integrated into NeutrinoReview[2] using a client-server model, where the server is a REST API implemented in FastAPI[3] and the client is a web application built with React[4]. The overall system architecture is depicted in Fig. 1.

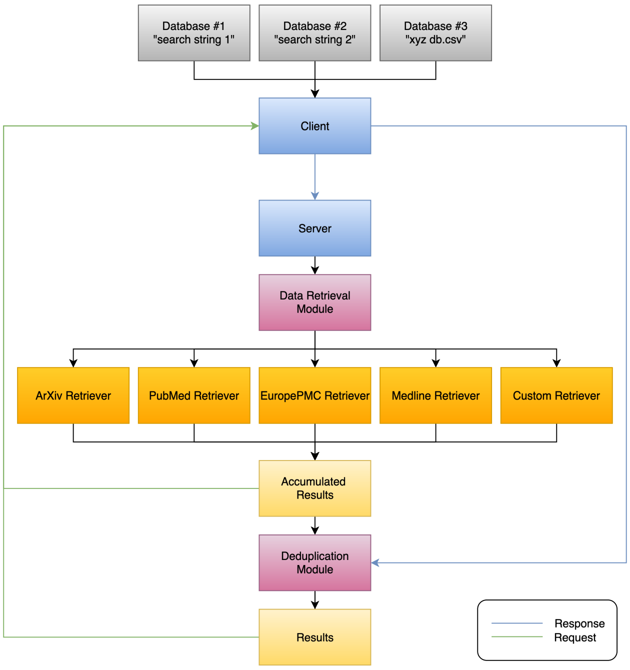


Figure 1: Data Retrieval and Deduplication Design

Data Retrieval

The data retrieval module streamlines an early step in systematic reviews: collecting literature from multiple databases. The module implements an automated pipeline that queries multiple databases in parallel, consolidates the retrieved records, and prepares the dataset for deduplication. The current implementation supports four databases selected for their public, keyless APIs: PubMed, Europe PMC, Medline, and ArXiv.

The module supports databases central to different scientific domains: PubMed, Europe PMC, and Medline for medicine, and ArXiv for disciplines across natural science and computer science. To ensure flexibility, sources that lack a supported API, the system provides a custom import function. This feature allows users to upload a CSV file of citations, which is then processed using the same pipeline. The retrieval logic for the four native databases and the CSV importer is implemented in five distinct classes, each inheriting from a common `BaseRetriever` class.

The data retrieval process is initiated with a user-provided search string and a selection of target databases. For each selected database, the `search` method of the corresponding retriever class executes the query. Subsequently, the raw

[2] `https://gitlab.cern.ch/caimira/caimira-wp4/neutrinoreview`

[3] `https://fastapi.tiangolo.com/`

[4] `https://react.dev/`

https://doi.org/10.5281/zenodo.17233862

data is parsed and normalized into a standardized format that abstracts away the heterogeneity of the various sources. By leveraging asynchronous programming, the system fetches and normalizes records concurrently, yielding a consolidated dataset ready for deduplication.

Deduplication

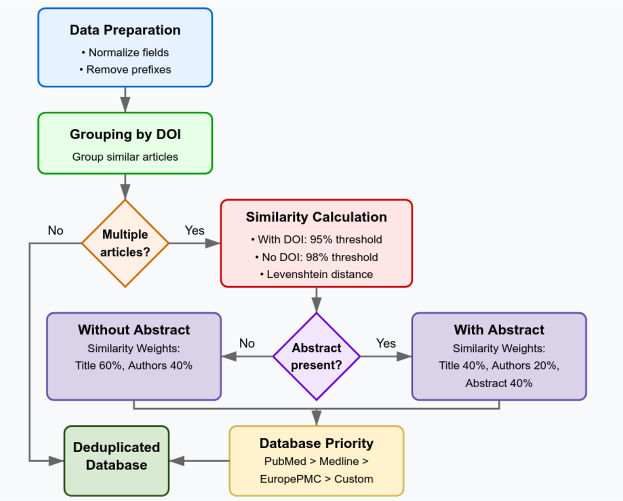


Figure 2: Deduplication Algorithm Design

This module implements a deduplication algorithm to resolve duplicate entries inherent in data aggregated from heterogeneous sources. The implemented process is based on a robust approach inspired by the Deduklick algorithm [18] and illustrated in Fig. 2.

The process begins with metadata normalization, followed by a primary grouping heuristic based on DOI. If a group contains a single article, it is marked as unique. Records in multi-member DOI groups, and all records without a DOI, undergo a pairwise similarity analysis. The algorithm applies a 95% similarity threshold for the former case and a 98% threshold for the latter.

The core of the algorithm is the similarity calculation, which computes a composite score from weighted metadata fields. For articles with an abstract, the weighting is as follows: Title (40%), Authors (20%), and Abstract (40%). For those without an abstract, the weighting is: Title (60%) and Authors (40%). As the specific Deduklick weights are not public, these values were determined empirically through rigorous testing. The similarity for each field is calculated using the Levenshtein distance, normalized by the maximum string length [19].

Finally, a database priority system resolves duplicate sets. When a duplicate is identified, the algorithm retains the record according to a predefined database ranking: PubMed is preferred over Medline, which has precedence over Europe PMC, followed by arXiv, and finally custom databases. This hierarchy was established based on consultation with domain experts. After the detection process is complete, the database is updated with the deduplicated records.

Evaluation

To evaluate NeutrinoReview's efficiency and usability, a user study was conducted comparing it against a widely used conventional tool called Rayyan[5]. The study involved seven participants, all of whom were novices with no prior experience using either system. Each participant carried out the complete data retrieval and deduplication workflow for four distinct medical systematic reviews using both tools. To ensure a standardized comparison, participants were provided with predefined search strings for PubMed, Medline, and EuropePMC, taken from original published reviews to closely mimic real-world scenarios. The primary performance metric was task completion time, measured from the creation of a new review project to the completion of deduplication in each system.

Each participant followed a standardized protocol to ensure procedural consistency. The search strings together with the reference to the corresponding published systematic reviews, as wel as the experiment protocol are provided in the supplementary material[6].

Initially, participants were given a written instruction detailing the steps to follow and a demonstration of the systematic review workflow in both NeutrinoReview and Rayyan to ensure a consistent baseline of understanding for all users. During the subsequent task-performance phase, a facilitator was present to observe and, upon request, provide clarification or assistance to prevent impasses. This 'assisted completion' protocol was designed to ensure that the recorded task times primarily reflect the tool's efficiency.

For a valid comparison, the deduplication process in Rayyan was configured to emulate NeutrinoReview's automated algorithm. Using Rayyan's 'AutoResolver' feature, a multi-step logic was established that mirrored the process in NeutrinoReview. The database priority (PubMed > Medline > EuropePMC) was replicated, and the conflict resolution rules were set to check for duplicates sequentially by DOI, normalized title and author, and a 95% similarity threshold. This methodological alignment ensured an equitable basis for benchmarking the deduplication performance.

RESULTS

Fig. 3 presents the time required by participants to perform data retrieval and deduplication using NeutrinoReview and Rayyan across four test reviews. The boxplots demonstrate that all participants completed the tasks more quickly with NeutrinoReview than with Rayyan. Notably, even the slowest participant using NeutrinoReview outperformed the fastest participant using Rayyan in 3 out of 4 test reviews. On average, task completion with NeutrinoReview required about 1.5 minutes, whereas the same tasks took more than 6 minutes with Rayyan. This corresponds to a 4-fold increase in speed and an overall time reduction of 75%. These findings indicate that the automated workflow implemented

[5] `https://www.rayyan.ai/`

[6] `https://zenodo.org/records/17075731`

https://doi.org/10.5281/zenodo.17233862

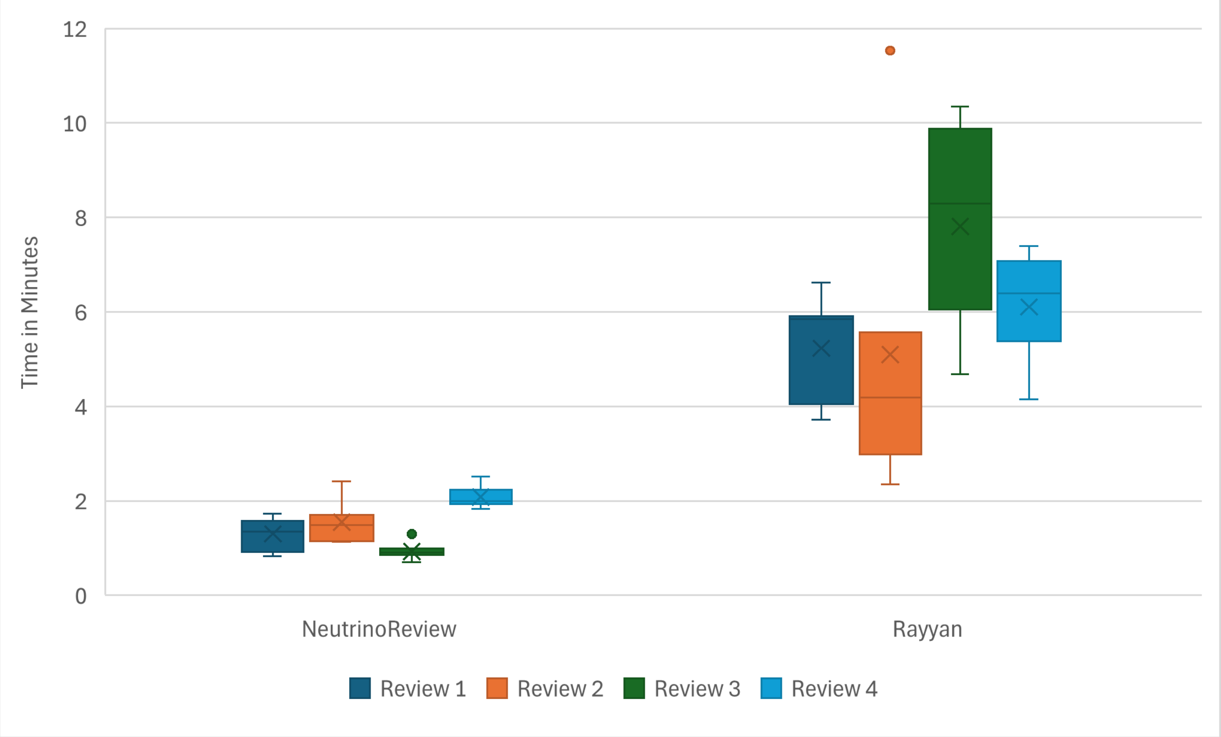


Figure 3: Comparison between NeutrinoReview and Rayyan

in the presented module substantially reduces the effort required for data retrieval and deduplication in systematic reviews. Integrating such a module into review management tools can streamline the process by minimizing the number of tools and user interfaces involved and by accelerating the initial stages of study selection.

It is important to emphasize that these results should not be interpreted as evidence that NeutrinoReview is the superior tool overall. NeutrinoReview is a proof of concept, while Rayyan is a well-established and robust review management platform. Rayyan provides considerably greater flexibility in deduplication settings and may have deliberately refrained from incorporating automated data retrieval due to robustness concerns associated with external dependencies or other considerations. The findings presented here should therefore be understood as a comparison between two fundamentally different approaches: one that emphasizes automation through integrated data retrieval and one-click deduplication, and another that prioritizes user control through customizable deduplication following manual data import. Which approach is more appropriate in practice will depend on factors beyond time efficiency alone, including robustness, flexibility, and the specific requirements of each review.

LIMITATIONS AND FUTURE WORK

This study was designed to evaluate the extent of workload reduction achieved through the implementation of automated data retrieval and deduplication modules. As such, other important aspects of system performance were not examined. In particular, the robustness of retrieval and deduplication processes—such as error rates, handling of incomplete or inconsistent metadata, and resilience to changes in source interfaces—remains unassessed. In addition, although the developed module supports automated data retrieval from arXiv, this functionality was not included in the experimental evaluation. The omission was due to technical constraints, specifically the absence of a bulk-export feature in the arXiv web search, which limited the feasibility of a systematic performance comparison. While this study demonstrated that integrating a data retrieval and deduplication module can further streamline the review process, future work should expand the range of supported data sources and systematically evaluate robustness metrics to provide a more comprehensive assessment of automated retrieval and deduplication workflows.

CONCLUSION

This study introduced and evaluated a proof-of-concept data retrieval and deduplication module designed to streamline the initial phases of systematic reviews. Integrated into the NeutrinoReview platform, the module automates multi-database retrieval, normalizes metadata, and applies a deduplication algorithm. In a controlled user study, the module reduced the time required for data retrieval and deduplication by more 75% compared with an established review tool. These findings demonstrate the potential of automation to reduce human workload and accelerate the review process.

The results underscore the importance of integrating robust retrieval and deduplication capabilities into review man-

https://doi.org/10.5281/zenodo.17233862

agement systems. By minimizing tool fragmentation and providing one-click functionality for otherwise repetitive tasks, such modules can help optimize the efficiency of systematic reviews and support timely evidence synthesis. While NeutrinoReview serves as a proof of concept rather than a production-ready system, the demonstrated workload reduction provides a strong argument for incorporating similar functionalities into existing or next-generation review management tools. Future research should extend the evaluation to cover robustness, scalability, and integration with a broader range of bibliographic databases. Ultimately, advancing such automation has the potential to not only accelerate systematic reviews but also to enhance their consistency, reliability, and impact on evidence-based research.

ACKNOWLEDGEMENTS

The joint CERN and WHO ARIA[7] project is funding the PhD project, in the context of which this paper was written. Furthermore, we greatly thank OpenWebSearch.EU[8] project and the members for their help and support with this publication.

REFERENCES

[1] Shekelle PG, Maglione MA, Luoto J, et al. *Global Health Evidence Evaluation Framework*. Rockville, MD: Agency for Healthcare Research and Quality (US); 2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK121300/table/appb.t21/. Table B.9, NHMRC Evidence Hierarchy: designations of 'levels of evidence' according to type of research question (including explanatory notes).

[2] Cook DJ, Greengold NL, Ellrodt AG, Weingarten SR. The relation between systematic reviews and practice guidelines. *Annals of Internal Medicine*, 1997;127(3):210–216.

[3] Clarke M, Hopewell S, Chalmers I. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. *The Lancet*, 2010;376(9734):20–21.

[4] Robinson KA, Brunnhuber K, Ciliska D, Juhl CB, Christensen R, Lund H. Evidence-based research series–paper 1: what evidence-based research is and why it is important? *Journal of Clinical Epidemiology*, 2021;129:151–157.

[5] Lund H, Juhl CB, Nørgaard B, Draborg E, Henriksen M, Andreasen J, et al. Evidence-based research series–paper 2: using an evidence-based research approach before a new study is conducted to ensure value. *Journal of Clinical Epidemiology*, 2021;129:158–166.

[6] Beller EM, Chen JK, Wang UL, Glasziou PP. Are systematic reviews up-to-date at the time of publication? *Systematic Reviews*, 2013;2:1–6.

[7] Demetres MR, Wright DN, Hickner A, Jedlicka C, Delgado D. A decade of systematic reviews: an assessment of Weill Cornell Medicine's systematic review service. *Journal of the Medical Library Association (JMLA)*, 2023;111(3):728.

[8] Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open*, 2017;7(2):e012545.

[9] S. Van der Mierden, K. Tsaioun, A. Bleich, C. H. C. Leenaars *et al.*, "Software tools for literature screening in systematic reviews in biomedical research," *ALTEX*, vol. 36, no. 3, pp. 508–517, 2019.

[10] B. Nussbaumer-Streit, M. Ellen, I. Klerings, R. Sfetcu, N. Riva, M. Mahmić-Kaknjo, G. Poulentzas, P. Martinez, E. Baladia, L. E. Ziganshina, et al., "Resource use during systematic review production varies widely: a scoping review", *Journal of Clinical Epidemiology*, vol. 139, pp. 287–296, 2021, Elsevier.

[11] J. Thomas, S. Graziosi, J. Brunton, Z. Ghouze, P. O'Driscoll, M. Bond, A. Koryakina, "EPPI-Reviewer: advanced software for systematic reviews, maps and evidence synthesis", *EPPI Centre, UCL Social Research Institute, University College London*, 2023, Accessed in July 2025. Available: https://eppi.ioe.ac.uk/cms/Default.aspx?tabid=2914

[12] DistillerSR Inc., "DistillerSR. Version 2.35", 2023, Accessed in July 2025. Available: https://www.distillersr.com/

[13] M. Ouzzani, H. Hammady, Z. Fedorowicz, A. Elmagarmid, "Rayyan—a web and mobile app for systematic reviews", *Systematic Reviews*, vol. 5, no. 1, p. 210, 2016, Springer.

[14] Veritas Health Innovation, "Covidence systematic review software", Melbourne, Australia, 2025, Accessed July 2025. Available: https://www.covidence.org/

[15] E. Sandner, I. Jakovljevic, A. Simiceanu, L. Fontana, A. Henriques, A. Wagner, C. Gütl, "NeutrinoReview: CONCEPT PROPOSAL FOR AN OPEN SOURCE REVIEW MANAGEMENT TOOL", *6th International Open Search Symposium #ossym2024*, p. 53, 2024.

[16] E. Sandner, B. Hu, A. Simiceanu, L. Fontana, I. Jakovljevic, A. Henriques, A. Wagner, C. Gütl, "Screening automation for systematic reviews: a 5-tier prompting approach meeting Cochrane's sensitivity requirement", *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pp. 150–159, 2024, IEEE.

[17] E. Sandner, M. Negovetić, K. Kothari, I. Taj, L. Fontana, A. Henriques, I. Jakovljević, A. Simniceanu, A. Wagner, C. Gütl, "Cal-X: Enhancing Systematic Review Screening with LLMs and Next-Token Likelihood Calibration", *2025 3rd International Conference on Foundation and Large Language Models (FLLM)*, IEEE, 2025, in press.

[18] N. Borissov, Q. Haas, B. Minder, D. Kopp-Heim, M. von Gernler, H. Janka, D. Teodoro, P. Amini, "Reducing systematic review burden using Deduklick: a novel, automated, reliable, and explainable deduplication algorithm to foster medical research", *Systematic Reviews*, vol. 11, no. 1, p. 172, 2022, Springer.

[19] Y. Yujian and L. Bo, "A normalized Levenshtein distance metric," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.

[7] https://partnersplatform.who.int/tools/aria

[8] https://openwebsearch.eu

https://doi.org/10.5281/zenodo.17220241

AUTOMATING LICENSE-AWARE FULL-TEXT RETRIEVAL FOR SYSTEMATIC REVIEWS: AN END-TO-END SCALABLE SYSTEM TO REDUCE REVIEWER WORKLOAD

D. Zhuk[*,†,1], E. Sandner[†,2,4], I. Jakovljevic[2], A. Simniceanu[3], L. Fontana[3], A. Henriques[2], A. Wagner[2], C. Gütl[4]
[1]University of Vienna, 1010 Vienna, Austria
[2]CERN, 1211 Geneva, Switzerland
[3]WHO, 1211 Geneva, Switzerland
[4]Graz University of Technology, 8010 Graz, Austria

Abstract

Systematic reviews are widely regarded as the most rigorous method for synthesizing scientific evidence, yet they remain highly labour-intensive. Full-text retrieval is a monotonous, repetitive, and time-consuming task that requires reviewers to locate and validate large numbers of articles. Existing tools only partially address this step, with limited support for automated, open-source, and legally compliant retrieval across heterogeneous repositories. To address this gap, a license-aware, open-source system was developed to automate full-text retrieval, extraction, and validation as part of the NeutrinoReview project. The system integrates open APIs (Unpaywall, PubMed, EuropePMC, Crossref) with a prioritized lookup strategy, browser-based PDF downloading, text extraction, and metadata-based validation. Performance was evaluated across 500 articles from five major scholarly repositories (PubMed, PMC, EuropePMC, IEEE Xplore, ACM Digital Library). Results show consistently high combined extraction rates (CER ≥ 0.800) and average processing times of 7–9 seconds per article. In realistic review scenarios, the system achieves a PDF retrieval rate of 82.68% and reduces manual retrieval workload by approximately 80%, corresponding to time savings of more than 3 hours in median sized SRs. These findings demonstrate the feasibility of automating a critical step in SR workflows, improving reproducibility and scalability while freeing researchers to focus on evidence synthesis.

INTRODUCTION

To gain a comprehensive understanding of a subject area, fragmented knowledge must be organized into structured information. Systematic review (SR) is a synthesis of identified and critically assessed evidence for topic understanding. This process is considered more rigorous and robust than a literature review as it follows a strict methodological framework, typically accompanied by predefined inclusion criteria [1]. Generally, SR consists of several phases that may vary depending on the methodology applied:

- Project Initiation and Data Retrieval – defining the research question, setting inclusion and exclusion criteria, and retrieving bibliographic metadata from relevant scholarly repositories.
- Screening – applying eligibility criteria to titles, abstracts, and subsequently full-text articles to ensure only relevant studies are included.
- Data Extraction – gathering relevant methodological details, outcomes, and contextual information from the included articles for further analysis and synthesis.

Full-text retrieval is the key part of the screening stage of SR, where reviewers must examine the complete text of articles to determine their eligibility [2]. Without access to the full-text data, important methodological details, outcomes, or context may remain hidden, leading to biased evidence synthesis. In particular, this paper addresses the following research question: *How can an open-source, legally compliant solution be developed to automate full-text retrieval, extraction, and validation across multiple scholarly repositories, reducing reviewer workload and improving reproducibility in SRs?*

BACKGROUND AND RELATED WORK

It is worth noting that overall performing a high-quality SR requires a lot of manual work and remains time-consuming, especially when following specific guidelines to be built upon (e.g., Preferred Reporting Items for SRs and Meta-Analyses) [3]. Conducting SR may take from 6 to 18 months [4]. In particular, full-text retrieval often represents a critical bottleneck – while bibliographic records are typically retrievable through Application programming interfaces (API) or structured search interfaces, access to corresponding full-text articles can be fragmented, license-restricted, or entirely unavailable without manual intervention. For instance, one of the most common problems include but are not limited to:

- Publisher paywalls and subscription barriers as many articles remain inaccessible without institutional access or individual payments.
- Licensing and copyright restrictions as even when access is granted, text-mining or bulk retrieval may be legally constrained.
- Heterogeneous platforms and formats as full texts are dispersed across multiple sources with inconsistent metadata, formats, and access protocols.

[*] a12446422@unet.univie.ac.at
[†] Both authors contributed equally to this work

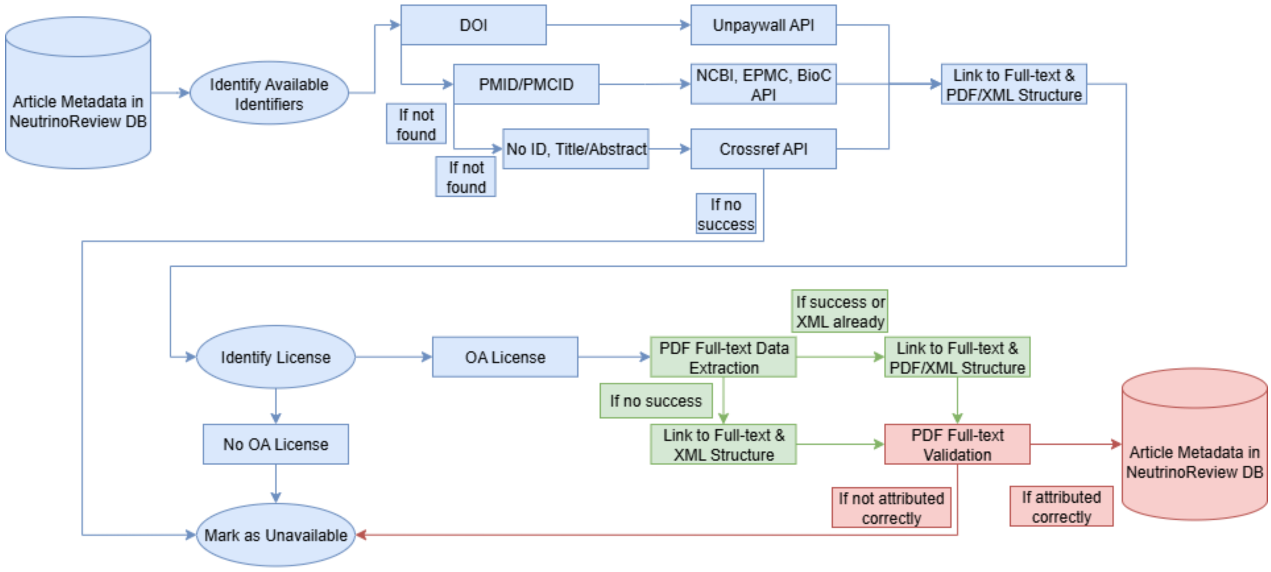

Article Metadata in NeutrinoReview DB
Identify Available Identifiers
DOI
PMID/PMCID
If not found
No ID, Title/Abstract
If not found
Unpaywall API
NCBI, EPMC, BioC API
Crossref API
If no success
Link to Full-text & PDF/XML Structure
Identify License
OA License
No OA License
Mark as Unavailable
If success or XML already
PDF Full-text Data Extraction
If no success
Link to Full-text & XML Structure
Link to Full-text & PDF/XML Structure
PDF Full-text Validation
If not attributed correctly
If attributed correctly
Article Metadata in NeutrinoReview DB

Figure 1: Conceptual architecture of full-text retrieval module for systematic review automation tools; Blue represents the Open-Source API Querying Component, Green represents PDF Full-text Data Extraction, and Red represents PDF Full-text Validation.

- Incomplete or unreliable linking as bibliographic metadata often lacks stable IDs or direct links to full-text sources, requiring manual searches.
- Limited API support as only some repositories (e.g., PubMed) provide open APIs, while others restrict programmatic access.

As a result, reviewers often spend substantial time locating, downloading, and verifying eligible resources and corresponding articles – an effort that detracts from the analytical phase of the review process. It is reported that resource-intensive task such as full-text retrieval can consume a minimum of 8,000 minutes of researcher time, depending on the screening approach used [5]. Despite recognition of the challenges described, existing tools (e.g., ASReview, Cadmus, Covidence) provide only partial solutions, with limited support for automated, open-source, and legally compliant full-text retrieval [6]. It is therefore worth emphasizing that the burden of effort in SR is still skewed toward labour-intensive steps, where retrieval, extraction, and validation of full-texts are performed manually, even when upstream bibliographic searches are automated. The imbalance overall not only slows down review production but also risks inconsistencies across respective research projects consuming valuable time resources [7]. Consequently, those persistent limitations highlight the need for new approaches that integrate metadata search with transparent full-text access. Such approaches will not only reduce manual workload but also promote consistency and reproducibility SR teams.

CONCEPTUAL DESIGN

The NeutrinoReview prototype[1] already supports automated bibliographic metadata retrieval from major sources such as PubMed, MEDLINE, and EuropePMC, as well as from user-supplied datasets. The metadata is stored in a structured database, providing a foundation for scalable and reproducible review workflows. However, the current implementation stops short of delivering full-text retrieval capabilities, leaving reviewers to perform this step manually.

Figure 1 illustrates the proposed system for full-text retrieval, featuring end-to-end, license-aware architecture designed for seamless integration into the NeutrinoReview project. It is organized into three key components, each addressing a critical step in the retrieval process, which are described subsequently.

Open-Source API Querying

The first component takes bibliographic inputs (DOI, PMID/PMCID, title, authors) and attempts to discover legally retrievable full-text artefacts (i.e., PDF URLs, XML structure) and accompanying license. It follows a prioritized lookup strategy designed to maximize accuracy and reproducibility:

- If DOI is present, the system queries Unpaywall [8] to obtain candidate open-access PDF URLs and license information. Unpaywall is preferred because it aggregates open-access locations and returns explicit license metadata when available.
- If only PMID/PMCID is supplied (or DOI lookup fails), the system queries PubMed/PubMed Central (PMC)/BioC endpoints to recover structured XML and any license statements embedded in repository metadata. When the result contains DOI, the DOI is rechecked against Unpaywall as a secondary source.
- As a final lookup, a metadata-to-DOI lookup against Crossref is attempted using title and author strings (also in case if DOI/PMID/PMCID lookups fail); any discovered DOI is then checked with Unpaywall.

All license strings returned by external services are normalized into a compact decision set used by downstream

[1] https://gitlab.cern.ch/caimira/caimira-wp4/neutrinoreview

https://doi.org/10.5281/zenodo.17220241

logic: permissive for text mining and storing (open) or not permissive (unavailable). The component logs each service query, timestamps, raw service responses, and the normalization rationale to create an auditable tracking record.

PDF Full-text Data Extraction

The second component is responsible for acquiring the canonical article file (i.e., PDF structure) from PDF URLs when available, converting that file into extractable text, and returning a normalized textual representation suitable for downstream parsing, validation, and screening. It is implemented by two cooperating routines: a browser-based downloader used as a fallback and primary PDF retrieval and text-extraction function. In operation, the extractor first prepares conservative browser-like HTTP headers and prefers structured or direct access. If that fails, it performs an HTTP GET and validates the response Content-Type before opening the bytes. When publishers serve PDF files dynamically or require JavaScript, the extractor falls back to a headless Chromium downloader that polls a temporary download directory for a completed .pdf file. Once valid PDF stream is obtained, the extractor iterates pages to collect page-level text and returns a single concatenated text (with page breaks preserved). Such failures as non-PDF responses, network timeouts, corrupted files, or images result in a None return and are recorded with standardized diagnostics.

PDF Full-text Validation

The third component verifies that the full-text extracted from a retrieved PDF corresponds to the expected bibliographic metadata and meets minimum quality criteria before the document is further processed. This validation is performed by two routines: a TF-IDF/cosine similarity scorer and a validator that applies heuristic thresholds.

The validator lowercases the extracted full-text data and uses the first 10,000 characters as the primary search window since titles, authors, and abstracts typically appear near the start. A conservative regex attempts to extract an “abstract” block from the text; pairwise similarities are then computed between the supplied title and the document start, the supplied abstract and the extracted abstract, and the supplied authors and the document start. Then, the validator returns a compact diagnostic object containing the three similarity scores and a boolean flag of validity; by default, a record is accepted if abstract similarity is greater than 0.20 or authors similarity is greater than 0.40 or title similarity is greater than 0.30. These thresholds are set up empirically by testing different ranges are tested for abstract, author, and title similarities, including 0.2 to 0.4, 0.3 to 0.5, 0.6 to 0.8, and 0.7 to 1.0. Given the low dimensionality of abstracts and article metadata, selected thresholds yielded the best results and are sufficient for robust validation. If no text is available or the checks fail decisively the function signals invalidity (i.e., False); all similarity scores are logged for tracking and threshold tuning.

Solution Outcome

The outcomes of the system are machine-readable tables that encode for each article: its licensing status, the canonical PDF link, PDF structure, XML structure, and PDF validation outcomes (if any). These outputs can be directly consumed by SR pipelines for automated or manual full-text screening, data extraction, or critical appraisal. Moreover, the solution is extensible, ensuring that new retrieval methods, content sources, or document validation formats can be incorporated without substantial redesign. Crucially, the approach adheres to applicable legal restrictions and does not depend on paid content providers.

EVALUATION METHODOLOGY

For evaluation, the proposed system was applied to multiple scholarly sources – PubMed, PMC, EuropePMC, IEEE Xplore, and ACM Digital Library – with a focus on medical literature. These tests are conducted in order to demonstrate the feasibility of significantly reducing reviewer workload while maintaining reproducibility and scalability. Each source is queried with domain-specific search strings designed to capture representative subsets of research articles relevant to airborne transmission, respiratory particle dynamics, or open-access retrieval architectures. From each source, the first 100 articles are returned by the queries selected, resulting in 500 articles in total. The query configurations are as follows:

- PubMed – sorted by Best Match, bibliographic metadata is PMID, in Summary (text) format.
- PMC – sorted by Default order, bibliographic metadata is PMCID, in PMCID list format.
- EuropePMC – sorted by Relevance, bibliographic metadata is PMCID, in ID list format.
- IEEEXplore – sorted by Relevance, bibliographic metadata is DOI, in Plain text format.
- ACM Digital Library – sorted by Recency, bibliographic metadata is DOI, in ACM Ref format.

Search strings for each of the sources can be found in Appendix 1.

The system performance is quantified using extraction and validation outcomes per 100-article from each source. Table 1 lists the considered metrics along with their definitions.

Table 1: Metrics Overview

Metrics Name and Description	Metrics Abbreviation and Equation
Open Articles – Count of articles determined to have an open license	$OA = N_{OA}$
PDF Retrieval Rate – Fraction of open articles with a canonical PDF link successfully retrieved	$PRR = \frac{N_{PDF}}{N_{OA}}$
PDF Extraction Rate – Fraction of open articles from which PDF structure is extracted	$PER = \frac{N_{PER}}{N_{OA}}$

XML Extraction Rate – Fraction of open articles with XML structure available	$XER = \frac{N_{XER}}{N_{OA}}$
Combined Extraction Rate – Fraction of open articles with either PDF or XML structure available	$CER = \frac{N_{PER} + N_{XER} - (N_{PER} \cap N_{XER})}{N_{OA}}$
Total Processing Time – Wall-clock time statistics, total runtime (in seconds)	$TPT = T_{proc}$

EVALUATION RESULTS

The system's ability to retrieve and extract full-text varied significantly across repositories, reflecting differences in openness, metadata availability, and source infrastructure, as detailed in Table 2.

Table 2: Benchmark Evaluation Results

Name	OA	PRR	PER	XER	CER	TPT
PubMed	90	0.589	0.478	1.000	1.000	707.971
PMC	100	0.620	0.410	1.000	1.000	926.470
EPMC	93	0.925	0.871	1.000	1.000	912.000
IEEE	10	1.000	0.800	0.000	0.800	246.39
ACM	57	1.000	0.860	0.000	0.860	841.220

In terms of availability, open-source articles coverage is the highest for PMC (100/100) and EuropePMC (93/100), reflecting their open mandates. The proposed solution also performes well on PubMed (90/100), while coverage on ACM Digital Library (57/100) and especially IEEEXplore (10/100) is much more restricted.

In terms of retrieval, the system achieves its strongest performance on EuropePMC, with PRR (0.925) and PER (0.871), complemented by perfect XML coverage. On ACM and IEEE perfect PRR (1.000) is reached, but the absence of XML fallback limits CER to 0.860 and 0.800, respectively. PubMed and PMC sources provide complete coverage through XML, though their respective PER scores are less reliable.

Processing times are generally consistent, averaging 7-9 seconds per article. IEEEXplore shows the fastest total runtime due to its small OA sample, whereas PMC required slightly longer because of additional fallback operations.

Overall, the system demonstrates strong performance across all sources, with CER never falling below 0.800, ensuring that full-text data is consistently available either in PDF or XML format.

DISCUSSION

The impact of automated full-text retrieval in systematic reviews becomes evident when considering its potential to reduce reviewer workload in real-world scenarios.

An analysis of 195 systematic reviews showed that between 0 and 4,385 studies (mean = 63) were included at the title and abstract screening stage and therefore had to be retrieved in full text [9]. When automation is not available, reviewers must perform this step manually, and retrieving a single full text is estimated to take an average of 4 minutes [5]. Consequently, manually retrieving full texts for a systematic review with the mean number of included studies (63) requires about 4 h 12 min, whereas the most exhaustive case (4,285 studies) would demand approximately 292 h 20 min.

By contrast, the proposed solution achieves an average PRR of 82.68%, implying that only 17.32% of articles require manual retrieval. The average processing time for 100 studies is 726.81 seconds, corresponding to 7.3 seconds per article.

Applying this solution to a systematic review requiring 63 full texts, about 52 can be retrieved automatically, while 11 must be retrieved manually. The system's processing time amounts to 6 min 20 s, with an additional 44 min of manual work, yielding a total retrieval time of 50 min 20 s. This corresponds to a workload reduction of 3 h 21 min 40 s for a mean-sized systematic review.

Based on the same assumptions, for a systematic review with 4,385 records, the system's processing time would be 7 h 21 min 10 s without any parallelization of the retrieval mechanism, whereas manual retrieval would require about 232 h 22 min 50 s. In this extremely large case, the system reduces the workload by 232 h 22 min 50 s.

Consequently, the system can reduce the time required for full-text retrieval by 80%.

LIMITATIONS

Each component of the solution proposed has practical constraints that may influence performance. Firstly, coverage depends heavily on source policies: repositories with restrictive access models (e.g., IEEEXplore, ACM Digital Library) yield fewer open articles, which reduces overall retrieval opportunities despite high PDF success rates when links are available. Secondly, PDF extraction remains fragile in cases of scanned documents, image-only pages, or publisher-specific encodings, where structured XML is not available as a fallback. Thirdly, metadata inconsistencies (e.g., variant author strings, missing abstracts) can lower validation scores and may exclude possible usable texts. Moreover, processing speed, while generally acceptable, is influenced by network conditions and the need for browser-based fallback routines, which may not scale well at very large volumes. Finally, the system is designed to operate within legal boundaries of open-access content – paywalled or license-restricted materials remain inaccessible by design, which can limit completeness for certain research domains.

https://doi.org/10.5281/zenodo.17220241

FUTURE WORK

Future development of the system will focus on three main directions. Improving robustness of PDF extraction by integrating Optical character recognition (OCR) pipelines for scanned or image-only documents, and experimenting with hybrid approaches that combine parsing with Machine Learning-based text recovery. Expanding source coverage by incorporating additional APIs and institutional repositories, thereby improving completeness in restricted domains. Refining validation by training domain-adaptive similarity models that go beyond heuristics, enabling more accurate alignment of metadata and full-text data. Additionally, efforts will be made to optimize processing speed and resource efficiency, ensuring the system remains scalable for large SR projects. Continuous feedback and more real-world testing will guide those iterative improvements.

CONCLUSIONS

Thus, in this paper, a license-aware, open-source solution for automated full-text retrieval, extraction, and validation across multiple major scholarly repositories is presented. Benchmarking against PubMed, PMC, EuropePMC, IEEEXplore, and ACM Digital Library demonstrates that the system consistently achieves high combined extraction rates (CER is greater than 0.800), ensuring reliable availability of either PDF or XML structures. The results confirm both the feasibility and scalability of automating a critical bottleneck in systematic reviews, reducing manual reviewer workload while maintaining reproducibility. At the same time, differences across repositories highlight the continued challenges of restricted access and heterogeneous infrastructures. By providing extensible components and transparent diagnostics, the system lays a foundation for future improvements, including expanded coverage, more robust extraction methods, and tighter integration with SR pipelines.

ACKNOWLEDGEMENTS

This research was conducted as part of the joint CERN–WHO ARIA[2] project, which funds Elias Sandner's PhD studies and within which this paper was prepared. We also gratefully acknowledge the OpenWebSearch.EU[3] project and its members for their valuable support with this publication.

REFERENCES

[1] R. Randles and A. Finnegan, "Guidelines for writing a systematic review", *Nurse Education Today*, vol. 125, p. 105803, June 2023. doi:10.1016/j.nedt.2023.105803

[2] L. Schmidt *et al.*, "Data extraction methods for systematic review (semi)automation: Update of a living systematic review", *F1000Research*, vol. 10, article 401 (version 3), Apr. 2025. doi:10.12688/f1000research.51117.3

[3] F.M. Delgado-Chaves *et al.*, "Transforming literature screening: The emerging role of large language models in systematic reviews", *Proc. Natl. Acad. Sci. U.S.A.*, vol. 122, e2411962122, Jan. 2025. doi:10.1073/pnas.2411962122

[4] V. Phillips and E. Barker, "Systematic reviews: Structure, form and content", *Journal of Perioperative Practice*, vol. 31, p. 349-353, Jan. 2025. doi:10.1177/1750458921994693

[5] I. Shemilt *et al.*, "Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews", *Syst. Rev.*, vol. 5, article 140, Aug. 2016. doi:10.1186/s13643-016-0315-4

[6] L. Affengruber *et al.*, "An exploration of available methods and tools to improve the efficiency of systematic review production: a scoping review", *BMC Med Res Methodol*, vol. 24, article 210, Sept. 2024. doi:10.1186/s12874-024-02320-4

[7] K.E.K. Chai *et al.*, "Research Screener: a machine learning tool to semi-automate abstract screening for systematic reviews", *Syst. Rev.*, vol. 10, article 93, Apr. 2021. doi:10.1186/s13643-021-01635-3

[8] Unpaywall API, https://unpaywall.org/products/api

[9] R. Borah *et al.*, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry", *BMJ Open 2017*, vol. 7, article e012545, Oct. 2016. doi:10.1136/bmjopen-2016-012545

[2] https://partnersplatform.who.int/tools/aria

[3] https://openwebsearch.eu

https://doi.org/10.5281/zenodo.17220241

APPENDIX

Source Name	Search String
PubMed	(airborne[tiab] OR aerosol*[tiab] OR "airborne transmission"[tiab] OR "air transmission"[tiab] OR inhalation[tiab]) AND (risk[tiab] OR "risk assessment"[tiab] OR exposure[tiab] OR hazard*[tiab]) AND (model[tiab] OR models[tiab] OR modelling[tiab] OR modeling[tiab] OR "mathematical model"[tiab] OR "computational model"[tiab] OR simulation[tiab] OR simulations[tiab])
PMC	(droplet* OR particle* OR aerosol*) AND (size OR diameter OR "particle size" OR "droplet size" OR volume* OR cm OR centimetre OR centimeter OR µm OR micron OR "micrometer" OR "micro-meter") AND ("expiratory activity" OR "expiratory activities" OR "respiratory activity" OR "respiratory activities" OR breath* OR speak* OR talk* OR shout* OR sing* OR cough* OR sneez*)
EuropePMC	((droplet* OR particle* OR aerosol*) AND (size OR diameter OR "particle size" OR "droplet size" OR volume* OR cm OR centimetre OR centimeter OR µm OR micron OR micrometer) AND ("expiratory activity" OR "expiratory activities" OR "respiratory activity" OR "respiratory activities" OR breath* OR speak* OR talk* OR shout* OR sing* OR cough* OR sneez*))
IEEEXplore	(open OR "open-source" OR "open access") AND (search* OR retrieval OR discovery OR "information retrieval") AND (architecture* OR framework* OR system* OR platform* OR infrastructure* OR toolkit*)
ACM Digital Library	(open OR "open-source" OR "open access") AND (search* OR retrieval OR discovery OR "information retrieval") AND (architecture* OR framework* OR system* OR platform* OR infrastructure* OR toolkit*)

Appendix 1: Search strings for benchmark evaluation

https://doi.org/10.5281/zenodo.17253014

KNOWLEDGE SOVEREIGNTY IN DISABILITY INFORMATION RETRIEVAL: ARCHITECTING PRIVACY-PRESERVING AND SUSTAINABLE INFRASTRUCTURE

Noor A. Fathima*[1], Noor K. Kubra[2], A. Wagner[1]

[1] CERN, Geneva, Switzerland
[2] University of Mysore, Mysore, India

Abstract

Knowledge ecosystems shaped by disabled people, caregivers, and other marginalized groups remain systematically underrepresented in mainstream search and AI systems [1–3]. These ecosystems contain embodied practices, ethnographies, and everyday adaptations that rarely surface online, leaving critical insights absent from datasets that inform decision-making in healthcare, employment, and social inclusion [4, 5]. Addressing this gap requires infrastructures that can surface and structure such knowledge in ways that are transparent, trustworthy, and open to community participation [6, 7].

The OpenWebSearch.eu (OWS) initiative is developing the Open Web Index (OWI) to provide reusable, open web data for research and innovation [8, 9]. Building on this foundation, we demonstrate how a vertical search engine can be prototyped by using the OWILIX tool [10] to extract targeted slices of the OWI and consuming them within CIFF-compatible search frameworks such as MOSAIC [9, 11] for interactive exploration. This workflow serves as a technical backbone for Nooon, a privacy-preserving search engine envisioned to surface disability-related knowledge through multimodal contributions [12, 13].

BACKGROUND

Underrepresented Knowledge Ecosystems

Mainstream search engines and AI systems prioritize large, standardized datasets [2]. As a result, knowledge ecosystems[1] shaped by marginalized groups, such as disabled people, their caregivers, and immediate families who are often the primary witnesses of daily life, remain largely invisible [1, 4]. These ecosystems are heterogeneous and context-dependent, encompassing both formal sources (e.g., policy documents, research reports) and informal contributions (e.g., lived experiences, workplace adaptations, family care practices) [5].

Within Nooon's design, these ecosystems are recognized as *microdata*: fine-grained, individual-level or community-level traces that reveal embodied practices, ethnographic insights, and everyday adaptations [3]. Making such microdata visible, trustworthy, and reusable is essential for addressing structural exclusion, where disability narratives are too often reduced to tokenized symbols (e.g., wheelchair icons) instead of nuanced, situated information [13].

Need for Open, Transparent Search

Commercial search engines provide little visibility into their data collection and ranking processes [14, 17]. This opacity limits the ability of researchers, advocacy groups, and smaller communities to shape how their knowledge is indexed and surfaced [15, 16]. Transparency and openness are therefore critical if search is to serve underrepresented groups fairly. Beyond technical access, open infrastructures also support *epistemic sovereignty*: the capacity of communities to contribute to and represent their knowledge on their own terms [18].

The OpenWebSearch Initiative

OpenWebSearch.eu (OWS) is a European research initiative building the Open Web Index (OWI), an open, reusable corpus of web data [8]. OWI provides a foundation for building alternative and domain-specific search engines without dependence on proprietary indices [9]. By making web data accessible to researchers, developers, and civil society, OWS aims to create a sustainable and pluralistic search ecosystem in Europe and beyond [6].

The OWILIX Tool

To make the OWI usable at scale, the project developed OWILIX, a command-line tool that allows targeted extraction of subsets (or "slices") of the index [10]. OWILIX enables developers to tailor datasets to specific domains by filtering based on URLs, keywords, or metadata [9]. This functionality is particularly valuable for building vertical search engines, where precision and domain relevance are more important than exhaustive coverage.

The MOSAIC Framework

MOSAIC is an open-source search and visualization framework that can integrate multiple indices, including slices created with OWILIX [11]. It provides a flexible frontend for exploring search results, experimenting with ranking strategies, and evaluating the user experience [9]. For vertical search prototypes, MOSAIC offers a ready-to-use environment to demonstrate the impact of targeted indexing.

* afshan.shokath@gmail.com

[1] In information science, a knowledge ecosystem refers to a dynamic and interconnected system of actors, practices, and artifacts that collectively produce, store, and circulate knowledge.

https://doi.org/10.5281/zenodo.17253014

The Nooon Project

Building on these infrastructures, Nooon is conceptualized as a privacy-preserving search engine designed to surface disability-related knowledge. It envisions combining slices of OWI with additional indices created through multimodal knowledge contributions (e.g., text, gestures, sensory adaptations) provided directly by disabled people and caregivers [12,13]. Nooon's design extends the open, transparent ethos of OWS by embedding privacy, trust, and sustainability into its core [6], ensuring that sensitive disability narratives can be surfaced without fear of exploitation or misrepresentation.

Scope of This Work

In this paper, we focus on constructing a first proof-of-concept slice of the OWI that we refer to as the `Nooon` index. The slice is obtained by selecting web documents whose HTML content contains the keyword 'disability', providing a domain-specific corpus for further exploration. While this approach is deliberately simple, it establishes a reproducible baseline for evaluating how knowledge related to disability, often hidden or tokenized in mainstream systems, can be isolated as a distinct corpus. By extracting and surfacing a slice of web data where the term "disability" appears in the underlying HTML, we begin to make visible a knowledge ecosystem that is otherwise fragmented or invisible. The subsequent sections describe how this slice, referred to as the `Nooon` index, is prepared, structured, and made accessible as a first step toward building search infrastructures that respect and sustain marginalized knowledge.

SYSTEM DESIGN & ARCHITECTURE

Our design follows a modular workflow: first, identifying and extracting relevant subsets of the OWI using OWILIX; second, preparing the resulting data in formats compatible with CIFF-based search libraries; and finally, deploying the data through MOSAIC for interactive exploration. Each stage of the workflow emphasizes transparency, reproducibility, and sustainability, ensuring that the resulting slice not only serves as a technical proof of concept but also as a first step toward surfacing disability-related knowledge ecosystems in open search infrastructures. The following subsections describe this workflow in brief, from dataset access to frontend deployment.

Dataset Access, Querying and Slice Selection with OWILIX

To construct domain-specific corpora for our prototype, we accessed the Open Web Index (OWI) through the OWILIX command-line interface. OWILIX enables remote inspection, synchronization, and querying of OWI datasets in a manner similar to Git, providing versioned, auditable workflows.

Our environment was deployed on a dedicated bare-metal server (open-science-search.ch), configured with Python 3.11 using pyenv. We created an isolated virtual environment (owi) to install the required dependencies, including py4lexis and owilix, directly from the project's PyPI mirror. This setup ensured reproducibility and avoided dependency conflicts. The full configuration is provided in Appendix A (Puppet manifests).

We began by listing available datasets to identify suitable partitions. Using the command:

```
owilix remote ls it4i:latest
files=**/language=eng/*
```

we retrieved the most recent English-language snapshot from the IT4I data center. This confirmed the availability of a daily dataset containing over 5.6 million documents (24.3 GiB across 1,037 files).

After verifying scope, we synchronized the relevant subset locally with:

```
owilix remote pull it4i:latest
"files=*/language=eng/*"
```

This operation pulled 184 Parquet files containing both metadata and plain text. OWILIX's incremental synchronization ensured that only new or changed files were downloaded, reducing both storage requirements and network load. By selecting only English-language partitions rather than mirroring full daily datasets (600 GB/day), we further minimized environmental and computational costs. This selective approach aligns with the project's sustainability goals: smaller slices not only reduce energy intensity but also lower barriers for replication by other researchers and advocacy groups.

Once the dataset was locally available, we generated a focused slice for disability-related content using a SQL-like filter:

```
owilix query slice --local all:latest \
  "where=main_content like ’%disability%’" \
  collection_name="disability" \
  creator="NoorAF"
```

This created a named slice (disability) annotated with provenance metadata, including a creator tag and timestamp. Preserving these slice specifiers (it4i:latest files=…, where=…) makes the workflow reproducible and auditable, ensuring that future researchers can regenerate the exact same dataset.

Together, the listing, pulling, and querying steps formed a reproducible workflow for constructing thematic corpora tailored to underrepresented knowledge ecosystems. In this study, the resulting slice served as the foundation for Nooon, which further aims to have indices with multimodal lived-experience contributions as microdata.

Consuming Datasets with CIFF-compatible search frameworks

Once slices of the Open Web Index (OWI) were retrieved locally, the next step was to make them consumable within

https://doi.org/10.5281/zenodo.17253014

CIFF-compatible search frameworks, enabling interactive exploration. For this purpose, we first experimented with MOSAIC, the reference open-source framework developed in the OpenWebSearch.eu project. MOSAIC integrates Lucene-based search indices with Parquet-formatted metadata, providing a unified environment for retrieval and visualization. This made it well suited for our initial experiments with disability-focused slices.

To prepare the OWILIX-generated slice, we exported the data into a CIFF index and corresponding Parquet files. MOSAIC requires these to be arranged in a specific directory hierarchy, with Lucene indices and metadata separated into dedicated folders. In our case, the OWILIX export produced a compressed Lucene index (index.ciff.gz) alongside Parquet metadata. These were organized under a serve/ directory, mirroring the structure expected by MOSAIC.

Preparing Data for MOSAIC

After constructing the disability-focused slice, we exported it from OWILIX into a format consumable by external search frameworks. OWILIX provides a local export function that outputs both a CIFF file (compressed Lucene index) and associated Parquet metadata. This export was performed as follows:

```
cd ~/tmp/
mkdir data
owilix local export all/id=id outdir=$(PWD)
/data
```

The resulting directory contained an index.ciff.gz file together with metadata files prefixed metadata_. These artifacts formed the raw input for MOSAIC.

To align with MOSAIC's expected structure, we converted the CIFF file into a Lucene index using the official converter image:

```
mkdir -p data/serve/lucene
podman run \
  --rm \
  -v "$PWD/data":/data:Z \
  opencode.it4i.eu:5050/
  openwebsearcheu-public/mosaic/lucene-ciff \
  /data/index.ciff.gz \
  /data/serve/lucene/disability-index
```

The metadata was then organized into a corresponding folder:

```
mkdir -p data/serve/metadata/
disability-index
mv data/*metadata_* data/serve/metadata/
disability-index
```

This resulted in a directory hierarchy under /tmp/data/serve/ containing both a Lucene index (disability-index) and the associated Parquet metadata. At the file level, the Lucene directory contained the expected segment files (e.g., _0.fdt, _0.fdx, _0_Lucene90_0.doc, segments_1), confirming that the CIFF-to-Lucene conversion had succeeded.

At this stage, the slice was fully prepared for consumption by MOSAIC.

```
lucene/
  disability-index/     # Lucene index files
  (_0.fdt, _0.fdx, segments_1, etc.)
metadata/
  disability-index/     # Parquet metadata
  files
```

Here our domain-specific slice, labeled disability-index is served. This ensured compatibility with MOSAIC's configuration while preserving semantic clarity about the dataset's thematic scope.

Deployment of MOSAIC

With the Lucene and Parquet structures in place, we deployed MOSAIC as a containerized service. Rather than building from source, we relied on pre-built images from the OpenWebSearch.eu GitLab registry, executed via Podman for compatibility with our infrastructure. The backend service was launched as follows:

```
podman run -d --name mosaic --network host \
  -v "$PWD/data":/data:Z \
  localhost/mosaic:latest \
  --lucene-dir-path $PWD/data/serve/lucene/ \
  --parquet-dir-path $PWD/data/serve/metadata/
```

This deployment was hosted on the bare-metal server, integrated into CERN's internal infrastructure. For security, the host is not directly exposed to the internet; instead, inbound traffic is routed via a load balancer running Nginx as a reverse proxy. This design ensured reliable user access while isolating the backend from direct external exposure.

The backend search service (`mosaic`) consumed the Lucene and Parquet directories prepared in the earlier step and exposed them via a JSON API. Verification at the `/mosaic/index-info` endpoint confirmed that the `nooon` index (our disability-focused slice) was successfully loaded, containing over 2.59M English-language documents. Additional indices such as `simplewiki` and `unis-graz` were also present, demonstrating that OWILIX-exported slices can be integrated as first-class indices within MOSAIC.

Deploying the MOSAIC Frontend

Following backend validation, we deployed the frontend container (`mosaic-fe`), which serves a lightweight web interface built on nginx. The frontend connects directly to the backend API, enabling interactive queries and faceted exploration of multiple indices. Running backend and frontend as separate services provided flexibility: indices could be updated or swapped at the backend without interrupting the user interface, while alternative frontends could be deployed

https://doi.org/10.5281/zenodo.17253014

if needed. This modularity highlights MOSAIC's suitability for multi-tenant deployments, where diverse domain-specific corpora can be hosted and accessed through a common interface.

Figure 1 shows the frontend deployment with the `disability-index` slice active. The interface exposes filters for language, query limits, and geographic boundaries, enabling targeted exploration of the corpus.

We then tested the interface with a query for the term *"disability"*. As shown in Figure 2, results were retrieved from the `disability-index`, including documents such as legal advice directories, government manuals, and parliamentary committee reports. Each result is enriched with metadata such as language, word count, index date, and links to the original source, allowing tailored exploration of the corpora.

This demonstrates full end-to-end functionality: user input via the frontend, retrieval from the Lucene/Parquet backend, and structured result presentation to the user.

CONCLUSION

In this paper, we demonstrated how slices of the Open Web Index (OWI) can be extracted, prepared, and consumed within the MOSAIC framework. Using the OWILIX tool, we generated a disability-focused slice, exported it into Lucene and Parquet formats, and validated it through both backend API inspection and frontend deployment. Our deployment on bare-metal infrastructure behind a secure reverse proxy confirmed that OWILIX-exported slices can be hosted, queried, and visualized interactively using MOSAIC.

Beyond the technical workflow, this case study illustrates a reproducible and sustainable model for constructing *vertical search engines*. By embedding provenance metadata, minimizing unnecessary data transfers, and reusing open-source frameworks, we show how targeted corpora can be surfaced in ways that are transparent and resource-efficient. Furthermore, the ability to host multiple corpora concurrently demonstrates MOSAIC's scalability for multi-domain or multi-tenant search applications.

Our prototype index, `Nooon`, exemplifies how these infrastructures can serve underrepresented knowledge ecosystems. Disability-related corpora, often absent or misrepresented in mainstream search and AI systems, can be surfaced as first-class indices within an open, auditable framework. This approach aligns with broader goals of privacy, trust, and inclusion, ensuring that minority knowledge is not erased but made visible in ethically responsible ways.

FUTURE WORK

Several extensions are planned. First, we will explore topic-level filtering using `curlielabels` derived from Curlie.org, enabling semantic slice construction by domain hierarchy (e.g., *Society* → *Disability*, *Health* → *Conditions*) rather than keyword alone. This would yield more robust corpora aligned with specific communities of knowledge.

Second, while MOSAIC provides a flexible reference framework, we anticipate that many organizations will want to augment their *internal search systems* with OWI data. Since internal search is often based on ElasticSearch or OpenSearch, we will extend our pipeline to support direct export from OWILIX to JSON and ingestion into OpenSearch. This requires either developing a converter or adapting OWILIX's export functions to natively push slices into OpenSearch indices. Demonstrating this workflow would broaden adoption, as it allows OWI slices to be embedded directly into enterprise search infrastructures.

Third, improvements to the MOSAIC frontend are envisaged. Current functionality supports keyword queries, language filters, and metadata inspection; future iterations will explore category-level browsing, faceted search by Curlie labels, and cross-corpus comparison (e.g., *Disability in Employment* vs. *Disability in Education*). Enhancing the frontend will make vertical search applications more intuitive for non-technical users, which is critical for adoption by advocacy groups and organizations.

Finally, we plan to extend beyond English-only text to multilingual and multimodal corpora, integrating images, video, and lived-experience narratives contributed directly by users. Combined with client-side preprocessing for privacy preservation, this will bring us closer to the vision of Nooon as a privacy-first, inclusive search engine for minority knowledge ecosystems.

ACKNOWLEDGEMENTS

This work was supported through guidance and supervision from the Open Search Foundation, in particular its working group on Ethics [19]. Additional feedback and community input were provided by the Disability Network, an informal group within the Diversity and Inclusion programme at CERN, where the first author has served as a contact point [20].

REFERENCES

[1] World Health Organization. *World Report on Disability*. WHO Press, 2011. `https://www.who.int/publications/i/item/9789241564182`.

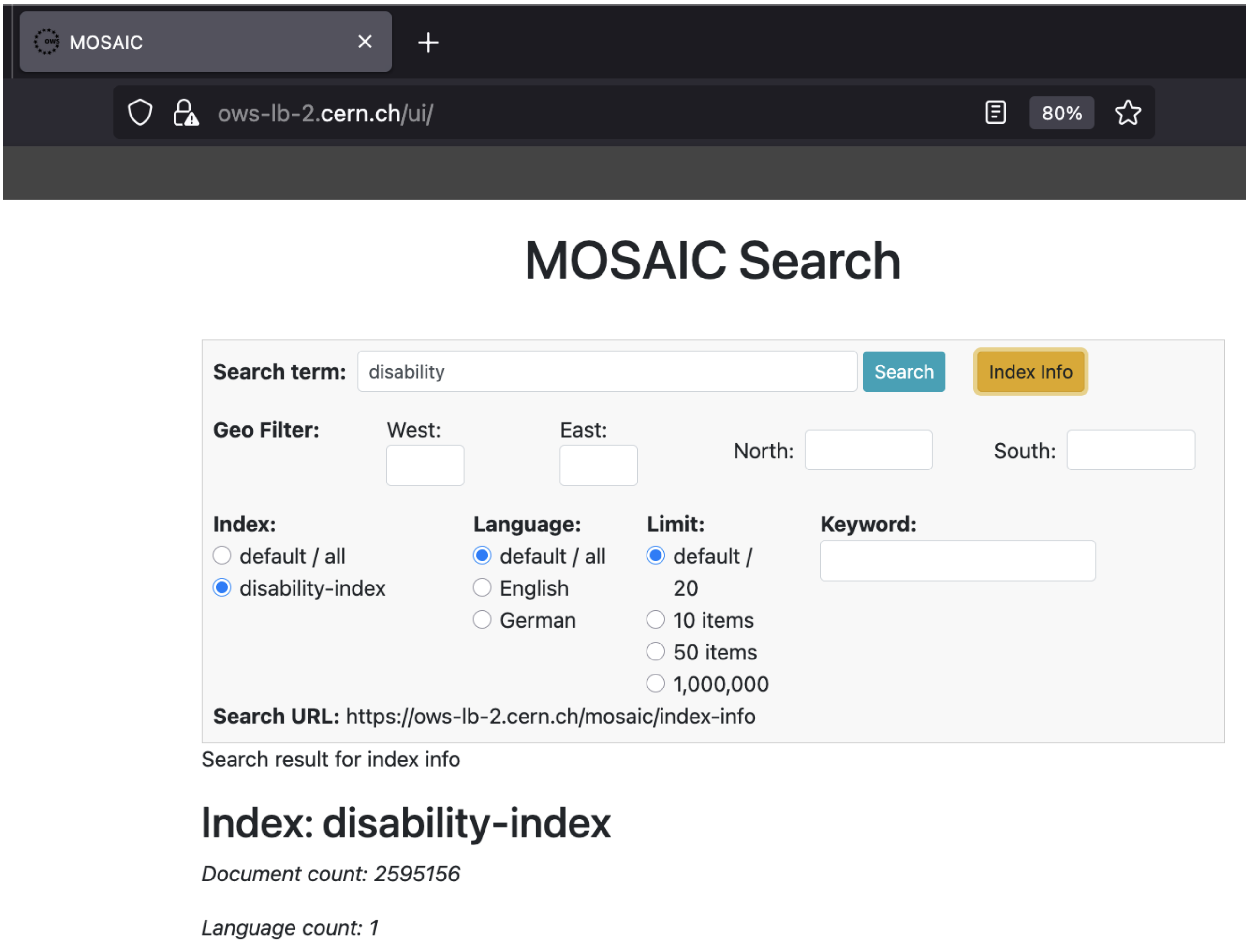
[2] Jacob Metcalf, Emily F. Keller, and danah boyd. Perspectives on Big Data, Ethics, and Society. *Council for Big Data, Ethics, and Society*, 2016. `https://bdes.datasociety.net/council-output/perspectives-on-big-data-ethics-and-society/`.

[3] Chris Williams. *Disabled Data*. Routledge, 2021. ISBN: 978-0-367-12345-6 (Placeholder, no confirmed ISBN found).

[4] International Labour Organization.
Labour market outcomes of persons with disabilities.
ILO, 2023.
`https://ilostat.ilo.org/`
`new-ilo-database-highlights-labour-market-challenges-of-p`

[5] Gary L. Albrecht, Katherine D. Seelman, and Michael Bury (eds.). *Handbook of Disability Studies*. Sage, 2001. ISBN: 9780761916529. DOI: 10.4135/9781412976251. `https://uk.sagepub.com/en-gb/eur/book/handbook-disability-studies#description`.

https://doi.org/10.5281/zenodo.17253014



MOSAIC Search

Search term: disability [Search] [Index Info]

Geo Filter: West: East: North: South:

Index:
○ default / all
● disability-index

Language:
● default / all
○ English
○ German

Limit:
● default / 20
○ 10 items
○ 50 items
○ 1,000,000

Keyword:

Search URL: https://ows-lb-2.cern.ch/mosaic/index-info

Search result for index info

Index: disability-index

Document count: 2595156

Language count: 1

Figure 1: MOSAIC frontend deployed via the `mosaic-fe` container on bare-metal infrastructure. The screenshot shows the `disability-index` slice, containing 2.59M English-language documents.

[6] OpenWebSearch.eu Consortium. D4.2: Report of privacy, transparency, and trust models for search applications. OpenWebSearch.eu, 2023. DOI: 10.5281/zenodo.placeholder (Placeholder, no confirmed DOI found). `https://openwebsearch.eu/deliverables/` (General project deliverables page, specific report not confirmed).

[7] Paula Helm and Selin Gerlek. Empirical AI Ethics: Reconfiguring Ethics Towards a Situated Plural and Transformative Approach. Preprint, submitted to *Cambridge Forum on AI: Culture and Society*, 2025.

[8] OpenWebSearch.eu Consortium. Annex 1 – Description of the Action. European Commission Horizon Europe Grant Agreement, 2022. `https://cordis.europa.eu/project/id/101070014` (General project page, specific annex not publicly available).

[9] OpenWebSearch.eu Consortium. *The OpenWebSearch Book*. 2024. `https://openwebsearcheu-public.pages.it4i.eu/ows-the-book`.

[10] OpenWebSearch.eu Consortium. OWILIX Command Line Interface. 2024. `https://opencode.it4i.eu/openwebsearcheu-public/owi-cli/`.

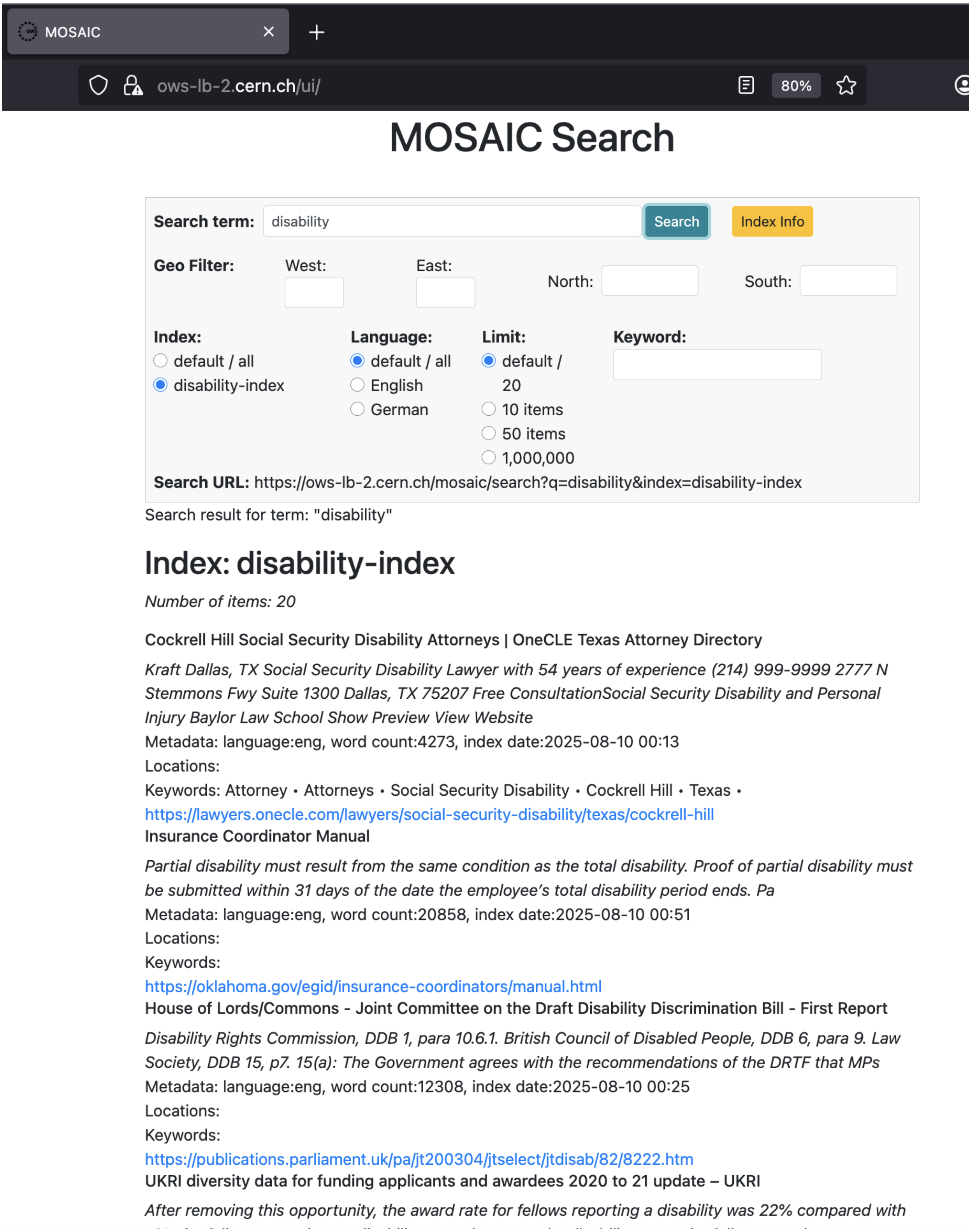
[11] OpenWebSearch.eu Consortium. MOSAIC Search Framework. 2024. `https://openwebsearcheu-public.pages.it4i.eu/ows-the-book/content/howto/c_mosaic.html`.

[12] Noor A. Fathima, A. Wagner, and The OSF Working Group Ethics. Knowledge Sovereignty: Empirical Ethics in Privacy-Preserving, Sustainable Disability Information Retrieval Infrastructure. Preprint, submitted to *Cambridge Forum on AI: Culture and Society*, 2025. OpenWebSearch.eu / CERN, Geneva, Switzerland, 2025. `https://cds.cern.ch/record/2932730/files/CERN-OPEN-2025-004.pdf`.

[13] B. Blaser and R. E. Ladner. Why is Data on Disability so Hard to Collect and Understand? In: *2020 Research on Equity and Sustained Participation in Engineering, Computing, and Technology (RESPECT)*, Portland, OR, USA, pp. 1-8, 2020. DOI: 10.1109/RESPECT49803.2020.9272466. `https://ieeexplore.ieee.org/document/9272466`.

[14] Frank Pasquale. *The Black Box Society: The Secret Algorithms That Control Money and Information*. Harvard

https://doi.org/10.5281/zenodo.17253014



MOSAIC Search

Search term: disability [Search] [Index Info]

Geo Filter: West: East: North: South:

Index:
○ default / all
● disability-index

Language:
● default / all
○ English
○ German

Limit:
● default / 20
○ 10 items
○ 50 items
○ 1,000,000

Keyword:

Search URL: https://ows-lb-2.cern.ch/mosaic/search?q=disability&index=disability-index

Search result for term: "disability"

Index: disability-index

Number of items: 20

Cockrell Hill Social Security Disability Attorneys | OneCLE Texas Attorney Directory

Kraft Dallas, TX Social Security Disability Lawyer with 54 years of experience (214) 999-9999 2777 N Stemmons Fwy Suite 1300 Dallas, TX 75207 Free ConsultationSocial Security Disability and Personal Injury Baylor Law School Show Preview View Website

Metadata: language:eng, word count:4273, index date:2025-08-10 00:13
Locations:
Keywords: Attorney • Attorneys • Social Security Disability • Cockrell Hill • Texas •
https://lawyers.onecle.com/lawyers/social-security-disability/texas/cockrell-hill

Insurance Coordinator Manual

Partial disability must result from the same condition as the total disability. Proof of partial disability must be submitted within 31 days of the date the employee's total disability period ends. Pa

Metadata: language:eng, word count:20858, index date:2025-08-10 00:51
Locations:
Keywords:
https://oklahoma.gov/egid/insurance-coordinators/manual.html

House of Lords/Commons - Joint Committee on the Draft Disability Discrimination Bill - First Report

Disability Rights Commission, DDB 1, para 10.6.1. British Council of Disabled People, DDB 6, para 9. Law Society, DDB 15, p7. 15(a): The Government agrees with the recommendations of the DRTF that MPs

Metadata: language:eng, word count:12308, index date:2025-08-10 00:25
Locations:
Keywords:
https://publications.parliament.uk/pa/jt200304/jtselect/jtdisab/82/8222.htm

UKRI diversity data for funding applicants and awardees 2020 to 21 update – UKRI

After removing this opportunity, the award rate for fellows reporting a disability was 22% compared with

Figure 2: MOSAIC frontend in use. A query for the term "disability" against the `disability-index` returns diverse results, including attorney directories, insurance manuals, and parliamentary reports. Metadata (language, word count, index date) and original source links are also displayed, demonstrating end-to-end functionality.

https://doi.org/10.5281/zenodo.17253014

University Press, 2015. ISBN: 9780674970847. https://www.hup.harvard.edu/books/9780674970847.

[15] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018. ISBN: 9781479837243. https://nyupress.org/9781479837243/algorithms-of-oppression/.

[16] Cathy O'Neil. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown Publishing Group, 2016. ISBN: 9780553418835. https://www.penguinrandomhouse.com/books/531209/weapons-of-math-destruction-by-cathy-oneil/.

[17] Luke Stark and Kate Crawford. The Work of Art in the Age of Artificial Intelligence: What Artists Can Teach Us About the Ethics of Data Practice. *Surveillance & Society*, 2019. DOI: 10.22215/surjsoc.2019.10821. https://ojs.library.queensu.ca/index.php/surveillance-and-society/article/view/10821.

[18] C. Estelle Smith, Avleen Kaur, Katie Z. Gach, Loren Terveen, Mary Jo Kreitzer, and Susan O'Conner-Von. What is Spiritual Support and How Might It Impact the Design of Online Communities? *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 2021. DOI: 10.1145/3449117. https://doi.org/10.1145/3449117.

[19] Open Search Foundation. Ethics in Search Working Group. Available at: https://ethicsinsearch.org/en/.

[20] CERN Diversity and Inclusion. Disability Network. Available at: https://diversity-and-inclusion.web.cern.ch/networks/disability-network.

https://doi.org/10.5281/zenodo.17229191

THE PROJECT KISU: AI-BASED SEARCHING AND FINDING – EASY, INCLUSIVE, AND SELF-DETERMINED

K. Altmeyer, M. Hladky, S. Malone, M. Platz, K. Reese, L. Schick, V. Wolf, Saarland University, Saarbrücken, Germany
T. Gottsmann, M. Wiesner, fragFINN e.V., Berlin, Germany
C. Plote, Open Search Foundation, Starnberg, Germany

Abstract

Many users, especially children, older people, or people with special needs, find it difficult to formulate precise search queries that lead to relevant results. This can lead to frustration, limited use of the internet, and ultimately, a feeling of digital exclusion. A lack of understanding of the search logic and an inappropriate interpretation of the search results harbor the risk of users consuming unreliable information or even being exposed to disinformation.

In the KIsu project, we use an LLM that analyses and interprets the search terms entered. This AI model is trained to recognize and specify the user's actual search intention from incomplete or imprecise queries. We are also developing workshops and information materials to promote the intelligent use of search engines and AI.

KIsu is a cooperative project between Saarland University and fragFINN e.V. (as well as the German Research Center for Artificial Intelligence and Open Search Foundation). In this paper, the current status of the project is presented.

INTRODUCTION

Google dominates the search engine market, followed by Bing, which also provides the search results of most so-called alternative search engines and is gaining popularity thanks to the AI chatbot ChatGPT. These dominant providers do not disclose their algorithms and search indices, leading to restrictions in transparency, accessibility, data protection, and security, among other things. The consequences can be bias, information asymmetries, and data-driven discrimination (Bobic, Platz & Gütl, 2021 [1]; Galindo & Garcia-Marco, 2017 [2]). Although search engines play a significant role in our everyday lives, most people use them without knowing or questioning how they work. Effects such as a lack of transparency and immaturity are increasing massively due to generative AI language models. In addition, commercial providers have little interest in researching and offering low-barrier access.

To enable all people – regardless of their age or gender, with or without disabilities – to use internet searches and search engines in an informed, self-confident, and self-determined way and thus facilitate their participation in the digitally pervaded world, we are developing learning and information materials and organizing workshops in the KIsu project. We want to support citizens in overcoming the black box effect of searches and analyzing and criticizing search results. Additionally, we aimed to specifically identify the challenges encountered by particular target groups during internet searches and to develop methods for supporting these groups both pedagogically and technologically.

The learning materials on search engine literacy (e.g., Platz et al., 2023 [3]) and the Open Search Foundation (https://opensearchfoundation.org/en/children-and-internet-search/) served as a starting point. They are co-creatively adapted with various groups from civil society and further developed concerning the use of AI frontends and chatbots. We focus on children, young people and older people.

All materials developed in the project are published as Open Educational Resources (OER). Design principles are derived and implemented for designing AI frontends for search engines that are orientated towards the common good based on the previously developed learning materials and explorative user workshops.

This paper describes the objectives and steps of the KIsu project. Furthermore, the research design and first results of the project are presented.

OBJECTIVES

With KIsu, we want to strengthen the digital sovereignty and participation of people with special needs, children, and older people. We aim to enable as many user groups as possible to use search engines efficiently and inclusively.

With the help of AI-supported language models, accessible search interfaces, and training, we create the conditions for using the internet as a source of information safely, confidently, and competently. In addition, our accompanying studies provide new insights into how generative AI models can be used to reduce existing barriers to internet use. In summary, KIsu pursues the following goals:

- Promoting the competent and responsible use of Internet searches through developing and testing learning and information materials.
- The identification of factors and specific needs in dealing with search engines and the derivation of implications for the further development of learning materials and for designing low-barrier AI search engine frontends.
- Fine-tuning a needs-based AI language model and developing a model-agnostic data pipeline.
- Provision of AI frontends for optimized search (technology maturity level 5).
- Provision of the data collected in the project (the intervention and effectiveness study).

https://doi.org/10.5281/zenodo.17229191

PROJECT STEPS

The project duration is 19 months (01.06.2024-31.12.2025). The project consists of four steps:

- *Step 1 – Material and front-end development:* learning, training, and information materials are designed co-creatively with various groups from civil society to promote the competent use of internet searches. The following group compositions and sizes were envisaged: 15 children, including children with various special educational needs (e.g., emotional-social development, mental development, hearing, physical and motor development, learning, vision, and language), 15 young people, including young people with various special educational needs, 15 older people, including people with various special educational needs. Care is taken to ensure a balanced gender composition. For the development of the first intuitive frontends for the search engines, a modular design is being sought that enables simple rule-based interactions and can be expanded later. Survey tools are selected and developed to identify factors and specific needs in dealing with search engines. A kick-off workshop with all project participants took place. *Milestone: first version of materials and front end.*
- *Step 2 – Material and front-end optimization:* materials and frontends are tested with a controlled intervention study (n=45) and optimized based on the results. *Milestone: Optimized version of materials and the front end.*
- *Step 3 – Dissemination of the materials developed in the project and front-end development:* A training concept and OER will be prepared in cooperation with various groups from civil society and published design principles for the design of low-barrier AI search engine front-ends are derived and implemented in a final application (training of a needs-based AI language model). The design principles for developing AI frontends to remove barriers to use and facilitate access to civic data will be made available. In addition, a conference will be held with all project participants. A multiplier network will be established. *Milestone: training concept, OER, optimized version of the frontend, design principles, frontends.*
- *Step 4 – Provision of data and publication of research results:* The data collected will be processed and passed on to a research data repository. The research work and study results will be published for open access by the relevant specialist audience. *Milestone: Open data, journal publications.*

RESEARCH DESIGN AND RESULTS

An Action Design Science Research (ADSR) approach is pursued (Mullarkey & Hevner, 2019 [4]). Design Science Research (DSR) is a paradigm rooted in the philosophy of pragmatism. DSR involves problem-solving research to answer research questions related to human problems and produces valuable artefacts. ADSR centers on co-creative collaboration between scientists and users. The goals of the first phase (*diagnostic phase*) are to analyze the problem space and the solution space (here: the identification of factors and specific needs in dealing with search engines and the derivation of implications for the development of learning materials and for the design of low-barrier AI search engine front-ends) for research and practice and their relevance in mutual agreement between the researcher-user team. A mixed methods approach (e.g., Kuckartz, 2014 [5]) is pursued in which quantitative data collection methods such as questionnaires are combined with qualitative data collection methods such as interviews and observations in co-creative workshops. The sample comprises children, young people, and older people (see subsections below). The quantitative data is analyzed descriptively and inferentially, the qualitative data is analyzed using Design thinking methods, such as developing Personas (Uebernickel et al., 2015 [6]) combined with qualitative content analysis (Mayring, 2015 [7]), and the results are correlated.

Then follows the *design phase*, in which the artefact is identified and conceptualized (here: learning materials and low-barrier AI search engine frontends). Design principles are (further) developed through several iterative cycles within the design phase. Collaborative activities with co-creative activities are essential here, as the researcher-user team aims to create artefacts that incorporate innovative ideas for solving the given problems. In the implementation phase, concepts are developed to use the artefact. An actual application offers the opportunity to evaluate the efficiency and effectiveness of the proposed design in practice.

We are currently in the design phase (*Step 2* of the project). We have already outlined detailed learning materials and selected specific tools to test the effectiveness of these. We will soon be testing these on an initial sample. At the same time, we are working on implementing the AI-supported search frontends and optimizing their functionalities based on user feedback. A central problem is that there has been very little sound research into the heterogeneous target groups and their behavior when searching the internet. We must, therefore, first create a solid empirical basis for further development steps. The project elements, co-creative workshops, a controlled intervention study, and the AI frontend are described below.

Co-creative Workshops

In line with Ind & Coates (2013 [8]), end-users are involved, which leads to more relevant and usable products and services while reducing risk. Participatory design is used to develop iterative prototypes to test user reactions. The workshops were conceptualized using Design thinking (e.g., Uebernickel et al., 2015 [6]). In the workshop the participants design their own digital assistant, that can help them to find what they search on the internet. In order not to tempt the workshop participants to reproduce existing solutions, but to become creative themselves, the word 'digital assistant' was used instead of 'search engine'.

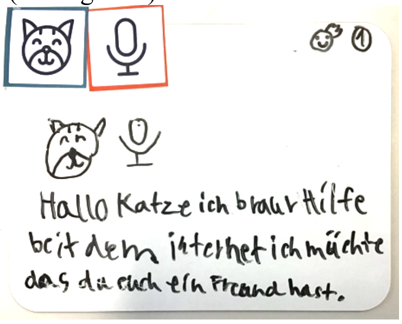
The following key- questions guided the workshops:

- What are the features of your digital assistant?

https://doi.org/10.5281/zenodo.17229191

  - What does your digital assistant look like?
  - What should your digital assistant be able to do?
- Input method:
  - How do you want to tell your digital assistant what you are searching for?
  - What elements do you need on the screen to start your search?
- Output method:
  - How should your digital assistant present the search results?
  - How should it tell you what it has found?
  - How should the search results be presented?
- What happens if you have (not) found what you were looking for?
  - How do you tell your digital assistant?
  - How does the digital assistant react to this?
  - What should the digital assistant do?

The participants designed the interface of their assistant in small groups of 3-4 individuals, using small whiteboards and whiteboard markers as well as icons that could be stuck to the board (see Figure 1).



Hallo Katze ich braur Hilfe beit dem itterhet ich müchte das du auch ein Freund hast.

Figure 1: Design of the digital assistant and input method by a child

Four co-creative workshops were organized with the following user groups:

- Children attending the 2nd grade in primary school (ca. 7 years old)
  - 8 girls, 5 of whom speak German as a second language
  - 7 boys, 3 with German as a second language and 1 with German as a foreign language
- Mathematical interested children attending the 3rd or 4th grade in primary school (between 9 and10 years old)
  - 5 girls
  - 9 boys
- Young people attending the 7th grade in grammar school (ca. 13 years old)
  - 9 female
  - 10 male
- Older people (between 64 and 87 years old)
  - 10 female
  - 4 male

During the workshops, qualitative interviews were performed with the participants. The workshops and interviews were videographed, and key scenes were transcribed. For analysis, personas are derived. Personas are descriptive models of users. They are archetypes with a set of properties of different but – concerning defined aspects – comparable persons (Uebernickel et al., 2015 [6]).

Initial results indicate that children prefer having social and friendly interactions with their digital assistant, while trustworthiness is particularly important for older users. Speech input and output seem to be suitable across all user groups. Adolescents expressed a clear preference for tailored assistance, meaning the digital assistant should precisely match the complexity of the provided information to their specific needs and sensitively adjust its conversational tone – such as adopting a humorous style when searching for entertainment content.

Controlled Intervention Study

In the co-creative workshops, we observed that the group comprising slightly older children and teenagers demonstrated the highest level of prior knowledge regarding internet search strategies, use of information technology, and AI. Consequently, this group also provided the most substantial input for the co-creation of informational materials and the design of a suitable frontend for internet searches.

Based on these insights, we will conduct an initial intervention study specifically targeting this age group, using tailored learning materials and a customized frontend designed for internet search activities. The study will address three research questions:

1. Does the use of specifically developed learning materials and the customized frontend lead to measurable improvements in aspects of search engine literacy and proficiency in AI-supported internet searches?
2. How do children evaluate the usability and usefulness of the developed frontend?
3. How do the designed informational materials impact children's interaction with the frontend during internet searches?

The study will involve a minimum of 45 participants from grades 5 to 8 attending a Montessori school. The participants will work individually but will be organized into small groups for practical purposes. The study begins with all children completing pre-tests assessing their knowledge and attitudes toward internet searching and AI. Subsequently, children will be randomly assigned to one of two intervention groups:

- Intervention Group 1 (IG 1) will first engage with the learning materials covering general information as well as input and output processes of AI-supported internet searches. They will then complete a structured internet search task using the newly developed frontend. Following this, participants will evaluate the frontend's usability

https://doi.org/10.5281/zenodo.17229191

and usefulness. Their interactions with the frontend will be recorded. Finally, the initial tests on search engines and AI will be repeated to measure learning outcomes and attitude changes.

- Intervention Group 2 (IG 2) differs only in the sequence of tasks: children in this group will first work with and evaluate the frontend, followed by the study of informational materials.

Overall, we expect both groups to benefit from our intervention by gaining essential knowledge related to critical aspects of AI-supported internet searches and, if present, correcting uncritical attitudes towards AI. Comparing the interactions between IG 1 and IG 2 will reveal the extent to which participants benefit from the learning materials during actual internet searches. It is hypothesized that participants in IG 1 will apply more of the principles covered in the materials compared to IG 2, leading them to perceive the developed frontend as more useful and usable.

AI frontend

A central aspect of our project is the development of innovative, low-barrier user frontends. The user interface should be able to recognize and understand the user's questions, especially those of people with special needs. By integrating an LLM, the user query is analyzed, optimized, and converted into a suitable search query. This improved query is then forwarded to a search engine, e.g., FragFINN.de. There are also plans to include other (alternative) search engines (such as Ecosia). Despite the inherent opacity of large language models (LLMs), we would, like to use them specifically to dialogue with the user to better understand their actual informational needs and generate a more precise and relevant search request as part of our project. This process should enhance the understanding of users' actual informational needs of search results and accessibility and user-friendliness for all user groups. The ability to fine-tune LLMs for inclusive language or to adapt them to the unique search queries of children, older adults, or people with disabilities is central to us. Open-source models like Llama 3 or Mistral enable flexible adaptation through methods such as LoRA or QLoRA, while proprietary models like GPT-4-Turbo allow fine-tuning via API. Open-source models also have the advantage that they are more cost-effective, there is no direct dependency on individual companies, the dynamics of the models can be fully controlled, and data protection mechanisms are easier to implement and review. Through workshops and educational programs, we provide information about the data protection practices of search engines and the generative AI systems we use, which promote a more conscious and secure handling of personal data.

The explainability and interpretability of the responses generated by LLMs are currently essential research topics. Although the presently available LLMs are not yet fully interpretable, we intend to actively follow the latest advances in this area and, where possible, integrate them into our project. This includes the evaluation and potential implementation of methods to increase the transparency and traceability of AI-supported processes.

CONCLUSION

The project KIsu aims to make digital information services accessible and understandable for everyone and thus strengthen digital participation in the long term. The project unfolds its impact through

- the use of the training modules in training courses and train-the-trainer courses,
- utilizing and making available the design principles and study results for the development of AI front-ends to reduce barriers to use and facilitate access to civic data,
- building a multiplier network through training and OER,
- raising awareness through public relations work.

In the long term, we hope that AI will recognize users' search intentions and proactively support them in specifying their queries, critically evaluating information, and making informed decisions. We also hope that the training concepts and technologies developed in our we recognize their potential to improve search queries significantly project will be integrated and used by educational institutions, social institutions, and advice centers to promote digital skills throughout society in the long term.

ACKNOWLEDGMENTS

We thank all the participants in the co-creative workshops and surveys, our student assistants and the German Federal Ministry for Family Affairs, Senior Citizens, Women, and Youth for funding the project, funding code 3924406K05.

REFERENCES

[1] Bobic, A., Platz, M. & Gütl, C. (2021). Towards Open Search Applications for the broader Community. In *Proceedings of the 3rd International Symposium on Open Search Technology,* 11-13 October 2021, CERN, Geneva, Switzerland.

[2] Galindo, F. & Garcia-Marco, J. (2017). Freedom and the internet: empowering citizens and addressing the transparency gap in search engines. European journal of law and technology, 8(2),1–18.

[3] Platz, M., Decker, A. & Plote, C.(2023). Förderung eines verständigen Umgangs mit Suchmaschinen im Elementarbereich. M. R. Textor & A. Bostelmann (Hrsg.), Das Kita-Handbuch.

[4] Mullarkey, M. T., & Hevner, A. R. (2019). An elaborated action design research process model. European Journal of Information Systems, 28(1), 6–20

[5] Kuckartz, U. (2014). Mixed methods: methodologie, Forschungsdesigns und Analyseverfahren. Springer-Verlag.

[6] Mayring, P. (2015). Qualitative Inhaltsanalyse: Grundlagen und Techniken. Beltz Pädagogik

[7] Ind, N., & Coates, N. (2013). The meanings of co-creation. European business review, 25(1), 86-95.

[8] Uebernickel, F., Brenner, W., Pukall, B., Naef, T., & Schindlholzer, B. (2015). Design Thinking: Das Handbuch. Frankfurter Allgemeine Buch.

https://doi.org/10.5281/zenodo.17251788

A CHARTER FOR PUBLIC INTEREST INTERNET SEARCH

C. Plote, Open Search Foundation e. V., Starnberg, Germany,
A. Nussbaumer, Technical University Graz, Graz, Austria

Abstract

Digital web search systems are key infrastructures of the information society. They influence access to knowledge, public discourse and political participation. Yet their development is mostly driven by commercial interests and opaque decision-making. To address growing concerns about transparency, fairness and democratic control, the Open Search Foundation initiated a collaborative process to explore the ethical dimensions of web search. This article presents one result of this process – a Charter of ten guiding principles that aim to articulate normative reference points for a Public-Interest Web Search Infrastructure.

INTRODUCTION: WHY WEB SEARCH MATTERS FOR THE COMMON GOOD

Search engines and AI-based search applications are much more than technical platforms. As a critical infrastructure, they fundamentally shape what information becomes visible on the Internet, which perspectives enter public discourse, and how people make decisions. They influence access to education, democratic participation and the formation of public opinion.

At present, this central function is concentrated in the hands of a few global platform companies. Their interests are often opposed to the public good, democratic values and the roots of the Internet as an open platform. These systems, based on proprietary indices and opaque algorithms, prioritise commercial goals over the public interest. At the same time, the same few companies systematically collect sensitive user and behavioural data at scale. This concentration of power undermines transparency, pluralism and democratic control – and endangers fundamental rights such as privacy and freedom of information.

The common good is being neglected – an unacceptable situation from an ethical and societal point of view. Already back in 2000, Lucas D. Introna and Helen Nissenbaum [5] argued that “web-search mechanisms are too important to be shaped by the marketplace alone” and that search engines “must work in the greater public interest.”

With increasing geopolitical tensions, the risks posed by this imbalance are growing. Political manipulation, disinformation, and government surveillance are becoming more likely, while smaller or alternative systems struggle to gain visibility. Legal and regulatory efforts have not yet succeeded in changing this structural imbalance.

Therefore, it seems necessary to establish a common set of rules that includes guidelines, recommendations and measures for an web search that is oriented towards the common good, and as such not only serves society and puts people first, but also returns more control to society.

Background and Development

The Open Search Foundation (OSF), an independent non-profit organisation based in Germany, has initiated a multi-stakeholder process to develop an ethical framework for web search aligned with the common good. Supported by the Stiftung Mercator as part of the #EthicsInSearch project, this collaborative effort brought together experts from academia, civil society, and technology – most notably through the OSF’s Ethics Working Group and the project OpenWebSearch.eu that is building the prototype for an Open Web Index (OWI).

Starting with the question of what constitutes a public good web search, a series of workshops in different constellations explored its ethical and societal foundations. While various ethical frameworks for digital technologies already exist, a deliberately exploratory 'blank page' approach was adopted – seeking to identify relevant values, risks and mitigation strategies specific to the context of web search, rather than adapting existing models.

The process involved several steps: identifying and classifying values central to web search, mapping ethical risks, and developing concrete strategies to mitigate ethical and societal risks. The overall aim was to formulate guiding principles that define the normative, technical and institutional conditions for a search landscape that upholds individual rights, enables participation and strengthens both personal and collective autonomy.

Based on these findings, the OSF drafted a Charter for Web Search for the Common Good, the #FreeWebSearch Charter, which outlines ten guiding principles. In the following, the authors – both members of the OSF Ethics Working Group – briefly present these principles and their associated building blocks, and explain their specific relevance to web search in the public interest.

Defining ‘web search’ and ‘web search systems’

In this article we focus on web search related to the retrieval of different types of information (e.g. politics, news, history, science, art, etc.) from the Internet. As ‘web search’ no longer refers exclusively to classical search engines – it also includes AI-supported tools such as chatbots and generative models, platform-internal search functions, and hybrid systems that retrieve, rank and generate answers – this article uses the term ‘web search systems’ to encompass all those technologies that influence the visibility, accessibility and prioritisation of information in the digital sphere, whereas ‘web search’ means the whole societal and technological system in general.

https://doi.org/10.5281/zenodo.17251788

THE TEN PRINCIPLES

Guarantee transparency and explainability

Transparency is essential for accountability, trust and participation. Without insight into how web search systems work, what data sources they use and what factors influence their results, it is impossible to identify bias, discrimination or undue influence. Transparency enables oversight, prevents abuse, and empowers users to make informed decisions and critically engage with digital processes. It is a prerequisite for digital sovereignty.

To support this principle in practice, four concrete dimensions of transparency have been identified:

Algorithmic and functional transparency: Research shows that the ranking of search results 'has a dramatic impact on consumer attitudes, preferences and behaviour', and in particular primacy effects influence the formation of attitudes and beliefs and can even have a significant impact on voting behaviour [1]. To counteract these effects, users need to understand how results are generated and what factors influence ranking and visibility. Key criteria such as ranking logic, moderation rules and personalisation mechanisms must be disclosed. Easy-to-understand explanations (e.g. 'Why am I seeing this?') and interfaces for user feedback and independent research are essential. Trade secrets should not be used to avoid scrutiny.

Transparency of sources and index data: Web search systems display only selected content. The criteria and processes used to include or exclude information must be transparent. Users and auditors need to understand which sources are included, how they are weighted, and where results and AI-generated answers come from.

Transparency of business models and influence: Vendors must disclose who controls the system, how it is funded, and how these structures may influence results. Advertising, political funding or other (financial) dependencies must be made visible. Paid content must be clearly identified.

Foster competition and technological independence

Without viable alternatives, dominant providers are likely to further cement their control over access to information – shaping public discourse, limiting choice and undermining transparency. 'As any ranking per se prefers some items over others, the real problem comes when one search engine with a huge market share dominates what is shown in response to user queries.' [6] Thus, a diverse search ecosystem is essential for competition, democratic participation and digital sovereignty. Users must be able to freely choose which search systems to use; this requires alternatives to dominant platforms, fair conditions for competition, and the development of independent infrastructures.

The realisation of this principle requires structural, regulatory and technical conditions, including the following core aspects:

Limit monopolies and organise the market fairly: Effective regulation is needed to address market concentration and ensure fair competition. Governments must not only establish binding rules for data access, public accountability and plurality, but also secure rigorous enforcement.

Promoting open standards and interoperability: Open standards encourage technological independence and innovation. This also applies to web search technologies, which consequently need to be open and interoperable. This includes open protocols and APIs, independent servers and shared infrastructure. The decoupling of indexing, ranking and user interfaces could contribute to technological sustainability and enable a more diverse and sustainable search ecosystem. [6]

Fostering alternative search technologies: A pluralistic search ecosystem depends on visible and viable alternatives. However, public-interest, emerging or niche search systems face structural disadvantages in terms of visibility, funding, and infrastructure. Pluralism in search cannot be left to the market alone. Alternative public interest systems need long-term funding, visibility and support. Decentralised and open technologies need to be promoted through public and private investments in interoperable infrastructures, open indexes [6] and open source tools. Educational and public institutions should actively adopt and promote these alternatives.

Build sustainable and resilient infrastructures : Technological sustainability is vital for long-term autonomy and public-interest orientation. A pluralistic search ecosystem requires deliberate, sustained support for alternatives that align commercial success with democratic values. This includes establishing viable economic foundations and shared technical resources that ensure long-term independence.

Strengthening the protection of privacy

Search behaviour reveals highly sensitive information about individuals – their thoughts, concerns, intentions and identities. If this data is collected, analysed or monetised without consent, it poses serious risks to privacy, autonomy and personal safety. The growing power imbalance between users and providers becomes especially problematic when behavioural data is used to personalise results, manipulate attention, or target users.

Protecting privacy in search is therefore not merely a technical or legal task – it is a matter of individual freedom, democratic integrity and public trust.

To implement this principle effectively, three areas of action are particularly relevant:

Privacy and data minimisation as a basic principle: Digital search systems must respect the right to informational self-determination and apply data minimisation by default. This includes minimising data collection, applying privacy by design, anonymising behavioural data and treating user data as highly sensitive. Privacy settings must be set to maximum protection by default.

https://doi.org/10.5281/zenodo.17251788

Returning data sovereignty to the users: Users must have full control over their data. This includes simple to use access, deletion and opt-out options, as well as protection against manipulative 'dark patterns'. Tools such as one-click anonymisation, data dashboards and complaint mechanisms are essential.

Ban on tracking, profiling and monitoring: Tracking and profiling must be prevented. This applies to all behavioural data, not just traditional identifiers. Profiling for advertising or political targeting must be prohibited. Non-profiling alternatives must be easily accessible. Aggregated, anonymous data can be used for evaluation and research, as suggested by Granitzer et. al. [3].

Giving users, content creators, and advertisers more control.

When Web Search Systems dominate the entire information ecosystem, they become gatekeepers that control all stakeholder relationships, unilaterally determining how information can be found, presented, or monetised. Empowering users, creators and advertisers restores balance, and creates a fairer digital marketplace where each stakeholder can align the search ecosystem with their legitimate needs and values.

Empowering stakeholder in web search systems involves the following three key elements:

Increase influence on the selection of sources, search processes and functions: Users should be able to influence the selection and weighting of sources. The way sources are prioritised shapes which perspectives users encounter. If this process remains opaque, critical engagement and informed assessment are hindered. Customisable filters, visualised metadata (e.g. origin, relevance, timeliness) and adjustable source preferences can help users navigate search more independently and pluralistically.

Facilitate control of search functions: Currently, providers determine all key parameters – from indexing and crawling to algorithmic sorting and display logic. Thus, users need influence over the search logic itself. Ranking settings, personalisation options, and content types should be adjustable. Interfaces must be transparent and allow non-personalised, ad-free or thematic search modes.

Improve control for content providers: Furthermore, content providers must be granted enhanced control, as current standards (e.g. robots.txt) are deemed inadequate [3]. Tools such as dashboards and machine-readable standards should facilitate fine-grained control over indexing, snippets and metadata.

Empowering advertisers and reduce platform dependencies : Advertisers depend on a few platforms for customer reach while platform corporations control both the marketplace and its rules. Advertisers must regain autonomy over their advertising strategies without being locked into monopolistic systems and non-ethical practices.

Preventing discrimination and enabling participation for all

Inclusive, non-discriminatory and accessible search is a prerequisite for education, participation and opinion formation in a democratic society. If search systems reinforce existing inequalities, they marginalise vulnerable groups and undermine digital justice. Thus, web search systems should connect people, not exclude them. They must be designed to ensure equal access to information – irrespective of factors such as background, language, disability, or social conditions.

Three main operational dimensions can be distinguished to address discrimination and foster inclusion:

Preventing discrimination through data and algorithms: The presence of biases in data and algorithms has the potential to perpetuate social inequalities; as such, it is imperative that systems are designed to detect and mitigate such effects, ensuring that no group is systematically marginalised.

Promoting fair treatment of content and content providers: Rankings must not favour commercial or dominant content at the expense of smaller, non-commercial or public-interest information. Providers must ensure equitable conditions and refrain from privileging (their own) content without transparency.

Ensuring universal accessibility and cultural diversity strengthens inclusion and equity. Search interfaces should follow accessibility standards (e.g. WCAG, EN 301 549) and support content in multiple languages, including regional dialects and minority languages. Without this, large parts of the population are denied access to digital knowledge.

Information plurality and diversity of perspectives

Democratic discourse, public knowledge and social cohesion depend on diverse viewpoints. If information is filtered in a way that marginalises certain views or reinforces dominant narratives, this distorts public debate, excludes minority groups and undermines equal access to knowledge.

In shaping which perspectives are visible, web search systems, including AI-powered applications, play a key role. They carry a special responsibility to reflect the complexity of public discourse and to actively counter one-sided representations.

Ensuring a diverse and pluralistic information environment involves the following fields of intervention:

Inclusion and promotion of a wide range of sources and viewpoints: web search rankings have a significant impact on consumer choices, mainly because users trust and choose higher-ranked results more than lower-ranked results [1]. Algorithms must not favour only dominant narratives or commercially strong sources. Rankings should not rely solely on popularity or technical relevance, but also reflect journalistic quality, pluralism

https://doi.org/10.5281/zenodo.17251788

and social relevance. Smaller and independent sources must be included, especially on sensitive or controversial issues.

Prevention of algorithmic filter bubbles and echo chambers: While tailored results can increase short-term relevance, they risk creating echo chambers that limit exposure to alternative views. Providers must offer tools to manage or disable personalisation and ensure exposure to diverse viewpoints.

Equitable treatment of non-commercial content : Commercial dominance in search results creates an information landscape shaped by profit rather than the public good, while providers of non-commercial content often lack the financial capacity and knowledge to optimise their content for web search, resulting in their content becoming invisible. Equitable treatment of non-commercial content – from public health information to community resources – ensures that public interest is not subordinated to commercial interests.

Strengthening local and cultural content: Global platforms often prioritise dominant languages and centralised content, while regional or minority perspectives become invisible. To preserve cultural and linguistic diversity, regional and minority content should be supported through local weighting, multilingual support and targeted promotion.

Protect access to vital public information: Content of public interest, such as educational resources, health information and emergency services, is often vital and must be remain visible and easily accessible, rather than being crowded out by commercial search results. Providers should promote open access and work with trusted institutions to maintain quality and relevance.

Responsibility for environmental and social impact

Web search systems shape not only what we find, but also how societies evolve. Their development, however, often follows commercial or efficiency-driven logics that overlook broader societal and environmental impacts. As their influence grows, so does the need to align them with public-interest goals such as environmental and ethical responsibilities.

The social and ecological responsibility of search systems is reflected in the following areas of concern:

Technology impact assessments: Technology impact assessments should be mandatory to anticipate the risks of web search systems to democracy, public discourse and the environment. Trade-offs – such as speed versus energy consumption – must be critically assessed.

Minimising the environmental footprint: Search systems consume considerable energy and resources across their lifecycle. Search technologies need to minimise their environmental footprint. This includes efficient coding, durable hardware and transparency about resource use such as energy and water usage including transparent environmental metrics (e.g. CO_2 per search). Public funding should support research on resource-efficient search.

Human rights and social responsibility: Providers of web search systems must uphold human rights throughout their supply chains, reject unethical industries and commit to inclusive, socially responsible development.

Industry-wide responsibility beyond legal minimums: A voluntary code of conduct can promote higher ethical standards, complement legal norms and serve as a benchmark for responsible practice. Independent assessments and transparency tools enhance credibility – a chance to gain market advantages especially for smaller providers.

Ensure and strengthen integrity and trustworthiness of search results

The reliability of search results is essential for informed decision-making, social cohesion and democratic discourse. The manipulation of search results on the other hand can distort public perception [4], mislead users and undermine trust in digital information.

Providers of web search systems must ensure that their systems are tamper-proof, unbiased and transparent in order to justify users' trust and safeguard the societal function of search. To safeguard the integrity of digital search, the following four aspects should be considered:

Protection against manipulation: Search providers need to protect against manipulation, such as bought rankings, politically biased autosuggestions or SEO spam. Clear rules, public safeguards and independent audits must be in place.

Protection against disinformation and harmful content: Disinformation and harmful content undermine trust and can threaten public health, safety and democracy. Providers must build in mechanisms to identify and flag misleading content and prioritise science-based, verifiable sources – especially on sensitive issues such as climate, health and elections.

Transparent and controllable autocomplete functions: Features such as Google's autocomplete that provide potential queries in the search box while a user is typing, not only 'frame how to consider particular ideas and their associated values', but also 'have the power to shape the terms of a user's enquiry, and consequently, direct wider public discourse' [2]. To reduce bias and manipulation, autocomplete mechanisms must be transparent, explainable, and under user control. Users should understand how suggestions are generated, be able to customise them, and have clear options to disable or report problematic entries.

Accurate presentation and interface design: The way search results are presented affects perception and decision-making. 'Current SERPs exhibit a very complex structure. They contain organic results, ads (search-based advertising; for example, text ads, shopping ads), verticals (e.g. news, images), direct answers and knowledge graph results'[8]. Interfaces must be clear, understandable, truthful and accurate. Search providers must not distort,

https://doi.org/10.5281/zenodo.17251788

misrepresent or manipulate how information is displayed to users. Results must reproduce content correctly and not distort it.

Strengthen search competences and critical awareness

Confident and informed use of web search systems requires more than access – it requires competence. Yet studies show that this competence is often lacking. While users rate their search skills as high, they have limited understanding of how search engines work or are funded Schultheiß & Lewandowski [8]. Platz et al. [7] observe immaturity in user behaviour and low awareness of privacy risks among students. "Search literacy" and "search engine literacy" help users assess credibility, interpret results, and navigate complexity – supporting participation, protecting against manipulation, and enabling democratic engagement. Understanding the ethical and societal dimensions of search technologies also fosters ethical awareness and design.

This principle translates into three main fields of action, all addressing different actor groups:

Promote search skills specifically and at an early stage: Search skills need to be developed early and systematically. Curricula should integrate digital search skills from kindergarten to higher education. Tutorials and public resources can complement this in lifelong learning.

Anchoring ethical awareness in technology and companies: Those who design and operate web search systems also need to understand their impact. Ethics training, interdisciplinary exchange and clear guidelines will help ensure that technology serves the public good.

Foster critical research and public scrutiny : Independent research on Web Search Systems is essential for understanding their societal impact and ensuring accountability, yet independent researchers cannot adequately study them due to lack of data access, legal barriers, and insufficient funding.

Enforcing democratic control and binding rules

When the reach of web search systems becomes systemic, voluntary commitments are no longer sufficient. Without enforceable accountability, risks such as discrimination, disinformation, or abuse of power remain unchecked. Providers must be held accountable for how their systems affect individuals and society. Effective accountability requires binding rules, independent control, and societal participation from local levels through national regulation to international standards across national borders.

A rights-based and enforceable accountability framework requires attention to the following areas:

Mandatory obligations and independent oversight: Trust cannot be based on goodwill alone. Accountability requires binding commitments, independent oversight and introduce sanctions for harm. Without external oversight and real consequences, even well-intentioned ethics codes fail. Enforceable standards protect against abuse of power and ensure providers are held responsible for the consequences of their actions. Users affected by biased or harmful results need clear avenues for redress.

Distributing power by participatory and democratic governance: Today, key decisions in the search landscape are made by a small number of commercial actors. Democratic governance can address this imbalance by distributing power and ensuring that oversight is independent, multidisciplinary and open to public debate. It can minimise one-sided influence, strengthen trust in digital infrastructures, and help incorporate different perspectives to align Web Search Systems with society's needs.

In this sense, a search infrastructure in the public interest should not only be accessible to the public – it must be shaped by the public.

International frameworks and shared standards: Free access to trustworthy information is a universal right – not a privilege of origin or market position. Search technologies are global, but rights and protection structures often end at national borders. Without international cooperation, powerful actors will remain highly influential, while local alternatives have little chance. Users in vulnerable regions are particularly at risk – from censorship, surveillance or exclusion from access to reliable information.

International cooperation and coordinated frameworks that work across borders are needed to reduce digital dependencies and enforce equitable standards worldwide, regardless of geography or market power. Human rights, transparency and public interest principles should be embedded in multilateral agreements.

CONCLUSION: ETHICAL ORIENTATION FOR THE FUTURE OF WEB SEARCH

The societal significance of web search has grown steadily over the past decades. As a central interface for access to knowledge, education, and public discourse, it shapes not only what individuals find, but also what perspectives are made visible, and how collective decisions are informed. Despite this relevance, the development and governance of web search systems has largely remained in the hands of private actors – driven by commercial incentives, technological optimisation and opaque decision-making structures.

This article outlined ten guiding principles for a public-interest-oriented web search as formulated in the #FreeWebSearch Charter[1]. They are the outcome of an exploratory and interdisciplinary process initiated by the Open Search Foundation, building on the work of its Ethics Working Group. The principles are not yet a comprehensive ethical framework, but they aim to provide orientation for the further ethical, regulatory and institutional discussion and development of web search systems.

[1] The #FreeWebSearch Charter was made public at the end of September 2025. https://charter.freewebsearch.org

https://doi.org/10.5281/zenodo.17251788

While the challenges addressed in the Charter affect society as a whole, the responsibility to act lies particularly with those in positions of influence: policymakers, public institutions, search system providers, researchers, educators, civil society organisations and the media. As web search increasingly functions as a societal infrastructure, responsibility must be shared across sectors and national borders.

The principles also serve as a reference point for ongoing practical work – for example, within the European OpenWebSearch.eu project, where selected aspects are being explored and implemented.

Further development of a comprehensive framework for ethical web search remains an open and ongoing process. As a living system, it depends on the critical reflection, uptake and discussion by various stakeholders. The Charter, as a first step, aims to stimulate this debate and to provide a foundation for more in-depth analyses, mitigation and implementation strategies and concrete guidelines in the future.

Acknowledgements:

The authors are members of the Open Search Foundation's Ethics Working Group. This work draws on the group's collaborative efforts over the past few years. The views expressed are those of the authors.

The authors would like to thank the Stiftung Mercator and the CAIS Center of Advanced Internet Studies for facilitating workshops.

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

REFERENCES

[1] Epstein, R., & Robertson, R. E. (2015). The search engine manipulation effect (SEME) and its possible impact on the outcomes of elections. Proceedings of the National Academy of Sciences, 112(33), E4512-E4521. https://doi.org/10.1073/pnas.1419828112

[2] Graham, R. (2023). The ethical dimensions of Google autocomplete. Big Data & Society, 10(1). https://doi.org/10.1177/20539517231156518

[3] Granitzer, M., Voigt, S., Fathima, N. A., Golasowski, M., Guetl, C., Hecking, T., Hendriksen, G., Hiemstra, D., Martinovič, J., Mitrović, J., Mlakar, I., Moiras, S., Nussbaumer, A., Öster, P., Potthast, M., Senčar Srdič, M., Sharikadze, M., Slaninová, K., Stein, B., & Wrigley, S. N. (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. https://doi.org/10.1002/asi.24818

[4] Haider, J., Rödl, M. (2023). Google Search and the creation of ignorance: The case of the climate crisis. Big Data & Society, 10(1). https://doi.org/10.1177/20539517231156518

[5] Introna, L. & Nissenbaum, H. (2000)]. Shaping The Web: Why The Politics Of Search Engines Matters. The Information Society. 16. 169-185. 10.1080/01972240050133634.

[6] Lewandowski, D., & Sundin, O. (2021). How an open web index could help solve the „search problem". https://medium.com/in-search-of-search/how-an-open-web-index-could-help-solve-thesearch-problem-41363b96158e. Accessed: March 31 2025.

[7] Platz, M., Klan, F., Decker, A. (2022): Developing and promoting search engine literacy in primary education. Proceedings of 4th International Open Search Symposium #ossym2022, Online, hosted by CERN, Geneva Switzerland, 10-12 October 2022, M. Granitzer, C. Gütl, C. Plote, S. Voigt, A. Wagner (eds). https://10.5281/zenodo.8399978. Pp 43-48.

[8] Schultheiß, S., & Lewandowski, D. (2023). Misplaced trust? The relationship between trust, ability to identify commercially influenced results and search engine preference. Journal of Information Science, 49(3), 609–623. https://doi.org/10.1177/01655515211014157

https://doi.org/10.5281/zenodo.17229609

LEGAL CHALLENGES FOR OWI AND SEARCH ENGINE USE[*]

P. C. Johannes[†], H. Koulani[‡], ITeG, University of Kassel, Germany
L. Beer[§], Open Search Foundation e.V., Munich, Germany

Abstract

Opening up the current search engine market, would create diversity and freedom of choice when searching the internet, thus strengthening overall informational self-determination. The European Open Web Index (OWI) aims to serve as an alternative to the closed, non-transparent systems of large platforms and search engines. It would provide a basis for the services of a large number of (new) search engines and other application developers. This article focuses on the applicable European legal framework. After a short introduction of the OWI and it's business uses, it concentrates on the use case "vertical search engines". Addressed are the most important legal challenges to this use case from the perspective of the OWI as well as the vertical search engine developers. Based upon this, the article investigates the user perspective, i.e. what is required for an OWI to be trusted and used. It concludes with an overall assessment and future outlook.

OWI BUSINESS MODELS

The business models of search engines are based on the targeted marketing of user data, for example for personalised online advertising. The collection and analysis of user data makes people very predictable for advertisers. This makes the users of search engines vulnerable to manipulation, for example by online advertising that is specifically tailored to the individual surfing behaviour, interests and circumstances of the internet users. The search engine market is currently dominated worldwide by only four providers who have their own so-called web indexes. An OWI as an alternative to the closed, non-transparent systems of the large platforms will provide a basis for the services of a large number of (new) search engines. It can also be used for research and innovation, for example in the field of artificial intelligence (see [1] for more details). The OWI is intended to strengthen the EU's digital sovereignty by reducing dependence on the search engine monopolists through a sustainable, freely accessible web index [2]. Similar to other indexes, the OWI is being created by systematically crawling the web, analysing the crawled content and storing it with metadata in a database [3].

In order for this emerging open search infrastructure to be fully effective, it must be designed in a way that is compatible with fundamental rights and can be operated within the current European legal framework[4]. For example questions arise in the context of the 'right to de-indexing' and the effects and application of the European legal framework on data, digital services and online platforms. The aim is to align the design of an open search infrastructure with the fundamental rights and principles that the European Commission has also declared as the benchmark for the 'Digital Decade' [5]. The PriDI project [6] is researching how an OWI can be designed in a way that is compliant with fundamental rights of users and operators and is protective of privacy.

To achieve this, possible use cases of an OWI were developed and examined. This allowed to analyse legal challenges from the perspective of the users of an OWI. Use cases were initially based on the study by *Nowakowski/Zimmermann* [7]. That study groups possible use cases into the categories web search, enterprise search, information portals, value-adding services, content management and e-commerce. Web search and value-adding services are by far the largest categories. The study contains a list of possible applications, although it should be noted that this is not comprehensive. These applications are mostly only tagged and briefly described in the study, but not defined in detail. The PriDI study builds upon those and thus closes the gap for certain use cases and lays the foundation for further legal evaluation [8]. The detailed use cases always consist of the goal of the use case and the relevant actors within the specific use case. In this context, a distinction is also made between a user and an actor.

When assessing the OWI from a legal and user acceptance perspective, a distinction must be made between different stakeholders. A wide range of people and entities can be considered as actors. In the case of the OWI, this might be an organisation that acts as a data retriever/consumer or the institutions that develop the OWI as index developers. For the purposes of this study the following roles were defined:

First of all, there are the *data subjects*. This role describes persons or companies whose personal data and intellectual property (IP) are stored in the OWI and are used or can be found by tools or search engines based on the OWI.

[*] Based on research of "Privacy-enhancing digital infrastructures" (PriDI), funded by the German Federal Ministry of Education and Research (BMBF). The responsibility for the content of this publication lies with the authors. This papers contains parts that have already been submitted for publication to CPDP.AI 2025.

[†] paul.johannes@uni-kassel.de

[‡] koulani@uni-kassel.de

[§] lb@opensearchfoundation.org

https://doi.org/10.5281/zenodo.17229609

The index itself is developed and maintained by the *OWI developer*. The OWI developers have by nature of collaboration alone formed an operator consortium. For the purposes of this study it is assumed they have formed a legal entity of some kind.

The data retrievers or data consumers of the information contained in the OWI are referred to as *application developers*. They are persons or organisations that request the retrieval of web data from the index in order to create and develop various tools and models based on the retrieved data for their own applications and services, e.g. for a search engine.

Lastly, *end users* also come into contact with the index and the applications built on it's use. These are the natural or legal persons who use the tools and systems developed by the application developers.

USE CASE “VERTICAL SEARCH ENGINE”

The OWI can serve as a basis for the development of new (vertical) search engines. For example, search engines could be created for specific user groups, geographical areas or topics. The OWI creates the basis for this by providing a comprehensive database. Developers of specialised search engines can access the index and filter out exactly the data that is relevant for the respective purpose. This means that smaller organisations and companies can also develop search engines and thereby increase diversity in the search engine market. Organisations and companies do not have to devote enormous resources to crawling and indexing the entire or a part of the web, but can deliver precise and relevant results for specific topics based on the OWI. As the source code of the OWI is publicly available, developers can understand the underlying algorithms and adapt them to their specific needs. For example, the OWI can be used to create news search engines that provide individuals with news about both current and past events. In addition, press reviews on various topics, events or even companies could be compiled.

For the purpose of this study, the following assumptions for the use case were made:

The primary objective of this use case is to develop a news search engine tailored for individuals and businesses by retrieving web data on current news. This initiative involves several key actors: application developers, data subjects (such as article authors, newspapers, and individuals mentioned in news articles), OWI developers, and end users.

It is also assumed that the application developers are a sole proprietorship whose clients include companies and authorities seeking to monitor media coverage of individuals and businesses. Additionally, non-governmental organizations (NGOs) interested in specific political topics, such as legislation, climate protection, and study results, also utilize their services.

The interaction process begins with the application developer engaging with the OWI system to request access to a pre-built news index. The developer specifies that the data will be used for developing a news web search engine . Upon receiving the request, the system processes it and, if technically feasible, notifies the data subjects about the intended use of their data. Once processed, the application developer receives the requested data, concluding the interaction.

End users interact with the news search engine by submitting queries related to events, persons, or companies. The system displays relevant results along with their respective sources. Users may also be informed that the search engine relies on OWI data and the time period from which the data is sourced, ensuring transparency in the data usage process. This interaction concludes once the user has reviewed the search results.

LEGAL CHALLENGES

The complexity of the OWI and it's use by prospective vertical search engine application developers raises many questions as to how it falls under the law of the European Union and the laws of it's member states.

European law on data and online services has undergone major changes in recent years. The General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) lays down detailed rules on the handling of personal data. Now a regulatory network of more or less specialized, directly applicable legal acts for data and digital services has emerged. Among others the Digital Services Act (DSA, Regulation (EU) 2022/2065), the Data Governance Act (DGA, Regulation (EU) 2022/868), the Data Act (DA, Regulation (EU) 2022/868), the Digital Markets Act (DMA, Regulation (EU) 2022/1925) and the AI Act (AIA, Regulation (EU) 2024/1689) form a legal framework for digital services and on data (see more detailed [[9]]). Regulations are directly and uniformly applicable in all member states. Complexity is added by the various laws of member states, establishing national frameworks for implementing these regulations, e.g. in Germany the Digitale Dienste Gesetz (DDG) for the DSA and the Bundesdatenschutzgesetz (BDSG) for the GDPR.

Others questions arise from the legal framework for copyright. In the EU this is primarily governed by a combination of EU Directives, international treaties, and national laws of member states (also see [1]).

This article focuses on legal challenges faced by the OWI developer and the vertical search engine application developers across the regulatory domains of data protection and digital services. The initial assessment of the use case highlights the complexities and the need for ongoing studies.

In the case that the OWI serves as a basis for the development of new (vertical) search engines the OWI developer provides a comprehensive database to the application developer of this new search engine. The ensuing relationship as well as the operation of the new service, have to be scrutinized in light of the aforementioned regulations. This article focuses on the GDPR and the DSA.

https://doi.org/10.5281/zenodo.17229609

RULES ON DATA PROTECTION

The OWI will, intentionally or unintentionally, contain personal data of natural persons (data subjects) within the meaning of Article 4 No. 1 GDPR [10][11]. This data will be contained in the crawled data. It may even be special category data within the meaning of Article 9 para. 1 GDPR. Personal data could also be part of the metadata collected and enriched by the OWI (e.g. owners of websites, contact details). The protection of personal data is a fundamental right (see Article 8 Charter of Fundamental Rights of the EU (CFR)) and is governed by the GDPR. The entity operating the OWI is a controller with all responsibilities under the GDPR in relation to the personal data it processes.

LAWFUL PROCESSING

The processing of personal data is only lawful if it can be justified on a legal basis within the meaning of Article 6 GDPR. The creation and operation of an independent, freely accessible web index is in the public interest, if only to reduce the dependence of Internet users on foreign search engines. For the lawfulness of processing in the public interest pursuant to Article 6 para. 1 subpara. 1 lit. e GDPR requires a legal basis in Union law or in the law of the Member States in accordance with para. 3. Such a basis is not yet apparent, so that a public body or public bodies processing the OWI would have to fall back on a general purpose authorisation in law of the member states. For example: A data centre, organized as a public body in Bavaria in Germany is likely to process personal data for an OWI, e.g. crawling, enriching and sorting websites and indexing them, on the basis of its own purpose (i.e. statutes and/or establishment act) in conjunction with Article 4 Bavarian Data Protection Act (BayDSG) on the basis of and in conjunction with Article 6 para. 1 subpara. 1. lit. e and para. 2 and 3 GDPR.

For private entities, the creation and operation of the OWI can only be based on the protection of legitimate interests of the controller or a third party in accordance with Article 6 para. 1 subpara. 1 lit. f GDPR. This requires a balancing of the interests, fundamental rights and freedoms of the data subject. The legitimate interest of the OWI operator as the controller is to create and maintain an OWI. The legitimate interest of the general public as a third party is to have access to an independent, freely accessible web index. The interest of the data subjects in not being included in a web index and thus protecting their informational self-determination does not outweigh these legitimate interests. The OWI only crawls data that is freely accessible to everyone and has already been crawled many times by search engine operators such as Google, Bing, Yandex or Baidu. It is also common knowledge that this happens technically, so that data subjects can protect themselves against it by taking appropriate measures against those who publish their data. Therefore, the interests of the data subjects do not outweigh the legitimate interests of the OWI developer and the general public in the creation of an OWI. This is also true for the application developers using the OWI's data. The news search engine developers as defined in the use case can rely their processing of personal data (OWI data as well as the queries of it's end users) on lit. f as well.

The constellation in which an OWI crawls special categories of personal data that are publicly accessible but have not been made publicly accessible by the data subject itself remains problematic and has not yet been fully clarified. This also applies to application developers who use this data. According to Article 9 para. 1 GDPR, the processing of special categories of personal data like health data or data regarding sexual orientation, is prohibited unless there is an exception according to Article 9 para. 2 GDPR, especially making that data available to the public by the data subject itself (lit. e).

If a third party has made this data accessible on the internet instead of the data subject itself, there might be no legal basis for the OWI to crawl this data and the search engine developers to use this data. It is doubtful that the OWI provider and search engine developer could face requests for deletion in accordance with Article 17 GDPR. Theoretically, fines could be imposed in accordance with Article 83 GDPR, whereby many factors would play a role in the assessment of a possible fine, in particular whether the controller is responsible for disclosing the specific data in question. However, in light of the fact that major search engines have been exposed to this risk for many years and infringements have not yet been prosecuted, a fine is rather unlikely. Still, a risk remains. To address this, it could be argued that the OWI and the search engines operate under the assumption, that information contained in crawled webdata is not processed within the meaning of Article 9 para. 1 GDPR, since it does no take action to identify this information as such. It can also be argued, that the OWI and the search engine can assume that, since all data crawled is publicly available, it was made publicly available by the data subject itself within the meaning of Article 9 para. 2 lit. e GDPR. Further processing to identify or ascertain this are not necessary because of Article 11 GDPR, which states that the controller shall not be obliged to maintain, acquire or process additional information in order to identify the data subject for the sole purpose of complying with the GDPR. Furthermore, in the case that the controller of the OWI is a public body, it is very likely that it is allowed to process special categories of personal data in accordance with member state law on the basis of a general purpose authorisation statute in conjunction with its own governing statues on the basis of Article 9 para. 2 GDPR.

OBLIGATIONS

Both the OWI developer and application developers like search engine providers would have to comply with GDPR principles (Article 5 GDPR) when processing personal data included in the in the database of the OWI. As controllers they must fulfil a number of technical and organizational obligations in accordance with the GDPR in order to mitigate the risks for data subjects and their rights and fundamental freedoms (e.g. Chapter IV

https://doi.org/10.5281/zenodo.17229609

GDPR). On the technical and operational level these include, but are not limited to the creation and maintenance of records of processing activities (Article 30 GDPR), the obligation to implement appropriate technical and organisational measures to ensure security of processing (Article 32 GDPR) and the implementation of appropriate technical and organisational measures to ensure data protection by design and by default (Article 25 GDPR). Other obligations of OWI and search engine providers are the need to undertake a Data Protection Impact Assessment (DPIA, see Article 35 GDPR) and to designate a Data Protection Officer (see Article 37 GDPR).

The OWI developer as well as the application developers should also create information policies: The OWI developer as well as the search engine developer are not obliged to inform the data subjects about the data collection or processing proactively. According to Article 14 para. 5 lit. b sentence 2 GDPR, however, the controller must take appropriate measures to protect the fundamental rights and freedoms as well as the legitimate interests of data subjects. Sentence 2 mentions the provision of information to the public as an example. The OWI developer as well as the search engine developer should therefore provide transparent, clear and comprehensible information on their website about the manner in which (personal) data is collected and the rights of the data subjects as well as the reasons why they respectively, as the controllers, are relying on Article 14 para. 5 lit. b GDPR. They should therefore draft and make publicly available a privacy policy or data privacy statement. Furthermore, the GDPR contains a number of data subject rights about which the controller must provide full information and which it must comply with at the request of the data subject. These include the right to access under Article 15 GDPR, the right to rectification under Article 16 GDPR, the right to erasure under Article 17 GDPR, the right to restriction of processing under Article 18 GDPR and the right to object under Article 21 GDPR. In addition, every data subject has the right to lodge a complaint with a supervisory authority in accordance with Article 77 GDPR. The OWI developer as well as the search engine developer should as organisational measures implement each concepts to receive and act upon queries of a data subject pertaining to their rights.

The OWI developer shares data with the application developers e.g. the search engine provider. If there are several controllers who jointly determine the purposes and means of processing, they are jointly responsible. This applies if several organizations operate the OWI together. As such, they are joint controllers and obliged to define in an agreement in accordance with Article 26 GDPR how the obligations under the GDPR are to be implemented in detail. The joint controller agreement shall be made available to the data subjects. If infringements occur during the processing of personal data, both joint controllers can be held liable. That the OWI developer and search engine developer would also be joint controllers within the meaning of Article 26 GDPR is unlikely. They might be for certain aspects of their relationship and data processing operations, like the sharing of the index data or of parts of the crawled data.

RULES ON DIGITAL SERVICES

In its current form, the OWI as well as the (news) search engine will each fall under the regulation of digital services, in particular the DSA. The DSA regulates the liability and scope of due diligence obligations for providers of digital intermediary services. The aim of the regulation is to provide an appropriate framework for the digital space, to create a safe, predictable and trustworthy online environment and to protect fundamental rights, see Article 1 para. 1 DSA. According to Recital 29 DSA intermediary services “cover a wide range of economic activities that take place online and are continuously evolving to enable the fast, safe and secure transmission of information and to provide convenient solutions for all stakeholders in the online ecosystem”.

INTERMEDIARY SERVICES

The classification of the OWI as an intermediary service under the individual characteristics of the DSA is difficult, especially when it comes to the qualification as a hosting service, online platform or search engine. The OWI can only be classified as an intermediary service within the meaning of Article 3 lit. g No. iii DSA and as an online search engine within the meaning of Article 3 lit. j DSA with some assumptions or leaps of thought [[12]]. It is obvious that the legislator had use cases such as social media and market portals in mind when designing the DSA, while higher-level applications such as a web directory or index were not even considered. Also, the rules on online search engines were not thought through to the end. Nevertheless, the functions and possibilities of the OWI are often comparable with typical use cases covered by the DSA, so that according to the purpose of the DSA, the OWI should fall under it and thus a legal-teleological argumentation (purpose of the law) can be made in this direction.

The OWI developer would not itself fall under the rules of the DSA concerning search engines, since it does not provide a search interface. The developer of the news search engine on the other hand would. The search engine operator would likely be considered an intermediary and could benefit from limited liability for content hosted by third parties, provided they comply with their obligations under the DSA (e.g. content moderation, transparency reports).

In terms of both, the purpose of the DSA and general considerations, it would make sense to allow the OWI developer to benefit from the exemption from liability of the DSA or comparable privileges. At the same time, however, the OWI developer should also be subject to corresponding duties of care, which are considerable. These obligations are even greater when the OWI would be categorized as a very large online search engine.

The legislator should clarify this accordingly and include both online search engines and indexing services

https://doi.org/10.5281/zenodo.17229609

in the definition of Article 3 lit. g DSA and name them accordingly in the obligations and exemptions from liability. In lieu of such clear-cut rules as of now, an OWI operator should assume the OWI is as search engine and hosting service within the meaning of the DSA and regulate its operation accordingly.

LIABILITY UNDER DSA

As providers of hosting services, the OWI developer as well as the search engine developer would be exempt from liability in accordance with Article 6 para. 1 lit. a and b DSA, provided that they have no knowledge of illegal content in the index data and take immediate action to block access to this content as soon as they become aware of it. According to Recital 22 DSA, knowledge cannot be assumed solely from the general awareness that the service can be used to store illegal content. Recital 22 DSA also emphasizes that automatic indexing of illegal content is not sufficient to establish specific knowledge. According to Article 7 DSA, the exemption of liability also applies if the provider, on its own initiative, carries out voluntary investigations in good faith and diligently or undertakes other measures to detect, identify and remove illegal content or to block access to illegal content or takes the necessary measures to comply with legal requirements.

A general obligation of the OWI developer or search engine developer to monitor and actively investigate the index data for illegal content does not exist in accordance with Article 8 DSA. However, pursuant to Article 6 para. 4 DSA, at the request of a judicial or administrative authority, they would have to comply with an order pursuant to Article 9 para. 1 DSA to take action against illegal content and must, pursuant to Article 10 DSA, inform the issuing authority immediately of the receipt and implementation of the order.

GENERAL PROVISIONS UNDER DSA

According to Article 11 DSA, the OWI operator is obliged to designate an easily accessible central point of contact for the authorities of the Member States, the Commission and the panel within the meaning of Article 61 DSA. The point of contact should enable smooth electronic communication. In addition, a central point of contact for users of the service must be designated in accordance with Article 12 DSA, which enables users to communicate directly, quickly and effectively.

In accordance with Article 14 DSA, both the OWI developer and the search engine developer are obliged to provide clear and comprehensible information in its General Terms and Conditions (GTC) about restrictions on the information provided and the use of the services. According to Article 3 lit. u DSA, the GTC are all clauses that govern the contractual relationship between the developer and the users or end users of the respective services. According to Article 14 para. 1 sentence 2 DSA, the information obligation covers all guidelines, procedures, measures and tools used to moderate the content. The OWI developer provides services to the application developers. It should therefore provide the necessary information at least to the search engine developer. The search engine developer renders it's service to end users and should address them. Furthermore, as discussed before it is conceivably to consider all natural persons whose data are indexed as users of the index within the meaning of the DSA. Therefore, the OWI developer should provide the information publicly towards all, data subjects and end users, not only in the interest of transparency but also to ensure compliance with the DSA.

The information should explicitly mention algorithmic decision-making, human review and the procedural rules of the internal complaints management system pursuant to Article 20 para. 1 DSA. Therefore, the OWI developer in particularly is obliged to explain transparently and in detail whether and how user content is crawled and indexed and how it is used in the index and made accessible to third parties. In order to comply with this obligation, the OWI developer could disclose the crawling methods and technologies used, as well as provide a list of the excluded index terms and domains.

In addition, in accordance with Article 14 para. 1 DSA, the procedural rules of the complaints management system and measures against abusive use of the system must be specified. Furthermore, in the case of very large online platforms, the GTC must be supplemented with content on available legal remedies in accordance with Article 14 para. 5 and 6 DSA and must be available in the official languages of all Member States in which the service is offered.

Article 15 DSA specifies special transparency obligations: according to this, the OWI developer must make a report on the content moderation carried out publicly available at least once a year. This would also apply to the search engine developer. Content moderation is defined in Art. 3 lit. t DSA as the activities of intermediary service providers aimed at identifying and combating illegal content provided by users that is incompatible with the provider's GTC, including measures relating to the accessibility of illegal content or information. Article 15 para. 1 lit. a to e DSA lists various aspects that need to be addressed.

ADDITIONAL PROVISIONS FOR HOSTING SERVICES

According to Article 16 para. 1 DSA, the OWI developer and search engine developers would be obliged to introduce an easily accessible and digital reporting procedure so that persons or entities can report illegal content. In order to facilitate a sufficiently accurate and reasoned report, Article 16 para. 2 sentence 2 lit. a to d DSA state that reports should contain (1) a sufficiently reasoned explanation as to why the person concerned considers the information to be illegal content, (2) a clear indication of the exact electronic location of this information, such as the URL address or, if necessary, the name of the website where the information is stored, (3) the name and email address of the reporting party, unless the information relates to a criminal offense, and (4) a

https://doi.org/10.5281/zenodo.17229609

statement that the reporting party has a good faith belief that the report is accurate and complete.

Such reports must be processed promptly, carefully, free of arbitrariness and objectively and inform the reporting party of the use of automated means for decision-making within the meaning of Article 16 para. 6 DSA; in addition, in accordance with Article 16 para. 3 and 5 DSA, an acknowledgement of receipt and the decision must be issued to the reporting party without delay and the possible legal remedy must be explained.

In addition, the OWI developer and search engine developers need to inform the law enforcement authorities of the respective Member State according to Article 18 para. 1 DSA or, in accordance with Article 18 para. 2 DSA, a representative or Europol immediately as soon as they become aware of information that gives rise to suspicion of a committed or possible criminal offence that poses a threat to the life or safety of a person.

USER ACCEPTANCE AND TRUST

Although legally compliant design and considerations like data protection and privacy hold great importance in Europe, user behaviour suggests that these factors are often secondary when selecting digital services. In practice, other aspects tend to influence decisions more strongly. Consequently, integrating essential elements of user acceptance into the development of OWI and related tools is crucial. User perception and their intention to use new technological solutions are essential concepts researched over the last decades. Many theories and models were introduced providing insights into factors that affect user acceptance of technology, such as TAM [13], UTAUT [14] and their extended versions [15, 16, 17]. Furthermore, trust in IT-artifacts has become an important construct to be considered while designing new technologies [18]. For the practical application of user acceptance and trust considerations for OWI, a combination of two theoretical theories is drawn upon, Trust-TAM [19] and UTAUT2 [17]. The latter introduces seven factors, such as performance and effort expectancy, social influence and facilitating conditions. Our study aims at adapting each factor to the OWI context and introducing suitable user acceptance considerations under each factor. In particular, the factor performance expectancy could be expanded to include two considerations, reliability and speed. These concepts encourage developers of OWI tools to leverage the database of OWI to implement search engines and services since reliability and speed of OWI responses matter to the tools' results. Considering the end users as actors in these use cases, promoting reliability and speed in the functionality and structure of OWI indirectly affects the perception of end users since these aspects would be mirrored in the developed search engines and local services for the end usage.

For a European web index to be competitive on a global scale—or at least to challenge international rivals within Europe—it must balance European principles of privacy and data security with key user preferences, ensuring that design choices and requirements are well aligned.

CONCLUSIONS AND OUTLOOK

This analysis demonstrates that the development and implementation of an OWI and applications based on it, such as search engines, presents a unique set of challenges and opportunities for the protection of fundamental rights. While the OWI holds the potential to enhance access to information (Article 11 CFR), foster competition in the digital market, and promote innovation, it also raises concerns regarding data protection (Article 8 CFR), the freedom to conduct a business (Article 16 CFR), and the protection of intellectual property (Article 17 CFR).

The current legal framework of GDPR and DSA guides and restricts the development of an OWI and it's application such as search engines in many ways. The requirements of both legal acts must be observed. However, it should also be emphasized that neither the GDPR nor the DSA fundamentally prevent or ban the business model presented.

Further studies and assessments of the OWI, it's use cases and their relationships with each other are ongoing and necessary. The comprehensive legal framework of the EU addresses some of the underlying concerns for rights and freedoms. Ongoing monitoring of the legal framework, coupled with adaptation, is crucial to ensure that the OWI contributes to a more open, inclusive, and rights and freedoms respecting European digital ecosystem.

REFERENCES

[1] L. Beer, P. C. Johannes and H. Koulani, "Legal aspects of AI training and retrieval augmented generation", presented at OSSYM 2025 - 7th International Open Search Symposium, Helsinki, Finland, Oct. 2025, paper #####, this conference – *submitted.*

[2] L. Beer *et. al.*, "An Open Web Index - opportunities and risks for fun-damental rights and freedoms", presented at CPDP.ai 2025, Brussels, Belgium, May 2025 – *submitted.*

[3] G. Hendriksen *et al.,* "The Open Web Index. Crawling and Indexing the Web for Public Use." in *Advances in Information Retrieval: 46th European Conference on Information Retrieval. ECIR 2024,* Glasgow. UK. Mar. 2024. pp. 130-143

[4] L. Beer *et. al.*, "Ein offener Webindex: Chancen und Risiken für Grundrechte und Grundfreiheiten", in Tagungsband Plattform Privatheit 2024, Berlin, Germany, Oct. 2024 – *forthcoming.*

[5] European Commission, COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS State of the Digital Decade 2024, COM(2024) 260 final.

[6] PriDI, https://pridi-projekt.de/.

[7] D. Nowakowski, N. Zimmermann, "Market potential assessment of OpenWebSearch.eu - Exploring the economic and societal impact of an Open Web Index", Mücke/Roth, Germany, 2024.

[8] H. Koulani, L. Beer, P. C. Johannes, C. Geminn, M. Söllner and S. Voigt, "Anwendungsfälle eines offenen Webindex", PriDI, Mar. 2025, published as white paper on https://pridi-projekt.de/.

[9] C. Geminn and P. C. Johannes (Eds.), *Handbuch Europäisches Datenrecht*, Baden-Baden, Germany: Nomos, 2025, to be published.

[10] C. Geminn, K. Erenli and L. Pfeiffer, "Legal Challenges of an Open Web Index", International Cybersecurity Law Review, vol. 2, no. 2, pp. 183-94, 2021. https://doi.org/10.1365/s43439-021-00017-8.

[11] C. Geminn, "Rechtsfragen eines offenen Web-Index - Infrastrukturen für die digitale Gesellschaft", Multimedia und Recht, no. 12, pp. 16-19, 202

[12] M. Nebel and P. C. Johannes, "Open Web Index im Lichte des Digital Services Act – Voraussetzungen – Grenzen – Rechtsfolge", Multimedia und Recht, no. 12, pp. 1010-1016, 2024.

[13] F. D. Davis, "User acceptance of information technology: system characteristics, user perceptions and behavioral impacts," International Journal of Man-Machine Studies, vol. 38, no. 3, pp. 475–487, Mar. 1993, doi: 10.1006/imms.1993.102

[14] V. Venkatesh, M. Morris, G. Davis, and F. D. Davis, "User acceptance of information Technology: toward a unified view," *MIS Quarterly*, vol. 27, no. 3, p. 425, Jan. 2003, doi: 10.2307/30036540.

[15] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: four longitudinal field studies," Management Science, vol. 46, no. 2, pp. 186–204, 2000, [Online]. Available: https://www.jstor.org/stable/263475

[16] V. Venkatesh, J. Thong, and X. Xu, "Consumer Acceptance and use of Information technology: Extending the unified theory of acceptance and use of technology," MIS Quarterly, vol. 36, no. 1, p. 157, Jan. 2012, doi: 10.2307/41410412.

[17] M. Söllner, A. Hoffmann, and J. M. Leimeister, "Why different trust relationships matter for information systems users," European Journal of Information Systems, vol. 25, no. 3, pp. 274–287, Dec. 2015, doi: 10.1057/ejis.2015.17.

[18] M. Söllner, A. Hoffmann, and J. M. Leimeister, "Why different trust relationships matter for information systems users," European Journal of Information Systems, vol. 25, no. 3, pp. 274–287, Dec. 2015, doi: 10.1057/ejis.2015.17.

[19] D. Gefen, E. Karahanna, and D. Straub, "Trust and TAM in online shopping: an integrated model," MIS Quarterly, vol. 27, no. 1, p. 51, Jan. 2003, doi: 10.2307/30036519.

[1] L. Beer, P. C. Johannes and H. Koulani, "Legal aspects of AI training and retrieval augmented generation", presented at OSSYM 2025 - 7th International Open Search Symposium, Helsinki, Finland, Oct. 2025, paper #####, this conference – *submitted.*

[2] L. Beer *et. al.*, "An Open Web Index - opportunities and risks for fun-damental rights and freedoms", presented at CPDP.ai 2025, Brussels, Belgium, May 2025 – *submitted*.

[3] G. Hendriksen *et al.,* "The Open Web Index. Crawling and Indexing the Web for Public Use." in *Advances in Information Retrieval: 46th European Conference on Information Retrieval. ECIR 2024,* Glasgow. UK. Mar. 2024. pp. 130-143.

[4] L. Beer *et. al.*, "Ein offener Webindex: Chancen und Risiken für Grundrechte und Grundfreiheiten", in Tagungsband Plattform Privatheit 2024, Berlin, Germany, Oct. 2024 – *forthcoming*.

[5] European Commission, COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS State of the Digital Decade 2024, COM(2024) 260 final.

[6] PriDI, https://pridi-projekt.de/.

[7] D. Nowakowski, N. Zimmermann, "Market potential assessment of OpenWebSearch.eu - Exploring the economic and societal impact of an Open Web Index", Mücke/Roth, Germany, 2024.

[8] H. Koulani, L. Beer, P. C. Johannes, C. Geminn, M. Söllner and S. Voigt, "Anwendungsfälle eines offenen Webindex", PriDI, Mar. 2025, published as white paper on https://pridi-projekt.de/.

[9] C. Geminn and P. C. Johannes (Eds.), *Handbuch Europäisches Datenrecht*, Baden-Baden, Germany: Nomos, 2025, to be published.

[10] C. Geminn, K. Erenli and L. Pfeiffer, "Legal Challenges of an Open Web Index", International Cybersecurity Law Review, vol. 2, no. 2, pp. 183-94, 2021. https://doi.org/10.1365/s43439-021-00017-8.

[11] C. Geminn, "Rechtsfragen eines offenen Web-Index - Infrastrukturen für die digitale Gesellschaft", Multimedia und Recht, no. 12, pp. 16-19, 2021.

[12] M. Nebel and P. C. Johannes, "Open Web Index im Lichte des Digital Services Act – Voraussetzungen – Grenzen – Rechtsfolge", Multimedia und Recht, no. 12, pp. 1010-1016, 2024.

[13] F. D. Davis, "User acceptance of information technology: system characteristics, user perceptions and behavioral impacts," International Journal of Man-Machine Studies, vol. 38, no. 3, pp. 475–487, Mar. 1993, doi: 10.1006/imms.1993.1022.

[14] V. Venkatesh, M. Morris, G. Davis, and F. D. Davis, "User acceptance of information Technology: toward a unified view," *MIS Quarterly*, vol. 27, no. 3, p. 425, Jan. 2003, doi: 10.2307/30036540.

[15] V. Venkatesh and F. D. Davis, "A theoretical extension of the technology acceptance model: four longitudinal field studies," Management Science, vol. 46, no. 2, pp. 186–204, 2000, [Online]. Available: https://www.jstor.org/stable/2634758

[16] V. Venkatesh and H. Bala, "Technology Acceptance Model 3 and a research Agenda on interventions," Decision Sciences, vol. 39, no. 2, pp. 273–315, May 2008, doi: 10.1111/j.1540-5915.2008.00192.x.

https://doi.org/10.5281/zenodo.17229609

[17] V. Venkatesh, J. Thong, and X. Xu, “Consumer Acceptance and use of Information technology: Extending the unified theory of acceptance and use of technology,” MIS Quarterly, vol. 36, no. 1, p. 157, Jan. 2012, doi: 10.2307/41410412.

[18] M. Söllner, A. Hoffmann, and J. M. Leimeister, “Why different trust relationships matter for information systems users,” European Journal of Information Systems, vol. 25, no. 3, pp. 274–287, Dec. 2015, doi: 10.1057/ejis.2015.17.

[19] D. Gefen, E. Karahanna, and D. Straub, “Trust and TAM in online shopping: an integrated model,” MIS Quarterly, vol. 27, no. 1, p. 51, Jan. 2003, doi: 10.2307/30036519.

https://doi.org/10.5281/zenodo.17229684

LEGAL ASPECTS OF AI TRAINING AND RETRIEVAL AUGMENTED GENERATION*

L. Beer[†], Open Search Foundation e.V., Munich, Germany
P. C. Johannes[§], H. Koulani[¶], ITeG, University of Kassel, Germany

Abstract

This paper explores legal aspects of the development of the Open Web Index (OWI), a publicly funded European initiative designed as an alternative to proprietary web indexes. It examines how the OWI supports the training of Large Language Models (LLMs) and enhances Retrieval Augmented Generation (RAG) systems. The discussion covers the OWI's architecture, including its distributed crawling and indexing methods, which allow for the collection of vast amounts of web data. By making high-quality, accessible data available, this open infrastructure could benefit smaller companies and research institutions that might otherwise struggle to compete with larger players. The paper delves into the regulatory landscape within the European Union, particularly in relation to the AI-Act and copyright law. It considers the legal challenges surrounding the OWI's use in LLM training and RAG, emphasizing the importance of data quality, legal compliance, and public trust. The conclusion highlights key areas for future research, including the need to clarify frameworks for rights of use, consent and processing authorisation for data to addresses legal uncertainties.

INTRODUCTION

Artificial intelligence (AI) has become widely adopted in a variety of areas at lightning speed and has become an integral element of science, business and society. RAG, i.e. the combination of search and a generative component, is no longer a term that only experts understand. In fact, there are numerous RAG systems on the market that are used by millions of people in the EU and elsewhere every day [1].

At the same time, the European Union (EU) has adopted a large number of regulations and directives in the area of digital governance that impose numerous obligations on the developers and users of such applications. Further legislation is also planned for the future to ensure fairness in the digital space and guarantee the competitiveness of European companies. In his report for the EU Commission on the future of European competitiveness, *Mario Draghi* brings up the problem of extensive regulation: "innovative companies that want to scale up in Europe are hindered at every stage by inconsistent and restrictive regulations" [2]. Deregulation is often a part of the demands by economic interest groups as the many legal requirements are difficult to keep track of, especially for small and medium-sized companies, and therefore might hinder innovation.

The PriDI (Privacy enhancing digital infrastructures) research project, has set itself the goal of making the complex digital legislation at national and EU level easy to understand for developers of digital applications [3]. To this end, the consortium of the University of Kassel and the Open Search Foundation is analysing user-related and legal requirements for the training of LLM's and RAG systems if these applications are based on data from the so-called OWI. Such an index is currently being developed by the EU-funded research project OpenWebSearch.EU [4].

This paper briefly describes the creation and curation of the OWI. It then provides key information on the processes of LLM training and the operation of RAG systems. The authors then shed light on the legal framework within the EU in which the developers of such applications operate. A particular focus is put on the obligations arising from the regulation of AI and the intellectual property rights of the owners of website content. In addition, the paper analyses requirements that arise from the user's perspective with regard to trust and acceptance of AI systems that are based on the OWI. It concludes with an outlook on further research work.

THE OPEN WEB INDEX

Similar to the Google or Bing indexes, the Open Web Index is being created by systematically crawling the web, analysing the crawled content and storing it with metadata in a database [5]. The OWI is intended to strengthen the EU's digital sovereignty by reducing dependence on the search engine monopolists through a sustainable, freely accessible web index.

The researchers have set up a distributed crawling, indexing and hosting architecture for the OWI. This consists of the combination of a frontier crawler, that basically charts the web along embedded links and collects URLs, and distributed worker crawlers, that later on fetch the websites and store the content in so called web archive (WARC) files. Later on, the "raw" web data is further processed, cleaned, filtered, enriched with metadata, classified according to

* Based on research of project Privacy-enhancing digital infrastructures (PriDI), funded by the German Federal Ministry of Education and Research (BMBF). The responsibility for the content of this publication lies with the authors. Parts of tthe findings presented in this paper have already been submitted for publication to CPDP.AI 2025.

[†] lb@opensearchfoundation.org
[§] paul.johannes@uni-kassel.de
[¶] koulani@uni-kassel.de

https://doi.org/10.5281/zenodo.17229684

language and web genre and stored as web index charts following the common index file format (CIFF) and additional metadata sets.

The system is designed in a way that it federates storage and computing capacities across several high performance computing centres across Europe and can be dynamically extended with additional computing centres being added to the federation. To access the index, providers of LLMs and RAG systems or other scientific users of the OWI can authenticate themselves via a public system and can access and retrieve parts of the index via a command line tool. Currently the web data is made available under a research licence, but the research team is also working to grant access to the system for commercial purposes.

The public accessibility of the index is intended to strengthen freedom in internet searches and to form a basis for innovations in science and economy. The researchers have now crawled around 2.23 billion URLs in 185 different languages. The Open Web Index currently has a volume of around 14 TB and is already available to interested developers for initial tests. However, Google's index with a volume of around 100.000 TB is much larger as it includes also thumbnails and other data, whereas the OWI currently includes text data only [6].

ACTORS

When assessing the Open Web Index and its use cases for LLM training and RAG from a legal and user acceptance perspective, a distinction must be made between different stakeholders. Firstly, there are the data subjects. This role describes persons or companies whose personal data and intellectual property are stored in the OWI and are used or may be accessed by tools based on the OWI. The index itself is developed and maintained by the OWI developer. The OWI developers have joined together to form an independent legal entity, the operator consortium. The data retrievers or data consumers of the information contained in the OWI are referred to as application developers. They are persons or systems that request the retrieval of web data from the index in order to create and develop various tools and models based on the retrieved data. They can be individuals, organisations, companies, public institutions or start-ups that use OWI's open data to develop their own applications and services. Finally, end users also come into contact with the index. They are the natural or legal persons who use the tools and systems developed by the application developers.

USE CASES OF THE OWI

The Open Web Index can be used in various ways , e.g. as a basis for search engines (see [7] and [8] for more details). This paper focuses on the use of the index's data to train AI-Systems, i.e. the training of LLM and the development of RAG systems.

Training of LLMs

For the training of LLMs, comprehensive and high-quality pre-filtered web data are essential. A common example of such an LLM is Mistral Large 2 by the French company Mistral AI, on which the company's chatbot is based. These models are trained with large amounts of text data, which mainly come from online sources. With the help of machine learning using neural networks and deep learning methods, LLMs learn to recognise statistical relationships between words and sentences in order to understand and generate texts.

The OWI contributes to the development of new LLMs by providing smaller companies with a sufficient amount of training data at low cost. By offering an open and transparent alternative to proprietary datasets, the OWI enables start-ups and research institutions to train their own models without relying on a few dominant players in the field. This not only fosters innovation and diversity in AI development but also promotes fair competition. As a result, smaller companies can develop high-quality products that can compete with the leading language models currently available on the market.

Retrieval Augmented Generation

RAG is an advanced approach in AI that enhances text generation models by integrating an information retrieval component. This method combines the generative capabilities of language models with the precision of retrieving relevant data from external knowledge sources such as databases, documents, or the web.

The first key component of RAG is the language model, which is trained on vast amounts of text data to understand language and generate coherent responses. This model serves as the foundation for answering queries and can be trained with OWI data. The second component, retrieval, dynamically accesses a knowledge base to fetch relevant information in real time. The OWI can serve as such knowledge base. By combining these two elements, RAG enables AI systems to produce responses that are not only contextually accurate but also based on the most current available data.

This approach is particularly valuable in areas where precise and up-to-date information is crucial, such as customer support, medical consultation, legal research, and knowledge management. By overcoming the limitations of static training data, RAG ensures that AI-driven solutions remain relevant and reliable, even in rapidly evolving fields.

LEGAL ASPECTS

Relevant legislation

The complexity of the Open Web Index raises many questions as to how the index and its applications fall

https://doi.org/10.5281/zenodo.17229684

under Union and member state law. European law on data and online services has undergone major changes in recent years. Where initially mainly the General Data Protection Regulation (GDPR, Regulation (EU) 2016/679) laid down detailed rules on the handling of personal data, now a network of more or less specialized, directly applicable legal acts has emerged. The Digital Services Act (DSA, Regulation (EU) 2022/2065), the Data Governance Act (DGA, Regulation (EU) 2022/868), the Data Act (DA, Regulation (EU) 2022/868), the Digital Markets Act (DMA, Regulation (EU) 2022/1925) and the AI Act (AIA, Regulation (EU) 2024/1689) form a legal framework for digital services and business models [9]. These regulations are directly and uniformly applicable in all member states.

At the same time, there is a harmonized copyright law framework in the European Union. It is primarily governed by a combination of EU Directives, international treaties, and national laws of member states. While the EU aims to harmonize copyright laws across its member states, variations still exist at the national level.

This article focuses on legal challenges faced by the OWI developer and the LLM and RAG application developers across the regulatory domains AI regulation and copyright. The initial assessment of the use case highlights the complexities and the need for ongoing studies.

Regulation of AI-Systems

The AIA aims to regulate systems and practices in the field of artificial intelligence. It was adopted in order to create a robust and flexible legal framework that makes the use of AI and automated decision-making systems trustworthy and secure. The AIA introduces a uniform framework for AI systems based on a risk-based approach, see Recital 26 AIA. AI systems are in Article 3 No. 1 defined as a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. The higher the risk, the more substantial are the obligations put on operators (see Article 3 No. 8 AIA for definition) of AI systems. AI systems with unacceptable risks, e.g. systems that allow “social scoring” by governments or companies, are considered a clear threat to people's fundamental rights and are therefore banned pursuant to Article 5 AIA. To address their specific transparency risk, AI systems like chatbots must clearly inform users that they are interacting with a machine, while certain AI-generated content must be labelled as such, see Article 50 AIA. Only a few AI systems with limited risk face no obligation under the AIA. High-risk AI systems according to Article 6 AIA and Annex I and II AIA on the other hand, such as AI-based medical software or AI systems used for recruitment, must comply with strict requirements, including risk-mitigation systems, high-quality of data sets, clear user information, human oversight.

Direct applicability of AIA The developer of the OWI would have to examine to what extent the provisions of the AIA directly apply to the technologies used to facilitate the OWI. It’s plausible that the algorithms utilized by the OWI to assist and coordinate web crawling will be categorized as minimal risk, as they primarily focus on internal operations. However, to conclusively establish this, a thorough risk assessment is required. Following the risk-based approach, AI systems “that may have a significant adverse impact on the health, safety and fundamental rights of persons” (Recital 46 AIA) are classified as high-risk AI systems in Article 6 AIA, whereby a distinction is made between high-risk AI systems in connection with product regulation (para. 1) and stand-alone high-risk AI systems (para. 2).

AI systems that are safety components of products (Article 3 No. 14 AIA) or are themselves products covered by the harmonization legislation in Annex II (e.g. machinery, toys, elevators, radio equipment, cableways, medical devices, motor vehicles and aircraft) are deemed high-risk systems. The OWI would probably not be classified as a high risk systems pursuant to Article 6 para. 1 AIA, since it is not used as a safety component covered by Annex I of the AIA.

Article 6 of the AIA also designates high-risk AI systems as those enumerated in Annex III. This includes AI systems in biometrics, critical infrastructures, education, employment, basic services, law enforcement, migration, asylum and border control, as well as the administration of justice and democratic processes. AI used in indexing, as well as determining the exclusion or inclusion of certain website content, generally have tangible external impacts on the index usage by third parties. However, it remains unlikely that these operations, or the entirety of the index, might be classified under one of the sectors specified in Annex III, with the exception of the "critical infrastructures" listed as No. 2.

Indirect applicability Furthermore, it must be asked if the AIA contains regulations that influence the use of OWI data for specific AI applications. For example, certain AI systems are banned under Article 5 AIA. The OWI developer might therefore seek to prohibit the use of its index for the training of such AI systems with unacceptable risks. This could be achieved by means of the index licence or terms of conditions for using the OWI as a service. The OWI operator would have certain leeway, since the prohibition clause does not prevent scientific research into the use of AI system merely capable of prohibited practices. Furthermore, according to Article 2 para. 6 AIA, the regulation does not apply to AI systems or AI models, including their outputs, which are developed

https://doi.org/10.5281/zenodo.17229684

and put into operation for the sole purpose of scientific research and development, see Article 3 No. 11 AIA. This is intended to promote innovation and protect scientific freedom, see Recital 25 AIA. In accordance with Article 13 Charter of Fundamental Rights of the European Union, scientific research includes activities with the aim of “gaining new knowledge in a methodical, systematic and verifiable manner”. This includes basic research and applied research in the public (e.g. universities) and private (e.g. industrial research) sectors. Development includes the application and implementation of the knowledge gained through research. Still, the exception is to be interpreted narrowly in terms of wording and well as meaning and purpose.

Another example would be, that pursuant to Article 10 para. 2-5 AIA in conjunction with Article 10 para. 1 AIA high-risk AI systems must be developed with training, validation and test data sets that meet the certain quality criteria. Article 10 para. 3 of the AIA stipulates that training, validation and test data sets must be relevant, sufficiently representative and, as far as possible, error-free and complete with regard to the intended purpose. Among other things, it is questionable whether legally erroneous data (e.g. data obtained in violation of data protection or copyright law) or data anonymized or pseudonymized for data protection reasons (e.g. due to added noise) can still be considered error-free and complete [10]. The data records must also have the appropriate statistical characteristics, if necessary also with regard to the persons or groups of persons for whom the high-risk AI system is to be used as intended. OWI data provides a massive and diverse source of information, including text and links. This diversity is crucial for training AI models. In order to be usable under the quality criteria for high-risk AI systems, the OWI developer should use specific techniques to ensure data quality, e.g. data cleaning, augmentation, balancing or annotation. At the same time it could try to create its datasets in a way, subsequent application developers could use or build on to ensure data quality for their specific use case.

Copyright

Within the European Union, intellectual property rights are mainly determined by European law, but are implemented mostly at national level. For copyright law, which is particularly relevant in the context of LLM training and RAG, the EU legislator has adopted provisions in the Copyright Directive (2001/29/EC) and the Directive on Copyright in the Digital Single Market (2019/790), which have been implemented into national law by the member states. In the following, the legislation in Germany (mainly the German Act on Copyright and Related Rights – UrhG) is taken as an example.

The UrhG defines the extent to which copyright-protected content may be indexed in the OWI and used by applications based on the index. The OWI contains a large amount of data, most of which are protected works under Section 2 UrhG. These works are reproduced regularly as part of their inclusion in the index. However, the right of reproduction defined in Section 16 UrhG is, in principle, granted to the author of the work and not to the OWI or application developers in accordance with Section 15 para. 1 No. 1 UrhG. Copyright-inducing, at least temporary, reproductions cannot be avoided when creating the index and training LLMs with the index data.

However, the Copyright Act contains various exceptions that can justify acts of reproduction. In 2021, the German legislator created the exception rule of Section 44b UrhG for general text and data mining in implementing Article 4 of the Directive No. 2019/790. This is designed to make it possible to analyse large amounts of digital information [11]. According to the legal academia [12-20], the training of AI models can usually be justified by Section 44b UrhG and case law also shows a slight tendency in this direction [21]. However, any reservations of the creator pursuant to Section 44b para. 3 UrhG must be taken into account. If the rights holder opposes to the use of their website content for LLM training or RAG development, the application and OWI developers must adhere to the content owner’s reservations.

In the case of acts of reproduction created by crawlers during the indexing of web content, Section 44a UrhG also comes into question. In this, the legislator provides for an exception for acts of reproduction that are only of a temporary nature and part of a technical process, have no independent economic significance and serve a purpose of Section 44a UrhG.

USER ACCEPTANCE AND TRUST

While legal compliance, data protection, and privacy are highly valued in Europe, user behaviour suggests that these factors often take a backseat when choosing digital services. In practice, other aspects tend to have a stronger influence on users to adapt new technologies. Many research studies focused on investigating concepts and aspects of user perception and their intention to use and trust technologies [22-26]. Therefore, incorporating key elements of user acceptance into the development of OWI and related tools is essential. To apply user acceptance and trust principles effectively to OWI, our study combines insights from Trust-TAM [26] and UTAUT2 [24]. The latter identifies seven key factors—such as performance and effort expectancy, social influence, and facilitating conditions—all of which we aim to adapt to the OWI framework. In terms of establishing trust with developers as actors in these use cases, knowledge-based trust could be conceived introducing familiarity as an antecedent of this trust. Supporting the developers’ familiarity with the structure and functionalities of OWI promote their confidence and

https://doi.org/10.5281/zenodo.17229684

trust in the interaction with OWI. Particularly, this familiarity could be implemented in terms of providing web data for LLM training in a format that is standard in these contexts and thus reducing cognitive load required to acquire web data from OWI for the discussed matters. This adaptation is still in its early stages and represents a work in progress, with the mentioned ideas serving as an initial foundation for further development.

FUTURE WORK

Regardless of the previously conducted analysis of the legal framework, some issues in the context of the use of the OWI for LLM training and the development of RAG are still open. In future works, the interaction, including contractual relationships, between the developer of the OWI and the developer of LLMs and RAG would need to be clarified. Any legal loopholes or unintended consequences d*e lege lata* should be addressed by further developing the law, either on the Union level, or where possible, on the national level. In this context, the focus should be on creating simple and concise provisions that are easy for developers to implement. The PriDI project will also focus on this in its future work and specify the legal requirements outlined above in the form of requirement and design patterns.

The OWI can also be used to train other web data-based AI applications, such as knowledge representation and reasoning (KRR) systems. In contrast to LLMs, which use statistics to produce texts, KRR systems represent information in a way that a computer can understand it and solve complex problems like a human. The aim is to create intelligent machines that learn from human knowledge and act in the same way. KRR systems are used, for example, in quality management to monitor product quality or to prevent fraud in the insurance industry. Future publications will need to consider whether the above legal requirements also apply to KRR systems.

CONCLUSION

The OWI data can be used for the training of LLMs and RAG development, since its data would be comprehensive and high-quality, pre-filtered web data. While the OWI operator conceivably would not fall under the AIA, the LLM and RAG developer most likely would. Depending on its specific use case, the LLM or RAG system could even be classified as a high-risk AI system. Either way, the quality of the provided data as well as the legality of its content and its provision are paramount. Even if the AIA would not be applicable for the specific use case (e.g. because of Article 2 para. 6 or 12 AIA), data protection law as well as copyright law most certainly still would.

In regards to data protection law, the OWI developer would have to make sure that it is allowed to share or publish the personal data it has collected. The LLM and RAG developer would have to make sure, that it is allowed to process the personal data on the basis of one of the authorisations in Article 6 GDPR. For example: The AI or RAG developer most likely would be allowed to process publicly available personal data of persons linked to a business under Article 6 para. 1 subpara. 1 lit. f GDPR.

Likewise, in regards to copyright, the LLM and RAG developer would still have to make sure, it could use the provided data. Ideally it could rely on a legal limitation to the copyright, like Section 44b UrhG.

REFERENCES

[1] According to DemandSage, `https://www.demandsage.com/perplexity-ai-statistics/` the RAG-systems Perplexity AI has about 2 million daily active users.

[2] European Commission, `https://commission.europa.eu/topics/eu-competitiveness/draghi-report_en`, p. 6.

[3] For ore information on the research project see its website `https://pridi-projekt.de/home-en/`.

[4] For more information on the OWS.EU project see its website `https://openwebsearch.eu/`.

[5] More details on the creation and operation of the OWI can be found in G. Hendriksen *et al.*, “The Open Web Index. Crawling and Indexing the Web for Public Use”, in *Advances in Information Retrieval: 46th European Conf. on Information Retrieval. ECIR 2024*, Glasgow. UK. March 2024. pp. 130-143.

[6] Google, `https://www.google.com/intl/en_us/search/howsearchworks/how-search-works/organizing-information/`.

[7] Several other applications are listed in D. Nowakowski, N. Zimmermann and L. Kerner, “Market potential assessment of OpenWebSearch.eu - Exploring the economic and societal impact of an Open Web Index”, Mücke/Roth, Germany, Rep., 2024, `https://openwebsearch.eu/wp-content/uploads/2024/09/MarketAssessmentOfOWI-Report-V1.pdf`

[8] P. C. Johannes, L. Beer and H. Koulani, “Legal challenges of using the OWI for search engines”, presented at OSSYM 2025 - 7th Int. Open Search Symposium, Helsinki, Finland, Oct. 2025, paper #####, this conference, submitted.

[9] Also see C. Geminn and P. C. Johannes, *Handbuch europäisches Datenrecht.* Baden-Baden, Germany: Nomos, in preparation.

[10] I. Vogel *et al*, “Natural Language Processing (NLP) und der Datenschutz - Chancen und Risiken für den Schutz der Privatheit”, in *Informatik 2022. Lecture Notes in Informatics (LNI), p.* 659. `doi: 10.18420/inf2022_24`

[11] German Bundestag `https://dserver.bundestag.de/btd/19/274/1927426.pdf`

[12] E.g. D. Bomhard and J. Siglmüller, “AI Act - das

https://doi.org/10.5281/zenodo.17229684

Trilogergebnis", *Recht Digital*, vol. 5, no. 2, p. 50, 2024.

[13] M. Dregelies, "KI-Training unter dem AI Act", *Gewerblicher Rechtsschutz und Urheberrecht*, vol. 126, no. 20, p. 1484, 2024.

[14] R. Heine, "Generative KI: Nutzungsrechte und Nutzungsvorbehalt", *Gewerblicher Rechtssschutz und Urheberrecht in der Praxis*, vol. 16, no. 4, p. 88, 2024.

[15] F. Hofmann, "Retten Schranken Geschäftsmodelle generativer KI-Systeme?", *Zeitschrift für Urheber- und Medienrecht*, vol. 68, no. 3, p. 166, 2024.

[16] L. Kaede, "Training generativer KI-Modelle ist (auch) Text- und Data-Mining. Anwendbarkeit der TDM-Schranke des § 44b UrhG", *Künstliche Intelligenz und Recht*, vol. 1, no. 5, p. 162, 2024.

[17] N. Maamar, "Urheberrechtliche Fragen beim Einsatz von generativen KI-Systemen", *Zeitschrift für Urheber- und Medienrecht*, vol. 67, no. 7, p. 481, 2023.

[18] K. Wagner, "Generative KI: Eine "Blackbox" urheberrechtlicher Haftungsrisiken?. Balanceakt zwischen Innovationsförderung und effektivem Rechtsschutz für Werke Dritter", *Zeitschrift für IT-Recht und Recht der Digitalisierung*, vol. 27, no. 4, p. 298, 2024.

[19] Contrary opinion: T. W. Dornis and S. Stober, *Urheberrecht und Training generativer KI-Modelle: Technologische und juristische Grundlagen.* Baden-Baden, Germany: Nomos 2024. doi: 10.5771/9783748949558

[20] Contrary opinion: T. W. Dornis, "Generatives KI-Training und Text- und Data-Mining. Eine funktionale Unterscheidung", *Künstliche Intelligenz und Recht*, vol. 1, no. 5, p. 156, 2024.

[21] LG Hamburg, Judgement of 27 September 2024, Reference 310 O 227/23, https://www.itm.nrw/wp-content/uploads/2024/09/2495651-en.pdf

[22] F. D. Davis, "User acceptance of information technology: system characteristics, user perceptions and behavioral impacts" *International Journal of Man-Machine Studies*, vol. 38, no. 3, pp. 475–487, Mar. 1993. doi: 10.1006/imms.1993.1022

[23] V. Venkatesh, M. Morris, G. Davis, and F. D. Davis, "User acceptance of information Technology: toward a unified view," *MIS Quarterly*, vol. 27, no. 3, p. 425, Jan. 2003. doi: 10.2307/30036540

[24] V. Venkatesh, J. Thong, and X. Xu, "Consumer Acceptance and use of Information technology: Extending the unified theory of acceptance and use of technology" *MIS Quarterly*, vol. 36, no. 1, p. 157, Jan. 2012. doi: 10.2307/41410412

[25] M. Söllner, A. Hoffmann, and J. M. Leimeister, "Why different trust relationships matter for information systems users" *European Journal of Information Systems*, vol. 25, no. 3, pp. 274–287, Dec. 2015. doi: 10.1057/ejis.2015.17

[26] D. Gefen, E. Karahanna, and D. Straub, "Trust and TAM in online shopping: an integrated model" *MIS Quarterly*, vol. 27, no. 1, p. 51, Jan. 2003. doi: 10.2307/30036519

https://doi.org/10.5281/zenodo.17229737

IN-BROWSER AGENTIC WEB: A DECENTRALIZED APPROACH TO INFORMATION ACCESS

S. Zerhoudi, M. Granitzer, University of Passau, Passau, Germany

Abstract

The centralization of web search raises critical concerns regarding privacy protection and user autonomy in information access. While advancements in web technologies offer new possibilities for personal information management, current search systems typically process user data on external servers with limited personalization options. This paper introduces a conceptual methodology for browser-based web indexing that processes and stores data locally, addressing these privacy and control limitations. Our approach implements targeted crawling mechanisms aligned with individual user interests and maintains all operations within the browser environment. The technical framework converts web content into dense vector representations through semantic embedding techniques, enabling efficient storage and retrieval within browser constraints. The architecture features: (1) an in-browser language model for semantic search and context-aware content generation, (2) adaptive crawling algorithms that adjust parameters based on storage limitations and user behavior, and (3) incremental updating mechanisms to maintain index freshness. Evaluation approaches using both simulation-based assessment and human participant validation are proposed. This work encourages research on privacy-preserving web search technologies and establishes a foundation for developing user-controlled information retrieval systems.

INTRODUCTION

The digital ecosystem's rapid growth in web content creates both opportunities and challenges for information retrieval. Current web search services, controlled by a few major corporations like Google and Microsoft, typically employ user tracking, centralized indexing, and undisclosed algorithmic methods, raising concerns about privacy, data sovereignty, algorithmic transparency, and offline access [Granitzer et al.(2024), Hendriksen et al.(2024a)].

These centralized search providers rely on collecting and analyzing user data to improve search relevance and advertising revenue, which raises ethical questions regarding privacy and control. Their algorithmic processes often lack transparency, potentially enabling manipulation or biased results influenced by commercial or political factors [Granitzer et al.(2024)]. This opacity undermines user trust and may compromise information reliability. Additionally, the market dominance of a few search providers has also led to practices like Search Engine Optimization (SEO), where content creators prioritize algorithmic visibility over informational quality and user value.

In response to these limitations, research interest has shifted to decentralized, transparent, and privacy-conscious alternatives. The Open Web Index (OWI) initiative promotes openly accessible indexing infrastructures and standards, emphasizing transparency, collaboration, and open data principles [Hendriksen et al.(2024b)]. OWI addresses centralized indexing challenges by creating public data structures that democratize search engine development. This project employs extensive indexing operations supported by high-performance computing (HPC) resources across Europe, aiming to diversify the digital information ecosystem.

Concurrent with these large-scale efforts, advances in browser-based AI inference technologies have created new possibilities for privacy-focused and personalized web indexing. Recent developments, such as WebLLM [Ruan et al.(2024)], demonstrate the feasibility of running sophisticated AI models directly within browsers. These technologies leverage WebGPU [Kenwright(2022)] and WebAssembly [Haas et al.(2017)] to enable efficient local processing without external cloud services. By processing data locally, these browser-based approaches inherently enhance privacy and user autonomy.

Adaptive web crawling techniques driven by semantic modeling of user interests have emerged as essential components for personalized retrieval [Durga et al.(2024)]. Unlike traditional fixed crawling algorithms, user modeling approaches that adapt to browsing patterns and real-time interactions improve retrieval accuracy and relevance. This adaptive methodology ensures content remains specific and current, enhancing user experience.

Our research proposes an in-browser web indexing approach that integrates targeted, adaptive crawling and content acquisition based on user-defined interests, local indexing using compressed vector embeddings, and semantic search powered by browser-based language models. This methodology addresses limitations of centralized systems by prioritizing privacy, personalization, offline functionality, and user control.

The approach centers on creating a localized, browser-contained semantic index using compressed dense embeddings, providing contextual understanding beyond keyword-based techniques. This allows the system to deliver personalized search results within the user's local environment. Our work extends principles from the OWI initiative but adapts them to browser environments. Rather than employing collaborative indexing at scale, our approach focuses on localized data organization, efficient embedding methods, and streamlined inference capabilities suitable for resource-limited personal computing contexts.

Our contributions include: (1) proposing a conceptual design for a decentralized, privacy-preserving browser-based web indexing approach that addresses privacy, autonomy, and offline access challenges; (2) defining theoretical adap-

https://doi.org/10.5281/zenodo.17229737

tive crawling and content acquisition methods based on semantic user-interest models that align content retrieval with preferences; (3) outlining efficient semantic embedding techniques optimized for browser-based storage and computation constraints; and (4) describing potential integration of browser-based language model capabilities supporting semantic search and retrieval-augmented generation for personalized content.

RELATED WORK

Browser technologies have advanced substantially from basic rendering to sophisticated local computation capabilities. Extending WebLLM's work [Ruan et al.(2024)], researchers have further optimized on-device language model inference, reducing memory requirements and improving execution speed. These technical advances complement privacy-enhancing technologies research, where [Kumar et al.(2025)] developed frameworks for evaluating privacy preservation in AI applications without functionality compromises.

Vector space representation of web content has enhanced information retrieval beyond keyword matching. Recent embedding techniques capture semantic relationships and contextual nuances that keyword approaches cannot address. Embedding compression methods have reduced storage requirements by up to 75% while maintaining 90% of semantic integrity [Li et al.(2024)]. These efficiency improvements are particularly valuable for browser environments with storage constraints. Research shows that optimized quantization and dimension reduction techniques maintain retrieval quality while reducing computational demands, balancing semantic precision with resource limitations.

Adaptive crawling methodologies have proven effective beyond basic personalization. Building on [Durga et al.(2024)]'s user modeling work, subsequent studies have measured benefits showing up to 40% improvement in content relevance through dynamic crawling parameter adjustment. These approaches combine user interaction signals with content classification to create refined interest models. By analyzing content consumption patterns, dwell time, and explicit preferences, these systems develop accurate representations of user information needs that evolve over time. This adaptability particularly benefits specialized knowledge domains where standard crawling often misses relevant but less-connected content.

Distributed indexing system architecture has evolved beyond simple centralized/decentralized divisions. The European OpenWebSearch.eu [1] project demonstrates how federated approaches can distribute computational workloads while maintaining consistent access patterns. Their federated storage approach separates crawling, indexing, and retrieval components, allowing specific optimization of each element [Granitzer et al.(2025)]. This architectural pattern informs our browser-based approach, though we adapt these principles to operate entirely within the client environment.

Content freshness maintenance in limited-resource environments represents another relevant research direction. Traditional search engines use continuous crawling with extensive server infrastructure, but resource-constrained systems require more strategic approaches. Recent research shows that selective recrawling based on content volatility prediction can maintain index freshness with reduced computational requirements [Gossen et al.(2015)]. These predictions use content type, historical update patterns, and domain characteristics to prioritize recrawling for rapidly changing content while conserving resources for stable information.

While these research areas provide valuable foundations, integrating them into a cohesive browser-based indexing system presents unique challenges that remain insufficiently addressed. Current approaches tend to focus on individual components—either optimizing language models [Ruan et al.(2024)], improving vector representations [Li et al.(2024)], enhancing crawling strategies [Durga et al.(2024)], or developing distributed architectures [Hendriksen et al.(2024b)]—without fully considering how these elements interact within browser constraints. Our work synthesizes these advances into a comprehensive framework specifically designed for browser environments, addressing the technical limitations and privacy concerns inherent in centralized search systems. By combining adaptive crawling, efficient semantic indexing, and local retrieval augmentation, we propose a system that balances performance requirements with privacy preservation. The following sections detail our conceptual architecture and operational workflow, demonstrating how these components work together to enable personalized web indexing directly within the browser.

CONCEPTUAL ARCHITECTURE

This section outlines a conceptual approach to browser-based web indexing designed to enhance privacy and personalization. The methodology addresses constraints of centralized search systems through client-side processing, storage, and retrieval techniques that function within web browser limitations while enhancing user control. The methodology enables localized information management that reduces dependency on external search providers while maintaining search functionality.

Content Acquisition

The foundation of effective personalized indexing begins with selective content acquisition based on user interests. Unlike conventional web crawlers that aim for comprehensive coverage, this approach employs targeted crawling to retrieve only content aligned with individual user preferences, thereby reducing storage requirements while enhancing relevance.

The system would construct dynamic user interest profiles through multiple mechanisms. Building on [Durga et al.(2024)]'s user modeling approach, the profile would incorporate both explicit inputs (user-specified topics, do-

[1] `https://openwebsearch.eu/`

https://doi.org/10.5281/zenodo.17229737

mains, and keywords) and implicit signals (browsing patterns, bookmarking behavior, and content interaction histories). These profiles would continuously evolve through adaptive algorithms that detect shifts in interests and adjust accordingly.

Guided by these profiles, the crawling component would assign priority scores to potential URLs based on semantic alignment with user interests. This prioritization mechanism would consider both content similarity to established interests and exploration potential for adjacent topics. The crawler would maintain compliance with web standards and site policies, respecting robots.txt directives and implementing appropriate rate limiting to ensure responsible resource utilization.

Beyond crawling methods, the system would offer alternative content acquisition pathways. Users can leverage the OWI Python client (owilix) developed by [Granitzer et al.(2025)], which provides sophisticated dataset management capabilities specifically designed for OWI environments. This tool enables efficient pushing and pulling of datasets and supports remote SQL query execution, allowing users to retrieve daily index slices precisely tailored to their interests without the overhead of full crawling operations.

For users with private document collections, the system would implement a secure, privacy-preserving ingestion pipeline. This process begins with the secure parsing of personal documents stored in a self-hosted cloud solution, extracting valuable textual content and metadata. The extracted information is then normalized and loaded into DuckDB [Raasveldt and Mühleisen(2019)], a lightweight analytical database deployed within the user's private infrastructure. This embedded database efficiently indexes the content, creating optimized structures for rapid querying. To enable seamless integration with client-side applications, the indexed content can be exported from DuckDB in JSON or similar serializable formats and imported into a compressed browser database. This final step bridges server-side indexing with client-side storage, providing users with efficient offline search capabilities while maintaining end-to-end privacy protection throughout the entire pipeline.

Semantic Indexing

Once content is acquired, the system would transform it into optimized representations suitable for browser-based storage and retrieval. The primary mechanism for this transformation would be dense vector embeddings that capture semantic relationships between content items beyond simple keyword matching.

These embeddings would map textual content into multidimensional semantic spaces where proximity indicates conceptual similarity. Drawing inspiration from techniques described by [Li et al.(2024)], the system would generate embeddings at multiple granularity levels, from document-wide representations to sentence-level encodings. A key feature would be adjustable dimensionality, allowing dynamic balancing between semantic precision and storage efficiency. This adaptability would enable the system to operate effectively across devices with varying resource constraints.

The processed content would reside in compressed browser databases utilizing technologies like IndexedDB [Al-Shaikh and Sleit(2017)]. To maximize storage efficiency within browser constraints, the system would implement structured data partitioning inspired by larger-scale approaches from the Open Web Index initiative [Granitzer et al.(2025)]. Content would be organized into logical segments based on source domains, temporal factors, and thematic categories, enabling efficient query processing. Additionally, metadata elements such as titles, content acquisition dates, and language indicators would be integrated directly alongside semantic representations to facilitate rapid filtering and result refinement during retrieval operations.

Interactive Retrieval

The retrieval process would begin with query encoding, transforming user information needs into the same semantic vector space used for content representation. These query embeddings would then undergo similarity comparison against the indexed content using established metrics such as cosine similarity, identifying the most relevant matches from the local database.

Building on recent advances in browser-based AI frameworks demonstrated by WebLLM [Ruan et al.(2024)], the system would incorporate a locally executed language model for advanced retrieval and content synthesis. This model would implement retrieval-augmented generation (RAG) techniques, using the locally indexed content to ground its responses in user-specific information sources. The browser-native execution would leverage technologies like WebGPU [Kenwright(2022)] and WebAssembly [Haas et al.(2017)] to optimize performance within client-side constraints.

User control would remain central to the retrieval process through customizable search parameters. These would include domain-specific weightings (prioritizing preferred sources), temporal filters (focusing on recent or historical content), and adjustable balance between semantic similarity and metadata matching. These customization options would allow users to tailor the system's behavior to specific information-seeking contexts, from exploratory research to targeted fact-finding.

Index Freshness Management

Maintaining relevance over time requires mechanisms for content refresh and index optimization. The proposed system would implement context-aware scheduling for recrawling operations, prioritizing sources based on factors including update frequency, user engagement patterns, and content volatility.

Instead of complete reindexing, the system would employ incremental processing techniques that efficiently integrate new content into existing indices. This approach would minimize computational overhead while ensuring the index remains current. The scheduling mechanism would balance

https://doi.org/10.5281/zenodo.17229737

multiple factors: user preferences, connectivity conditions, and device resource availability, preferentially performing intensive operations during optimal conditions (e.g., during low-activity nighttime hours).

Content pruning strategies would prevent unbounded index growth by identifying and removing outdated or low-relevance items from the database. These decisions would consider multiple signals including recency, access frequency, and semantic redundancy with newer content. This comprehensive maintenance approach would ensure the system remains responsive and resource-efficient over extended usage periods while adapting to evolving user interests.

OPERATIONAL WORKFLOW

This section describes the conceptual workflow and component interactions in the proposed browser-based indexing approach. The design integrates various processes to enable personalized information access while maintaining user privacy and control throughout the operational cycle. Figure 1 shows an overview of the in-browser approach architecture and workflow.

User Modeling Initialization

The proposed system would begin with minimal setup requirements, avoiding intrusive information gathering during initialization. Instead of demanding extensive upfront configuration, the system would gradually build user interest profiles through two complementary mechanisms.

The passive observation component would analyze content from past conversational search activities and pages visited during normal browsing in accordance with user privacy preferences. This lightweight semantic analysis would extract key concepts, entities, and topics without disrupting user experience. The extracted information would populate an initial interest model that evolves over time as the user continues browsing.

Complementing passive observation, the system would provide explicit feedback mechanisms through which users could review, modify, or remove interests identified by the system. These controls would be prominently accessible within the browser extension, ensuring users maintain awareness and control over their interest profiles.

Adaptive Crawling Strategy

Once user interests are established, the content acquisition process would begin. The crawling component would employ a dynamic prioritization mechanism that evaluates potential URLs based on multiple factors: semantic alignment with identified interests, browsing frequency and historical engagement patterns. This prioritization would optimize resource allocation by focusing on content most likely to provide value to the specific user.

To operate effectively within browser constraints, the crawler would implement adaptive resource management techniques. These would include adjustable parameters for concurrent requests, crawling depth, and scheduling frequency based on device capabilities and connection status. During active browsing sessions, the crawler would reduce its activity to minimize impact on performance, while potentially increasing activity during idle periods.

The crawler would respect robots.txt directives, implement appropriate rate limiting, and follow standardized crawling policies. These practices would ensure the system behaves responsibly within the broader web ecosystem while gathering personalized content.

In-Browser Indexing

The indexing process would operate entirely within the browser environment, transforming retrieved content into searchable representations. Content processing would begin with semantic embedding generation, converting textual content into dense vector representations using locally stored or dynamically loaded models. These embeddings would capture semantic relationships between content items, enabling meaning-based rather than keyword-based retrieval.

Following embedding generation, the system would extract and integrate metadata elements including titles, content acquisition dates, source information, and language indicators. This structured approach would enable efficient filtering during search operations. The indexed content would be organized using partitioning strategies inspired by the OWI project [Granitzer et al.(2025)], dividing the index logically by content origin, topical domains, or temporal factors.

To maintain index freshness while minimizing computational demands, the system would implement incremental updating mechanisms. Rather than rebuilding the entire index when new content is acquired, only changes would be processed and integrated. A local changelog would track modifications enabling efficient updates. The system would also employ intelligent pruning algorithms to remove outdated or low-relevance content, preventing unbounded index growth over time.

Retrieval-Augmented Search

When users initiate a search query, the in-browser language model would process the input to understand the information need. The query would be encoded into the same vector space used for content representation, enabling direct comparison between the query and indexed content. The retrieval engine would identify relevant content based on semantic similarity measurements, returning results ranked by relevance to the user's query.

For complex information needs, the system would implement retrieval-augmented generation as described by [Ruan et al.(2024)]. This approach would ground language model outputs in the user's personal index, combining the flexibility of generative AI with the accuracy of retrieved information. By leveraging locally stored content, responses would reflect the user's specific knowledge base rather than generic information.

The search interface would provide interactive refinement options, allowing users to adjust result presentation based on

https://doi.org/10.5281/zenodo.17229737

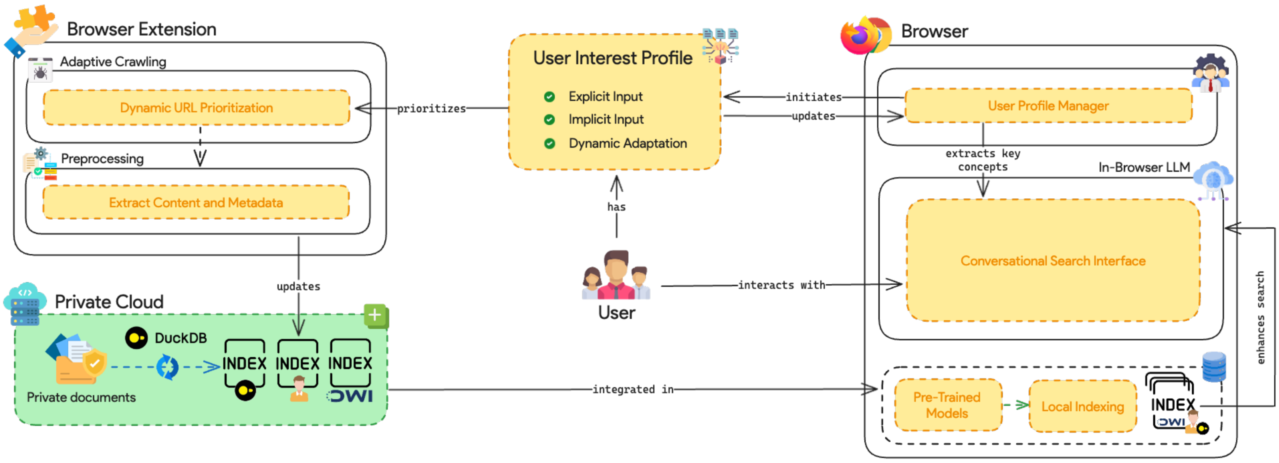


Figure 1: An overview of the In-Browser indexing and personalized content retrieval approach.

various parameters. These adjustments might include source preferences, recency requirements, or topic emphasis. Each interaction would feed back into the system's understanding of user preferences, gradually improving retrieval accuracy through ongoing learning from user behavior patterns.

User-Controlled Privacy

Privacy protection would be fundamental to the system architecture, with all data processing occurring exclusively within the browser environment. This localized approach would ensure sensitive information remains under user control rather than being transmitted to external servers. The design would collect only information necessary for system functionality.

The system would provide comprehensive transparency regarding data usage through an accessible browser extension interface. This interface would display the current interest model, crawling activities, and index contents in user-friendly formats. All aspects of the system would remain user-modifiable, with options to edit, export, or delete any stored information.

Control granularity would extend to operational parameters, allowing users to adjust the balance between personalization depth and resource utilization. Users could configure crawling schedules, storage limitations, and embedding dimensions based on their preferences and device capabilities. This flexibility would enable the system to accommodate diverse usage patterns and hardware constraints while maintaining core functionality.

Through this integrated operational flow, the proposed system would create a self-contained information ecosystem within the browser environment. By combining interest modeling, adaptive content acquisition, semantic indexing, and retrieval-augmented search, it would offer personalized information access while preserving user privacy.

EVALUATION APPROACH

Evaluating a browser-based indexing system presents specific challenges requiring careful methodological planning. This section outlines some possbile research-based approaches to assess such conceptual architectures.

Technical performance evaluation requires adapting standard information retrieval metrics to the browser context. Measures such as precision, recall, and mean reciprocal rank must be applied within personal indexing constraints, where corpus size varies between users and changes over time. These metrics should assess retrieval effectiveness relative to indexed content rather than global repositories. Browser-specific indicators including memory usage, storage efficiency, and interface responsiveness are crucial for evaluating client-side feasibility.

Simulation-based assessment offers valuable insights for conceptual architectures before full implementation. User simulation methods described by [Balog and Zhai(2025)] can be adapted to model various user interests, browsing patterns, and information needs. This enables systematic testing across different user profiles without extensive development resources. By creating synthetic browsing histories and interest profiles, researchers can generate representative personal indexes for testing. Simulated queries with predetermined relevance judgments provide measurable performance metrics while allowing parameter variation.

LLM-based agents, following methods proposed by [Lu et al.(2025)], offer an effective evaluation strategy. These agents can simulate different user personas with varying information needs, technical expertise, and privacy concerns. This facilitates assessment of both technical performance and user experience aspects, including interface usability and perceived utility. While LLM agents cannot completely replicate human behavior, they provide cost-effective initial evaluation before human participant testing.

Scientific validity requires careful **benchmark development**, including curated web content with predefined relevance judgments, standardized browsing profiles, and consistent query sets. Such benchmarks enable reproducible comparisons between implementation approaches and help assess design decisions regarding embedding dimensions, crawling strategies, and index partitioning methods.

https://doi.org/10.5281/zenodo.17229737

Human participant validation remains essential for thorough evaluation. Well-designed user studies employing mixed methods can assess both objective performance and subjective experience. For this purpose, frameworks like SearchLab [Zerhoudi and Granitzer(2025)] offer valuable capabilities as a modular web-based platform specifically designed for search behavior studies. Participants should engage with the system over extended periods to capture realistic usage and allow natural interest profile development. Performance evaluation should combine logged interaction data and structured tasks with defined success criteria. Qualitative methods such as think-aloud protocols, interviews, and usability questionnaires complement quantitative measures by revealing user perceptions. The comprehensive data collection capabilities of SearchLab reduce the need for custom application development, allowing researchers to focus on study design rather than technical implementation.

IMPACT AND RESEARCH DIRECTIONS

The browser-based indexing approach we propose has implications beyond individual search experiences. This section examines potential effects on web ecosystems, user autonomy, and technological synergies, while outlining future research paths.

Web Information Ecosystems

Decentralizing web indexing through personal browser-based systems could alter web information dynamics. Current indexing power concentration among few corporations has created an environment where content discovery is mainly controlled by proprietary algorithms optimized for advertising revenue rather than information diversity. As [Granitzer et al.(2024)] note, this centralization introduces systematic biases that may homogenize content and favor commercial interests.

A distributed approach where users maintain personal indexes could reduce these concentrating effects. Content creators might respond by producing more specialized material for niche audiences instead of optimizing solely for dominant search algorithms. Publishers currently invest in search engine optimization techniques that often prioritize algorithmic visibility over content quality. When discovery becomes more personalized through browser-based indexing, these incentives may shift toward content that serves user interests rather than algorithmic preferences.

The proposed browser-based indexing system would function alongside broader open web initiatives. Users could optionally contribute anonymized, aggregated indexing data (with explicit consent) to collaborative projects like OpenWebSearch.eu [Granitzer et al.(2024)], creating a mutually beneficial relationship between personal and collective indexing efforts. This arrangement could address a limitation of purely personal indexing: reduced content discovery breadth. By voluntarily participating in federated efforts, users could maintain privacy advantages while contributing to and benefiting from collective knowledge organization.

User Autonomy

The architecture we propose improves user control over personal data and search experiences in several ways. By processing and storing data locally, the system removes the external data transfers found in centralized indexing systems. Users would gain protection from external data collection and clarity about what information their system has captured and how it affects their search results.

The adaptive user-interest model offers another aspect of user empowerment. Unlike fixed indexing approaches that treat all users identically, the proposed system would refine its understanding of individual interests through browsing patterns and explicit feedback. This responsiveness allows search results to reflect actual user needs rather than general assumptions or commercial priorities. The system could show users visualizations of their interest profiles, allowing them to adjust or correct misinterpretations, enhancing both control and system accuracy.

Clarity extends beyond data collection to the search process itself. Commercial search engines typically provide minimal insight into result selection for queries. A locally managed index could give users clear explanations of ranking factors, potentially building trust in the system. This clarity could help users develop better search strategies and understand the connection between their browsing behaviors and search outcomes.

Leveraging AI Models

Recent advancements in language models and generative AI create valuable opportunities for browser-based indexing systems. Local language models could improve multiple system aspects, from interest profiling to search query processing. By analyzing content semantics more deeply, these models could build more nuanced representations of user interests than conventional keyword-based approaches. This capability could help the system differentiate between temporary information needs and enduring interests, adjusting crawling priorities accordingly.

The development and evaluation of such systems present distinct challenges that AI could help address. Language models could simulate various user behaviors to test system responsiveness across different usage patterns. While [Lu et al.(2025)] caution about limitations in AI-based simulation, such approaches could still provide useful insights during early development stages. These simulations could help identify weaknesses in crawling strategies or interest modeling before deployment with actual users.

As browser-integrated language models like WebLLM become more capable, the system could implement proactive indexing based on anticipated information needs. The model might identify concepts related to current browsing activities and index relevant content in advance. However, such capabilities raise important questions about resource usage and user consent that would require careful consideration in any implementation.

https://doi.org/10.5281/zenodo.17229737

Technical Challenges

This proposal faces several implementation challenges. Browser memory and processing limitations represent a primary obstacle. Research is required to develop compact vector databases suitable for browser environments. Current embedding methods are typically designed for server environments with greater computational resources, requiring adaptation for client-side use. Techniques such as quantization [Li et al.(2024)] that reduce storage requirements while preserving semantic information could enhance the feasibility of the system.

Adaptive interest models represent another research challenge. User modeling implementations must balance complexity with computational efficiency. Research into incremental model updates could improve user experience and resource use. Incorporating explicit feedback and implicit signals while maintaining model coherence presents a machine learning challenge requiring further study.

As web content spans multiple modalities, research into efficient multimodal indexing becomes crucial. Extending browser-based systems to represent and search across text, images, audio, and video presents technical challenges. Unified embedding spaces that capture cross-modal relationships while remaining compact would advance the field.

Evaluating personalized, decentralized search systems presents methodological challenges. Developing standardized benchmarks that accommodate individual differences while allowing systematic comparison would facilitate progress. Such frameworks need to address search quality, resource efficiency, and user satisfaction.

These technical challenges highlight how browser-based indexing intersects information retrieval, machine learning, and human-computer interaction, requiring solutions that consider social and ethical implications of distributed information access.

ACKNOWLEDGEMENTS

This research was funded by the European Union's Horizon Europe research and innovation program under grant agreement No 101070014 (OpenWebSearch.EU, `https://doi.org/10.3030/101070014`).

REFERENCES

[Al-Shaikh and Sleit(2017)] Ala'a Al-Shaikh and Azzam Sleit. 2017. Evaluating IndexedDB performance on web browsers. In *2017 8th International Conference on Information Technology (ICIT)*. IEEE, 488–494.

[Balog and Zhai(2025)] Krisztian Balog and ChengXiang Zhai. 2025. User Simulation in the Era of Generative AI: User Modeling, Synthetic Data Generation, and System Evaluation. *arXiv preprint arXiv:2501.04410* (2025).

[Durga et al.(2024)] Csl Vijaya Durga, RJ Anandhi, Saloni Bansal, Navdeep Singh, Ravi Kalra, and Nabaa M Bader. 2024. Adaptive Web Crawling Strategies Based on Ontological User Interest Modeling for Personalized Content Retrieval. In *2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies*. IEEE, 1–5.

[Gossen et al.(2015)] Gerhard Gossen, Elena Demidova, and Thomas Risse. 2015. iCrawl: Improving the freshness of web collections by integrating social web and focused web crawling. In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*. 75–84.

[Granitzer et al.(2025)] Michael Granitzer, Mohamad Hayek, Sebastian Heineking, Gijs Hendriksen, Martin Golasowski, Michael Dinzinger, and Saber Zerhoudi. 2025. OpenWebSearch. eu-Building an Open Web Index on EuroHPC JU Infrastructures. *Procedia Computer Science* 255 (2025), 43–52.

[Granitzer et al.(2024)] Michael Granitzer, Stefan Voigt, Noor Afshan Fathima, Martin Golasowski, Christian Guetl, Tobias Hecking, Gijs Hendriksen, Djoerd Hiemstra, Jan Martinovič, Jelena Mitrović, et al. 2024. Impact and development of an Open Web Index for open web search. *Journal of the Association for Information Science and Technology* 75, 5 (2024), 512–520.

[Haas et al.(2017)] Andreas Haas, Andreas Rossberg, Derek L Schuff, Ben L Titzer, Michael Holman, Dan Gohman, Luke Wagner, Alon Zakai, and JF Bastien. 2017. Bringing the web up to speed with WebAssembly. In *Proceedings of the 38th ACM SIGPLAN conference on programming language design and implementation*. 185–200.

[Hendriksen et al.(2024a)] Gijs Hendriksen, Michael Dinzinger, Sheikh Mastura Farzana, Noor Afshan Fathima, Maik Fröbe, Sebastian Schmidt, Saber Zerhoudi, Michael Granitzer, Matthias Hagen, Djoerd Hiemstra, Martin Potthast, and Benno Stein. 2024a. The Open Web Index. In *Advances in Information Retrieval*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer Nature Switzerland, Cham, 130–143.

[Hendriksen et al.(2024b)] Gijs Hendriksen, Michael Dinzinger, Sheikh Mastura Farzana, Noor Afshan Fathima, Maik Fröbe, Sebastian Schmidt, Saber Zerhoudi, Michael Granitzer, Matthias Hagen, Djoerd Hiemstra, Martin Potthast, and Benno Stein. 2024b. The Open Web Index - Crawling and Indexing the Web for Public Use. In *Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 14612)*, Nazli Goharian, Nicola Tonellotto, Yulan He, Aldo Lipani, Graham McDonald, Craig Macdonald, and Iadh Ounis (Eds.). Springer, 130–143. `https://doi.org/10.1007/978-3-031-56069-9_10`

[Kenwright(2022)] Benjamin Kenwright. 2022. Introduction to the webgpu api. In *Acm siggraph 2022 courses*. 1–184.

[Kumar et al.(2025)] Priyanshu Kumar, Elaine Lau, Saranya Vijayakumar, Tu Trinh, Elaine T Chang, Vaughn Robinson, Shuyan Zhou, Matt Fredrikson, Sean M Hendryx, Summer Yue, et al. 2025. Aligned LLMs Are Not Aligned Browser Agents. In *The Thirteenth International Conference on Learning Representations*.

[Li et al.(2024)] Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2024. 2d matryoshka sentence embeddings. *arXiv preprint arXiv:2402.14776* (2024).

[Lu et al.(2025)] Yuxuan Lu, Bingsheng Yao, Hansu Gu, Jing Huang, Jessie Wang, Laurence Li, Jiri Gesi, Qi He, Toby

https://doi.org/10.5281/zenodo.17229737

Jia-Jun Li, and Dakuo Wang. 2025. UXAgent: An LLM Agent-Based Usability Testing Framework for Web Design. *arXiv preprint arXiv:2502.12561* (2025).

[Raasveldt and Mühleisen(2019)] Mark Raasveldt and Hannes Mühleisen. 2019. Duckdb: an embeddable analytical database. In *Proceedings of the 2019 international conference on management of data*. 1981–1984.

[Ruan et al.(2024)] Charlie F Ruan, Yucheng Qin, Xun Zhou, Ruihang Lai, Hongyi Jin, Yixin Dong, Bohan Hou, Meng-Shiun Yu, Yiyan Zhai, Sudeep Agarwal, et al. 2024. WebLLM: A High-Performance In-Browser LLM Inference Engine. *arXiv preprint arXiv:2412.15803* (2024).

[Zerhoudi and Granitzer(2025)] Saber Zerhoudi and Michael Granitzer. 2025. SearchLab: Exploring Conversational and Traditional Search Interfaces in Information Retrieval. In *Proceedings of the 2025 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '25), March 24–28, 2025, Melbourne, VIC, Australia*. ACM. https://doi.org/10.1145/3698204.3716475

ISBN: 978-92-9083-705-3 ISSN: 2957-4935

https://doi.org/10.5281/zenodo.17234055

ASSESSING THE RELIABILITY OF HUMAN AND LLM-BASED SCREENING IN SYSTEMATIC REVIEWS: A STUDY ON FIRST-TIME REVIEWERS

E. Sandner*[1,3], D. Scharf[3], T. Wautischar[3], I. Jakovljevic[1], A. Simniceanu[2],
L. Fontana[2], A. Henriques[1], A. Wagner[1], C. Gütl[3]
[1]CERN, 1211 Geneva, Switzerland
[2]WHO, 1211 Geneva, Switzerland
[3]Graz University of Technology, 8010 Graz, Austria

Abstract

Systematic reviews (SRs) rely on rigorous study selection to ensure methodological quality. Double-blind screening by two independent reviewers is considered the gold standard, but it is also time-consuming and resource-intensive. Large language models (LLMs) have recently been proposed as a means of reducing this workload, yet their acceptance remains limited because their performance is insufficiently benchmarked against the standards applied to human screeners. To address this gap, this study investigates the screening behavior of novice reviewers and compares their performance with that of an LLM. In a graduate-level course, 54 students conducted title and abstract screening across ten information retrieval topics. Each record was independently screened by four students and by an LLM applying a five-tier classification approach. Inter-rater reliability was measured using Fleiss' κ and Cohen's κ, while sensitivity and specificity were calculated under different screening configurations. The results show that novice screeners achieved only fair to moderate agreement (overall Fleiss' $\kappa = 0.386$), while the LLM's agreement with the human consensus (Cohen's $\kappa = 0.516$) was higher than the average human–human agreement, yet still within a similar range. Performance analysis revealed that single human screening (sensitivity 84.03%, specificity 90.36%) outperformed the LLM (80.30% / 85.50%). Double human screening achieved near-perfect sensitivity (99.18%) at the cost of lower specificity (82.10%). A hybrid setting of one human plus the LLM improved sensitivity (94.98%) relative to a single human. These findings highlight both the variability of novice human decisions and the potential role of LLMs as complements, but not replacements, in screening workflows.

INTRODUCTION

During the study selection phase of a systematic review (SR), hundreds or even thousands of abstracts must be manually assessed for eligibility. To ensure methodological rigor, high-quality reviews require double-blind screening [1], meaning that two independent reviewers evaluate each paper in parallel. This labor-intensive process is designed to minimize human bias and error, thereby meeting the strict quality standards of SR methodology.

Several approaches have been proposed to automate this time-consuming and repetitive task. However, automation methods designed for broad applicability across domains and eligibility criteria face limited acceptance within the evidence synthesis community. This is largely because they fail to demonstrate compliance with the methodological standards of SRs, particularly the requirement to safeguard against the exclusion of relevant studies, a criterion measured by sensitivity [2].

While some studies consider a sensitivity of 95% to be sufficient [3, 4], Cochrane[1], a research organization renowned for its high-quality SRs and for setting standards in the SR process, requires a sensitivity of 99% for any tool intended to replace human screening [5].

Evaluation of an automation system is typically performed by comparing its outcomes against a human-annotated gold standard, ideally derived from double-blind screening. Although automation must not be introduced into the SR process at the cost of reduced quality, one may question whether human screeners reliably satisfy the quality requirements imposed on automation systems, especially when training and domain expertise are limited.

Therefore, this study investigates the screening behavior of 54 students as they conducted their first SR exercise. Students were assigned one of ten topics in the domain of 'Information Search and Retrieval' and asked to screen 30 records. Each record was independently evaluated by four students. Furthermore, each record was subjected to large language model (LLM) screening using the 5-tier algorithm with a permissive setting [6]. Based on the collected data, inter-rater agreement was analyzed using Cohen's and Fleiss' kappa. Furthermore, by assuming the consensus of all four screeners as the gold standard, double-blind screening could be compared with alternative scenarios. Building on these results, this paper addresses the following research questions:

RQ1: What is the inter-rater agreement beyond chance among first-time screeners?

RQ2: How does the performance of single screening compare with double-blinded screening?

RQ3: How does the performance of LLM screening compare with that of human screening?

* elias.sandner@cern.ch

[1] `https://www.cochrane.org/`

https://doi.org/10.5281/zenodo.17234055

The paper proceeds with background and related work to situate the study in context. It then outlines the methodology, including data collection and analysis, before presenting and interpreting the results. This is followed by a discussion of limitations and directions for future work. The closing section summarizes the main contributions and conclusions.

BACKGROUND AND RELATED WORK

By synthesizing findings from potentially all relevant studies on a given research question, a SR represents the most reliable research methodology for evidence-based conclusions [7]. Therefore, SRs play a crucial role in the medical field, guiding decision-making and shaping clinical practice guidelines [8]. However, the rigorous nature of the process makes SRs highly time-consuming and resource-intensive. Completing a single SR often requires several months, and in some cases, even years [9–11].

When conducting a SR, an initial database query is typically designed to be broad rather than highly specific, ensuring comprehensive coverage of relevant studies. The retrieved candidate studies then undergo human screening, which is considered one of the most time-consuming stages of the SR process [12].

Numerous efforts have aimed to reduce the human workload through automation. For instance, Cochrane has developed a machine learning classifier to identify candidate studies and exclude those with study designs other than Randomized Controlled Trials (RCTs) [5]. However, its applicability is limited to reviews that include only RCTs. Additionally, all other criteria remain unaddressed, thus limiting workload reduction.

While training a classifier for broad applicability is impractical, classification systems leveraging the reasoning capabilities of LLMs have attracted increasing attention in recent years. However, their acceptance for real-world application within the evidence synthesis community is limited, as exhaustive evaluation and validation are lacking to demonstrate that these approaches achieve the expected sensitivity and do not miss relevant studies. [2]

However, the performance of human screeners is also subject to limitations, with errors and disagreements observed even under double-blinded conditions that are designed to minimize bias.

Gartlehner et al. [13] analyzed human screening performance in a crowd-based, parallel-group randomized controlled trial. All 280 participants had prior experience in abstract screening and were required to pass an initial task demonstrating that they could correctly label at least 80% of a test set. Each participant was assigned up to 100 candidate studies, which together resulted in 24,942 screening decisions on 2,000 randomly selected abstracts. On average, each abstract was screened 12 times. The study found that single screeners achieved a sensitivity of 86.6%, while double-blinded screening reached 97.5%. The respective specificities were 79.2% and 68.7%.

Issaiy et al. [14] compared the screening performance of GPT-3.5 Turbo with that of three general physicians. LLM screening was conducted by instructing the model to assign each study a numerical value between 1 and 5 based on its relevance, with categories 1–3 considered as include decisions. In this prospective simulation study, both the LLM and each participant screened 1,198 records spanning different subject areas within radiology. The ground truth was established based on the screening decisions of two expert researchers with 5 and 20 years of experience. The results showed moderate agreement among the three general physicians ($\kappa = 0.45$) and substantial agreement between the two expert researchers ($\kappa = 0.79$). In contrast, agreement between the LLM and the general physicians was lower, with a mean κ of 0.27. Screening performance of single screeners, consensus, and the LLM is detailed in Table 1. The findings demonstrate that neither individual human screeners nor the consensus of three screeners, whether based on majority voting or a sensitive consensus, meets the sensitivity requirements established for automated solutions.At the same time, the results indicate that the automated approach missed fewer relevant studies than the human screeners. However, the authors also noted that the group of general physicians consisted of relatively young researchers without specialized training in radiology and therefore may not adequately reflect the population that typically carries out screening in evidence-based medicine.

Table 1: Screening performance of general physicians, consensus methods, and ChatGPT, based on data from [14].

Screener	Sensitivity	Specificity
General Physician 1	0.55	0.94
General Physician 2	0.55	0.99
General Physician 3	0.74	0.94
Voting Consensus	0.62	0.98
Sensitive Consensus	0.90	0.89
ChatGPT	0.95	0.65

Relying on well-trained human researchers and the resource-intensive double-blinded screening process is undoubtedly required in SRs that underpin evidence-based practice, shaping clinical guidelines, healthcare policies, and treatment decisions. However, SRs also play a crucial role in the context of primary research. Performing a SR of existing evidence before initiating a new study is essential to ensure both its quality and relevance [15]. A thorough understanding of prior studies helps identify research gaps, formulate meaningful questions, and inform the design of new studies [16, 17].

When SRs are conducted for this purpose in domains outside medicine, training requirements for human screeners may be less stringent. Furthermore, in such cases, decisions regarding the use of automation may be more flexible, and resource savings from automation could enable SRs that would otherwise be infeasible due to resource constraints. More generally, in settings where automation achieves per-

https://doi.org/10.5281/zenodo.17234055

formance equivalent to the human screeners who would otherwise conduct the task, its use may be justified. However, the literature lacks studies on inter-rater agreement or screening performance of first-time screeners in computer science.

METHODOLOGY

To address the defined research questions, a structured methodological framework was applied that links data collection with subsequent analysis. This chapter describe how screening data were generated in the course setting, how they were transformed into a comparable dataset for humans and the LLM, and which metrics were used to assess reliability and performance.

Data Collection

The data collection was conducted as part of the graduate-level university course “Information Search and Retrieval” held during the winter term 2024/25 at Graz University of Technology. This course focuses on the key concepts of information retrieval (IR) and web search systems, while also introducing the core principles of SRs. All students enrolled on the course were tasked with conducting a SR project in the second half of the term, covering the entire SR pipeline and including additional exercises to gain practical experience. At the start of the term, 60 students were enrolled on the course and organised into ten groups of six students each. By the time the SR part of the course began, the number of enrolled students had decreased to 54 due to students dropping out for various reasons.

During the group selection process, each team had to choose a topic that would form the basis of their work on various tasks throughout the course, and, most importantly, their main focus for the SR project. Those ten topics were pre-selected by the course instructors in order to achieve two objectives: to ensure diversity between the topics and to highlight current research trends in the IR domain. To achieve the latter, the course instructors referred to the main topics in the proceedings of the 47th International ACM SIGIR Conference on Research and Development in IR when selecting the topics [18]. The selected topics were:

1. Neural IR
2. Retrieval Augmented Generation
3. GenIR and Search with LLMs
4. Evaluation with and for LLMs
5. Multilingual Retrieval
6. Question Answering and Summarisation
7. Conversational IR and Recommendation
8. Explainability in Search and Recommendation
9. Privacy and Security in Recent IR Systems
10. Users and Simulations in IR Systems

The entire SR project was divided into individual stages and tasks. These had to be completed either individually, in subgroups (smaller groups formed from the original group), or in the original groups. The first stage consisted of the creation of eligibility criteria, the identification of seed papers, the creation of the search string, the data retrieval, and the deduplication of the retrieved records. For each group, members first defined their eligibility criteria individually using the PICO (Participants, Intervention, Comparison, and Outcomes) framework as suggested in SR guidelines [1, 19]. Next, the group reached a consensus by comparing and refining these individual criteria. These eligibility criteria were then used to create the search strings for the two literature databases, ACM [2] and IEEE [3]. The course instructors selected these two literature databases as the primary sources for this SR project because they are widely used in the field of IR and share a similar, user-friendly search interface that enables students to apply their own search strings for data collection. To align with the lecture’s teaching objectives while maintaining sufficient variability for meaningful analysis, the number of studies per topic was limited to a manageable sample size of 200±20 papers per literature database. Because the same papers could appear in both literature databases, each group was required to perform a deduplication step when combining the two retrieved result sets. These deduplicated result sets marked the conclusion of the first stage and were subsequently submitted to the course instructors.

Conducting a comprehensive SR of hundreds of papers would not only be impractical in the context of a university course but also offer little educational value. Therefore, the number of papers assigned to each student for title and abstract screening (TiAb-screening) was reduced to align with the time constraints and learning objectives of the course. To guarantee that each student has a good variety of papers ranging from unsuitable to suitable in their screening set regarding their chosen topic, a LLM based screening algorithm (5-Tier Prompting Approach [6]) was utilized to select a subset of the retrieved papers. This algorithm uses the eligibility criteria previously created by each group and does an automated LLM screening of the retrieved papers, assigning each of them to one of five classes, indicating how likely the respective paper is to meet those criteria. Based on these classes, a subset of the original retrieved papers was created for each group, containing both papers that were deemed suitable (classes 1-3) and papers that were not suitable (classes 4-5) based on the respective eligibility criteria, as determined by the LLM screening. These subsets and subsequent tasks were structured so that every record in each subset was screened by exactly four different group members, with each of them screening approximately 30 papers. At this stage, the members of each group were unaware of the LLM screening process or the intentional overlap introduced

[2] `https://dl.acm.org/search/advanced`
[3] `https://ieeexplore.ieee.org/search/advanced`

https://doi.org/10.5281/zenodo.17234055

in their assigned paper sets. They were simply instructed to screen the papers they had been given, decide if they want to include or exclude the paper at this stage, and in the case that they want to exclude it, give a reason why.

Only afterwards were they tasked to combine their individual screening result and resolve potential conflicts. Conflict resolution was performed based on majority voting, and in case of a tie, through discussion. In instances where a paper was ultimately excluded, teams were required to provide a reason for exclusion. As a result, each group produced a document compiling all individual TiAb-screening decisions alongside the consolidated group decision. These documents indicated which papers from the subsets were ultimately included or excluded in the TiAb-screening phase, along with the corresponding justifications. They were uploaded for the course instructors and were the basis for the future tasks, including a full-text screening phase of the papers included after the TiAb-screening phase. This, however, is no longer relevant for the scope of this paper. The collection of these consolidated TiAb-screening documents forms the basis for the dataset of this analysis.

Data Processing and Analysis

The resulting dataset contains screening decisions from 54 individual human screeners alongside consolidated, conflict-resolved decisions provided by ten teams.

These team-level decisions were treated as the ground truth against which individual human screeners and the LLM system were evaluated. The evaluation followed the 5-tier framework proposed in [6], which classifies papers on a relevance scale from 1 (highly relevant) to 5 (not relevant). Since this framework does not yield a direct binary include/exclude outcome, the LLM outputs were converted to binary for comparability. Specifically, papers assigned a score of 4 or 5 by the LLM were classified as “excluded,” while papers assigned a score from 1 to 3 were treated as "included."

For data processing and evaluation, a Python pipeline was developed. In the first step, screening decisions from the two sources were combined into a Pandas DataFrame using a shared internal ID. Column names were standardized, and the resulting merged data was exported into team-specific CSV files.

A second script of the pipeline re-imports these files and transforms all decisions into binary format, applying the predefined thresholding strategy for the LLM outputs. For the human screeners, decisions are directly converted into binary values. This dataset then served as the basis for calculating the evaluation metrics described below.

To assess inter-rater reliability both among human screeners and between humans and the LLM, Cohen’s kappa and Fleiss’ kappa as defined in (1) and (2) are computed.

$$\text{Cohen's } \kappa = \frac{p_0 - p_e}{1 - p_e} \tag{1}$$

where p_o denotes the observed proportion of agreement between two raters and p_e the proportion of agreement expected by chance.

$$\text{Fleiss' } \kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e} \tag{2}$$

where $\bar{P}$ is the mean observed agreement across all subjects and raters, and $\bar{P_e}$ the mean expected agreement by chance, based on the marginal proportions of each category.

Cohen’s kappa measures the level of agreement between two raters, whereas Fleiss’ kappa extends this measure to situations involving more than two raters. Importantly, both metrics assess reliability rather than validity; they indicate how reliable the raters agree in their screening decisions among raters but do not evaluate whether they align with the chosen ground truth. A key advantage of these metrics is that they correct for the level of agreement that might be expected to occur purely by chance, making them more robust than raw agreement percentages. The values of kappa are bounded between –1 and 1, where 1 represents perfect agreement, 0 reflects agreement at the level of chance, and negative values suggest systematic disagreement.

Initially both, Cohen’s kappa and Fleiss’ kappa were computed to assess inter-rater reliability. However, during the analysis it became clear that Cohen’s kappa did not provide substantial additional insight when applied to human screeners, except in the specific case of comparing the consolidated team decision with the LLM. Consequently, Fleiss’ kappa was adopted as the main measure of inter-rater reliability among human raters.

For the performance analysis, several experimental conditions were designed and evaluated against the ground truth:

- Single human screener
- LLM only
- Two human screeners
- Single human screener and LLM

In scenarios involving two screeners, conflicts could occur regarding inclusion decisions. To resolve such disagreements, a paper was classified as “included” if at least one screener selected it. This approach was purposefully chosen to ensure that potentially relevant papers were less likely to be missed. Each screening setup was subsequently evaluated using sensitivity and specificity, as defined in (3) and (4).

$$\text{Sensitivity} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{3}$$

$$\text{Specificity} = \frac{\text{True Negative}}{\text{True Negative + False Positive}} \tag{4}$$

Sensitivity quantifies the proportion of truly relevant papers that were correctly identified as “included,” while specificity measures the proportion of truly irrelevant papers that were correctly identified as “excluded”.

https://doi.org/10.5281/zenodo.17234055

Although additional metrics were calculated, such as positive predictive value (precision) and negative predictive value, sensitivity and specificity emerged as the most informative and widely recognized metrics in the context of the research questions. Consequently, these two measures were prioritized in the evaluation, whereas the raw results of additional metrics are presented in the supplementary material[4].

The resulting data was exported to a CSV file containing all metrics generated by the pipeline. Subsequently, the file was imported into a spreadsheet, where overall average and median values are calculated for further analysis. Since every screener made their decision completely independent from each other, no weighting was introduced per team.

RESULTS

This section first reports the reliability between individual human screeners as well as between human screeners and LLM decisions. It then presents screening performance in terms of sensitivity and specificity across different settings, before discussing the results with respect to the research questions.

Inter Rater Reliability

As outlined in the Methodology section, Fleiss' kappa was used to assess inter-rater reliability among the human screeners. The resulting scores are presented in Table Table 2. Each column corresponds to a team score, while the rows represent specific subsets of four screeners (A–F). The penultimate row reports the mean score across all subgroups within a team, and the final row provides the overall mean across all teams.

Because not every screener evaluated every paper, certain combinations of screeners yield no κ values. In these cases, there were no papers jointly assessed by all members of the respective subgroup. This limitation is particularly evident in teams consisting of four screeners, where only one subgroup produces a valid value, as illustrated in Table 2 for Team 01. A similar situation arises in six-member teams, where certain subgroups lack overlap in screened papers.

After averaging the subgroup scores within each team, kappa values range from $\kappa = 0.24$ to 0.65, with an overall mean of $\kappa = 0.39$. According to the interpretation proposed in [20], 6 out of 10 teams fall into the category of "fair agreement." Three teams achieve scores between 0.40 and 0.60, corresponding to "moderate agreement," while Team 5 attains "substantial agreement" with a $\kappa = 0.65$. These findings show high variability across teams and subgroups, suggesting that the limited prior experience of the screeners with the SR process contributed to inconsistent decision-making within the teams.

When assessing inter-rater reliability between the human consensus and the decisions generated by the LLM, the resulting agreement levels are comparable to those observed among human raters themselves. Because the analysis is based on pairwise agreement between two raters, Cohen's κ was used as the reliability measure. As detailed in Table 3, the observed values ranged from minimal agreement ($\kappa = 0.26$) to strong agreement on a single occasion ($\kappa = 0.80$). On average, the agreement between human raters and the LLM was $\kappa = 0.52$, which is classified as weak agreement according to [21]. These results indicate that the level of agreement between human raters and the LLM is similar to the degree of agreement observed among human raters.

Performance Evaluation

In the performance analysis of this study, the classification accuracy of human screeners and LLMs was evaluated under different conditions and configurations, benchmarked against the ground truth described in the Methodology section.

The corresponding results are summarized in Table Table 4, which reports the sensitivity and specificity across the four defined experimental setups. The first row reflects the ground truth, which is the consensus of four independent human screeners, with conflicts resolved through discussion. The following rows show the performance of the defined experimental setups.

The findings demonstrate that human screeners consistently outperform the LLM with respect to both sensitivity and specificity, irrespective of whether single or double screening is applied. A single human screener achieved a sensitivity of 84.03% and a specificity of 90.36%. In contrast, the LLM archived 80.30% sensitivity and 85.50% specificity.

When double screening is used, sensitivity increased to 99.18%, but at the expense of specificity, which decreased to 82.10%. This trade-off shows the primary advantage of double screening, where the likelihood of omitting relevant studies decreases, but in return more irrelevant studies are included, leading to a lower specificity.

Another configuration involved pairing a single human screener with an LLM in a double-screening arrangement. In this setting, the combined system achieved a sensitivity of 94.58% and a specificity of 79.22%, representing an approximate 10% increase in sensitivity relative to a single human screener. As expected, specificity was reduced compared to either the human or the LLM alone.

Discussion

The inter-rater reliability results, as measured by Cohen's κ values reported above, demonstrate that novice reviewers show considerable inconsistency in their screening decisions. This outcome is not unexpected, as the reviewers were required to evaluate pre-defined research questions in domains that were, at times, unfamiliar to them. Moreover, the fact that each team produced the eligibility criteria by themselves, introduced additional interpretive flexibility, which in turn contributed to disagreement regarding the inclusion of specific studies. Only one team achieved a level of "substantial agreement," suggesting higher consistency within this group. Possible explanations include prior domain knowledge, prior

[4] https://zenodo.org/records/17113018

Table 2: Fleiss' κ for four human screeners across teams.

Team Number	01	02	03	04	05	06	07	08	09	10
A, B, C, D	0.25	0.73	-	-0.03	0.66	0.25	0.65	-	0.76	0.30
A, B, E, F	-	-	0.43	-	0.42	-	0.03	-	0.59	0.32
C, D, E, F	-	-	0.63	-	0.54	-	0.67	-	0.43	0.59
A, B, C, F	-	-	-0.07	-	0.76	-	0.56	-	0.43	0.09
B, C, D, E	-	0.45	-0.04	1.00	0.85	0.51	0.76	-	0.11	0.01
A, C, D, E	-	0.13	-	0.43	-	0.22	-	-	-	-
A, B, C, E	-	-0.04	-	0.26	-	0.37	-	-	-	-
A, B, D, E	-	0.29	-	-0.07	-	0.56	-	0.45	-	-
Team average	0.25	0.31	0.24	0.32	0.65	0.38	0.53	0.45	0.46	0.26
Average						0.38				

Table 3: Cohen's κ between human consensus and LLM across teams.

Team Number	**01**	**02**	**03**	**04**	**05**	**06**	**07**	**08**	**09**	**10**
Human Consensus, LLM	0.52	0.26	0.33	0.77	0.80	0.35	0.36	0.69	0.54	0.54
Average						0.52				

Table 4: Screening performance: sensitivity and specificity across scenarios.

Screening Scenario	**Sensitivity**	**Specificity**
Consensus of 4 human screeners (Ground-truth)	100.00%	100.00%
Single human	84.03%	90.36%
LLM only	80.30%	85.50%
Double human	99.18%	82.10%
Single human and LLM	94.98%	79.22%

experience with the SR process, or potential bias introduced by working together in this individual task. Importantly, the non-random nature of these findings is supported by the fact that Cohen's kappa explicitly accounts for agreement beyond chance.

A similar pattern appears when comparing team-level screening decisions with those generated by the LLM. The corresponding κ values vary considerably across teams, ranging from almost no agreement to very strong agreement. Notably, one team that demonstrated only moderate internal agreement suddenly aligned more strongly with the LLM at the team level. As mentioned in the Methodology, however, these results must not be interpreted as evidence of validity, since the applied κ values reflect relative agreement rather than correspondence to a definitive ground truth. The only conclusion that can be drawn from these results is that the screeners agree with the LLM to about the same extent as they agree with one another.

When looking at the performance analysis of this work, it becomes clear that double blinded screening clearly outperforms single screening. This is the standard in most SRs, since especially in medical domains this is a critical step to ensure that no relevant studies are missed. Nevertheless, lower specificity can substantially increase the workload during the full-text screening stage, which remains an important consideration, as this phase is very time-intensive as well. The observed reduction in specificity of approximately 8% is explainable with the fact that any screening conflicts are resolved in favor of inclusion. Under this rule, specificity cannot mathematically increase, as the number of exclusions in the double-screening setting is necessarily less than or equal to that in single screening.

These findings are in line with observations reported in [13], where single screening showed comparable levels of performance, and the introduction of double screening led to a significant improvement in sensitivity. At the same time, a similar trade-off was observed in the form of reduced specificity, which shows the inherent balance between maximizing the detection of relevant studies and managing the additional workload created by lower exclusion rates.

When comparing the performance of the LLM to the ground truth, human screeners demonstrated higher sensitivity, both under single screening and double screening conditions. This difference may be partly due to the baseline design, which was intentionally constructed to avoid bias in favor of the LLMs. When a single human screener was combined with the LLM, effectively creating a double screening scenario, sensitivity improved, although it did not reach the level achieved by double human screening. While the evaluated LLM configurations are not yet suitable as replacements for human screeners, they may serve as valuable additions in situations where resources for a second human screener are not available, enhancing the overall quality of the screening process.

While studies such as [14] report lower performance of human screeners, it is important to note that screening com-

https://doi.org/10.5281/zenodo.17234055

plexity strongly depends on factors such as the domain or the complexity of eligibility criteria. Furthermore, human screening performance likely decreases with an increasing number of studies to be screened.

LIMITATIONS AND FUTURE WORK

The present study was designed with deliberate caution to avoid overstating the capabilities of LLM-based screening. Its purpose was not to promote automation, but to transparently explore under which conditions LLMs may substitute or complement human screeners. This cautious design led to several limitations that must be considered when interpreting the findings.

First, the gold standard was derived from the consensus of human screeners, which inherently biases the evaluation in their favor. This effect is particularly pronounced for double human screening, where the consensus process directly shaped the reference against which alternative scenarios were judged. Relatedly, the eligibility criteria were defined by the student teams themselves and expressed in natural language. Such criteria allow interpretive flexibility, and individual screeners likely applied a consistent internal interpretation that aligned with the team consensus. The LLM, by contrast, had to infer the intended meaning without this implicit alignment, which disadvantaged its performance.

Second, the study setting differs substantially from real-world SRs. Each student screened only about 30 records, far fewer than the hundreds or thousands typically encountered in practice. As a result, important factors such as fatigue or consistency over time could not be captured. Furthermore, the course-based environment meant that screening carried no real-world consequences, unlike professional SR projects where errors may directly impact research or policy decisions.

Finally, the evaluation of the LLM was restricted to a single prompting configuration and a single model version. This narrow setup does not capture the potential variability in performance that may arise from alternative prompting strategies, different thresholds for inclusion, or comparisons across multiple LLMs.

Building on these limitations, several directions for future work emerge. Expert-annotated gold standards should be established to reduce bias in favor of human screeners and enable more rigorous benchmarking. The robustness of eligibility criteria could be tested by comparing team-defined formulations with expert-reviewed standards. Larger-scale studies are needed to assess long-term screening behavior, while experiments should include participants with different levels of SR expertise. Finally, more advanced evaluations of LLMs are warranted, exploring diverse prompting strategies and multiple models to determine the extent to which automation can complement or substitute human screeners.

CONCLUSION

This study examined the reliability and performance of novice human screeners compared with a LLM in the context of literature screening. Based on data from 54 students across ten IR topics, this study provides empirical evidence on how first-time screeners behave when tasked with TiAb-screening and how their outcomes compare to those of an automated system.

The findings demonstrate substantial variability among novice reviewers. Inter-rater reliability ranged from fair to moderate, with only one team achieving substantial agreement. These results confirm that human screening is not immune to inconsistency, particularly when reviewers lack domain expertise and prior experience with SR methodology. At the same time, the LLM achieved an agreement level with human consensus that was similar to the agreement observed among humans themselves, suggesting that automation can emulate human-like decision variability.

In terms of performance, single human screeners outperformed the LLM, but the gap was modest. Double human screening remained the most effective approach, achieving nearly complete sensitivity. This reinforces the importance of double-blind screening in settings where missing relevant studies carries significant consequences. However, the hybrid configuration of one human and one LLM also proved promising, reaching a sensitivity level close to double human screening while using fewer human resources. Such a setup may offer a practical compromise in resource-constrained environments.

Overall, the study underscores two key points. First, the reliability of novice screeners is limited, highlighting the importance of training and quality assurance in SRs. Second, while the considered LLM based algorithm cannot replace human screeners, it may serve as a valuable complements that increases sensitivity and reduces the risk of overlooking relevant studies when additional human reviewers are not available. These insights contribute to a more balanced perspective on the role of automation in evidence synthesis, providing a foundation for cautious but progressive integration of LLMs into the SR process.

ACKNOWLEDGEMENTS

The joint CERN and WHO ARIA[5] project is funding the PhD project, in the context of which this paper was written. Furthermore, we greatly thank OpenWebSearch.EU[6] project and the members for their help and support with this publication.

REFERENCES

[1] Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., and Welch, V. A. (eds.). *Cochrane Handbook for Systematic Reviews of Interventions*, version 6.5 (updated August 2024). Cochrane, 2024. Available from www.cochrane.org/handbook.

[2] E. Sandner, L. Fontana, K. Kothari, A. Henriques, I. Jakovljevic, A. Simniceanu, A. Wagner, and C. Gütl, "Evaluating

[5] https://partnersplatform.who.int/tools/aria

[6] https://openwebsearch.eu

https://doi.org/10.5281/zenodo.17234055

Large Language Models for Literature Screening: A Systematic Review of Sensitivity and Workload Reduction," 2025.

[3] W. M. Bramer, M. L. Rethlefsen, J. Kleijnen, and O. H. Franco, "Optimal database combinations for literature searches in systematic reviews: a prospective exploratory study," *Systematic Reviews*, vol. 6, pp. 1–12, 2017.

[4] M. W. Callaghan and F. Müller-Hansen, "Statistical stopping criteria for automated screening in systematic reviews," *Systematic Reviews*, vol. 9, pp. 1–14, 2020.

[5] J. Thomas, S. McDonald, A. Noel-Storr, I. Shemilt, J. Elliott, C. Mavergames, and I. J. Marshall, "Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews," *Journal of Clinical Epidemiology*, vol. 133, pp. 140–151, 2021.

[6] E. Sandner, B. Hu, A. Simiceanu, L. Fontana, I. Jakovljevic, A. Henriques, A. Wagner, and C. Gütl, "Screening Automation for Systematic Reviews: A 5-Tier Prompting Approach Meeting Cochrane's Sensitivity Requirement," *Proceedings of the 2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pp. 150–159, 2024.

[7] P. G. Shekelle, M. A. Maglione, J. Luoto, et al., *Global Health Evidence Evaluation Framework*, Agency for Healthcare Research and Quality (US), Rockville, MD, 2013. Available at: `https://www.ncbi.nlm.nih.gov/books/NBK121300/table/appb.t21/`.

[8] D. J. Cook, N. L. Greengold, A. G. Ellrodt, and S. R. Weingarten, "The relation between systematic reviews and practice guidelines," *Annals of Internal Medicine*, vol. 127, no. 3, pp. 210–216, 1997.

[9] E. M. Beller, J. K.-H. Chen, U. L.-H. Wang, and P. P. Glasziou, "Are systematic reviews up-to-date at the time of publication?," *Systematic Reviews*, vol. 2, pp. 1–6, 2013.

[10] M. R. Demetres, D. N. Wright, A. Hickner, C. Jedlicka, and D. Delgado, "A decade of systematic reviews: an assessment of Weill Cornell Medicine's systematic review service," *Journal of the Medical Library Association: JMLA*, vol. 111, no. 3, p. 728, 2023.

[11] R. Borah, A. W. Brown, P. L. Capers, and K. A. Kaiser, "Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry," *BMJ Open*, vol. 7, no. 2, p. e012545, 2017.

[12] J. C. Carver, E. Hassler, E. Hernandes, and N. A. Kraft, "Identifying barriers to the systematic literature review process," in *Proceedings of the 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pp. 203–212, 2013.

[13] G. Gartlehner, L. Affengruber, V. Titscher, A. Noel-Storr, G. Dooley, N. Ballarini, and F. König, "Single-reviewer abstract screening missed 13 percent of relevant studies: a crowd-based, randomized controlled trial," *Journal of Clinical Epidemiology*, vol. 121, pp. 20–28, 2020.

[14] M. Issaiy, H. Ghanaati, S. Kolahi, M. Shakiba, A. H. Jalali, D. Zarei, S. Kazemian, M. A. Avanaki, and K. Firouznia, "Methodological insights into ChatGPT's screening performance in systematic reviews," *BMC Medical Research Methodology*, vol. 24, no. 1, p. 78, 2024.

[15] Clarke M, Hopewell S, Chalmers I. Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting. *The Lancet*, 2010;376(9734):20–21.

[16] Robinson KA, Brunnhuber K, Ciliska D, Juhl CB, Christensen R, Lund H. Evidence-based research series–paper 1: what evidence-based research is and why it is important? *Journal of Clinical Epidemiology*, 2021;129:151–157.

[17] Lund H, Juhl CB, Nørgaard B, Draborg E, Henriksen M, Andreasen J, et al. Evidence-based research series–paper 2: using an evidence-based research approach before a new study is conducted to ensure value. *Journal of Clinical Epidemiology*, 2021;129:158–166.

[18] *SIGIR '24: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, New York, NY, USA, 2024. ISBN: 9798400704314.

[19] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, and G. Lasa, "How-to conduct a systematic literature review: A quick guide for computer science research," *MethodsX*, vol. 9, p. 101895, 2022.

[20] Hartling L, Hamm M, Milne A, Vandermeer B, Santaguida PL, Ansari M, Tsertsvadze A, Hempel S, Shekelle P, Dryden DM. "Validity and Inter-Rater Reliability Testing of Quality Assessment Instruments [Internet]." *Rockville (MD): Agency for Healthcare Research and Quality (US); 2012 Mar. Report No.: 12-EHC039-EF. PMID: 22536612.*

[21] McHugh ML. "Interrater reliability: the kappa statistic." *Biochem Med (Zagreb). 2012;22(3):276-82*

https://doi.org/10.5281/zenodo.17209496

ADDING RETRIEVAL AUGMENTED GENERATION TO THE MOSAIC FRAMEWORK

F. Holz[1], D. Scharf[1], A. Nussbaumer[1], S. Gürtl[1]
[1]Graz University of Technology, Graz, Austria

Abstract

This paper presents a concept for adding Retrieval-Augmented Generation (RAG) features to the MOSAIC framework. MOSAIC enables web search in segments of the Open Web Index (OWI), in order to establish a special-purpose search engine. An extension, MOSAIC-RAG, has been developed that adopts a RAG approach. It is designed as a modular framework that has integrated a set of processing modules built on generative AI models, such as a module for re-ranking the search result, a module for summarising the full texts of the search result, or a module for summarising all search results. These modules can be ordered in an arbitrary sequence, in order to configure an overall process to improve the search result. Such configurations can be adapted for specific purposes and saved for later reuse.

INTRODUCTION

Recently, Large Language Models (LLM) have become very popular, because humans can interact with them in natural language when requesting information. They are capable of generating texts in various contexts, such as answering questions, providing extensive information, or summarising texts. In contrast to traditional search engines, they do not deliver original web documents, but generate responses based on a vast amount of information that has been used to train them. Though this type of information searching might be attractive for many people, there are also problems such as the phenomenon of hallucinations, outdated information, and missing information sources.

Retrieval-Augmented Generation (RAG) seeks to combine LLMs with traditional search engines. Different techniques have been proposed explaining how search engines are enriched with LLM functionalities[1]. A simple technique consists of the use of text chunks retrieved from a search engine for feeding and prompting an LLM. More advanced features include the improvement of the search query, as well as the re-ranking or summarisation of the results with the help of an LLM. Such an integration has several advantages and partially overcomes the aforementioned problems of LLMs. A web index with current data can inject up-to-date information into LLMs, and also provide original web documents on demand. Thus, hallucination is mitigated by providing factual knowledge in combination with generated texts.

This paper presents a RAG approach that is based on the Open Web Index (OWI). A special-purpose search engine created with data from the OWI is integrated with a framework that processes the retrieved data using different kinds of AI models. The next section describes the overall concept of this framework. This section is followed by a more detailed description of the modules used to improve the search process. Finally, an application is presented that showcases how a RAG system can be set up with our approach.

CONCEPT AND MODULAR FRAMEWORK

The overall aim of MOSAIC-RAG[1] is to enrich search engines using the Open Web Index (OWI) with features provided by Large Language Models (LLMs). The enrichment is mainly performed by further processing the search result, such as providing summarisations, re-rankings, or conversational search. The result is delivered to the end-user via a built-in web interface or an API that can be used by external applications. The overall concept is depicted in Figure 1 and explained in more detail in this section.

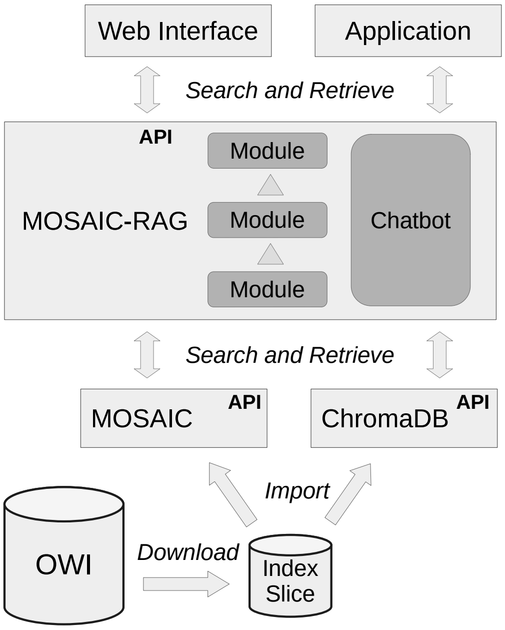


Figure 1: Conceptual design of MOSAIC-RAG.

The first step of creating a MOSAIC-RAG application consists in the creation of an index slice that serves as the underlying database for the search engine. Index slices are small- or medium-sized indices containing web documents related to a certain topic or a particular purpose. More precisely, they contain an inverted index represented in CIFF format[2] and metadata of each web document represented in Parquet format[3]. The metadata include the title, full text,

[1] https://opencode.it4i.eu/openwebsearcheu-public/mosaic-rag
[2] https://github.com/osirrc/ciff
[3] https://parquet.apache.org/

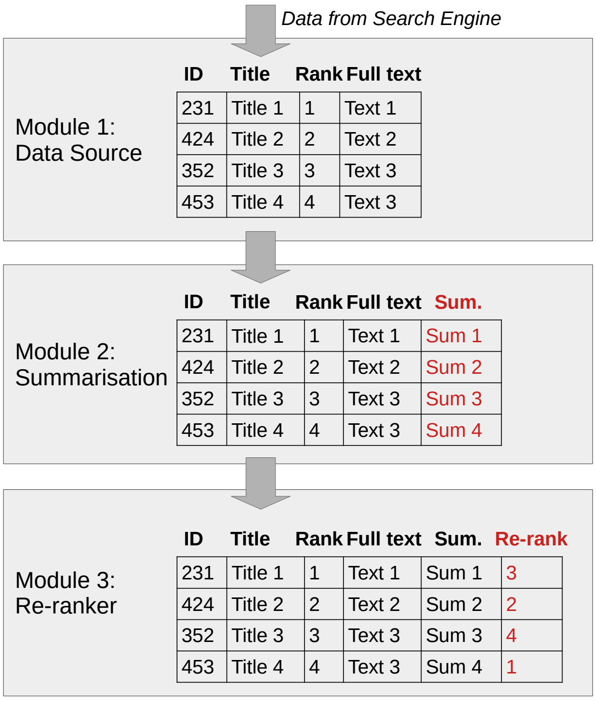
https://doi.org/10.5281/zenodo.17209496

URL, language, geo-coordinates, topic, and other information of the web document. Such slices can be downloaded from the OWI using queries that specify the domain and content of the index slice [2]. For example, index slices can contain web documents related to a certain topic, such as science news, a specific language, such as Finnish, or are part of a certain top-level domain.

The second step consists of the preparation of the search engine that delivers search results using the index slice. There are two options that are compatible with MOSAIC-RAG. First, MOSAIC is a framework and generic search application that makes index slices searchable [3]. Second, Chroma[4] is a vector database that allows to search documents using vector embeddings. Both search engines provide an API that allows the search for web documents and delivers lists of web documents including their metadata and full text.

Ingesting data slices works different for each of these search engines. MOSAIC is designed to easily integrate index slices by just copying them into a resource directory. Each index slice is represented as an index in MOSAIC and can be searched individually. Importing index slices into Chroma needs some pre-processing, as it requires vector embeddings for each web document, that can be created with suitable models, such as the Jina Embeddings 2 Model[4]. In Chroma, each web document is represented as a triple consisting of an ID, the vector embedding, and the metadata from the Parquet file. Later the search query is also represented as vector embedding using the same model, which allows Chroma to retrieve matching documents. In the future, the vector embedding will also be part of the OWI, which simplifies the importing procedure.

The core of MOSAIC-RAG is a modular pipeline that enriches the search result retrieved from the search engine. It includes a suite of processing modules that can perform various transformations of the search result. The currently available modules are described in the next section. For example, the full text of each result item (web document) can be summarised, the list of result items can be re-ranked, or an overall summary can be created out of the search result. The set of currently available modules is extensible and new modules can be added by implementing a base class that manges a data frame consisting of the search result. The rows of the data frame consist of the individual web documents and the columns comprise their metadata. Each module can manipulate the data frame in any way. Typically, a row with newly calculated information is added, for example with summarisation of the full text or by computing a new ranking (see Fig 2). Each module that uses an LLM to process the data can either chose to run the LLM locally, i.e., directly from the Python code, or use a remote inference point. The remote inference point can be configured globally for the whole MOSAIC-RAG instance. For this purpose, either a LiteLLM[5] or OpenAI compatible endpoint is required.



ID	Title	Rank	Full text
231	Title 1	1	Text 1
424	Title 2	2	Text 2
352	Title 3	3	Text 3
453	Title 4	4	Text 3

ID	Title	Rank	Full text	Sum.
231	Title 1	1	Text 1	Sum 1
424	Title 2	2	Text 2	Sum 2
352	Title 3	3	Text 3	Sum 3
453	Title 4	4	Text 3	Sum 4

ID	Title	Rank	Full text	Sum.	Re-rank
231	Title 1	1	Text 1	Sum 1	3
424	Title 2	2	Text 2	Sum 2	2
352	Title 3	3	Text 3	Sum 3	4
453	Title 4	4	Text 3	Sum 4	1

Figure 2: Modular Pipeline with data frames

The modules can be sequenced in any order depending on the purpose how the results should be processed. Thus a user can create a certain sequence of processing modules, in order to define the behaviour of MOSAIC-RAG (see also next section). Such a configuration is ephemeral, only lasting for the duration of the browser session. However, MOSAIC-RAG provides two ways of loading and saving the full configuration, i.e., custom color theme, custom titles, and the pipeline configuration. First, configuration can be downloaded in JSON format. Users can upload this JSON file to the frontend to restore a saved configuration. Second, this configuration can be stored on the server under a unique ID to be retrieved using a custom URL. Each module has also a few parameters to steer their behaviour, such as the selection which LLM should be used for the summarisation.

In addition to the modular pipeline, MOSAIC-RAG also supports a conversional search functionality. In a chat box, the user can ask an LLM questions about the current set of search results. The conversational search agent is instructed to only give answers based on the actual search results, not based on its own world knowledge.

In order to interact with MOSAIC-RAG, a web interface is provided that enables both the search and the configuration of the modular pipeline. The web interface uses MOSAIC-RAGS fully documented API. This allows other applications to use the full functionalities of the service.

RAG MODULES

This section describes the 18 currently implemented pipeline modules. These modules are organised in five groups, depending on their functionality: data source, sum-

[4] https://www.trychroma.com/
[5] https://www.litellm.ai

https://doi.org/10.5281/zenodo.17209496

marisation, re-ranking, pre-processing, and metadata analysis.

The data source modules deal with retrieving search results from external search engines when a user starts a query. Currently two search engines are supported, MOSAIC and Chroma. Details can be configured, such as the index used by MOSAIC or the embedding model used by Chroma. Furthermore, the number of search results can be limited. The data source module converts the data gathered in those search engines into a dataframe. This dataframe gets passed through the configured pipeline modules sequentially. After the final module, the dataframe gets sent to the user according to the API specification. Multiple data source modules can also be added to the same pipeline, allowing for the aggregation of data from different sources (e.g. multiple MOSAIC instances).

The pre-processing modules mainly deal with text cleaning and organising of the result set. There are modules to remove HTML tags and stop words, or to perform stemming operation on the text. These functions might not be needed in every case, as the search results may already be cleaned by the original search engine. As computing power is often limited, the Reduction Module is important because it reduces the size of the internal data frame based on a condition (usually the ranking). When processing large result sets in a pipeline containing at least one LLM module, such as a LLM Summarizer or an Embedder, the execution time of the total pipeline can be greatly reduced by decreasing the number of processed documents. Therefore, after performing some re-ranking, it might be sufficient to keep the best few documents and discard the rest.

The re-ranking modules change the ranking of the result set. Currently, four re-ranking modules are implemented by default in MOSIAC-RAG. The embedding re-ranker performs a new ranking based on the similarity of embedding vectors. Those embedding vectors will be created using the SentenceTransformer Python library if they do not already exist in the data frame. The TF-IDF (term frequency - inverse document frequency) [5] re-ranker is among the simplest and fastest approaches, allowing documents to be re-ranked based on their TF-IDF vector representations and a chosen similarity metric. Currently, MOSAIC-RAG supports the similarity metrics Euclidean distance, Manhattan distance, cosine similarity, and BM25. The latter differs slightly from the others, as it does not rely on the full TF-IDF vector representation. A BM25 ranking algorithm is also used by MOSAIC for its search. The two other re-ranking modules are based on the principle of large-language-model-re-ranking [6]. Here large language models (LLMs) are used to identify which document fits the given query best in a set of given candidate documents. The Group-Style LLM re-ranker module ranks documents by comparing a set of candidate documents against a given query and allowing the LLM to determine which document best matches the query. For each comparison, a score is assigned to the document that fits best. This process is repeated across all possible document combinations, given both the size of the candidate set and the total number of documents [7, 8]. The language model and the size of the candidate set can be configured. The final pre-implemented re-ranking module is the Tournament-Style LLM Re-ranker. Like the Group-Style variant, it relies on an LLM for re-ranking, but it reduces the number of required document comparisons, the most time-consuming step, by leveraging an existing ranking and refining it locally. The process follows the structure of a tournament tree, where the winning document advances while the losing one is eliminated. This approach requires significantly fewer LLM comparisons, improving efficiency. However, it functions more as a ranking enhancement than a full re-ranking. As it depends heavily on the initial ranking used as the seed, its effectiveness is greatest for identifying the top-ranked documents relevant to a query, while ranking quality tends to diminish further down the list. It is important to note that all pre-implemented re-ranking modules operate solely on the documents retrieved in the initial stage and do not perform any additional retrieval themselves [9].

There are two types of summarisation modules. The first one summarises the full text of each web document in the result set, while the second one generates one summary of all the documents in the result set. In both cases an LLM with targeted prompts is employed for these tasks.

Finally there are three metadata analysis modules. The first one is a simple word counter that counts the number of words in a web document. The Sentiment Analyser calculates a sentiment score for each web document. Based on six output scores for each sentiment the highest score is taken and stored in the dataframe. The relevance marking module marks parts of the full text that are most relevant. Both the sentiment analysing module and the relevance marking module use an LLM for their task.

APPLICATION CASE

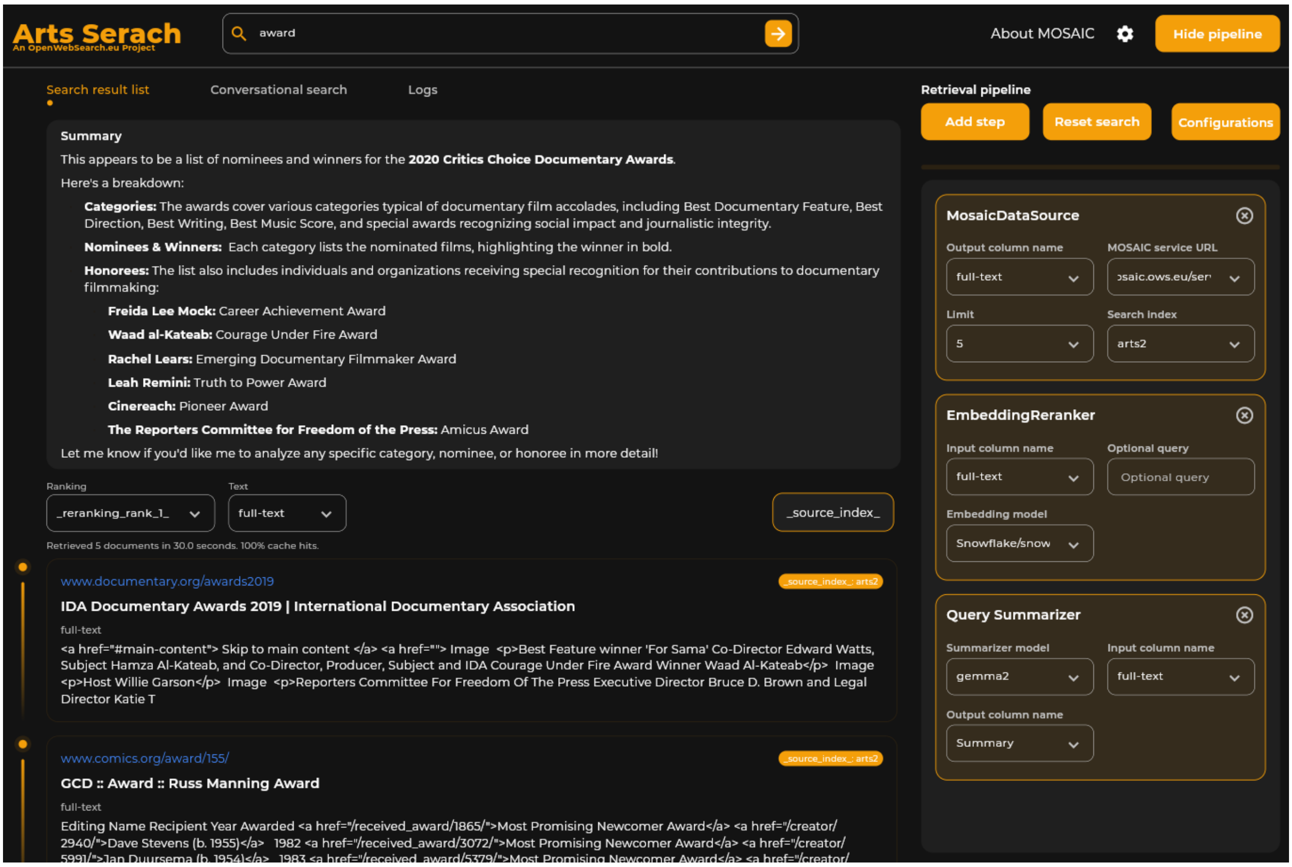
For demonstrating how a RAG system can be set up and configured with MOSAIC-RAG, an application has been created that enables search in the domain of arts. This arts search engine is depicted in Fig. 3.

First, an index slice has been created that only includes web documents related to arts. This was achieved by selecting web documents in the Open Web Index that are tagged with the Curlie label *Arts*. After integrating this index slice in MOSAIC, the service is started.

Second, MOSAIC-RAG is set up to act as an arts search engine. Hence, a MOSAIC-RAG data source module is configured to use the data from the arts index of the previously started MOSAIC service. Then an Embedding Re-ranking Module is added to improve the search result. Finally, a summarisation module is added that provides an overall summary of the search result on top.

Finally, the appearance of the web interface is configured. The title is changed to *Arts Search* and the colour scheme is set to *dark-orange*. The whole configuration is saved and an ID is automatically created, which allows to share this configuration via a single URL.

https://doi.org/10.5281/zenodo.17209496



Summary

This appears to be a list of nominees and winners for the **2020 Critics Choice Documentary Awards**.

Here's a breakdown:

Categories: The awards cover various categories typical of documentary film accolades, including Best Documentary Feature, Best Direction, Best Writing, Best Music Score, and special awards recognizing social impact and journalistic integrity.

Nominees & Winners: Each category lists the nominated films, highlighting the winner in bold.

Honorees: The list also includes individuals and organizations receiving special recognition for their contributions to documentary filmmaking:

Freida Lee Mock: Career Achievement Award

Waad al-Kateab: Courage Under Fire Award

Rachel Lears: Emerging Documentary Filmmaker Award

Leah Remini: Truth to Power Award

Cinereach: Pioneer Award

The Reporters Committee for Freedom of the Press: Amicus Award

Let me know if you'd like me to analyze any specific category, nominee, or honoree in more detail!

Ranking: _reranking_rank_1_
Text: full-text
_source_index_

Retrieved 5 documents in 30.0 seconds. 100% cache hits.

www.documentary.org/awards2019 _source_index_: arts2

IDA Documentary Awards 2019 | International Documentary Association

full-text

<a href="#main-content"> Skip to main content </a> <a href=""> Image <p>Best Feature winner 'For Sama' Co-Director Edward Watts, Subject Hamza Al-Kateab, and Co-Director, Producer, Subject and IDA Courage Under Fire Award Winner Waad Al-Kateab</p> Image <p>Host Willie Garson</p> Image <p>Reporters Committee For Freedom Of The Press Executive Director Bruce D. Brown and Legal Director Katie T

www.comics.org/award/155/ _source_index_: arts2

GCD :: Award :: Russ Manning Award

full-text

Editing Name Recipient Year Awarded <a href="/received_award/1865/">Most Promising Newcomer Award</a> <a href="/creator/2940/">Dave Stevens (b. 1955)</a> 1982 <a href="/received_award/3072/">Most Promising Newcomer Award</a> <a href="/creator/5991/">Jan Duursema (b. 1954)</a> 1983 <a href="/received_award/5379/">Most Promising Newcomer Award</a> <a href="/creator/

MosaicDataSource

Output column name: full-text
MOSAIC service URL: osaic.ows.eu/ser
Limit: 5
Search index: arts2

EmbeddingReranker

Input column name: full-text
Optional query: Optional query
Embedding model: Snowflake/snow

Query Summarizer

Summarizer model: gemma2
Input column name: full-text
Output column name: Summary

Figure 3: The Web Interface of MOSAIC-RAG configured as arts search engine. The summarisation and search results are on the left side and the processing pipeline on the right side.

CONCLUSION AND OUTLOOK

The main contribution of this paper consists of a Retrieval-Augmented Generation approach in the context of the Open Web Index. A special purpose and vertical search engine created with data from the OWI is integrated with a framework that processes the retrieved data using different kinds of LLMs.

Future work will include user studies that investigate the usefulness and acceptance of this approach. Furthermore, different configurations will be created and tested, in order to better understand their benefits for the user. In particular, the benefit for end-users of summarisations and re-rankings will be investigated.

ACKNOWLEDGEMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

REFERENCES

[1] Y. Gao *et al.*, *Retrieval-augmented generation for large language models: A survey*, 2024. https://arxiv.org/abs/2312.10997

[2] M. Granitzer *et al.*, "Impact and development of an open web index for open web search," *Journal of the Association for Information Science and Technology*, 2023. 10.1002/asi.24818

[3] S. Gürtl, *MOSAIC: Empowering a Modular Framework for Configurable and Tailored Web Search based on an Open Web Index*, 2024. https://diglib.tugraz.at/diplomaTheses

[4] M. Günther *et al.*, *Jina embeddings 2: 8192-token general-purpose text embeddings for long documents*, 2023.

[5] K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[6] Y. Zhu *et al.*, "Large language models for information retrieval: A survey," *ACM Transactions on Information Systems*, 2025. 10.1145/3748304

[7] W. Sun *et al.*, *Is chatgpt good at search? investigating large language models as re-ranking agents*, 2024. https://arxiv.org/abs/2304.09542

[8] J. Sun, X. Zhong, S. Zhou, and J. Han, "Dynamicrag: Leveraging outputs of large language model as feedback for dynamic reranking in retrieval-augmented generation," *arXiv preprint arXiv:2505.07233*, 2025.

[9] M. Rathee, S. MacAvaney, and A. Anand, "Guiding retrieval using llm-based listwise rankers," in *European Conference on Information Retrieval*, Springer, 2025, pp. 230–246.

https://doi.org/10.5281/zenodo.17237756

FUSION OF RETRIEVAL, GRAMMAR RULES AND DECISION TREES FOR TEXT GENERATION

E. Niehaus[†], W. Kühn, S. Müller
University of Kaiserslautern-Landau (RPTU), Landau, Germany

Abstract

The generation of scientific documents is accompanied by decisions of the author, e.g. the type of paper, publishing journal, and selection of an appropriate methodology. This generates a decision tree. Algorithmic support can provide options that the author selects. Combining this approach with Generative Artificial Intelligence (GenAI) , that is applied to trees and rules, results in a syntactic and a semantic structure. This conceptual paper discusses the fusion of Retrieval Augmented Generation (RAG), grammars, and decision trees with citable tree nodes. Tree nodes are defined by grammar rules with or without decision options and a user driven or randomized selection. The transparency of text generation is supported by a version control for the documentation of the paper evolution.

INTRODUCTION

The increasing application of GenAI [1] in the scientific domains requires transparency for the document evolution. Citations are used as a standard method to refer to publications and new scientific results based on the existing scientific knowledge. In the evolution of scientific publications, it is necessary to have transparency about the algorithmic support of Large Language Models (LLMs) and the human collaborative effort of the involved scientists. Publications in proceedings share requirements for the scientific structure of papers. The conceptual design of algorithmically generated syntactical options and decisions of an author generates a decision tree and simultaneously an Abstract Syntax Tree (AST) of the generation process. Decisions might be reverted, so tracking the changes is required for a transparent generation process. In this concept, the generation process can include algorithmic steps and steps created under control of the author (e.g., the modification of the generated text and the adaptation to the requirements and constraints of the paper). An alternating version in a version control system (VCS) [2] can be used to validate a collaborative modification history. Submissions of papers can be accompanied by a repository with version control. To be compliant with a transparent history of the evolution, citable grammar rules can be referred to in the document, similar to citations, as a grammar list. A tree node that was populated by GenAI requires the prompt, the used LLM, and the product of the LLM as part of the Version Control Repository. Even a more advanced use, like in software development, can be used with citable references in the version history. Branching from a specific version towards a new objective or an advancement of the current state of the art can be used with version control, which accelerate document development and text production.

In graph theory, the concept of a graph can be applied to an AST of the document structure (e.g., “PAPER” expands to 5 sections: “INTRODUCTION”, “METHODOLOGY”, “DATA ANALYSIS” “RESULTS” and “CONCLUSION”). The bottom-up strategy in computer science is used in a compiler that parses the code of a library and checks if the code is syntactical correct. For this paper we follow a top-down strategy, starting with start node as the root (“DOCUMENT”) and creating decision options for the author, e.g. “Write a scientific paper” or “Write a poem” or “Write a novel”. The decision option at the root node of the graph creates a version in the VCS, providing all decision options and saving the selected option. Reverting changes allows going back in the decision tree and changing the option. Version control supports this work-in-progress through a transparent evolutionary protocol. This is especially in the educational system an approach for assessment because it allows tracking the changes and the contribution (especially in a collaborative work).

Decision options applied grammar rules, and applications of LLMs to populate a node with text paragraphs can be complemented with RAG. The retrieval with keywords provided by the author or by a keyword extraction out of the generated paragraph or tree node descriptions creates, for example references retrieved from a local database of references. Matching references are provided to the authors as a decision option. Any new retrieval with a selection process can create a new version in the VCS. The classical compiler involves a bottom-up process for parsing and a top-down process to generate from a program written in source language A into a functional equivalent code into a programming language B. A LLM can be used to extract parts of the syntactical structure of documents. This can be applied to sections in novels where the protagonist's sections are responsible for a particular activity that can be summarized or even tokenized by a particular grammar rule. Comparisons between novels of the same structure can be performed as preliminary steps for a scientific analysis by a literary scholar. This shows the link between a bottom-up and a top-down application of decision and ASTs for generation and analysis.

Considering the technical process, a compilation of source code requires a tokenizer first to convert the given source code into an array of tokens. The list of tokens is transformed into an AST [3]. A parser reduces the string of the code, which is written in language A, to a start symbol S. A specific set of grammar rules is implemented for this

[†] email: niehaus@rptu.de

https://doi.org/10.5281/zenodo.17237756

purpose. In the conceptual design, the parsing of existing documents can be performed bottom-up while generation is top-down from the basic idea written in the first prompt towards a complete document an author wants to create. This selection process can also be influenced by the system prompt in GenAI, making it an integral part of the version control history.

The fully controlled part of the document generation can be accomplished by searching in a database for an appropriate grammar that serves a specific document generation task. In general, a grammar consists of a set of non-terminal symbols *N*, a set of terminal symbols *T*, a system of rules *R*, and a start symbol *S* as an element of *N*.

An AST represents the parsing process of the source code after the analysis of syntax and semantic. In the next step, the generated (and semantically attributed) AST is used to create an output source code in language B. The output language B can be binary/executable or a low-level programming language. Thus far, the parsing of source code in A and generation in code B can be distinguished conceptually.

In contrast to a bottom-up process, the top-down process for text generation creates decision options for a non-terminal symbol S_0, that can be replaced by S_1, S_2, S_3 or S_4. In the graph structure of the generated document, the selection of the author is documented in the VCS. After the selection, the other non-terminal symbols are not considered for the next steps of the document generation. The selection process is a difficult decision. A probability distribution over the decision options determined by empirical data for an organisation can support automated selection by a Monte Carlo approach [4] in conjunction with a fuzzy logic approach for describing the acceptance of natural language elements and their matching with grammar rules [5].

BASIC EXAMPLE OF CHOICES

As a example, it uses a non-terminal symbol *S*, that can be replaced by one of two decision options, as rule (1)

$$S \rightarrow SCI \mid NOV \quad (1)$$

The decision process allows the selection of a scientific paper *SCI* or a novel as a cultural contribution. The VCS stores the selection of rule (1) and the decision of the author to write a scientific paper. So, in the generation process, *S* is replaced by *SCI* and not by *NOV*. *SCI* represents a scientific article, and *MA* is a manual, e.g. for application of a workflow. This second rule (2) is branching into a sequence of five non-terminal symbols without a choice for the author. This replacement can be automated without user interaction.

$$SCI \rightarrow S_A S_I S_M S_R S_C \quad (2)$$

Such a deterministic expansion will not create a version in the VCS until another author decision is required, or GenAI is used to populate the non-terminal symbol with a LLM and a prompt. Rule (2) defines the grammar structure of a scientific paper. A scientific paper in this definition of grammar consists of five sections represented by the non-terminal symbols:

- S_A is the section *Abstract*,
- S_I the section *Introduction,*
- S_M the section *Methodology,*
- S_R the section *Results* and
- S_C represents the section *Conclusion.*

For example, above the abstract of the paper denoted by the non-terminal symbol S_A can be generated by a summary of the other section by using a LLM. So, summarizing the basic example above, it is necessary to distinguish four types of non-terminal symbols.

- **(ONT) optional non-terminal symbols** allow user selection of provided choices in the rule – see rule (1),
- **(DNT) deterministic non-terminal symbols** have no options and can be applied directly expanding a given non-terminal symbol SCI by a sequence of symbols - see rule (2),
- **(PNT) probabilistic non-terminal symbols** have a list of options with a probability distribution over the decision options, and one of the options is selected by a Monte Carlo approach,
- **(RNT) retrieval non-terminal symbols** have a query call to a search engine and return a list of choices. The author is asked to select one for further text generation and inclusion of one or more selected options. The selection process is similar to rule (1), but it dependents on a retrieval call,
- **(LLNT) Large Language non-terminal symbols** have a prompt and/or system prompt with a selected LLM. If the generation process reaches such a tree node, the prompt submits the current generation context to the LLM and the call of the LLM populates the tree node with generated text.

The non-terminal symbol of the type LLNT might also call other types of non-terminal symbols, such as PNT, DNT, or RNT. The generation process stops when author interaction is required. This is the case, when editing or reviewing of a section is required (such as the methodology section of a scientific paper) needs a quality assurance. Validation and reviewing of sections or generation steps will be documented in the VCS by adding a validation flag for a single author within a collaborative team.

CITATION OF GRAMMAR RULES

The different types of non-terminal symbols require different levels of citation. According to transparency, it is necessary to distinguish between a Monte Carlo selection of choices where an intellectual knowledge of the author is

https://doi.org/10.5281/zenodo.17237756

not involved and an ONT or RNT where the author has to select from a given number of choices. An author might want to annotate a tree node with a comment, stating why a specific choice is the most appropriate. It might also happen that none of the choices provided by the grammar rule ONT or RNT is appropriate, and the author provides in a traditional manner a section for the generated document. This step requires the most knowledge of the author, because the support was not useful. At the same time, edited options might be added to the local RAG system as an option because they might be reused in further documents in the authors' domain. This can be helpful for a transparent version history of reusable non-terminal symbols and tree node options as choices for upcoming text generations of the collaborative team. For PNT the probability distribution for an OR-expression, as in rule (1), can be defined as follows: The option SCI is assigned a probability of 0.7, while the replacement of S by NOV has the probability of 0.3. A generated random number less than 0.7 will result in the replacement of S by SCI, whereas a random number greater than or equal to 0.7 will lead to the replacement of S by NOV.

TEXT GENERATION – PROBABILITY

Text generation can be dependent on random experiments in an AST, where the rule allows choices. A probability distribution on a finite set of options describes this mathematically. The text generation creates decision numbers that are used in the generative process when optional cases in a rule are possible. For the rule $S \rightarrow SCI \mid MA$ a random number r between 0 and 1 with a uniform distribution on the interval [0,1] determines the replacement for a start symbol S. SCI is selected if r is smaller than 0.7, and MA is selected otherwise. The interval [0,1] is decomposed into n subintervals for n different options in the grammar rule. To ensure transparency in the derivation process from a tree node in an AST to child nodes, these random numbers ($r_1, r_2, \ldots, r_n$ for n) for n different random experiments should be explicitly assigned to the corresponding rules ($R_1, R_2, \ldots, R_n$) of the grammar.

DOCUMENT OBJECT IDENTIFIER

The different types of grammar rules extend the classical concept of a grammar. For a transparent version history, all non-terminal rules are regarded as a digital object that can be selected for application in text generation. Generalizing the approach conceptually, a digital object identifier (DOI) can be used to create a unique and persistent identifier for a rule and/or the corresponding subtree of an AST. Similar to page references in books, we extend probabilistic or decision-making processes of the author. Selecting an option is part of the citation process for grammar rules.

Due to the fact, that DOI can handle various digital objects [6], the generation of multimedia documents, that include digital objects like audio, animation, data, etc., a rule for text generation can be referred to consistently. DOI is standardized by the International Organization for Standardization (ISO) and is an existing implementation of the Handle System, so a reference for grammar rules follows an established work. It ensures a transparent history of a grammar rule by providing a unique identifier that allows retrieval of the rule and its corresponding subtree. Since the DOI functions consistently within the Uniform Resource Identifier framework, it enables reliable referencing and access to these elements. This paper allows generative models to be transparent if they are applied on the root file of academic, professional, and government documents and also on the decomposition of the documents reflecting generative processes transparently.

CONTEXT DEPENDENT GRAMMAR

Context-free or -dependent grammars represent the syntactical structure of the document or a language [7]. In compiler theory, an application of grammars are relevant for the replacement and application of a rule. In this context, we consider the current state of the decision tree or AST as the context in which the next steps for the text generation are performed. The existing validation of parts of the tree nodes in the AST or the decision tree affects the priority and relevance of the automated prompt selection for LLNT or RNT.

APPLICATION OF THE CONCEPT IN LITERARY STUDIES

In recent years, Literary Studies have transitioned from a model of individual research to interdisciplinary collaboration. The Digital Humanities demonstrated that traditional humanities methodologies can be expanded through the use of algorithmic processes. The integration of RAG and decision trees offers new possibilities for text generation and structuring, documentation, and analysis of scholarly arguments.

A central challenge in literary scholarship is ensuring the transparency and versioning of argumentative processes. Interpretations and methodological choices emerge through an iterative process of evaluating theories and readings. The fusion of retrieval grammar rules and decision trees could serve as a model for a documented research practice. Versioned decision trees can enhance transparency in scholarly work by making interpretive pathways explicit. Automated source retrieval through RAG could facilitate this process by suggesting relevant scholarly texts and integrating them into the decision-making structure. Additionally, the structuring of literary arguments can be improved by formalizing academic texts through ASTs and decision rules, thereby making recurring argumentative patterns identifiable.

Furthermore, literary studies provide a compelling testbed for generative models. Algorithmic versioning could, for example, increase transparency by tracking the evolution of research questions over time and highlighting paradigm shifts. Recent research has emphasized the role of human-machine collaboration in shaping new research methodologies [8]. These findings align with the applica-

https://doi.org/10.5281/zenodo.17237756

tion of retrieval-based decision trees, as they offer a framework where human interpretative agency and algorithmic support coalesce to foster more structured, transparent, and reproducible scholarly work.

The increasing collaboration between literary scholars, computer scientists, and data scientists underscores the relevance of these methodologies. The combination of humanities and computer science perspectives opens up new methodological possibilities and calls for a critical reflection on the epistemological foundations of both disciplines. While classical philology long relied on individual hermeneutic analysis, contemporary literary research is shifting towards collaborative, technology-assisted approaches. This development aligns with the broader "laboratory turn" in the humanities, where research environments are increasingly modelled after scientific laboratories, fostering interdisciplinary exchange and methodological innovation. As Pawlicka-Deger states: "The humanities lab does not simply imitate the science lab but adapts this new infrastructure for its own purposes and needs." [9] This shift also affects methodological frameworks: "The laboratory turn has emerged […] as a part of a wider process of the *laboratoriation* of social life, which has been occurring since the 1980s and with a significant intensification in the last ten years." [10] The combination of digital and traditional approaches requires technical proficiency and a reconsideration of fundamental scholarly paradigms. [11].

In this context, the application of retrieval-based decision trees to academic writing presents a transformative perspective: a structured, transparent, and interdisciplinary literary studies.

LIMITATIONS OF THE CONCEPT

According to the conceptual design, the context of the AST, and the decision, it requires additional conceptual work to handle a context as a tree structure and extracting the relevant context for a specific rule from a tree structure, e.g. given as JSON for a web-based generation on a client side. The current paper does not provide a conceptual design solution to derive the specific context from structure data given as decision tree or AST.

CONCLUSION

This conceptual approach describes the fusion of text generation by grammar rules of different types and using them in a VCS as an identifiable digital object. The main step is that the VCS can provide transparency for text generation. Recursive application of rules towards a final generated text creates an AST with tree nodes that represent an author decision, validation, or quality assurance in contrast to automated steps of document generation with deterministic or probabilistic grammar rules. The DOI serves as a mechanism to search, find, and identify these grammar rules and ASTs uniquely and the decision tree incorporated the work of humans with the document. This is a major step toward transparent separation of automated and human work on a product.

Quality assurance of authors and reviewers of the document might not be an automated part of the document evolution, but it changes the trust of the community in generated documents if the results have passed a human quality assessment. Reusability as part of FAIR data principle is supported for text generators with Uniform Resource Identifiers to fetch and apply the grammar rule for a specific generative task. For transparency, the DOIs for the grammar rules are accompanied by the author selection if the rules replace a non-terminal symbol of type RNT and ONT. Finally, by application of the proposed concept, a new grammar-driven document generation together with retrieval and GenAI. e.g., for scientific articles. Version control offers the possibility to share the generative steps with the version history. Beyond the final product, the version control offers transparency for "who did when what" in the evolution of the document. This is relevant for an educational system in which the contribution of students in collaborative learning should be separated from automated generation process of GenAI.

REFERENCES

[1] B. A. Nosek *et al.*, "Promoting an open research culture", *Science,* vol. 348, no. 6242, pp. 1422-1425, 2015. doi:10.1126/science.aab2374

[2] N. Nizamuddin *et al.,* "Decentralized document version control using ethereum blockchain and IPFS" *Comput Electr Eng,* vol. 76, pp. 183-197, 2019. doi:10.1016/j.compeleceng.2019.03.014

[3] C. Clark, "ASTs for optimizing compilers", *ACM SIGPLAN Notices,* vol. 36, no. 9, *pp. 25-30,* Sep. 2001. doi: 10.1145/609769.609773

[4] A. Gamba, "Real options valuation: A Monte Carlo approach" *Faculty of Management, University of Calgary WP*, 2002/3, 2003. doi:10.2139/ssrn.302613

[5] G. Satta and O. Stock, "Bidirectional context-free grammar parsing for natural language processing" *Artificial Intelligence*, vol 69, no. 1-2, pp. 123-164, 1994. doi: 10.1016/0004-3702(94)90080-9

[6] R. Chandrakar, "Digital object identifier system: an overview.", *The Electronic Library,* vol. 24, no. 4, pp. 445-452, Jul. 2006. doi:10.1108/02640470610689151

[7] R. Simmons and Y. Yu, "The acquisition and use of context-dependent grammars for English", *Computational Linguistics*, vol. 18, no. 4, pp. 391-418, 1992. doi:10.5555/176313.176314

[8] V. de Boer and L. Stork, "Hybrid Intelligence for Digital Humanities" *HHAI 2024: Hybrid Human AI Systems for the Social Good,* vol. 386, pp. 94-104, 2023. doi: https://doi.org/10.48550/arXiv.2406.15374

[9] U. Pawlicka-Deger, "The laboratory turn: exploring discourses, landscapes, and models of humanities labs." *Digital Humanities Quarterly*, vol. 14 no. 3, p. 1, 2020.

[10] Ibid., p. 63

[11]A. Lucke, "Methodologische Potenziale und Herausforderungen einer transdisziplinären Zusammenarbeit" in: A. Lucke and H. Johannes (ed.): Literaturwissenschaft und Informatik. Transdisziplinäre Perspektiven, digitale Methoden und selbstlernende Algorithmen. Bielefeld: transcript, pp. 7-34, 2024. doi: 9783839470039-001

https://doi.org/10.5281/zenodo.17238018

TOWARDS THE EXTRACTION OF LOCATION REFERENCES AND TOPICS FROM SEMI-STRUCTURED TEXTUAL DATA FROM THE OPEN WEB INDEX USING OPEN-SOURCE LARGE LANGUAGE MODELS

P. Gadziomski*[1], V. Rittlinger†[2], M. Pfeffer‡[1], S. Voigt§[2]

[1]Media University (HdM), Stuttgart, Germany

[2]German Aerospace Center (DLR), Earth Observation Center, Oberpfaffenhofen, Germany

Abstract

With the steadily growing relevance of the Web as information source and the associated increase in web content, the systematic extraction of structured information from it is becoming increasingly important. Every day, thousands of social media posts and news articles are published, containing not only thematic but also geo-spatial information such as location names and addresses. These data are relevant for numerous applications and research fields—including open-data projects like OpenStreetMap, the optimization of search engine indices, or their use for example in crisis management. The automated extraction of addresses from web sources—particularly from imprint pages—could efficiently capture legal information, such as compliance with the General Data Protection Regulation (GDPR) or improve deep learning models for Named Entity Recognition (NER). Despite the high relevance of this task, existing methods for address extraction from text have so far yielded only limited results due to inconsistent formatting of addresses, the ambiguity of words, and the embedding of addresses in unstructured texts. Since rule-based methods for address extraction achieve only limited quality, the use of Large Language Models (LLMs) is proposed as a promising alternative to specifically extract addresses from imprint pages. Since the release of GPT-3, LLMs have enabled significant advancements in various fields, particularly in automated text processing. Information extraction, as a subfield of Natural Language Processing (NLP), is gaining increasing relevance due to LLMs availability and functionality and becomes an active research topic. Given the continuous evolution of these models, this trend is expected to persist. This applies both to the technical developments of LLMs and to advancements in prompting methods, and refinement of model output through targeted inputs.

The data used in this study originates from the "Legal" datasets of the Open Web Index of the OpenWebSearch.EU projects, providing substantial amounts of imprint data. The data is restricted to German language. The extracted dataset is annotated using LLMs, followed by manual correction. The result is the creation of an annotated dataset with manually collected “gold standard” of geo-location samples for comparison and quality assessment. The hit rate of the LLM is documented to establish a well-founded basis for further work. The geo-localization results are documented to compare with different model outputs and the different applied prompting techniques. Following the extraction, an evaluation is conducted to determine at which level the models can extract relevant geo-information. Addresses consist of country, postal code, city, street name, and house number. A specific score is assigned to each of these elements. This metric is designed to assess the effectiveness of address extraction using LLMs. Additionally, it is examined whether the German language yields better results for German addresses or, if in general, the English language enables better extraction. To make optimal use of spatial data, the websites in the dataset are classified thematically e.g. by company type or a more diverse classification. For this classification task, the LLMs are provided with pre-existing thematic categories. The websites are classified according to the plain text and the URL of the website. The thematically classified data enables targeted evaluation of the geocoded data points through subsequent visualization and analysis.

The data obtained includes all addresses found in the imprint as well as the classification of the website. The output data from models with many parameters is expected to be complete and will likely surpass previous rule-based and data-driven approaches. However, there remains a possibility that the models, regardless of their parameter size, may not be able to perform address extraction and classification at sufficient quality. Extracting the spatial context in form of coordinates from the addresses enables a large-scale geographic analysis of imprint entries. The thematic context can indicate the type of institution on the site. Based on the achieved quality and hit rate, the computational power required by each model is analyzed to determine to optimize for the required computational resources. Therefore, the evaluation does not only capture the outputs but also records the number of generated tokens, the resulting costs, and the processing time.

LLMs are expected to achieve a significantly higher hit rate in address extraction than conventional methods through targeted prompting. The knowledge gained from this study can contribute to the improvement of data-driven geo-spatial text data analysis and can be used in areas such as geo-spatial search engines and many types of open data projects. It should be noted that this study is work in progress, and the results presented reflect first analysis results.

* pg058@hdm-stuttgart.de
† vanessa.rittlinger@dlr.de
‡ pfeffer@hdm-stuttgart.de
§ stefan.voigt@dlr.de

https://doi.org/10.5281/zenodo.17238127

LARGE-SCALE GRAPH VISUALISATION OF OPEN WEB INDEX AND ITS EVOLUTION IN TIME*

P. Smolková, K. Slaninová†

IT4Innovations, VSB - Technical University of Ostrava, Czech Republic

Abstract

Dynamic networks are models that describe the evolving relationships between real-world entities in various application domains such as social network analysis, communication, biological processes, or the Internet. Web networks are a special type of information network where nodes represent web pages, each other connected by hyperlinks.

Visualisation of complex networks is a key tool for their analysis and interpretation. The graphical representation allows intuitive recognition of structures such as communities, central nodes, or important connections between parts of the network. Well-designed visualisations make it easier to navigate the data, but also support understanding of dynamic changes in the network and enable effective presentation of results.

Visualisation and processing of (extreme) large-scale networks is challenging due to unique characteristics such as load imbalance, lack of locality, and access irregularity. Considering the possibilities offered by recent supercomputing power, we have revised current algorithms suitable for the visualisation of large-scale networks and were able to visualise networks in sizes ranging from hundreds of thousands to million nodes. The experiments were performed on the visualisation of the Open Web Index produced by the OpenWebSearch.eu project. The complexity of the problem is discussed in the context of performance and computation power needed for the visualisation of such (extreme) large-scale graphs.

[*] This publication has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, `https://doi.org/10.3030/101070014`).

[†] katerina.slaninova@vsb.cz

https://doi.org/10.5281/zenodo.17238236

EXPANDING THE LANGUAGE AND CULTURAL COVERAGE OF COMMON CRAWL

P. Ortiz Suarez*, G. Lindahl, T. Vaughan, S. Nagel
Common Crawl Foundation, Beverly Hills , United States of America

The Common Crawl Foundation is a nonprofit organization that has been operating since 2007. Its mission is to preserve and freely share samples of the public Internet. Common Crawl is a key partner to the AI community, as well as many other research communities. Our over ten-petabyte archive provides most of the web data used to train LLMs. Our crawling has always been polite and ethical, and strictly obeying `robots.txt`. We thus believe that improving Common Crawl's language diversity as well as its cultural and community diversity, will directly benefit everyone from the AI to the crawling and archiving communities.

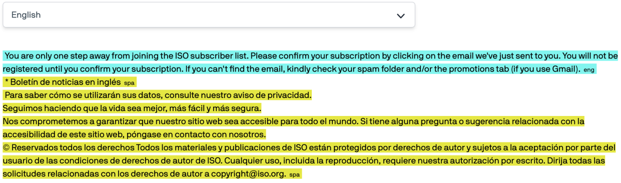
The Common Crawl Foundation has already been working on linguistic diversity with academic and industry partners, including Occiglot, HPLT, MLCommons, the Allen Institute (Ai2), the AI Alliance, the Linux Foundation AI and Data Foundation, and many more. However, while these efforts have already contributed to the cultural and linguistic coverage of our corpus, from our own statistics, we know that our data has always been biased towards English content making our dataset difficult to use for individuals and organizations from smaller linguistic communities.

We have always wanted to make Common Crawl as representative as possible of the Open Web, so we present here two projects on which we have been working and that we hope will allow us to expand the language and cultural coverage of our crawls, making it more representative of the actual linguistic and cultural diversity found on the web.

Both projects will require input from the community, as our team is small and we speak but a handful of languages, and as we also believe that the languages and the content written in them belong in the end to their respective linguistic communities.

The first initiative that we introduce here is the *Web Languages project*[1], which asks culturally-literate speakers to work together to make a list of important websites for different languages, cultures, and communities. We have asked for input for nearly 8,000 languages. These curated lists are then used by our web crawler to find clusters of linked websites which are important to the given culture or community. Even languages with very few web pages can be effectively crawled using this methodology.

This type of human collaboration and curation is a mature idea, and Common Crawl's team has successfully used this approach in the past. Success of this project depends upon collaborating with a wide range of people, recruited in collaboration with universities, companies, governments, and other organizations.


English

You are only one step away from joining the ISO subscriber list. Please confirm your subscription by clicking on the email we've just sent to you. You will not be registered until you confirm your subscription. If you can't find the email, kindly check your spam folder and/or the promotions tab (if you use Gmail). eng
* Boletín de noticias en inglés spa
Para saber cómo se utilizarán sus datos, consulte nuestro aviso de privacidad.
Seguimos haciendo que la vida sea mejor, más fácil y más segura.
Nos comprometemos a garantizar que nuestro sitio web sea accesible para todo el mundo. Si tiene alguna pregunta o sugerencia relacionada con la accesibilidad de este sitio web, póngase en contacto con nosotros.
© Reservados todos los derechos Todos los materiales y publicaciones de ISO están protegidos por derechos de autor y sujetos a la aceptación por parte del usuario de las condiciones de derechos de autor de ISO. Cualquier uso, incluida la reproducción, requiere nuestra autorización por escrito. Dirija todas las solicitudes relacionadas con los derechos de autor a copyright@iso.org. spa

Figure 1: Annotation interface for the LangID Project

The second project is an annotation campaign for Language Identification (LangID)[2] that we are conducting in collaboration with MLCommons. In this campaign we are asking participants to annotate a subset of Common Crawl data. We would like as many annotations, and cover as many languages as possible, in order to create the first web-based LangID dataset. Our goal is to train a small language classifier that would help us steer our crawling towards under-represented languages at crawl time. This concept is a mature idea with which we have already experimented on a small list of languages. Success of this project world-wide depends on collaborating with a wide range of people, who can be recruited in the same way as the first project.

These two projects are interconnected, and mutually complementary. The first finds communities, including regional communities that mainly have web content in a national language. The second project uses a different mechanism that can find web pages with under-represented language content wherever they are on the Internet, even if they are not connected to the main community and cultural clusters found by the first project.

With these two initiatives we hope to expand the access to research and technologies that our dataset has already made possible for high-resource linguistic and cultural communities, and make them available to communities all around the world. We will present the findings and challenges that we have encountered while conducting these two projects and how our crawl coverage has evolved since we started working on these initiatives.

* pedro@commoncrawl.org

[1] `https://github.com/commoncrawl/web-languages`

[2] `https://dynabench.org/tasks/text-language-identification`

https://doi.org/10.5281/zenodo.17238317

EFFICIENT SESSION SEARCH USING TOPICAL INDEX SHARDS

Gijs Hendriksen, Djoerd Hiemstra, and Arjen P. de Vries*
Radboud University, Nijmegen, The Netherlands

ABSTRACT

Retrieval is often considered one query at a time. However, in practice, queries regularly come in the context of sessions with coherent topics. By dividing a collection into topical index shards and matching the topical context of a session with the right shards, we may reduce the amount of resources required for answering each query. We consider two alternatives: (1) starting with exhaustive search and pruning unnecessary shards after each session turn, and (2) applying a resource selection algorithm to pre-select shards at the start of the session.

The first approach, which we call *shard pruning*, uses pseudo-relevance feedback of retrieved documents to select shards for later turns. It starts with exhaustive search for the first query in the session. After each query, we register which shards did not contribute any documents to the top 1500 retrieved documents, and remove them from consideration for subsequent session turns. In other words, each time we process a query, we prune the set of shards from which documents are retrieved. The rationale behind this approach is that the session topic should become more pronounced as the session proceeds, and thus the number of shards under consideration can be reduced as we go. The possible downside is that we prune shards that are not relevant to a specific query in the session but would be useful for a later one.

The second approach is based on the selective search setting, in which a *resource selection* algorithm is used to predict the set of relevant shards for a single query. In our case, we select the shards to use throughout the session after receiving user input at the start of the session. Note the difference with selective search: instead of performing resource selection for every query individually, we only perform it once for the first query in the session. The resulting list of shards is used for all queries in the session.

We evaluate both approaches on the TREC Conversational Assistance Track (CAsT) datasets, which contain conversational search sessions with coherent topics. We focus on CAsT 2019 and 2020 and use the manually rewritten queries provided by the task organizers to focus on retrieval effectiveness (instead of conversational query rewriting). We apply the QKLD-QInit clustering algorithm to partition the collection into a set of topical index shards.

Table 1 shows the mean recall (R@1000), cost-in-shards (how many shards were used) and cost-in-postings (how many postings were used) obtained by our shard pruning system on the CAsT 2019 dataset. Our setup is able to reduce overall cost by nearly 50%, while keeping recall within a 5% margin of exhaustive retrieval. Similar trends were observed on the CAsT 2020 collection.

Table 1: CAsT 2019 performance of systems that iteratively prune shards after each conversation turn.

	R@1000	CiS	CiP ($\times 10^3$)	
Exhaustive	0.84	94.0	1197.5	
Shard pruning	0.83	35.5	614.3	(–49%)

Table 2: CAsT 2019 performance of systems that select shards for the whole session using only the first query.

	R@1000	CiS	CiP ($\times 10^3$)	
Exhaustive	0.84	94.0	1197.5	
SRBR	0.81	5.7	206.9	(–83%)
CORI	0.83	58.0	877.6	(–27%)
ReDDE	0.82	32.0	596.4	(–50%)
Rank-S	0.82	36.8	642.1	(–46%)
Taily	0.82	51.0	778.6	(–35%)
L2R	0.82	25.0	516.7	(–57%)

Table 2 shows the same metrics for our system that preselects shards using the first query, using a number of popular resource selection algorithms: CORI, ReDDE, Rank-S, Taily and L2R. SRBR is an oracle method that ranks the shards based on the number of relevant documents they contain *for the whole session*. This setup is extremely effective when we use oracle resource selection. However, in practice, existing resource selection algorithms struggle to select the right shards using only the first query.

Our experiments show the viability of using topically partitioned document collections to make conversational question answering more efficient: high recall can still be achieved with a 50% reduction in costs. A system tuned for early precision requires even less resources.

Our work was accepted to ECIR 2025 [1] and our code is published to GitLab.[1]

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, `https://doi.org/10.3030/101070014`).

REFERENCES

[1] Gijs Hendriksen, Djoerd Hiemstra, and Arjen P. de Vries. Efficient Session Search Using Topical Index Shards. In *ECIR 2025*. To appear.

* {gijs.hendriksen, djoerd.hiemstra, arjen.devries}@ru.nl

[1] `https://gitlab.science.ru.nl/informagus/efficient-session-search`

https://doi.org/10.5281/zenodo.17238350

USING THE OPEN WEB INDEX TO CREATE NEW SEARCH APPLICATIONS FOR RESEARCH.FI

J. Theodoropoulos*, J. Kesäniemi, CSC – IT Center for Science

Abstract

We plan to utilize the Open Web Index to enhance the features of Research.fi, a Finnish portal showcasing national research related outputs. Our plans include improvements to the existing search functionality of the website and the introduction of fields for related publications. Our work aims to improve the accessibility of information related to research and provide a generalizable framework for similar open search projects.

EXTENDED ABSTRACT

OpenWebSearch.eu (OWS) is a project funded by Horizon Europe with the aim of building the European Open Web Index (OWI). The OWI provides an option for Internet search that is sovereign from the large technological companies dominating the field. While not targeting to build a single search engine to compete with the existing ones, the OWI provides a backbone to new search applications including search engine verticals. [1]

Research.fi is a Finnish service offered by the Ministry of Education and Culture, and developed by the CSC - IT Center for Science, also an participating organisation in OWS. The service contains information on research conducted in Finland including publications, grants, organizations and infrastructures. Information available in Research.fi is based on the National Research Information Hub, which acts as a national aggregator of research-related data in Finland. The purpose of the Research.fi is to improve the discovery of Finnish research, support the reuse of research-related information, and provide profiles for researchers for sharing their activities and outputs. [2]

In the first phase of the project, a prototype of the science index, i.e. a search vertical focusing on science related content, was created. First, crawled content was downloaded and filtered using the owilix tool [3] to include content only in the three languages supported by the Research.fi portal: Finnish, Swedish and English.

To create a dataset produced in a consistent manner, only data crawled in 2025 were used. Pages related to science were further filtered using the corresponding Curlie labels [4] provided in the metadata of the crawled websites. An additional need to filter out pages with low quality texts was noticed in initial testing, and a language model [7] was used to identify these pages for removal. These low quality pages included old message board threads but also pages from universities' websites that did not contain text that could be scraped in an efficient manner.

From the curated set of pages, the index was built with tools from OWS project [5], and MOSAIC [6] was utilized for queries. Although still in progress, multiple possible applications have been identified for using the curated science index as part of the Research.fi portal.

- Expanding the search results of Research.fi with sites from the OWI, thus containing information of science and research conducted outside of Finland.
- Creating a field of related content for entries present in Research.fi; e.g. using the OWI to identify research done using a specific grant.
- Utilizing the OWI to measure and provide insights of the impact of research output in metrics other than citations.
- Using a knowledge graph and topic modeling based on data from Research.fi to enhance queries made to the OWI.
- Leveraging the capabilities of large language models with a retrieval-augmented generation based search implementation.

This project is under continuous development with the goal to implement new features to the Research.fi portal by the end of this year. To maintain reliable features with data from the OWI, development is carried out in a way that enables reproduction of the results and updating them with more data from daily crawls.

ACKNOWLEDGEMENTS

The authors are employed by CSC – IT Center for Science and their work is partially funded by OpenWebSearch.eu.

REFERENCES

[1] OpenWebSearch.eu, `https://openwebsearch.eu`

[2] Research.fi, `https://research.fi`

[3] owilix, `https://opencode.it4i.eu/openwebsearcheu-public/owi-cli`

[4] Curlie.org, `https://curlie.org`

[5] Spark indexer, `https://opencode.it4i.eu/openwebsearcheu-public/spark-indexer`

[6] MOSAIC, `https://opencode.it4i.eu/openwebsearcheu-public/mosaic`

[7] llm-data-textbook-quality-fasttext-classifier-v2, `https://huggingface.co/kenhktsui/llm-data-textbook-quality-fasttext-classifier-v2`

[*] jason.theodoropoulos@csc.fi

https://doi.org/10.5281/zenodo.17238407

BUILDING THE OPEN WEB SEARCH COMMUNITY (AND KEEPING IT GROWING)

U. Gmelch[†], Open Search Foundation e.V., Starnberg, Germany

Abstract

This talk traces the history of the Open Web Search Community, its achievements and also throws a glance into the future. In particular, it addresses the different stakeholder groups in the Open Web Search community, their activities and the role of the Open Search Foundation inside the community.

The Open Web Search Initiative got started well before the OpenWebSearch.EU project. In fact, the community and particularly the Open Search Foundation were instrumental in inspiring the Next Generation Internet (NGI) section of the Horizon Europe program to raise its ambitions regarding the search component of NGI. Of this came the OpenWebSearch.EU project with 14 renowned European research partners and computing centres from 7 countries, currently in its final year.

The Open Search Foundation (OSF) [1] has been the central node in the Open Web Search community since its founding in 2018. It can be described as the ‘spider in the web’ but also as the catalyst for several projects in the Open Web Search field, such as OpenWebSearch.EU, the German PriDI project [2] in the legal domain or #ethicsinsearch [3], among others. Over the years, the Open Search Foundation has inspired big institutions, small organisations and individuals to join their efforts and jointly conceptualise, build and advance Open Web Search capacities in Europe and beyond.

The Open Web Search Community can be broadly categorized into six main stakeholder groups: compute providers, software & system developers, policy / oversight & funding organisations, democratic / public curation bodies, application & exploitation domain and data / distribution hubs. These can further be grouped by their proximity to the community, e.g. into core community members, contributors, active supporters and followers (or observers). Particular mention goes to the third-party partners of the OpenWebSearch.EU project who actively extend and enrich the existing community with further research, application development and legal guidance [4]. Beyond OpenWebSearch.EU, associated projects and activities in the Open Web Search field are KIsu [5], which builds an inclusive frontend for RAG systems, fit for use by children and the elderly, the three ethics workshops funded by the CAIS centre during 2024, where ethical questions around the Open Web Index were discussed and developed, and the yearly #FreeWebSearch Day [6], organised by OSF to promote free, open and transparent web search.

While the OpenWebSearch.EU project was, and is, very important in providing a core development capacity for Open Web Search, it was always important to ensure integration and active exchange with the broader community at the same time. One important connector is, of course, the Open Search Symposium (#ossym), already held for the seventh time this year. Another one is the monthly Community Update online call [7], that has recently featured webinars on the core R&D components of the OpenWebSearch.EU project. Another means for integrating the Open Web Search community are the OSF working groups [8] (tech, application, ethics, legal, education&literacy and economy) which hold regular meetings, online or occasionally in person, and advance the work on Open Web Search significantly. Overarching all of this is the Mattermost community platform on Open Web Search [9], where all contributors are welcome to share news, discuss ideas and collaborate.

The biggest achievement of the Open Web Search Community is probably the prototype of the Open Web Index [10], built within the OpenWebSearch.EU project and supported by the community. There are first use cases and cooperation with search engine providers, AI Factories and e.g. the OpenEuroLLM project building on the Open Web Index and the associated data products. Furthermore, through persistent communication and outreach, Open Web Search is now starting to be on the policy agenda in Europe, as evidenced for example by the session on Open Web Search organised at the NGI Forum in June, 2025 in Brussels [11].

In the near future, the Open Web Search Community very likely faces the challenge of not having a single / central follow-up activity or project to OpenWebSearch.EU. As a consequence, it will be important to build on the existing tools for integrating and communicating activities within the community: monthly community updates, working groups, Mattermost channels etc. An important factor in maintaining coherence in the Open Web Search community will be keeping the current infrastructure and the core contributors to the Open Web Index working together to enable further research and application development by the community, based on the Open Web Index.

REFERENCES

[1] https://opensearchfoundation.org/en/

[2] https://pridi-projekt.de/home-en/

[3] https://ethicsinsearch.org/en/home-e/

[4] https://openwebsearch.eu/third-party-projects/

[5] https://kisu.cs.uni-saarland.de/

[6] https://freewebsearch.org/en/

[†] community@openwebsearch.eu

https://doi.org/10.5281/zenodo.17238407

[7] https://openwebsearch.eu/community/owseu-community-updates/

[8] https://opensearchfoundation.org/en/working-groups/

[9] https://openwebsearch.eu/community/ows-eu-community-on-mattermost/

[10] https://openwebindex.eu/

[11] https://ngi.eu/ngi-forum25/

https://doi.org/10.5281/zenodo.17238455

EXPLORING TECHNOSCIENCE IN THE PUBLIC SPHERE: OPPORTUNITIES THROUGH OPEN WEB SEARCH

E. Di Buccio*, University of Padova, Padova, Italy

Abstract

Interdisciplinary research fields, such as Science and Technology Studies (STS), investigate the relationship between science, technology, and society, and how they influence each other. These investigations can benefit from the extraction and the analysis of different representations of science and technology issues — for example, comparing the perspectives on Nuclear Power of social media users or newspaper readers with those of scientists and other experts. The vast amount of digitized content available online enables the use of diverse sources for these representations, thus fostering the opportunity to consider a plurality of perspectives.

Examples of digitized informative resources include various media streams, such as social media posts and online newspaper articles. These types of sources are currently utilized in the TIPS project [6], an interdisciplinary initiative specifically aimed at studying science and technology discourse within the public sphere. To support researchers in fields such as the social sciences, a dedicated web platform has been designed and developed [2]. The platform is built on open-source libraries to promote reproducibility and is structured using a service-oriented architecture.

In addition to media streams, other web-based informational resources may serve to complement existing representations. As noted by Lewandowski [4], an Open Web Index can be a valuable asset for researchers across various disciplines, including Computational Social Science. In projects such as TIPS, the primary users are experts from research fields such as the Social Sciences, Humanities, and Communication Studies. These users often require access to and analysis of specific segments of the Web to support their investigations. This need aligns with the objectives outlined in [3] and the Open Web Search project, which aims to make the index openly accessible as data. Moreover, selected portions of this index may serve as the foundation for developing specialized search engines—such as the one designed and implemented in the TIPS project to assist researchers in these domains.

Those portions of the index might need to be enriched with specific metadata necessary for the experts analysis. Examples of such metadata include the “actors” mentioned in the informative resources, e.g., named entities; these are examples mentioned in [3] in the “semantic enrichment” step. However, other indicators are useful to investigate specific research questions in the considered application scenario. Examples of those indicators might be those devised to measure the degree to which a semantic dimension is present in a document, e.g., the presence of“risk”; those indicators might rely on controlled vocabularies [1] or more complex techniques relying on embedding-based representations. Other metadata might include linguistic properties or related measures, e.g., readability. Because of the efficiency constraints on processing huge amount of data, those metadata might be available only for specific portions and, for instance, accessible through specific verticals.

Furthermore, within the scope of a single research project, multiple verticals may be employed to address the diversity of public discourse sources. For example, in the case of the TIPS project, dedicated verticals can be developed to target specific types of content, such as online news, forums or scientific blogs. Drawing on the extensive body of literature in resource selection, particularly from the domains of Distributed Information Retrieval (IR) and Federated Web Search [5], future versions of the platform can not only leverage portions of the open web index, but also rank these portions according to their relevance to particular technoscientific issues. Resource selection algorithms might be also be the basis for novel indicators to support expert analysis, e.g., for measuring the prominence of technoscientific issue in the public discourse.

REFERENCES

[1] Di Buccio, E., Lorenzet, A., Melucci, M., Neresini, F. (2016). Unveiling latent states behind social indicators. CEUR Workshop Proceedings, 1831.

[2] Di Buccio, E., Cammozzo, A., Neresini, F., Zanatta, A. (2022). TIPS: Search and Analytics for Social Science Research. CEUR Workshop Proceedings, 3178.

[3] Granitzer, M., Voigt, S., Fathima, N. A., Golasowski, M., Guetl, C., Hecking, T., Hendriksen, G., Hiemstra, D., Martinovič, J., Mitrović, J., Mlakar, I., Moiras, S., Nussbaumer, A., Öster, P., Potthast, M., Srdič, M. S., Megi, S., Slaninová, K., Stein, B., 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. https://doi.org/10.1002/asi.24818

[4] Lewandowski, D. (2019). The web is missing an essential part of infrastructure. Communications of the ACM, 62(4), 24–24. https://doi.org/10.1145/3312479

[5] Shokouhi, M., Si, L. (2011). Federated Search. Foundations and Trends in Information Retrieval, Volume 5, Issue 1, 1–102. https://doi.org/10.1561/1500000010

[6] `https://www.tipsproject.eu/tips` [Last visited: April 30, 2025]

[*] emanuele.dibuccio@unipd.it

https://doi.org/10.5281/zenodo.17238499

LLM-ASSISTED EXPANSION OF PATENT AND SCHOLARLY LITERATURE KNOWLEDGE GRAPHS

André Rattinger *, ISDS, Graz University of Technology, Graz, Austria
Christian Gütl, ISDS, Graz University of Technology, Graz, Austria

Abstract

Patents and scholarly articles represent two deeply connected yet distinct sources of technical knowledge, each using specialized terminologies and referencing structures. Standard methods for connecting these sources—such as citation-based retrieval or classification overlaps—often miss nuanced or implicit relationships. To address this, we propose an LLM-assisted knowledge graph expansion pipeline, combining semantic embeddings and topological structures, validated through domain-specific constraints. Demonstrated initially within battery technology (CPC "H01M"), this pipeline generalizes effectively to other technical fields, enhancing knowledge discovery, prior art search, and strategic innovation analysis.

INTRODUCTION

Technological innovations documented in patents and scientific knowledge captured in scholarly articles represent two complementary yet distinct knowledge bases. Despite their interconnected nature, the integration of patents with academic literature is often minimal due to variations in language, structure, and referencing practices. Previous work has demonstrated the benefits of both semantic and topological graphs for patent retrieval and analysis [1,2], yet implicit, deeper connections between these domains frequently remain undiscovered. We propose to leverage Large Language Models (LLMs) to systematically uncover and validate latent semantic overlaps, connecting patent and publication knowledge graphs into a unified and enriched network.

Patents tend to emphasize legal and commercial aspects—claims, novelty, and scope of protection—while academic publications focus on rigor, reproducibility, and theoretical grounding. This disparity leads to variations in language (highly specialized or obfuscated legalese vs. structured academic prose), structure (claims vs. hypotheses), and referencing practices (formal classification codes vs. standard bibliographic citations). Consequently, direct links between these two corpora are frequently underexploited, hampering comprehensive knowledge discovery and prior art analysis [3].

Efforts to unify patents with academic literature typically rely on classification (e.g., CPC codes). While these methods excel at connecting documents in well-traversed or semantically obvious paths, they do not always capture deeper relationships—such as an unreferenced publication describing a method that closely matches a patented process [4]. Moreover, large-scale knowledge graphs, though promising, can be limited by the static nature of their input data; novel or implicit links remain hidden unless explicitly recorded [5].

Recent work in semantic patent graphs [1, 6] has shown that text-based embeddings (e.g., doc2vec or BERT) can reveal relationships overlooked by purely topological approaches. Simultaneously, topological graphs that rely on CPC co-classification or patent citations provide a robust, expert-assigned backbone [1, 7]. Despite these advances, the resulting graphs still tend to miss subtle overlaps across different classification categories or cross-disciplinary leaps between specialized research articles and patents.

In parallel, the rapid development of Large Language Models (LLMs) provides an opportunity to infer novel connections from natural language descriptions. LLMs can distill textual snippets—such as patent claims and scientific abstracts—into conceptual links, potentially labeling relationships with statements like "these two documents describe the same doping method" or "this publication could serve as potential prior art for that patent" [8]. However, LLMs are prone to overreach or "hallucinate," meaning any pipeline that leverages them must integrate domain-aware safeguards (e.g., checking chemical doping terms or verifying consistent references) to ensure correctness [4].

Given these convergent trends, we propose a unified pipeline that fuses semantic embeddings, topological structures, and LLM-based edge inference. Our pipeline begins by constructing a baseline knowledge graph from known links (citations, classifications, textual similarity), then solicits an LLM to propose new relationships where only moderate textual overlap is present. Potentially spurious suggestions are filtered by domain heuristics and textual consistency checks. The result is an expanded, domain-agnostic knowledge graph that better reflects the true scope of innovation and scientific discovery. Though illustrated using battery technology patents (CPC "H01M") due to its active and well-documented R&D environment [7], the approach naturally extends to other domains, such as AI (CPC "G06N") or semiconductor devices (CPC "H01L").

RELATED WORK

Early attempts to integrate patents and academic literature leaned primarily on citation extraction and classification overlaps, often treating each corpus separately or relying on manual heuristics. These methods excel at connecting well-documented prior art but often fail to expose deeper semantic relationships [4].

More recent approaches harness semantic embeddings to represent patents and publications in a shared vector space [6], enabling automated identification of potentially

[*] ...@protonmail.com

https://doi.org/10.5281/zenodo.17238499

related documents that lack direct citations. Parallel efforts focus on topological graphs built from co-classification or citation networks [1, 7], leveraging expert labeling to anchor patent knowledge bases. However, these purely topological approaches may overlook subtle or emerging links, especially across interdisciplinary boundaries [3].

A promising trend is LLM-assisted knowledge graph completion, where large language models infer new edges in graphs by understanding textual descriptions [5, 8]. While this has been demonstrated in general knowledge graphs and biomedical contexts [3], it remains less explored in patent–publication integration pipelines. Incorporating LLMs offers the possibility to bridge semantic gaps by interpreting specialized legal or scientific language and inferring relationships not explicitly stated [4]. The main challenge lies in mitigating LLM "hallucinations", underscoring the importance of domain-aware verification—such as chemical formula matching or consistency checks against domain ontologies.

Our work leverages these advances by introducing an LLM-driven layer to a base knowledge graph of patent–publication pairs. Through a combination of semantic embeddings, co-classification, citation data, and LLM-based link suggestions, we address both the missing explicit links and the subtler overlaps across conceptually adjacent technologies and research fronts. This approach is particularly apt for high-innovation domains like battery technology (CPC "H01M"), where rapid advancements outpace the coverage of static classification and citation systems [2, 7].

REFERENCES

[1] A. Rattinger *et al.*, "Semantic and topological patent graphs," in *SNAMS 2018: International Conference on Social Networks Analysis, Management and Security*, 2018, https://example.org/semantic-topological-patent-graphs.

[2] A. Rattinger, J.-M. Le Goff, and C. Guetl, "Semantic and topological graphs for patent retrieval," in *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS).* IEEE, 2019, pp. 175–180.

[3] J. Xu, C. Yu, J. Xu, Y. Ding, V. I. Torvik, J. Kang, M. Sung, and M. Song, "Pubmed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science," *arXiv preprint arXiv:2410.07969*, 2024.

[4] H. Aras, R. Dessi, F. Saad, and L. Zhang, "Bridging the innovation gap: Leveraging patent information for scientists by constructing a patent-centric knowledge graph," in *CEUR Workshop Proceedings*, vol. 3697, 2024, pp. 61–67.

[5] L. Yao, J. Peng, C. Mao, and Y. Luo, "Exploring large language models for knowledge graph completion," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2025, pp. 1–5.

[6] L. Siddharth, G. Li, and J. Luo, "Enhancing patent retrieval using text and knowledge graph embeddings: a technical note," *Journal of Engineering Design*, vol. 33, no. 8-9, pp. 670–683, 2022.

[7] H. Pohl and M. Marklund, "Battery research and innovation—a study of patents and papers," *World Electric Vehicle Journal*, vol. 15, no. 5, p. 193, 2024.

[8] B. Mo, K. Yu, J. Kazdan, P. Mpala, L. Yu, C. Cundy, C. Kanatsoulis, and S. Koyejo, "Kggen: Extracting knowledge graphs from plain text with language models," *arXiv preprint arXiv:2502.09956*, 2025.

https://doi.org/10.5281/zenodo.17238529

SCIENTIFIC QUESTION ANSWERING USING HYBRID RETRIEVAL AUGMENTED GENERATION

Roxanne el Baff, Tobias Hecking
German Aerospace Center (DLR), Institute of Software Technology, Cologne, Germany

Abstract

Large language models have strong capabilities for different purposes, such as searching and question-answering [1]. However, they hallucinate on domain-specific tasks. Recent research shows compound systems outperform standalone LLMs [2, 3]. A scientific domain such as Earth observation (EO), used by different fields such as oceanography and environmental science, requires a a tailored approach to deal with hallucinations to ensure in-depth answers [4]. This presentation introduces a system for EO that integrates LLM-based conversational search and question-answering by focusing on two components: a) data curation for a Retrieval-Augmented Generation (RAG)-based model and b) LLM-based evalu-

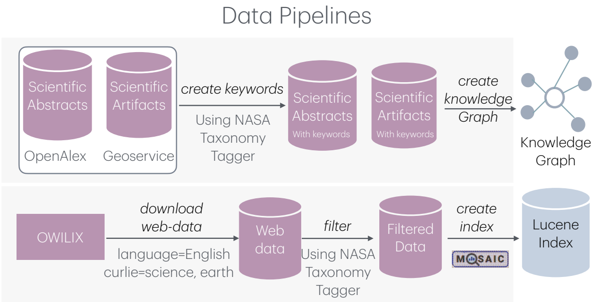


Figure 1: The two data pipelines: From Data Acquisition and Preprocessing to Knowledge-Graph creation (Top), and index creation (Bottom).

APPROACH

This section outlines our three-stage approach: 1) Data Pipelines, 2) RAG-based LLM, and 3) Evaluation.

1. Data Pipelines. We create an exhaustive dataset of earth observation, including three text genres [5] for two data pipelines. The first pipeline, **KG-pipeline**, (Figure1-top) creates a knowledge graph connecting scientific abstracts to scientific artifacts via keywords. The second pipeline, **INDEX-pipeline** (Figure1-bottom), uses crawled data from the Web, processes it, and generates a searchable index.

To filter and tag the data for the EO domain, we develop an EO Tagger, the TaxoTagger, based on the EO NASA taxonomy [6, 7] (e.g., earth storable), given a text, *n* keywords are returned with scores between zero and one. Below, we detail each pipeline:

- **KG-pipeline**: we download publication abstracts from OpenAlex [8], an open index of scholarly works tagged with topics across all scientific domains. We fetch abstracts with topics relevant to EO (e.g., Cosmic Evolution). Also, we download remote sensing and EO data from the DLR geoservice portal[1]. Then, we tag both genres using the TaxoTagger. We build a knowledge graph connecting scientific abstracts to artifacts (here, Geoservice data) via the keywords from the tagger.
- **INDEX-pipeline**: We download a search index shard from OpenWebIndex [9] using *owilix* [10, 11], restricted to English data tagged with *science* and *earth* (Curlie tags[2]). We filter the data with TaxoTagger, excluding those containing keywords below a specified threshold. Finally, we build a Lucene web index using MOSAIC [12].

2. RAG-Based LLM. The RAG-based LLM relies on the two pipelines described. When a user queries the LLM, the knowledge graph is queried for the top *k* results, where each result contains the top hits (nodes), along with their neighboring keywords and nodes. Simultaneously, the Lucene index is also queried for top *k* hits. These hits are incorporated within the LLM prompt context along with the query.

3. Evaluation For our evaluation, we first curate a set of prompts from domain experts. Due to the absence of reference data, we use Generative Pseudo-Labeling (GPL) [13] to create a weakly labeled dataset for training a response quality classifier. We conduct an automatic evaluation to show the impact of each data pipeline with an ablation study by comparing zero-shot prompting using one of the pipelines, both, or none. Subsequently, domain experts assess the best-performing model on criteria such as faithfulness, relevance, and completeness. Lastly, we compare human evaluations to LLM-based assessments to analyze their alignment.

CONTRIBUTIONS

Our contributions are the following:

- A reservoir of heterogeneous data sources for EO.
- An approach to evaluate the fusion between traditional retrieval (index-based) search with knowledge graph for a scientific domain with a RAG-based model.
- A comparative evaluation approach aligning LLM- and human-based evaluation for a scientific domain.

[1] `https://geoservice.dlr.de` accessed 25.03.2025
[2] https://curlie.org/

https://doi.org/10.5281/zenodo.17238529

REFERENCES

[1] C. Zhai, "Large language models and future of information retrieval: Opportunities and challenges," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 481–490.

[2] M. Zaharia *et al.*, *The shift from models to compound ai systems*, `https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/`, 2024.

[3] P. Lewis *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[4] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[5] J. Honeder, R. El Baff, T. Hecking, A. Nussbaumer, and C. Guetl, "A geo-contextualized multi-genre scientific search engine: A novel conceptual design and prototype evaluation," in *8th International Conference on Geoinformatics and Data Analysis*, Springer, 2025.

[6] J. Dutra and J. Busch, "Nasa technical white paper-enabling knowledge and discovery: Taxonomy development for nasa," *Retrieved January*, vol. 15, p. 2003, 2003.

[7] D. Miranda, "2020 nasa technology taxonomy," NASA, Tech. Rep., 2020.

[8] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *arXiv preprint arXiv:2205.01833*, 2022.

[9] M. Granitzer *et al.*, "Impact and development of an open web index for open web search," *Journal of the Association for Information Science and Technology*, vol. 75, no. 5, pp. 512–520, 2024.

[10] M. Granitzer *et al.*, "OpenWebSearch.eu - building an open web index on eurohpc ju infrastructures," *Procedia Computer Science*, vol. 255, pp. 43–52, 2025.

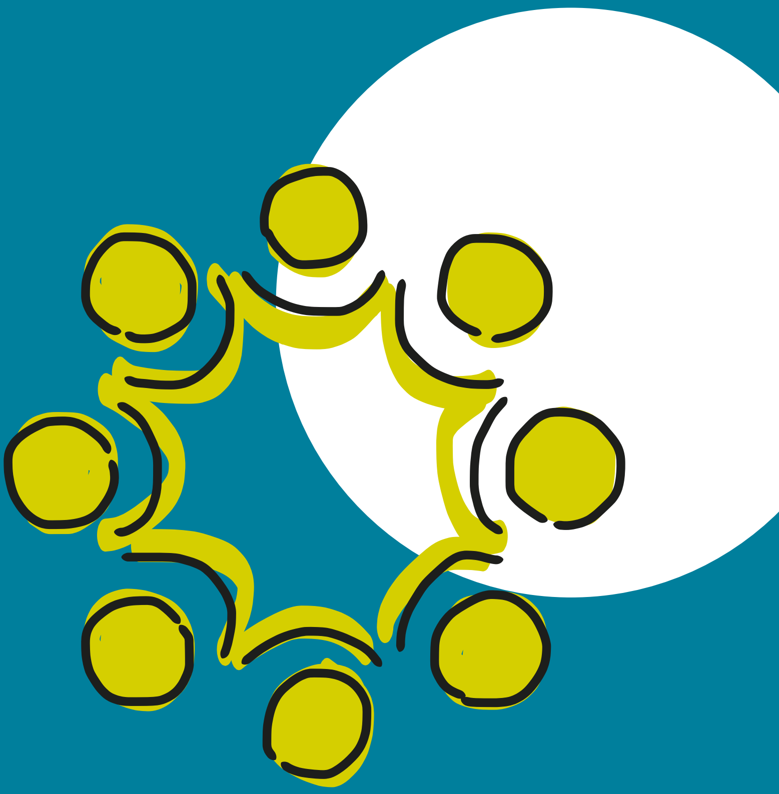
[11] M. Granitzer, *OWILIX - Open Web Index Client*, 2024. `doi:10.5281/zenodo.13833664`

[12] S. Gürtl and A. Nussbaumer, *MOSAIC Search Engine Framework*, version 0.1.0, 2024. `doi:10.5281/zenodo.13790237`

[13] K. Wang, N. Thakur, N. Reimers, and I. Gurevych, "Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval," *arXiv preprint arXiv:2112.07577*, 2021.

Appendix

List of Authors



open search
foundation

CSC

ISSN: 2957-4935
ISBN: 978-92-9083-705-3